China Perspectives

TOWARD MULTIMODAL PRAGMATICS

A STUDY OF ILLOCUTIONARY FORCE IN CHINESE SITUATED DISCOURSE

Lihe Huang

Toward Multimodal Pragmatics

Classic pragmatic theories emphasize the linguistic aspect of illocutionary acts and forces. However, as multimodality has gained importance and popularity, multimodal pragmatics has quickly become a frontier of pragmatic studies. This book adds to this new research trend by offering a perspective of situated discourse in the Chinese context.

Using the multimodal corpus approach, this study examines how speakers use multiple devices to perform illocutionary acts and express illocutionary forces. Not only does the author use qualitative analysis to study the types, characteristics, and emergence patterns of illocutionary forces, he also performs a quantitative, corpus-based analysis of the interaction of illocutionary forces, emotions, prosody, and gestures. The results show that illocutionary forces are multimodal in nature while meaning in discourse is created through an interplay of an array of modalities.

Students and scholars of pragmatics, corpus linguistics, and Chinese linguistics will benefit from this title.

Lihe Huang is Associate Professor at Tongji University and Humboldt Fellow of Germany-based Alexander von Humboldt Foundation. He is one of the leading young scholars in multimodal study and gerontolinguistics in China. His current research interest is utilizing the multimodal method to explore the linguistic behaviour of Chinese elders (visit him at: ageing.tongji.edu.cn).

China Perspectives

The *China Perspectives* series focuses on translating and publishing works by leading Chinese scholars, writing about both global topics and China-related themes. It covers humanities and social sciences, education, media and psychology, as well as many interdisciplinary themes.

This is the first time any of these books has been published in English for international readers. The series aims to put forward a Chinese perspective, give insights into cutting-edge academic thinking in China, and inspire researchers globally.

To submit proposals, please contact the Taylor & Francis Publisher for the China Publishing Programme, Lian Sun (Lian.Sun@informa.com).

Titles in linguistics currently include:

Teaching and Researching Chinese EFL/ESL Learners in Higher Education
Edited by Zhongshe Lu, Meihua Liu and Wenxia Zhang

Perception and Metaphor
A Comparative Perspective Between English and Chinese
Qin Xiugui and Tie Yi

New Research on Cohesion and Coherence in Linguistics
Zhang Delu and Liu Rushan

A Corpus-based Contrastive Study of the Appraisal Systems in English and Chinese Scientific Research Articles
Xu Yuchen, Yan Xuan, Su Rui and Kou Ying

Toward Multimodal Pragmatics
A Study of Illocutionary Force in Chinese Situated Discourse
Lihe Huang

For more information, please visit https://www.routledge.com/China-Perspectives/book-series/CPH

Toward Multimodal Pragmatics

A Study of Illocutionary Force in Chinese Situated Discourse

Lihe Huang

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

上海外语教育出版社
外教社
SHANGHAI FOREIGN LANGUAGE EDUCATION PRESS

First published in English 2022
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

and by Routledge
605 Third Avenue, New York, NY 10158

Routledge is an imprint of the Taylor & Francis Group, an informa business

Originally written in Chinese by Lihe Huang
Translated from Chinese into English by Lihe Huang

This English version is published by arrangement with Shanghai Foreign Language Education Press Co., Ltd, Shanghai, China.

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
Names: Huang, Lihe, author.
Title: Toward multimodal pragmatics : a study of illocutionary force in Chinese situated discourse / Lihe Huang.
Description: London ; New York : Routledge, 2021. | Series: China perspectives | Includes bibliographical references and index. |
Identifiers: LCCN 2021034252 (print) | LCCN 2021034253 (ebook) |
ISBN 9781032170916 (hardback) | ISBN 9781003251774 (ebook)
Subjects: LCSH: Chinese language—Discourse analysis. |
Pragmatics. | Modality (Linguistics) | Corpora (Linguistics)
Classification: LCC P302.15.C4 H83 2021 (print) | LCC P302.15.C4 (ebook) | DDC 495.101/45—dc23/eng/20211022
LC record available at https://lccn.loc.gov/2021034252
LC ebook record available at https://lccn.loc.gov/2021034253

ISBN: 978-1-032-17091-6 (hbk)
ISBN: 978-1-032-17092-3 (pbk)
ISBN: 978-1-003-25177-4 (ebk)

DOI: 10.4324/9781003251774

Typeset in Times New Roman
by codeMantra

Contents

Figures

Tables

Foreword

Multimodality and Pragmatics[1]

Prof. Dr. Yueguo Gu
Chinese Academy of Social Sciences
Beijing Foreign Studies University

This book aims to explore the interconnection between multimodality and pragmatics from the perspective of speech act. The speech act is not new in pragmatics. However, new knowledge can be brought to this traditional theory if multimodality is included. In this foreword, Prof. Gu would like to provide some background information about multimodality and pragmatics, which is the academic context of this book's discussion of multimodal pragmatics.

1 Preliminary Remarks

As multimodality has become a buzzword, it is of some urgency to explore the interconnection between multimodality and pragmatics. Multimodality, as Kress (2009: 54) points out, 'is not a theory even though it is often used as if it were. The term maps a domain of inquiry.' It is seen occurring in many disciplines, e.g. multimodal discourse analysis (MDA) in linguistics, multimodal corpus in language engineering, multimodal interface in human–computer interaction (HCI), and multimodal learning in education. There are two basic usages of the term emerging from these inquiries. One, as found in MDA, is in the sense of *multi* + *mode*, a mode being 'a socially shaped and culturally given resource for making meaning. *Image, writing, layout, music, gesture, speech, moving image, soundtrack* are examples of modes used in representation and communication' (Kress, 2009: 54; italics original). The other is in the sense of multi + modality, a modality being a sense organ and its interconnected neural networks.

> We receive information about the world through tactile sensations (body senses such as touch and pain), auditory sensations (hearing), visual sensations (sight), and chemical sensations (taste and olfaction). Each *sensory modality* has one or more separate functions.
>
> (Kolb & Whishaw, 2005: 135; italics added)

Multimodality in this foreword refers to the system of sensory modalities, which is found not only in humans but also in other animals, even in some plants. Issues concerning multimodality in MDA are best dealt with by MDA researchers.

Medical science and neuroscience in particular deal with neurophysiological structures and functions of multiple sensory modalities. The key issue we are concerned with here is: In what way does multimodality as such affect pragmatics? The answer clearly depends on how pragmatics is conceptualized that it incorporates multimodality as part of its theory building. If this is used as our yardstick, a survey of the existing literature shows a sparse picture. However, what is inspiring is that it is found in Morris's behavioural semiotics, which is acknowledged to be the original source for pragmatics. Other studies include Bates's study of developmental pragmatics (1976, 1979), Perkin's detailed investigation of pragmatic impairment (2007), and Gu's study of total saturated experience in situated discourse (Gu, 2009a, 2009b). The bulk of the foreword is to review these works.

As our review of Morris's behavioural semiotics will show, the pragmatics the founding father envisaged is much broader than the current mainstream linguistics pragmatics. The word multimodality being the latest coinage, the terms Morris uses for it are 'distance senses' (sight, hearing, and smell) and 'contact senses' (touch, taste) (see Morris, 1951 [1938]: 32). Since both types of senses play a vital role in living organisms' sign-making behaviours, Morris's theorization of sign behaviour naturally incorporates them as its intrinsic component. This multimodal component is distilled when linguistics pragmatics, formulated in the spirit of analytic philosophy of language, moves away from Morris's broader sign behaviour to focus on verbal behaviour only. The overall trend of pragmatics development, taking Morris's original vision as a reference framework, is reductionism in scope. The latest boom in multimodal studies, including embodiment in philosophy and sociology, encourages a growing current of bringing multimodality back to pragmatics.

2 Morris's behavioural semiotics: seeds for multimodal semiotic pragmatics

'The modern usage of the term **pragmatics**,' Levinson (1983: 1; bold original) observes, 'is attributable to the philosopher Charles Morris (1938).'

> By 'pragmatics' is designated the science of the relation of signs to their interpreters. [...] signs have as their interpreters *living organisms*, it is a sufficiently accurate characterization of pragmatics to say that it deals with the biotic aspects of *semiosis*, that is, with *all the psychological, biological, and sociological* phenomena which occur in *the functioning of signs.*
>
> (Morris, 1938: 30; italics added)

The italicized texts are worth stressing here, since they are the parts that are skewed in linguistics pragmatics. First Morris's 'living organisms' embraces both humans and animals, even plants included.[2] This cross-species scope of subjects is subsequently constricted to the human species only, and furthermore, it is the hearing members of the species that play a role in the conceptualization. Manual signers, Braille users – these visual and tactile users – are marginalized at the mercy of applicability of the theory conceptualized on the audial–oral modality only.

Second, Morris's 'all the psychological, biological, and sociological' is, however, found too wide to some philosophers and linguists, 'especially within analytical philosophy, the term *pragmatics* was subject to a successive narrowing of scope' (Levinson, 1983: 2; italics origin). The mainstream pragmatics driven by analytic philosophy is mono-modal in the sense that its conceptualization as well as practice is based on (1) viewing language as an object instead of viewing it as lived experience and (2) reduction of the language object to written form for visual access. Recall that even recorded live speech is transcribed into static symbols for the eyes to look at. In terms of multimodality, the original oral-auditory modality is transformed into visual modality. 'It is instructive,' Locke (2011: 6) observes, 'to consider the things that readers do not encounter on the printed page. They see no prosody, no voice quality, no tone of voice, no rate of speaking, no loudness, no vocal pitch, and no formant structure.'

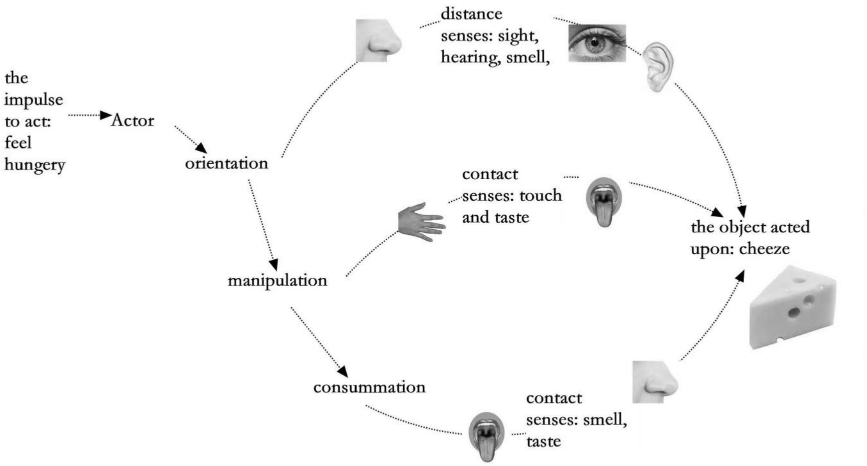
It is instructive to review the pragmatics envisaged by Morris in his series of works (1993 [1925], 1951 [1938], 1955 [1946], 1962, 1964; for a more detailed review, see Gu, 2019). Sign behaviour is the kernel of his theory building. It is rooted in Mead's theory of social behaviourism, which is distinctively different from behaviourism as substantiated in Watson (1998 [1924]) and Skinner (2005). Among the various differences, first, mind, to Mead (1962), was not to be reduced to non-mental behaviour, as Watson and Skinner did to it, but to be seen as a type of behavior genetically emerging out of non-mental types. Behaviorism accordingly meant for Mead not the denial of the private nor the neglect of consciousness, but the approach to *all experience in terms of conduct*. (Morris, 1962: xvii; italics added).

Second, human interaction with its environment is not like a 'puppet, whose wires are pulled by the physical environment' as assumed in Watson and Skinner, but is dynamic in that only those aspects of the world become stimuli when they affect the release of an 'ongoing impulse' (Morris, 1962: xvii).

Animals and humans both have impulses, which trigger behaviour of various kinds, of which sign behaviour is primary. In other words, sign behaviour is not the privilege endowed exclusively to the human species. Animals also are capable of engaging in sign behaviour. For instance, the impulse of feeling hungry is universal in the animal kingdom. It universally triggers the sign behaviour of searching and locating food. What makes the human species qualitatively different is the transition from impulse to rationality, the transition being inconceivable without the aid of human language, the most complicated form of sign behaviour ever seen.

From an impulse to its satisfaction is there an action process, which according to Mead (1962) displays a general pattern of three phases: (1) orientation, (2) manipulation, and (3) consummation. These serve as the bedrock for Morris to formulate his theory of sign behaviour. Let us continue the hunger impulse as an example. The actor, triggered by the impulse, launches the orientation phase of searching for food. There are a range of possibilities, one of which is a stimulus occurring in the environment and perceived by distance senses. Now the olfactory sense organ (=nose) registers an odour, which the brain-mind processes it as the odour of, say, cheese, an 'interpretant' in Peirce's terminology; thus, a sign is generated that, in the absence of an impulse-satisfying object, causes in an organism a disposition to a sequence of responses of the same type that the object itself would cause. The second phase of manipulation, following the disposition, involves contact senses, which are also the modalities invoked in the final consummation phase. The impulse-satisfying object (say, cheese) is the denotatum of the sign. It is worth stressing here that the properties of the sign denotatum – the cheese properties – are correlated with the perceptual sense organs and become part of the overall integrated experience of the whole semiosis, which can be graphically represented in Figure 0.1.

The essence of semiosis, as it will be shown below, lies in its conceptualization of sign behaviour as a process of living experience, unfolding over a here-and-now space–time. The role multimodality plays in such semiosis is twofold. First, it provides living experience with perceptually multimodal input and output. Second, it facilitates the construction of the subject's



the impulse to act: feel hungery
Actor
orientation
distance senses: sight, hearing, smell,
contact senses: touch and taste
manipulation
the object acted upon: cheeze
consummation
contact senses: smell, taste

Figure 0.1 Mead's action schema and Morris's adaptation in sign behaviour.

experienced environment. Such conceptualization is consistent with Uexküll‘s theory of Umwelt that is now acknowledged as one of the founding bedrocks of biosemiotics. As mentioned above, Morris’s behavioural semiotics is meant to embrace the sign behaviours of both humans and animals. So it is quite appropriate for us to demonstrate Morris’s semiosis with animal studies by Uexküll (2010 [1934]), although the two giant semioticians were unaware of each other’s works. The case to be cited here is Uexküll‘s famous study of tick in search of its prey in the wild.

Tick, according to Uexküll (2010 [1934]: 44–45), is an eyeless creature, but with general sensitivity to light in the skin; it is deaf; it has no sense of taste, and it becomes aware of the approach of its prey through the sense of smell. An adult female tick hangs inert on the tip of a branch in a forest clearing. Its position allows it to fall onto a mammal running past. Once a mammal happens to pass by, its skin glands send out butyric acid, which acts as a stimulus to the hanging tick, whose smell sense organ picks it up as a perception sign, and acts on it by dropping itself down to hit the mammal’s hairy warm skin. The temperature signals to the subject that it has spotted the right prey. The tick then uses ‘its sense of touch to find a spot as free of hair as possible in order to bore past its own head into the skin tissue of the prey.’ For the sake of comparison, a graphic is also drawn to illustrate Uexküll’s narrative (see Figure 0.2).

The similarity between the tick’s meaning-making behaviour and the sign behaviour of the hunger-cheese scenario is quite apparent. It is important to note that the tick‘s meaning-making behaviour is simultaneously both

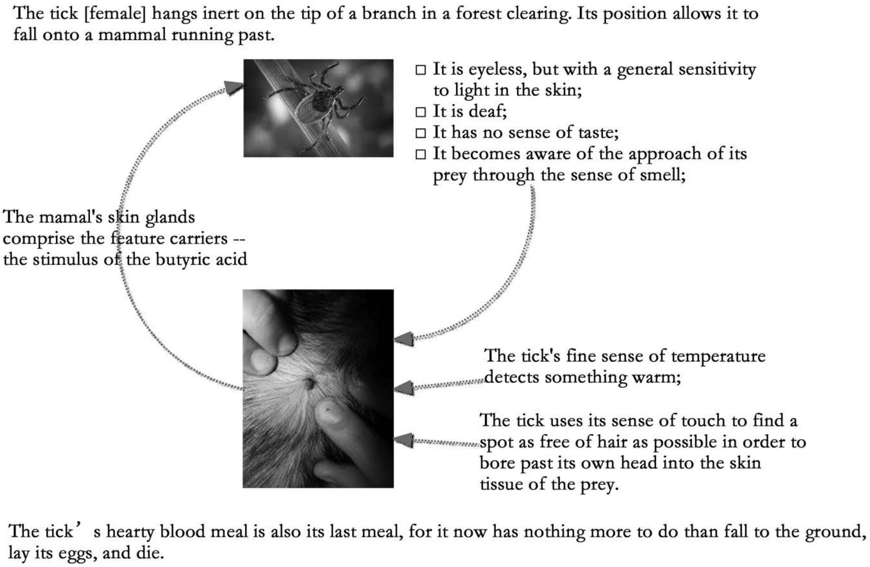

The tick [female] hangs inert on the tip of a branch in a forest clearing. Its position allows it to fall onto a mammal running past.
□ It is eyeless, but with a general sensitivity to light in the skin;
□ It is deaf;
□ It has no sense of taste;
□ It becomes aware of the approach of its prey through the sense of smell;
The mamal's skin glands comprise the feature carriers -- the stimulus of the butyric acid
The tick's fine sense of temperature detects something warm;
The tick uses its sense of touch to find a spot as free of hair as possible in order to bore past its own head into the skin tissue of the prey.
The tick’ s hearty blood meal is also its last meal, for it now has nothing more to do than fall to the ground, lay its eggs, and die.

Figure 0.2 Uexküll’s tick’s meaning-making.

framed and enabled by its perceptual sense organs. It is framed, for example, by it being eyeless, deaf, and tasteless. In other words, there are no such things as image, sound, or sweetness in the tick's experienced environment, i.e., the tick's Umwelt, in spite of the fact that the surrounding may be infiltrated with such things. The tick, however, is enabled by its skin's general sensitivity to light and its senses of smell and temperature. These modalities facilitate the construction of the tick's experienced environment in which it lives.

Semiosis is hierarchically organized in view of the complexity of sign behaviour. The semiosis, i.e., patterns of sign behaviour shared among living organisms demonstrated above, is referred to as *primary here-and-now semiosis*, to be contrasted with semiosis that transcends here-and-now space–time (see further discussion in Section 4). It is characterized by the fact that the subject's living experience and environment are constructed via multimodal interactions within a physical surrounding and that the subject's existence of living depends on a successful and continuous flow of such multimodal interactions. In an extraordinary case, the tick hanging above had waited for its prey to pass by for 18 years, during which time the tick had been kept in starvation (see Uexküll, 2010 [1934]: 52).

For ease of reference, we have, in the section title, 'smuggled' in the term 'multimodal semiotic pragmatics,' as a convenient label to refer to Morris's theorization of the primary here-and-now semiosis. When Morris deals with sign behaviour of language, he takes it for granted that verbal sign behaviour builds on the primary here-and-now semiosis. Deely (2001: 7–9) holds that animals remain in, and cannot transcend their simple Umwelts, whereas humans construct a 'linguistic Lebenswelt' on top of the simpler Umwelt. Deely's view is in total agreement with Morris's (see further discussion below).

3 Bates' study of developmental pragmatics: Piaget's sensorimotor intelligence

Developmental pragmatics by definition is concerned with how children acquire the use of language. Bates' series of seminal studies (collected in Bates, 1976, 1979) attempts to 'provide a broad ontogenetic view of the acquisition of pragmatics.' The sampling age period ranges between 9 and 13 months, 'a critical period in the emergence of communicative intentions, conventional signaling, and the idea that things have names' (Bates, 1979: 315). Intention, communication, convention, and name-object reference are traditional themes of linguistics pragmatics. Bates' approach to them is distinctive in two ways — incorporation of Peirce's semiotics (specifically Peirce's tripartite distinction of sign, see below) and Piaget's theory of sensorimotor intelligence (i.e., stage theory, see below) — lending support to multimodal semiotic pragmatics.

Bates (1976: 2) mentions Morris for his 'most widely cited definition of pragmatics' as a study of 'the relations between signs and their human users.' The definition is found to be flawed with 'some weaknesses,' for it 'misses the epistemological distinction between content and use, the psychological difference between objects and procedures.' Bates' critique of Morris is rather hasty, offhand. Morris's work cited by Bates is *Sign, Language and Behavior* (1946 edition by Prentice-Hall), but without indicating the page where the quote is taken. Bates seems to have missed the bedrock of Morris, viz. his theory of behavioural semiotics which in turn is intimately influenced by Mead's social behaviourism. When Morris draws the famous tripartite distinctions of syntactics (syntax), semantics, and pragmatics, he is not dividing 'linguistic science into three areas' as claimed by Bates (1976: 2). Morris is dealing with the division of labour in behavioural semiotics, of which linguistics is held to be a sub-part. By dismissing Morris, Bates adopts Peirce's theory of semiotics based on three types of signs: icons, indices, and symbols. These three types of signs become important metalanguage for Bates to construct her developmental pragmatics. This move no doubt is laudable and fruitful. Peirce's semiotics is logically founded (Hoffmeyer, 2008: 20–23),[3] whereas Morris's is social behavioural. The two are actually complementary (we cannot further elaborate here on this topic).

The second distinctive feature of Bates' approach is the incorporation of Piaget's sensorimotor intelligence. Piaget regards infant growth and maturation as a process of intelligence development, which is constructed incrementally through postnatal experience. He argues that development is structured with distinctive landmarks or stages. There are three major stages, each of which allows for further fine-grained identification of sub-stages. The first, being most relevant to discussion here, is the sensorimotor intelligence (from birth up to one and one-half to two years, see Piaget, 1971: 17). Piaget (1953) draws the following fine-grained distinctions:

Elementary sensorimotor adaptations

The first stage: the use of reflexes
The second stage: the first acquired adaptations and the primary circular reaction
The intentional sensorimotor adaptations
The third stage: The secondary circular reactions and the procedures destined to make interesting sights last
The fourth stage: the coordination of the secondary schemata and their application to new situations
The fifth stage: the tertiary circular reaction and the discovery of new means through active experimentation
The sixth stage: the invention of new means through mental combinations

Bates's study of infant pragmatic acquisition, in theory formation, embraces the whole range from 0 to 18 months, while in empirical investigation, her data set (9–13 months) only covers Piaget's Stages 4–6, the choice of which, as pointed out above, is believed to be associated with the 'dawn of language.'

What is the 'pragmatics' that infant babblers attempt to acquire during the sensorimotor period in preparation for late speech development? Bates and her associates focus mainly on the 'three major aspects of pragmatics – performatives, propositions, and presuppositions' (Bates, 1976: 113). As we all know, these pragmatic building concepts are all formulated on the basis of the full-fledged cognitively mature adult with a philosophically complicated mind. In what way is it justifiable to apply them to immature babblers? Bates's ingenuity lies in her interpretation of these concepts in *procedural terms*. Take performative for example. 'The term **performative** describes the organization of the child's communicative goal, e.g. to obtain an object through use of an adult, or to obtain adult attention through the use of an object' (1976: 113; bold original). Carlotta and Marta, Bates's two infant subjects, were found to have constructed such 'performative without words' prior to their referential use of words.

> We are led to the tentative conclusion that the *sensorimotor performative* is based on the cognitive developments of Stage 5, while the use of words with referents in such sequences is dependent upon the capacity for internal representation characteristic of Stage 6.
>
> (Bates, 1976: 77; italics added).

The sensorimotor performative is, in our view, the subject matter proper of multimodal semiotic pragmatics. Before looking at it more closely, we need to review Piaget's conceptualization of 'sensorimotor activity schema' that underpins Bates's sensorimotor performative and arguably lays the foundation for multimodal semiotic pragmatics.

The newborn's first cry declares the beginning of its exploration of the new world. As far as multimodality is concerned, embryological studies find that the somesthetic system (kinesthetic and cutaneous processes) is the earliest sensory system to develop in the human embryo (Stack, 2001: 351). The first indication of the developing ear can be found in embryos of approximately 22 days (Sadler, 2012: 321), and the auditory system becomes function to some extent by the sixth month (Fernald, 2001: 41). Karmiloff and Karmiloff-Smith remark:

> From the sixth month of gestation onward, the fetus spends most of its waking time processing these very special linguistic sounds, growing familiar with the unique qualities of its mother's voice and of the language or languages that she speaks.
>
> (2001: 1)

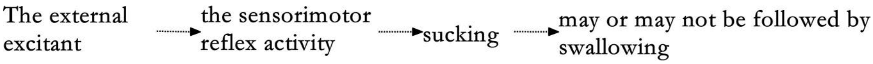

The external excitant
the sensorimotor reflex activity
sucking
may or may not be followed by swallowing

Figure 0.3 The global sucking reflex schema.

As for sight, at the time of normal birth, 'the peripheral retina of the eye is quite well developed, but the central retina … is poorly developed and undergoes considerable post-term changes' (Slate, 2001: 8). Unlike hearing, visual experience is impossible prior to birth. 'It is therefore not surprising to find that the visual world of the newborn infant is quite different from that of the adult.'

It is apparent now that the newborn's sensory modalities at birth are unevenly developed. Bearing this in mind, we are interested in how the newborn, equipped with immature, but growing multimodality, explores the new world through activities, such as sucking, kicking, touching, grasping, holding, and reaching, to name but only a few, which in Morris's terminology, are sign behaviours or to adopt Piaget's fine-grained terminology, *sensorimotor activities* or *behaviours*. It is important to note that sensorimotor behaviour is no less pragmatics than verbal behaviour, since it also involves a triadic relation between the subject, the sign, and the object. (Reminder in order here: *Symbol* in Piaget's works is the genus, while *sign* is the species. In Peirce and Morris, the usage of the two terms is reversed.)

Let us take the sucking reflex behaviour for example. It involves, among other things, the sensory tactile stimulus as 'external excitant' (Piaget's term), and the motor responses of mouth and tongue movements. The sucking reflexes show a behavioural pattern that Piaget calls the sucking reflex schema. The sucking reflex in the first instance may be activated to function when the newborn rubs the lips with its own hand, or when the mother's breast touches the lips. It can even be set in motion when the lips are touched by a cloth or an object like that. This 'global' sucking reflex schema can be captured graphically as shown in Figure 0.3.

Piaget characterizes this global schema as 'generalizing assimilation,' i.e., incorporating increasingly varied objects into the reflex schema, such as sucking his finger, mother's breast, pillow, quilt, and bedclothes. The newborn of course quickly updates the schema by 'recognitory assimilation,' i.e., by differentiating the nipple from non-nipple objects. The differentiation is associated with swallowing and satisfaction (i.e., 'consummation' in Mead and Morris). Persistent failures in achieving satisfaction lead to crying or rejection, which in turn result in the end of the reflex behaviour — when this happens, accommodation has taken place! The updated sucking reflex schema is shown in Figure 0.4.

What do we learn from this analysis? It is one of Piaget's fundamental tenets regarding the child's development, namely that *all mental objects are constructed from the child's sensorimotor activities upon the external world.*

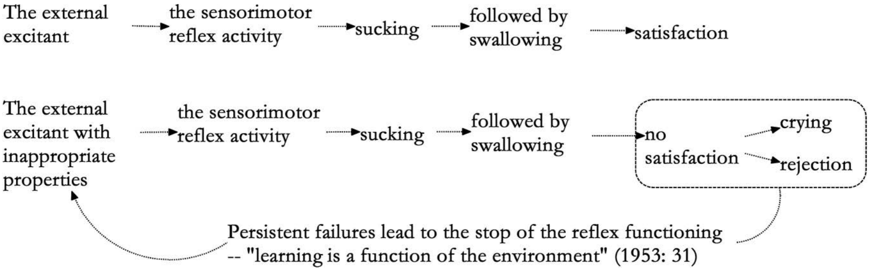

The external excitant
the sensorimotor reflex activity
sucking
followed by swallowing
satisfaction
The external excitant with inappropriate properties
the sensorimotor reflex activity
sucking
followed by swallowing
no satisfaction
crying
rejection
Persistent failures lead to the stop of the reflex functioning -- "learning is a function of the environment" (1953: 31)

Figure 0.4 The sucking reflex schema updated.

Real-world objects do not copy themselves onto the passive child. Rather, the child explores the world of objects through varied sensorimotor activities empowered by the framing and enabling, but maturing hence dynamic, sensorimotor capabilities. During this exploration, the child tries to impose its own 'sensorimotor schema' on the world (i.e., assimilatory adaptation) and revises those schemata when the child meets resistance (accommodatory adaptation).

Our analysis of Morris and Uexküll above has rendered it obvious that this fundamental tenet is shared by Morris's behaviour semiotics and Uexküll's biosemiotics. The infant constructs its Umwelt (i.e., experienced environment) through its sensorimotor activities (i.e., sign behaviour or meaning-making behaviour). The infant's Umwelt, due to its immature multimodality, is different from that of adults. What makes a human infant different from, say, a tick's baby, is that, as far as multimodality is concerned, the former's sensorimotor activity schema is highly plastic, capable of both assimilatory and accommodatory adaptations, whereas the latter's capacity is genetically fixed.

Now let us return to Bates. Bates's developmental pragmatics does not take the infant's language as something innately endowed, as assumed in the Chomskian paradigm, according to which all the postnatal experience does is simply 'trigger' the endowed I-language to grow (Cook and Newson, 2000: 106).[4] Bates, in contrast, adopts Piaget's epigenesis and shows that pre-speech sensorimotor activities lay a foundation for the development of language use. Bates's developmental pragmatics is rich, and full justice of her works has to be found elsewhere (see Tomasello and Slobin, 2005).

4 The primary here-and-now semiosis: total saturated experience with total saturated signification

The hunger-cheese sign behaviour from orientation to manipulation to consummation results in what Gu (2009b) proposes to call total saturated experience (TSE) with total saturated signification (TSS). Gu uses the scenario of enjoying the roast Peking duck to illustrate the concept. Given the goal, there are a range of possibilities: (1) going to a Chinese restaurant and eating a real roast Peking duck; (2) watching others eating it; (3) watching a video show of how people enjoy eating it; (4) listening to a talk about how a roast Peking duck is being made; and (5) reading a recipe about how to make a roast Peking duck. The first one in comparison with the remaining represents a total saturated experience of duck-eating. Its total saturated signification is typically associated with the qualities extracted from multimodal interactions with the real object, e.g. qualities such as its taste, colour, odour, crispiness, tenderness, and emotional states triggered by them.

In the primary here-and-now semiosis where the TSE-TSS is attained, the subject–object relation, in the subject's experienced Umwelt, is, first of all, a *consummatory relation* established through contact senses (taste, touch, etc.). In the Chinese theory of cuisine, sight and smell, Morris's distance senses are also regarded as contributing to the consummatory relation. The reference relation traditionally held in linguistics pragmatics between the subject and object is obviously inadequate. One can testify this by the fact that one can never enjoy eating duck by simply referring to or pointing at it!

The primary here-and-now semiosis characteristic of TSE-TSS has developmental significance at the individual personal level and at the national social level. Human languages in view of TSE-TSS experience develop unevenly. Each language is like what Wittgenstein calls an ancient city (1997 [1953]: 5^e), with old streets, lanes as well as new buildings, CBDs, etc. Take China for example. There are about 150 or more languages. Mandarin Chinese in comparison offers more modes of experience than the remaining ones. The primary here-and-now semiosis with TSE-TSS, i.e., multimodal semiotic pragmatics, is like the basement floor, on top of which there is what Gu (2009a) calls the land-borne situated discourse (LBSD), i.e., the sensorimotor-constructed Umwelt plus linguistic Lebenswelt – is the oldest in history, but the most dynamic, fleeting, saturated mode of experience. Mandarin Chinese (an inaccurate but convenient label here) over 3,000 years ago, on the other hand, invented writing script, thus creating another mode of experience, i.e., experience based on visual interpretation of written sign vehicles. This is the Written Word-borne discourse (WWBD). Of 150 or so languages, only a minority provides such a mode of experience for the users. Mandarin Chinese is further privileged to be able to provide two more modes of experience: (1) telephone calls, radio broadcasts, TV programmes, etc., that is, air-borne situated discourse (ABSD) for oral-auditory as well as visual consumption, and (2) the latest Web-borne situated discourse (WBSD), which transcends totally the primary here-and-now live semiosis into the domain of virtual reality.

Now it is beneficial to revisit Morris's definition of pragmatics cited above. Pragmatics is conceptualized to deal with the biotic aspects of semiosis encompassing *all the psychological, biological, and sociological* phenomena occurring in the functioning of signs. We have used two concepts

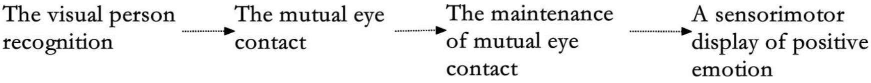

The visual person recognition
The mutual eye contact
The maintenance of mutual eye contact
A sensorimotor display of positive emotion

Figure 0.5 Sensorimotor schema of greeting.[5]

multimodality and *sign behaviour* as two exploratory probes to fathom the pragmatics initially envisaged by Morris. In this sub-section, we would like to highlight the interactions of sensorimotor mechanism, which are biological in nature, with the psychological and sociological via the subject's live sign behaviour. Take greeting for example. Greeting in the mode of primary here-and-now semiosis, if fully performed, is a total saturated experience with total saturated signification. The TSE-TSS greeting simultaneously displays multiple layered properties: (1) the well-integrated sensorimotor activity schema; (2) the gestural or verbal schema; (3) the emotional companionship; and (4) the sociocultural appropriateness. The two to four layers hardly need elaboration here. The first covers a complex phenomenon of bio-psychological nature. A well-integrated sensorimotor activity schema can be graphically shown in Figure 0.5.

The sensorimotor display of positive emotion is the externalization of the current thought and psychological emotion (e.g. feeling happy to see the greeting) that accompany the gestural and/or verbal behaviour of greeting. Gu (2013b) labels it as *mao* (貌, i.e., the *embodiment*) of thought and emotion, which is observable in facial expression, bodily posture, prosody of speech, etc. The four components are normally sequentially organized and well-coordinated. Malfunctions or hiccups will result, with everything else being equal, in failure.

Why bother about sensorimotor activity schema and well-integratedness? Does it have anything to do with pragmatics? Gu (2013b) shows that incongruence between what is said and what is emotionally displayed also triggers implicatures. Children with autism spectrum disorders (ASD) are particularly compromised in reaching the mutual eye contact, still less to maintain it for long. No better justification is given than those studies of pragmatic impairments provide, to which we turn.

6 Pragmatic impairment: Perkin's study

Apart from its bias on modelling the mature adult as pointed out above, the current mainstream pragmatics suffers from another limitation, namely that it is based on the assumption that linguistic communication 'typically appears to be a single, seamless process' (Perkins, 2007: 8). Little consideration is given to communication disorders in its theory formation. Moreover,

> Unlike clinicians, who need to understand a condition in its entirety in order to play appropriate intervention, pragmatic theorists have had the

> luxury of being able to focus only on the specific features which are of interest to them.
>
> (Perkins, 2007: 8)

When the current theories are applied to the clinical context, 'it is not always well suited to the needs of language pathologists and has led to a great deal of confusion in the clinical diagnosis of pragmatic impairment, and in regard to the nature of pragmatic impairment itself' (Perkins, 2007: 8).

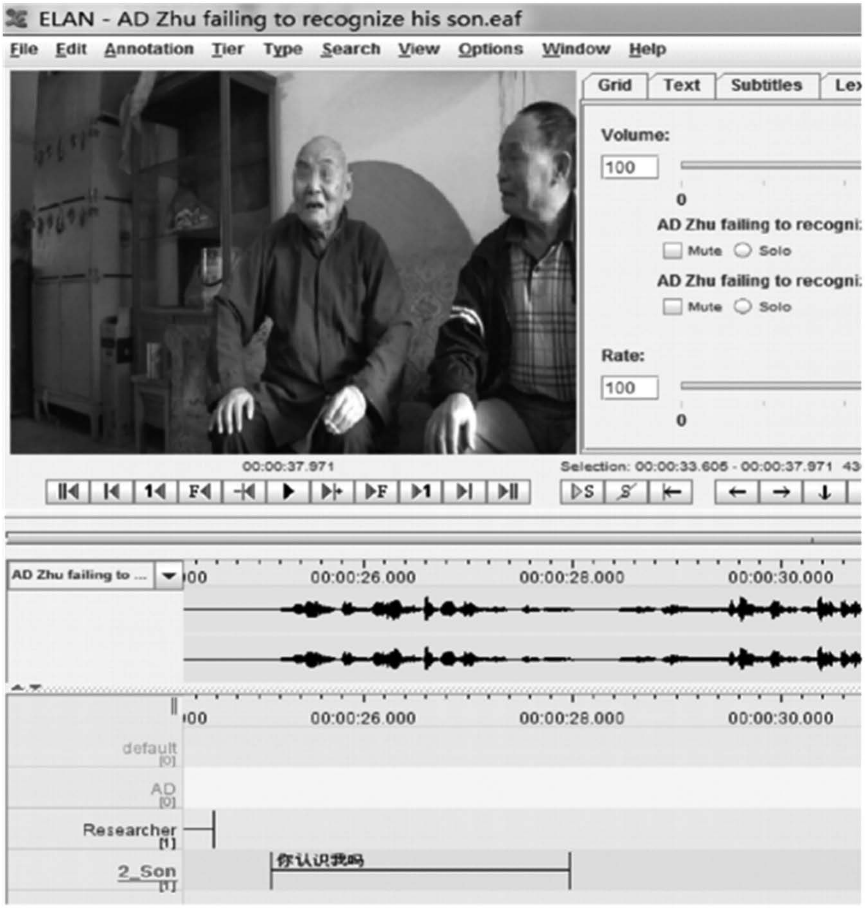
Hence, Perkins argues for a holistic model of pragmatics, namely 'emergentist pragmatics' (EP). Its holistic approach is particularly seen in the inclusion of 'sensorimotor systems' as one of the three 'elements' of theory construction (the other two being *cognitive systems* and *semiotic systems*). Both theoretical and practical aspects are compelling for Perkins to make this inclusion.

> Apart from obvious examples such as the use of gesture to compensate for linguistic output problems and the use of facial expression and tone of voice to interpret the attitudinal or emotional state of a speaker during comprehension, sensory input and motor output systems are rarely included in discussions of pragmatic ability and disability. However, once pragmatic functioning is seen as an emergent phenomenon it is clear that sensorimotor systems provide a range of communicative choices in the same way that cognitive and linguistic systems do; that restriction in choice as a result of impairment is pragmatically constraining and can have a knock-on effect both within the sensorimotor domain and in cognitive and linguistic domains; and that sensorimotor systems are as vulnerable as language and cognition to the effects of compensatory adaptation during interpersonal communication.
>
> (Perkins, 2007: 139)

The most devastating knock-on effect of sensorimotor impairment on pragmatics (and language in general), perhaps, is the loss of vision (acquired blindness) and hearing (acquired deaf). Alternative modality compensation has to be attained, e.g. tactile modality for sign behaviour and communication.

Pragmatic impairment due to malfunctions of sensorimotor systems presupposes knowledge about the system's normal functioning, i.e., knowledge of what Gu (2007, 2015) calls 'multimodal congruence.' Perkins touches upon the same phenomenon when he points out that visual and auditory perceptions play a key role in inferential processing, and misreading of facial expression or voice quality could result in failure to detect irony. Moreover, the 'expression of emotion and attitude is particularly multimodal, with meaning being conveyed via articulation, voice quality, prosody, facial expression, gesture, posture and gaze' (Perkins, 2007: 140).

As Perkins correctly points out, the coordination and integration of sensorimotor systems with cognitive systems and semiotic systems 'is a relatively unexplored area' (Perkins, 2007: 139). There are many hard and pressing



Volume: 100

Rate: 100

Selection: 00:00:33.605 - 00:00:37.971 43

2_Son: 你认识我吗

Figure 0.6 Visual agnosia of an Alzheimer's disease patient.

problems calling for solutions. Here is an instance taking from our corpus of Alzheimer's disease patients. As shown in the screenshot (see Figure 0.6), the left is the AD patient, and the right is his second son. There is a third person, i.e., the researcher, not shown in the picture. The conversation goes as follows:

> Researcher: (to the AD, while pointing at the son) Do you recognize him?
> AD: (Turn to look at his son) er …
> Son: (to the AD, while pointing at himself) Do you recognize me?
> AD: (Staring at his son) er …

Pragmatically speaking, the impairment can be categorized as person recognition impairment, or visual agnosia. But what is the cause of this impairment? Is it associated with the condition of cognitive memory retrieval? Or is it due to the compromise of associated visual neural pathways? Even if definite answers are found, clinicians are pressed by the patient for a therapeutic solution, which is yet another hard problem currently unsolvable.

7 Multimodal semiotic pragmatics: future directions for Morris's vision

It is worth reminding the fact that the term multimodal *semiotic pragmatics* does not refer to a mature established theory, out there, ready to pick up and use! It only points to a research direction initially envisaged by Morris, toward which other researchers, aware or unaware of Morris's works, happen to converge.

Multimodal semiotic pragmatics, as reviewed above, clearly holds a complementary relation with the current mainstream linguistics pragmatics. That is, it does not, and cannot replace the existing theories. Conversely, linguistics pragmatics does not, and cannot replace multimodal semiotic pragmatics either. Having said this, multimodal semiotic pragmatics seems to be 'more basic,' that is, it can serve as the 'ground floor' on which linguistics pragmatics is to be more securely situated. This is the direction toward which Bates's developmental pragmatics, Perkins's emergentist pragmatics, and Gu's study of four-borne discourses seem to lead us.

There are of course many research issues associated with multimodal semiotic pragmatics itself proper. As pointed out above, Morris would intend multimodal semiotic pragmatics (i.e., primary here-and-now semiosis) to include both animals and humans. So logically there will be multimodal semiotic pragmatics for animal sign behaviours. Very interestingly, the latest growing interest in biosemiotics (Barbieri, 2007; Hoffmeyer, 1996; Romanini and Fernandez, 2014) seems to move toward this direction. Biosemiotics also recognizes sign behaviour and sensorimotor mechanism as two domains of properties common to living organisms. Adopting these two as the point of departure to conceptualize pragmatics, the outcome is not just an issue of being broader or narrower but brings to the fore the fundamental question of where language comes from. The issue of evolution of language has been literally 'out of question,' for the mainstream linguistics pragmatics takes language as given, and pragmatics's business proper is to see how language is used in communication. This is particularly transparent when a componential view of pragmatics is adopted, with a division of labour between syntax, semantics, and pragmatics.

Ironically, the componential view is believed to have originated in Morris. However, a close reading of Morris will show that Morris is not at all responsible for such division of labour in linguistics pragmatics. This is because Morris, drawing the tripartite division, is concerned with behavioural semiotics, not linguistics! Moreover, 'syntactics, semantics, and pragmatics,' are 'subordinate branches' 'dealing, respectively, with the syntactical, the semantical, and the pragmatical *dimensions* of semiosis' (1951 [1938]: 8; italics added). Morris is not recommending dividing semiosis into three parts, but rather to look at it from three perspectives, as he takes it for granted that 'the various dimensions are *only aspects of a unitary process*' (1951 [1938]: 8; italics added). Later on, he emphasizes that, 'after making use of the abstractions involved in this treatment, we will specifically *stress the unity of semiotic*' (1951 [1938]: 13; italics added).

Of the three dimensions abstracted from semiosis, the pragmatic is, in actual use, primary and fundamental, this is because it operates at the level of behaviour making something become a sign in the first place. This is implicated in the remarks Morris makes:

> Syntactical rules determine the sign relations between sign vehicles; semantical rules correlate sign vehicles with other objects; pragmatical rules state the conditions in the interpreters under which the sign vehicle is a sign. Any rule when actually in use operates as a type of behavior, and in this sense there is a pragmatical component in all rules.
> (1951 [1938]: 35)

Put it differently, there is a sequential order implicit in semiosis, namely syntactics and semantics will have nothing to do until a sign is made, and it is the sign behaviour that fixes a sign vehicle and the object the sign is made to stand for.

Multimodal semiotic pragmatics does not take language for granted. Verbal behaviour is a species of sign behaviour. In the time of evolution, humans' verbal behaviour is a later comer. Stokoe observes:

> When gesture is defined as body movement that communicates more or less consciously, and Sign is taken as a generic term for natural sign languages, the progression from gesture to Sign is entirely natural. What makes it so is the nature of the hominid phenotype and its attributes – its vision and the physical structure and use of hands, arms, faces, and bodies.
> (Stokoe, 2000: 388)

If one accepts Stokoe's position, multimodal semiotic pragmatics will also have a role to play in the story of language evolution.

Finally, it is worth emphasizing the fact that Morris's conceptualization of semiotics is intended to have a unifying function:

> It is doubtful if signs have ever before been so vigorously studied by so many persons and from so many points of view. The army of investigators includes linguists, logicians, philosophers, psychologists, biologists, anthropologists, psychopathologists, aestheticians, and sociologists. There is lacking, however, a theoretical structure simple in outline and yet comprehensive enough to embrace the results obtained

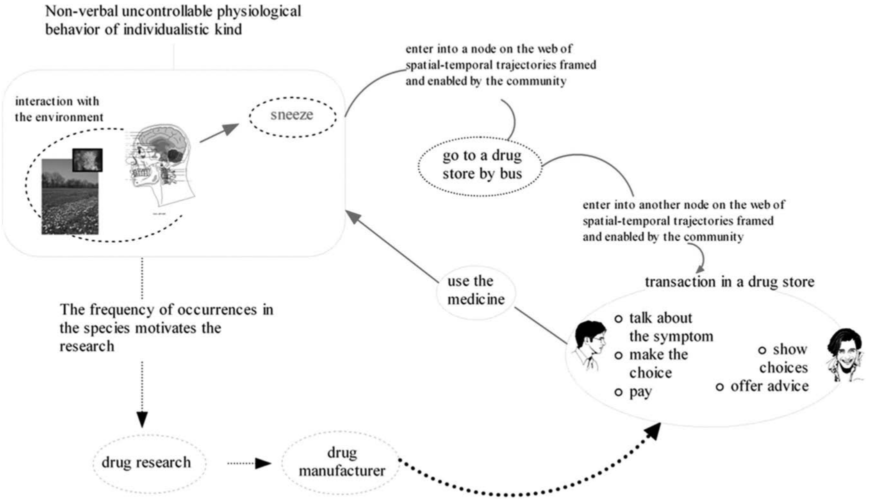

Non-verbal uncontrollable physiological behavior of individualistic kind
interaction with the environment
sneeze
enter into a node on the web of spatial-temporal trajectories framed and enabled by the community
go to a drug store by bus
enter into another node on the web of spatial-temporal trajectories framed and enabled by the community
transaction in a drug store
use the medicine
talk about the symptom
make the choice
pay
show choices
offer advice
The frequency of occurrences in the species motivates the research
drug research
drug manufacturer

Figure 0.7 An ecological chain of activities.

> from different points of view and to unite them into a unified and consistent whole. It is the purpose of the present study to suggest this unifying point of view and to sketch the contours of the science of signs.
> (1951 [1938]: 1)

Morris's 'unified and consistent whole' theory of signs is indeed solidly demanded in real-life semiosis, which can be demonstrated by what Gu (2012) calls the 'ecological chain' of activities. An individual Mr. Y suffers from hay fever as a result of his interaction with the physical environment. He sneezes nonstop. He takes a bus to go to a drug store – he enters into a node on the web of spatial-temporal trajectories framed and enabled by the community. In the drug store, i.e., another node on the web of spatial-temporal trajectories, he talks about his hay fever to a girl assistant, who shows him a few choices and offers him some advice. He makes a choice and pays. This whole transaction would have been impossible without a drug manufacturer producing the drug. The latter on the other hand would never have been produced without research on the drug. The drug research, in turn, is motivated by the fact that Mr. Y is not alone who sneezes at the exposure of flowering plant pollens. The whole series of activities, each seeming to happen separately and independently, actually form a coherent whole, as shown in Figure 0.7 (quoted from Gu, 2012: 550).

It would be no small achievement to account for this ecological chain in terms of the unifying metalanguage of behavioural semiotics as envisaged by the founding father!

Notes

[1] This foreword was an original paper in the book "Gerontolinguistics and Multimodal Studies" (Shanghai: Tongji University Press, 2020), edited by Yueguo Gu and Lihe Huang. The author and the press authorize the reprint.

[2] Nowadays, there are branches of semiotics for humans (i.e., anthroposemiotics), for animals (i.e., zoosemiotics), for plants (i.e., phytosemiosis), and even for physical objects (i.e., physiosemiosis). See Deely (1990) for introductory treatment.

[3] Peirce writes (1955: 98): "Logic, in its general sense, is, as I believe I have shown, only another name for semiotic … the quasi-necessary, or formal, doctrine of signs."

[4] Chomsky does not use 'trigger,' but verbs such as convert, map, select, etc., e.g. 'the innate component of the mind/brain that yields knowledge of language when presented with linguistic experience, that converts experience to a system of knowledge' (Chomsky, 1986: xxvi)

[5] The visual person recognition itself can be sufficient for the initiation of greeting. In this case, the greeting behaviour first of all functions as an attention attractor. It is not a proper greeting.

REFERENCES

Barbieri, Marcello, ed. (2007). *Introduction to biosemiotics*[M]. Berlin: Springer.

Bates, Elizabeth. (1976). *Language and context: The acquisition of pragmatics*[M]. New York: Academic Press.

Bates, Elizabeth. (1979). *The emergence of symbols*: *Cognition and communication in infancy*[M]. New York: Academic Press.

Bremner, Gavin & Fogel, Alan, eds. (2001). *Blackwell handbook of infant development*[M]. Oxford: Blackwell.

Chomsky, Noam (1986). *Knowledge of language: Its nature, origin, and use*[M]. Westport, CT: Praeger.

Cook, Vivian, & Newson, Mark. (2000). *Chomsky's universal grammar: An introduction*[M]. Beijing: Foreign Language Teaching and Research Press.

Deely, John. (1990). *Basics of semiotics*[M]. Bloomington: Indiana University Press.

Deely, John. (2001). *Four ages of understanding*[M]. Toronto: University of Toronto Press.

Fernald, Anne. (2001). Hearing, listening, and understanding: Auditory development in infancy. In Gavin Bremner & Alan Fogel (eds.), *Blackwell handbook of infant development* [C] (pp. 35–70). Blackwell Publishing Ltd.

Gu, Yueguo. (2007). Multimedia, multimodality and learning[J]. *Computer Assisted Foreign Language Learning, 114*, 3–12.

Gu, Yueguo. (2009a). Four-borne discourses: Towards language as a multi-dimensional city of history. In Li Wei & Vivian Cook (eds.), *Linguistics in the real world* [C] (pp. 98–121). London: Continuum.

Gu, Yueguo. (2009b). From the real-life situation to video stream data-mining[J]. *International Journal of Corpus Linguistics*, *14*(4), 433–466.

Gu, Yueguo. (2012). Discourse geography. In James Paul Gee & Michael Hanford (eds.), *The Routledge handbook of discourse analysis*[C] (pp. 541–557). London: Routledge.

Gu, Yueguo. (2013). The STFE principle in situated discourse: A multimodal corpus linguistics approach[J]. *Contemporary Rhetoric*, *6,* 1–19.

Gu, Yueguo. (2015). Multimodality and language[J]. *Contemporary Linguistics*, *4*, 448–469.

Gu, Yueguo. (2019). Morris' lost pragmatics: A plea for multimodal semiotic pragmatics[J]. *Chinese Semiotic Studies, 15*(2), 217–242.

Hoffmeyer, Jesper. (1996). *Signs of meaning in the universe*[M]. Bloomington: Indiana University Press.

Hoffmeyer, Jesper. (2008). *Biosemiotics: An examination into the signs of life and the life of signs*[M]. Scranton: University of Scranton Press.

Karmiloff, Kyra, & Karmiloff-Smith, Annette. (2001). *Pathways to language: From fetus to adolescent*[M]. Cambridge, MA: Harvard University Press.

Kolb, Bryan, & Whishaw, Ian Q. (2005). *An introduction to brain and behavior* (2nd ed.) [M]. New York: Worth Publishers.

Kress, Gunther. (2009). What is mode? In Carey Jewitt (ed.), *The Routledge handbook of multimodal analysis*[C] (pp. 54–67). London: Routledge.

Kress, Gunther. (2010). *Multimodality: A social semiotic approach to contemporary communication*[M]. London: Routledge.

Levinson, Stephen C. 1983). *Pragmatics*[M]. Cambridge: Cambridge University Press.

Mead, George H. (1962). *Mind, self, and society: from the standpoint of a social behaviorist* [M]. Charles W. Morris (ed.). Chicago. IL: The University of Chicago Press.

Morris, Charles W. (1951) [1938]. *Foundations of the theory of signs*[M]. Chicago, IL: The University of Chicago Press.

Morris, Charles W. (1955) [1946]. *Signs, language and behavior*[M]. New York: George Braziller.

Morris, Charles W. (1962). Introduction: George H. Mead as social psychologist and social philosopher. In George H. Mead (ed.), *Mind, self, and society: from the standpoint of a social behaviorist* [C] (pp. ix–xxxv). Chicago, IL: The University of Chicago Press.

Morris, Charles W. (1964). *Signification and significance: A study of the relations of signs and values*[M]. Cambridge, MA: The MIT Press.

Morris, Charles W. (1993) [1925]. *Symbolism and reality: A study in the nature of mind*[M]. Amsterdam: John Benjamins.

Peirce, Charles S. (1955). *Philosophical writings of Peirce*[M]. Ed. Justus Buchler. New York: Dover.

Perkin, Michael. (2007). *Pragmatic impairment*[M]. Cambridge: Cambridge University Press.

Piaget, Jean. (1971). *Biology and knowledge: An essay on the relations between organic regulations and cognitive processes*[M]. Chicago, IL: The University of Chicago Press.

Piaget, Jean. (1972). *The principles of genetic epistemology*[M]. Transl. Wolfe Mays. London: Routledge.

Priesler, Gunilla. (2001). Sensory deficits. In Gavin Bremner & Alan Fogel (eds.), *The Blackwell handbook of infant development*[M] (pp. 617–638). West Sussex: Blackwell.

Romanini, Vinicius, & Fernandez, Eliseo, eds. (2014). *Peirce and biosemiotics*[M]. Berlin: Springer.

Sadler, T. W. (2012). *Langman's medical embryology* (12th ed.) [M]. Baltimore, MD: Lippincott Williams & Wilkins.

Skinner, B. F. (2005). *Science and human behavior*[M]. Cambridge, MA: The B. F. Skinner Foundation.

Slater, Alan. (2001). Visual perception. In Gavin Bremner & Alan Fogel (Eds.), *The Blackwell Handbook of Infant Development*[C] (pp. 5–34). West Sussex: Blackwell.

Slater, P. J. B. (1999). *Essentials of animal behaviour*[M]. Cambridge: Cambridge University Press.

Stack, Dale M. (2001). The salience of touch and physical contact during infancy: Unraveling some of the mysteries of the somesthetic sense. In Gavin Bremner & Alan Fogel (eds.), *The Blackwell handbook of infant development*[C] (pp. 351–378). West Sussex: Blackwell.

Stokoe, William C. (2000). From gesture to sign (language). In McNeill, David (ed.), *Language and gesture*[M]. Cambridge: Cambridge University Press.

Tomasello, Michael, & Slobin, Dan Isaac, eds. (2005). *Beyond nature-nurture: Essays in honor of Elizabeth Bates*[M]. Mahwah, NJ: Lawrence Erlbaum Associates.

Uexkull, Jakob von. (2010) [1934]. *A foray into the worlds of animals and humans*[M]. Transl. Joseph D. O'Neil. Minneapolis: University of Minnesota Press.

Watson, John B. (1998) [1924]. *Behaviorism*[M]. New Brunswick, NJ: Transaction.

Wittgenstein, L. (1997) [1953]. *Philosophical investigations* (2nd ed.) [M]. G. E. M. Anscombe (transl.). Oxford: Blackwell.

Preface

Developing multimodal pragmatics from speech act study

Lihe Huang

For the long-term development of linguistics, the spoken language has been studied primarily as a static object or unichannel phenomenon, i.e., just speech or text outcomes. Furthermore, constrained by research perspectives and techniques, the so-called ‘non-linguistic’ aspects of communication – including gesture, facial expression, and body movement – have been ‘justifiably’ separated from language study. This traditional view of human language leads to linguists and relevant scholars focusing only on the use of language itself, ignoring the fact that multimodal resources convey meaning. However, today’s neuropsychology informs us that humans are born multimodal. That is to say, human interaction is multimodal, and meaning in discourse is created through an interplay of an array of modalities. We combine multimodalities in input and output when experiencing the world.

To respond to this new trend of advocating for a multimodal conception of language, most scholars agree that conducting linguistic research should use multimodality as the benchmark. As such, the object of linguistic study should consider all the information expressed in both vocal and other different channels, including prosody, gesture, facial expression, and body movement, which invariably accompany verbal expression in face-to-face situations. That is to say, linguistic phenomenon and human communication mechanism studies should apply a multimodal perspective and be based on various multimodal data collected through multiple techniques.

Multimodal research is a research paradigm that integrates multiple approaches. Different research fields and approaches define and interpret ‘modality’ differently. The concept of modality has three main definitions: (1) sense organs and interconnected neural networks, (2) semiotic resources for meaning construction, and (3) the way of representing information through some physical media. Accordingly, as an integrated paradigm, the multimodal study consists of several different approaches and fields, including (1) multimodal discourse analysis rooted in semiotics, (2) multimodal corpus-based study, and (3) multimodal study in neuroscience, human–computer interaction (HCI), and learning science. These approaches have been applied to many different fields, focusing on both fundamental research and technological development. The relevant fields include multimodal semiotics, multimodal corpus linguistics, neurolinguistics, HCI, and multimodal research in learning sciences and their related domains. The multimodal paradigm is extensively used in linguistics studies, which can involve many linguistics fields, including re-examining conventional topics and classic theories from the multimodal nature of daily human communication and then verifying, modifying, and developing traditional linguistic theories.

Fruitful achievements have been made in multimodal discourse analysis, multimodal metaphor research, multimodal stylistics, and multimodal pedagogy. Multimodal research in neurolinguistics, HCI, and learning sciences spans multiple disciplines, yielding impressive results in each domain. The development of modern information technology and artificial intelligence (AI) has continuously expanded the channels and methods of information interaction between humans and the outside world. In turn, the multimodal nature of human language communication holds the key to those pilot studies. Multimodal learning research aims to explore how to make full use of multimedia materials (e.g. the internet) for learning, to verify the benefits that multimodal learning brings to the internalization of acquired knowledge and the persistence of memory and learning efficiency, and to measure the impact that wearable devices and networks information platform have made on learning methods and effects. In contrast, studies based on multimodal corpus are relatively few in China, while flourishing abroad.

Compared with the above research outcomes, multimodal corpus studies of language use in situated discourse seem inadequate. For this, this book's author decided to adopt a multimodal corpus approach to explore pragmatic issues.

Language research should take multimodal interaction between people and the outside world as an essential data source. When dealing with situated discourse with saturated significance, the multimodal corpus approach can be ideal for quantitative analysis. Multimodal corpus refers to the corpus that integrates various information, such as texts, audio, and videos. Researchers can process, retrieve, and conduct statistical analysis of the corpus data in a multimodal way. The integration of textual, audio, and video records of situated discourse in the corpus can provide a good platform for further and more profound exploration of an extensive range of linguistic phenomena in real-life interaction.

This book considers language as the multimodal manifestation of people's face-to-face interaction. Both verbal and non-verbal channels, including speech, prosody, gesture, facial expression, and body movement, should be considered when a speaker's pragmatic interaction is being studied. Speech acts, regarded as a basic human communication unit, are a classic topic in pragmatics. Studies related to speech acts have broad application prospects in HCI, AI, and so forth, becoming a focus of the current linguistics community. Therefore, it is of great significance to apply a multimodal corpus approach to examine pragmatic topics. It is conducive to expanding the scope of pragmatic research while enriching related theories. This is why the author decided to re-examine speech acts from a multimodal perspective, rather than following the traditional path in pragmatics and collecting data just from the outcomes of speech or text. It is also an attempt to develop multimodal pragmatics (initially envisaged by Morris) by integrating multimodal corpus method.

In this book, the author adopts multimodal corpus linguistics and the primary thought of Simulative Modelling and takes speech acts performed by Chinese illiterate people in situated discourse as the research object. The goal is to discover how linguistic structures, prosodic features, and gestures, which are impacted by speakers' occurrent emotions and other factors, interact with each other to produce various live illocutionary forces. The study is also designed to describe the mechanism of the emergence of different illocutionary force indicating devices (IFIDs). Additionally, possible influential factors for such emergence are also explored. The whole study is based on the self-constructed multimodal corpus of Chinese situated discourse and adopts an interdisciplinary methodology.

More specifically, this study (1) conducted a conceptual analysis of different types of illocutionary force, and their ontological properties, conditions, and presented the rules for implementation in sets; (2) performed case analyses of each corresponding illocutionary force-token regarding its specific conditions, and presented their differences and shared characteristics; (3) devised a system of resources with a detailed explanation of their functions that the speaker uses to express illocutionary forces in different situations; (4) used statistical methods to describe the interactive mechanism among emotional states, prosodic features, and gestures when the speaker performs different illocutionary forces; and (5) made speculations about the multimodal means of illocutionary forces and the possible impact of the IFIDs.

Meanwhile, this book discusses the necessity of developing multimodal pragmatics to extend the pragmatics research horizon further. It concludes that multimodal paradigms could upgrade pragmatic research methods, which is crucial for researchers in interpreting the significance of language use more comprehensively and accurately from general human behaviour. Furthermore, this book elaborates on different multimodal research approaches used in multimodal language research and other related fields. The author also envisions some possible topics for the research of multimodal corpus approach to linguistic studies, which will provide profound theoretical significance and great applicable possibilities, and should become one of the emerging fields of language research in the future. Additionally, the author would like to point out that the speakers have authorized the use of the multimodal data in this book for academic purposes, and no third-party interest is involved.

Multimodal research is a comprehensive paradigm, integrating multiple approaches to various spheres, ranging from humanities and social sciences to natural sciences and engineering technologies. To further promote the development of multimodal research, the author co-founded and now co-chairs Tongji Institute of Linguistics and Multimodality, which is one of the first independent institutions of its kind in China. This institute specializes in several sub-fields, including multimodal discourse analysis and multimodal pragmatics. As an essential research output in the book series entitled "Multimodality and Special People Speech Studies Series" initiated by the institute, this book is suitable for researchers and students who wish to carry out frontier research in multimodality, pragmatics, corpora, and other fields.

1 Some preliminary remarks

1.1 Research review of speech act theory and illocutionary force

The study of **speech acts** is a critical area of pragmatics. It is vital to carry out an intensive investigation into how speakers use various expressive devices to perform speech acts and shape illocutionary forces in situated discourse.[1] As these researches present broad multidisciplinary application prospects in HCI, AI, etc., it has become one of the key research fields in today's linguistics community.

From a diachronic perspective, the study of speech acts can be divided into three stages: (1) establishing speech act theory and early research by analytical philosophers, (2) including linguists from a purely linguistic perspective, and (3) now expanding with the corpus-based approach.

1.1.1 The establishment of speech act theory and early studies by analytical philosophers

With the development of mathematical logic and linguistics, focusing on language has become a prominent feature of Western philosophy in the first half of the 20th century. Philosophers engaged in these studies are called linguistic philosophers, and their research belongs to the philosophy of analysis. Against the background of the 'linguistic turn' in Western philosophy, Austin proposed the Speech Act Theory (Austin, 1962) in the 1950s, which originated from his exploration of three philosophical issues, including the relationship between daily language and philosophical research, the methodology of behaviour research, and the distinction of constatives and performatives. Afterward, Austin rediscovered that there was no substantial difference between constatives and performatives, since as far as the speakers are sincere when they make the utterance, they are 'doing things with words' and producing the corresponding illocutionary force.[2] After this, Austin made a new leap in his exploration of speech act theory. Accordingly, the study of speech acts gradually became one of the core fields in pragmatics.

DOI: 10.4324/9781003251774-1

Adopting the abstract method, Austin extracted three types of acts from a complete speech act: Locutionary act, illocutionary act, and perlocutionary act. In speech act theory, a locutionary act is the act of making a meaningful utterance; an illocutionary act is the act that attaches certain illocutionary force to the meaningful utterance in a specific context; and a perlocutionary act is a speech act that produces an intended effect achieved in an addressee by a speaker's utterance (Gu, 1989: 30–31). However, it is worth noting that Austin's abstract method was not to divide speech acts into three separate parts, but to examine the same thing from different dimensions or perspectives; also, the relationship between abstracted acts is not compositional but inclusive (Gu, 1989: 32). Among these three abstracted speech acts, Austin shed most light on the illocutionary act; later, his speech act theory also entirely focused around the illocutionary act. After Austin, the illocutionary act/force became the top priority in the research of speech acts, which received extensive attention and in-depth discussion by scholars.

Austin claimed that speech act verbs/phrases are important cues for decoding speech acts. In this sense, he divided verbs in natural language into speech act verbs and perlocutionary act verbs, which can be regarded as one of the bases for distinguishing the two types of speech acts. There are thousands of speech act verbs; what does it mean for people accomplishing thousands of things 'using words?' In this case, it is necessary to bring up the taxonomy of illocutionary acts.[3] Since Austin utilized speech act verbs to distinguish illocutionary act/force, he believed that consulting dictionaries was the first useful step; five categories of speech act are listed: Verdictives, exercitives, commissives, expositives, and behabitives. However, Austin's classification has some problems. Firstly, the classification blurs the differences between illocutionary force and speech act verbs, and equates the two (see Gu, 2002a: F28). This action consequently misclassified the study of illocutionary force, which should be recognized as a behavioural study, as syntactic and semantic research on speech act verbs. Evidence shows that many scholars since Austin regarded speech act verbs as the core of the study of speech acts. Undeniably, such a tendency was influenced by Austin's approach, equating the classification of illocutionary force with that of speech act verbs. Secondly, unified standards are absent in Austin's categorization. For instance, the determining criteria of expositives are based on the speaker's attitude while exercitives are based on the speaker's status, power, and identity. Also, the contents of each category are somewhat confusing and sometimes overlapping. Nevertheless, though Austin's classification of illocutionary force seemed not to be that successful, it stimulated a heated discussion on how to classify speech acts, where scholars gradually deepened their understanding of speech act theory while recognizing Austin's classification deficiencies.

However, Austin was not the only scholar who regarded language as an act because phenomenological philosophers, including F. Brentano, E. Husserl, A. Reinach, and J. Daubert, long discussed the phenomena in words (Gu, 1994a: 2), and his understanding of the essence and taxonomy of speech act is also far from satisfactory. Despite that, scholars generally acknowledged that Austin's insights on speech acts and illocutionary force were valuable contributions to speech act theory (Gu, 1989: 36). Furthermore, it is of great illuminative significance to view language use as an act.

A series of subsequent studies by Searle et al. (Searle, 1969, 1976, 1979; Searle, Kiefer, & Bierwisch, 1980; Searle & Vanderveken, 1985, etc.) developed and enriched Austin's speech act theory. Gu (1994a: 2) proposed that Searle's work on speech acts could be summarized in three aspects: critiques of Austin's speech act theory, unique contribution to speech act theory, and logical analysis of the speech act presented in collaboration with Vanderveken.

In terms of Austin's speech act theory critiques, Searle mainly focused on the abstract segmentation and classification of speech acts. He argued that semantics and speech act are not two independent studies, but the same study from two different perspectives. The new recognition that the meaning of a sentence determines speech act in an appropriate context and basically defines illocutionary force makes the separation of the illocutionary act and locutionary act untenable. Searle therefore proposed 'the propositional act', i.e. different discourses can express the same proposition, but they have diverse illocutionary force (Gu, 1994a: 3). Searle (1976) pointed out that Austin's deficiencies in illocutionary acts are reflected in at least six aspects: (1) Explicit criteria for classification are absent; (2) it is a mistake to categorize illocutionary acts as speech act verbs; (3) speech acts overlap in different categories; (4) miscellaneous speech act verbs arise in a specific category of illocutionary act; (5) some speech acts do not match the definition of the category; and (6) there is a failure to recognize that not all verbs are speech act verbs. Therein, the lack of an explicit classification standard is the biggest drawback that gave birth to all the other defects.

As the first scholar to put forward a set of speech act theories and the concept of an 'indirect speech act,' Searle made a unique contribution to speech act theory. Following Searle (1969: 21), language communication belonged to the science of human behaviour, because its most basic unit is the speech act generated and restricted by a series of constitutive rules. Meanwhile, Searle determined the sufficient and necessary conditions together with the constitutive rules of speech acts and further explored 12 dimensions to classify speech acts. In terms of these dimensions, four of them are the most important: (1) The illocutionary point[4] of a speech act, (2) the direction of fit between discourse and the objective world, (3) the psychological states of discourse expression, and (4) the content of the proposition. On that basis, he divided speech acts into five categories: assertives, directives, commissives, expressives, and declarations. Then he employed formalized methods to describe the above five speech acts, respectively.

Moreover, Searle and Vanderveken (1985: ix) also sought to give a logical analysis of the speech act. In the preface of *Foundations of illocutionary* *logic*, they claimed that 'there have been few attempts to present formalized accounts of the logic of speech acts'. On this account, the book was written to establish a precise formalized theory of illocutionary acts under the guidance of modern logic resources. Vanderveken (1994: 100) pointed out that illocutionary logic can be integrated into general formal semantics (see Fu, 2005). Related studies mainly focused on the dimensions of speech act verbs, semantics, etc., restricting the study of speech acts and illocutionary force to the categories of logical philosophy and formal semantics to research by following the path of formal logic.

Grice's (1975, 1978) series of papers illustrated the cooperative principle and four subordinate maxims (maxim of quantity, quality, relation, and manner), which are under the postulate that the speaker and hearer are both rational beings.[5] These crucial axioms, initially proposed by Grice for conversational exchanges, were later applied to construe the speaker's sincere intention behind the illocutionary force in a specific situation. Likewise, Strawson (1964) also delved into the intentions entailed in speech acts, which is closely related to Grice's theory of meaning/intention. With the principle of politeness, Leech (1983) examined the motivation to use indirect speech acts and concluded that people use indirect speech acts out of politeness. Additionally, it should be noted that the 'utterance' proposed by Grice is a broad concept that includes both verbal and non-verbal acts. In other words, an utterance could be a string of words or a body movement. In previous research, pragmatics was tied in with the verbal communication of language; therefore, it excluded Grice's argument that body movements are utterances and Austin's theory of the perlocutionary act from pragmatic research. In fact, language use and social activities are intertwined, and hence a general theory of behaviour is needed as a more fundamental theoretical framework to deal with the complicated situations of language use (Gu, 2010: xv–xvi).

Bach and Harnish (1979) also made significant contributions to the development of speech act theory. They proposed that the speaker's intention and hearer's recognition are both crucial in language communication, that is, to understand the implications of the speaker's words, the hearer also has to make full use of the communicative intention and the hearer's inference in addition to the content and context of the discourse, so as to connect language structure and speech behaviour (Bach & Harnish, 1979: xi). Such an 'intention-and-inference' approach is widely divergent from Austin's argument – acting is totally conventional. Nevertheless, they did acknowledge that some illocutionary acts are conventional and do not need 'intention-reasoning' (Bach & Harnish, 1979: xvi).

Bach and Harnish (1979: xii) claimed that their methods almost rely upon analytical philosophy and linguistics and combine methods related to cognitive psychology and social psychology. Their explorations on pragmatics were already in accord with those of linguists, i.e. regarding language use as an activity in which people use language to exchange information. Building on that exploration, the illocutionary act is an act of linguistic communication, or rather, an act of expressing attitude through utterance. A successful illocutionary act (speech act) is achieved when the hearer understands the speaker's attitude through reflexive intention (Bach & Harnish, 1979: xv), and certainly, the speaker also 'intentionally' supposes that the hearer understands his/her intentions. Indeed, such a reflexive intention also serves as the distinguishing feature of different illocutionary acts. Bach and Harnish (1979) treated speech acts purely from the perspective of 'information communication,' which completely isolated illocutionary acts from perlocutionary acts, for the reason that as the latter is beyond the scope of linguistic communication, it should be excluded from the pragmatic field. Based on 'information and communication theory,' this pragmatic perspective influenced many linguists, including Leech and Levinson (Gu, 1993, 1994).

Additionally, the taxonomy of speech acts developed by Bach and Harnish (1979: 39–59) was quite distinctive. According to the attitudes of speaker and those of hearer, they categorized various speech acts into four groups, including: (1) constatives (that express a speaker's belief and his/her desire that the hearer forms a similar one), (2) directives (that express some attitude about a possible future action by the hearer and the intention that his/her utterance be taken as a reason for the hearer's action), (3) commissives (that express the speaker's intention to do something and the belief that his/her utterance obliges him/her to do it), and (4) acknowledgements (that express feelings toward the hearer, or the intention that the utterance will meet some social expectations regarding the expression of feelings).

At the early stage, many scholars explored the speech act theory regarding the relations and distinctions between intention, meaning, and illocutionary force. *Symposium on J. L. Austin* by Fann (1969) is an epitome of the fruitful achievements during this period. In the final analysis, as speech act theory was born under the influence of the historical background and ideological trend of modern philosophy's 'linguistic turn,' and as the early scholars engaged in the study of speech act theory were mainly analytical philosophers, the study of speech acts has long taken on the connotations of pronounced logical philosophy. Therefore, in the initial stage and for some time after that, mainstream research on speech acts in traditional British and American pragmatics mainly focused on the field of logical philosophy. Analytical philosophers, however, mostly researched speech acts based either on their introspective judgement or on their observations of written language and self-made examples, without noticing the influence of context (de Moraes & Rilliard, 2014: 233). The above phenomena are closely correlated to analytical philosophers' tasks in the early research of speech acts, for what they were considering was not only about linguistics but also about linguistic philosophy.

1.1.2 The participation of linguists

After the theoretical foundation was laid, speech acts, a consistent topic of pragmatics, received comprehensive and intensive research from scores of philosophers, and an abundance of academic achievements were made in this domain. Nevertheless, overall, the studies in question emphasize theories while overlooking the practical investigation of language (Gu, 1994b: 15). As pragmatics became an independent discipline, exploratory approaches and the research scope of speech act theory witnessed a gradual change after the inclusion of linguists.

When the speech act theory was established, Austin concurrently proposed taxonomies of speech act verbs and illocutionary acts; unfortunately, he failed to analyse them in detail. It is not until linguists were included in the research that the taxonomy, syntactic structure, and semantic dimension of the illocutionary act won more academic attention. However, many studies, affected by logical philosophy, were still confined to the field of formal semantics. Since the 1970s, the focus of pertinent studies was on decoding the syntactic structure of speech, or on developing the taxonomy of illocutionary acts and speech act verbs. For instance, Ross (1970) noticed that there are some shared syntactic properties between simple declarative sentences and explicit performatives; following semantic and syntactic cues, Verschueren (1980) systematically elaborated the definition, research values, and analytic methods of speech acts, and further labelled speech acts in different groups. As for adopted methodology, he no longer followed the traditional one originating from analytic philosophy, but preferred the practice of linguistics and anthropology as he borrowed some formalism in the description of speech act verbs. Furthermore, Verschueren (1999) also differentiated performative verbs from general speech act verbs through a detailed explanation of their performance.

At that time, Wierzbicka (1987) was recognized as the scholar who achieved happy outcomes in the study of speech acts. He believed that speech act verbs are essential for us to perceive the real world (consisting of interpersonal relationships and interactions). For a long time, however, a systematic study on speech act verbs was absent (Wierzbicka, 1987: 3). On the whole, previous studies only attached importance to the ‘exemplary’ exploration of speech act verbs whose definitions were also restricted to annotation in traditional dictionaries (Wierzbicka, 1987: 8). Not every speech act verb can be adequately explained with the help of ordinary dictionaries, as the biggest flaw of traditional dictionaries in defining verbs lies in ‘circularity’ (Wierzbicka, 1987: 4–5). For example, in the *Longman Dictionary of Contemporary English* (1978), multiple verbs are borrowed to interpret the speech act verb ‘ask,’ including answer, request, demand, call on, and invite, resulting in a concatenation of speech act verbs that also require further interpretation. Based on the *Longman Dictionary of Contemporary English*, Wierzbicka started a systematic investigation of English speech act verbs (1987). One of her findings shows that there are about 250–300 frequently used speech act verbs in everyday English, not counting marginalized or technical words. Also, she argued that there is no way to categorize speech act verbs unless each of their meanings is examined in detail. In light of the annotation of speech act verbs, Wierzbicka (1987: 1–35) explained it in four aspects (cf. Zhong, 2008): (1) Citations and metalanguage are adopted, while 'definition' is replaced by 'explication.' In her studies, it is common to see a speech act verb with a detailed exemplification without alluding to other speech act verbs. (2) Comparisons are made between the collocations of various speech act verbs in the pattern of negative structures (e.g. Wierzbicka, 1987: 46–47). (3) Pragmatic features are taken into account in the analysis of syntax. For example, to better illustrate the speech act verbs, the speaker's intentions, beliefs, and attitudes are canvassed in Wierzbicka's exemplification. (4) The ordering criteria of speech act verbs depend on the extent to which they are relevant. As is known to all, no dictionaries can skip the step of ordering words in the entry. In this case, Wierzbicka abandoned the traditional lexicographical method that puts words in alphabetical order, but turned to utilize the meaning of words as the ordering standard, especially their semantic similarities and differences. More specifically, she grouped closely related words into the same category, so as to reflect the relationship between those words.

Wierzbicka's pioneering exploration of speech act verbs has undoubtedly provided significant referential value for pertinent research. However, some deficiencies did arise in her studies. (1) There is room for expansion concerning the coverage of speech act verbs; (2) still, 'semantic primitives' used to interpret other speech behaviour verbs can be pared down and standardized; and (3) speech act verbs in different categories can be further analysed by comparison.

It should be pointed out that though scholars at this stage made significant achievements in investigating speech act verbs, the study of speech act verbs cannot be equated with that of speech acts. Tracing back to Austin's establishment of speech act theory and his early research, performative verbs were treated as an entrance or an analytic apparatus in examining speech acts, so it is apparent that the study of performative verbs is not equivalent to the study of speech acts. When exploring performative verbs, researches always looked into the grammatical dimension; but the fact is that only the performative functions of discourse were carefully studied, as they dealt with the pragmatic dimension (cf. Leech, 1983: 206–211). Although most scholars knew this well, there was no better access to investigating speech acts. They still prioritized the analysis of speech act verbs, which constitute a pivotal part of speech act study.

Many discussions on the classification of speech acts also emerged at this stage. Based on previous conclusions, Leech (1983: 206) added the category of '*rogative verbs*' as a novel type to speech acts. When it comes to inquiry or interrogation, observing the relationship between discourse and the objective world, it is the hearer (the person who is expected to answer the question) rather than the speaker that makes the discourse fit in the objective world. Levinson (1983: 239–242) argued that Seale's taxonomy was further improved compared with Austin's, but it could not cover all the types of speech acts due to the lack of a principal basis. Searle (1976: 10–16) divided speech acts into five groups: representatives, directives, commissives, expressives, declaration. In Verschueren's (1999: 24) view, though Searle fine-tuned the taxonomy of speech acts, the improvements he made are similar to those made by Austin, confined to three overlapping aspects, namely expected psychological state, direction of fit, and illocutionary point. If we reframe the way we approach speech acts, different taxonomies will arise. Developed from the individual perspective, Seale based the criteria of his taxonomy on the speaker's 'intentionality' in essence, overlooking the fact that speech acts are social behaviour in nature, and misclassifying speech acts as expressions of personal intentions labelled in the category of individual psychology, which is consistent with thought in his late works. Additionally, some scholars classified speech acts with the integration of discourse analysis. The following are some illustrative examples. Taking contextual components and syntactic-grammatical properties of discourse into account, Sinclair and Coulthard (1975) developed a taxonomy based on the interactive functions of speech acts. Hancher (1979) demonstrated his classification and the taxonomy of Vendler, Ohmann, and Fraser, which are both roughly similar to Seale's. Tsui (1994) creatively proposed the model of interaction, and three main moves of initiating, responding, and follow-up were included in her taxonomy of interaction analysis. In general, following Searle, many scholars only attempted to improve his taxonomy, restricting themselves to his classifying framework concerning the principle (cf. Wang, 2006: 86).

All this being said, researchers started to pay attention to the linguistic dimension of speech acts at this stage; however, pertinent studies failed to treat language as a kind of human behaviour for the reason that their focus was disappointingly confined to the syntactic structure of discourse and speech act verbs, especially their taxonomy. Correspondingly, the scope of the investigation was also confined to illocutionary force indicating devices (IFIDs) in terms of linguistic devices. For a long time, researchers used to pay more attention to performative verbs when it came to IFIDs, overlooking the importance of other IFIDs. Some researchers even borrowed some formalism to describe and analyse IFIDs (Zaefferer, 1981). As suggested by Searle (1969: 30) and Searle and Vanderveken (1985: 1–2), in English, inter alia, IFIDs range in their scope in mood, punctuation, word order, pitch, stress, etc. With increasing attention to colloquial and live speech in pragmatic studies, some researchers began to probe into the role of prosody in performing illocutionary force (de Moraes & Rilliard, 2014: 234).

Related studies in this field remained dissatisfying, albeit some linguists noticed the diversity of IFIDs and launched an investigation into a few non-verbal devices. The underlying cause was rooted in the scope and methodology of the research and the selected corpus. In traditional studies, many collected linguistic data by means of role-playing, discourse completion test, multiple-choice questions, rating scale, etc. (Felix-Brasdefer, 2010). Austin (1979: 237) once stated that there are a variety of non-verbal messages beneficial to the performance of felicitous speech acts. Such gestures as he alluded to, including winking, pointing, shrugging, and frowning, are considered to be crucial to the conveyance of pragmatic meaning even in the absence of speech (Austin, 1962: 76).

Consequently, a non-verbal apparatus involving tone, pitch, rhythm, stress, and accompanying gestures (e.g. winking, pointing, shrugging, frowning) constitutes alternative devices in performing the illocutionary act. Sometimes, behavioural information is conducive to getting rid of the ambiguity of speech. Taking an example given by Austin (1979: 245–246), imagining a person bowing deeply from the waist in front of you, various purposes or implications may be embedded in this action – he/she perhaps is going to observe the local flora on the ground ahead of you or to tie his/her shoelaces. Alternatively, maybe he/she is paying obeisance to you. Other IFIDs like 'taking one's hat off and speaking a greeting' may help eliminate such ambiguity. Thus, these movements are one of the indispensable components in performing a felicitous illocutionary act. Notwithstanding, Austin failed to explore more deeply in this regard; at the least, he enlightened the future research that physical movements should be an essential aspect of speech analysis.

Speech act theory is an incarnation of the infiltration and integration of language, behaviours, and society, through which Austin (1962) suggested that most discourse is imprinted with human behavioural properties, and language is indeed a unique human behaviour. As Levinson (1983: 246) pointed out, illocutionary force belongs to the category of behaviour, and appropriate instruments for its analysis should be sought in the theories of behaviour rather than in theories of meaning. Searle (1969; 2001) also insisted that language is part of theory of action and language is a form of intentional, rule-governed behavior. Though firmly claimed by some scholars, many previous studies were still unaware that speech acts are part of behavioural theory, and most scholars were always keen to explore the conditions for explanation and restriction within language (Fu, 2005). Under such circumstances, related studies made light of such social factors as the context in which speech acts are located. In Searle's late studies, he strayed from the main line adopted in his early studies that integrating language, behaviour, and society, attempting to seek a breakthrough in speech act from the intentionality[6] of the brain. Such practice linking language with the human brain's mechanism and ingraining speech act theory into the philosophy of mind disobeys the original intention when speech act theory was first established (cf. Gu, 1994a, 1994b).

Research at this stage laid particular stress on the illocutionary act while underestimating the vital role perlocutionary acts played in the whole pragmatic context. In the same vein, relevant research excluded perlocutionary acts from one of the organic components of 'complete speech acts,' as described by Austin from a pragmatic perspective. In this way, the initial purpose of pragmatic studies, that is, to investigate the actual usage of language, was undermined. Fortunately, in later studies, some scholars took discourse analysis and rhetoric into account, which proved to be intrusive for the ensuing research (cf. Gu, 2010: xiv–xvi; O'Keeffe, Clancy, & Adolphs, 2011: 96–98).

Along with research on speech acts, many scholars made objective reflections on speech act theory. For instance, holistic evaluations on the studies of speech acts and Searle's academic philosophy are illustrated in three critics; see Burkhardt (1990), Lepore and Gulick (1991), and Searle, Parret, and Verschueren (1992). Moreover, some scholars began to examine speech acts through interdisciplinary research from the dimensions of cognition and semantics, i.e. a comparative study of speech acts from a cross-cultural perspective thankfully arose in this field (Blum-Kulka, House, & Kasper, 1989). Another example is that considering speech acts as the basic unit of speech communication, scholars such as Dore (1973) began to study the mechanism of children's language acquisition from speech acts, yielding great achievements.

At a time when Western linguists were conducting intensive research on speech acts, the Chinese linguistic community began to introduce it. The pioneering introduction of exotic pragmatic theories can be traced back to 1979 when Xu GuoZhang, a well-known scholar, translated *How to Do Things with Words* by Austin into Chinese, which was enrolled in the *Journal of Linguistics* in the same year (Wang, 1990). Subsequently, other pragmatic research involving speech act theory was introduced to China by many Chinese scholars, such as Hu (1980), Gu (1989), and He (1989). Among them, Duan (1988) and Gu (1989, 1994a, 1994b) systematically introduced and critically analysed speech act theory. He (1984, 1988) discussed indirect speech acts, especially English ones. Some scholars also existed, Gu (2015) as one representative, who explored the taxonomy of the Chinese illocutionary act and Chinese speech act verbs with recourse to speech act theory (cf. Gao & Yan, 2004). In all, many Chinese scholars carried out a considerable amount of research on Chinese and English speech acts, and here we will not go into it.

1.1.3 Corpus-based speech act studies

As pragmatics evolved and research paradigms became more and more diverse, a new direction for the study of speech acts unfolded, which further enriched the research scope. Also, with researchers' efforts, the traditional meaning-oriented studies developed into multi-perspective ones that focused on the whole discourse and communicative process, unlocking pragmatics from the confines of linguistic philosophy and consistently reframing the research methods.

1.1.3.1 The rise of speech act study with text corpus approach

From the 1980s onwards, corpus-based language research started flourishing abroad. Viewing the development of corpus linguistics over the past 30 years, though accompanied by ups and downs, it has nowadays penetrated and influenced many linguistic fields. Today, Chinese scholars have utilized some fundamental methodologies in corpus linguistics to deal with linguistic issues, which also incorporated influence in other research fields concerning humanities and social sciences. Against such a background, corpus-based methodologies have been introduced to pragmatic research. Qian and Chen (2014) did a statistical survey on how many research papers in the pragmatic area were published with corpus-based approaches during the past ten years. Their conclusion showed that, among the studies that adopted the corpus approach, various research objects, perspectives, and analytical units could be found in overseas studies, while such resources appeared to be only in extant Chinese studies. Abroad, corpus-based papers accounted for nearly half (46.7%) of all pragmatic research papers, while the Chinese counterparts only accounted for 1.4% (taking the papers published in the *Journal of Pragmatics*, an authoritative journal of pragmatics research as an example). In 2004, a special issue on corpus methods for pragmatic research was published in the *Journal of Pragmatics*. Then some monographs and collections in the same domain (including synchronic and diachronic studies) successively came out, e.g. Adolphs (2008), Romero-Trillo (2008), Felder, Müller, and Vogel (2011), and Taavitsainen and Jucker (2014) (cf. Rühlemann & Aijmer, 2014). Since 2013, Springer has been publishing an annual *Yearbook of Corpus Linguistics and Pragmatics*, exploring the cross-relations between corpus linguistics and pragmatics, which provide plentiful examples of corpus-based approaches to pragmatics research.

The study of speech acts constitutes an important topic in corpus-based pragmatics research (Qian & Chen, 2014). Therefore, a number of corpus-based speech act studies boomed at home and abroad. In terms of overseas studies, some significant research outcomes on speech acts are listed as follows, to name but a few: Aijmer (1996) conducted an early study on the investigation of a few speech acts in English with a basis in the London-Lund Corpus of Spoken English; Cutting (2001) explored how speakers' attitudes affect the implementation of speech acts in authentic conversations; Adolphs (2008) paid keen attention to speech acts in spoken corpora; Carretero, Maíz-Arévalo, and Martínez (2013) analysed speech acts embodied in online discussions; and Kohnen (2000) and Rühlemann et al. (2010) discussed the application of corpus methods in pragmatic research, e.g. in the study of speech acts. As for Chinese research, Xiang (2007) evaluated the distinctive features of Chinese speech acts categorized in commissives from the perspective of semantic prosody; Kuang (2013) delved into Chinese speech acts expressing 'thanks' in the class of behabitives; and Liu (2013) gave an analysis of the characteristics of speech acts in requests performed by Chinese college students.

One of the difficulties in using corpus-based methods to study speech acts lies in the identification and annotation of their types. In this view, a series of annotation schemes were developed for pragmatic research (including those designed to mark speech acts), such as DAMSL (Dialogue Act Markup in Several Layers), SWBDD (Switchboard DAMSL), and DART (Dialogue Annotation & Research Tool) (Weisser, 2015). Plus, some pragmatically annotated corpora were constructed, among which speech acts were tagged in such corpora as the Corpus of Verbal Response Mode Annotated Utterances (Stiles, 1992) and a sub-corpus of the Michigan Corpus of Academic Spoken English (Maynard & Leicher, 2007). Some specific examples concerning the annotation of speech acts can be found in Stiles (1992), Leech and Weisser (2003), Garcia (2007), Kallen and Kirk (2012), and Weisser (2015) (see Rühlemann & Aijmer, 2014). However, the corpora mentioned exclusively deals with English discourse, and the corpora tagged with Chinese speech acts is still uncommon. In the natural language process, researchers created schemes for the automatic annotation and processing for conversational speech acts (Georgila, Lemon, Henderson, & Moore, 2009). Additionally, some software tools were developed to label speech acts automatically, such as the SPAACy (a semi-automated tool for annotating dialogue acts) designed by Weisser (2003), but the applicability to other languages still needs enhancement. Leech and Weisser (2003) pointedly asserted that currently, there are two kinds of annotation approaches to speech acts, namely, one is specialized for a single project, and the other is developed for broader, general purposes. Based on previous research, the SPAADIA (Speech Act Annotated Dialogues) Annotation Scheme was developed by Leech and Weisser (2014) to fulfil the goal of balancing both specificity and generality. That is to say, SPAADIA is compatible with a specific or universal task. As can be gleaned from the brief overview above, in recent years, improvements have been made in the annotation of speech acts, but a widely accepted annotation method or system is still absent (De Felice, Darby, Fisher, & Peplow, 2013: 78).

The crux of speech act tagging is that not all speech acts have a regularized and relatively fixed syntactic form. Rühlemann (2010: 290) refers to this as having no ‘lexical hook,’ because annotation must take contextual information in addition to words, phrases, etc. into account (Weisser, 2015: 84). At present, some researchers have adopted a line-by-line analysis method for all corpus texts, that is, every utterance is marked appropriately as a particular type of speech act, and then the marked utterances are further annotated in other languages or with other contextual information (McAllister, 2015). Usually, this method is accurate in judging and labelling speech acts, but it is time-consuming and labour-intensive when dealing with large-scale corpora. Thus, many studies prefer the joint method of computer keyword retrieval and one-by-one analysis in the annotation of speech acts (see, e.g., Adolphs, 2008; McAllister, 2015).

The study of speech acts with corpus-based approaches has benefited from the development of corpus linguistics, making it an increasing interest in the research of speech acts. Overall, although much progress has been made in this domain, a few drawbacks still arose in those studies.

Firstly, the corpora consulted are all text corpora. Even the study of speech in natural spoken language is based on transliterated text. So, it is not surprising that researchers attach much attention to language forms such as performative verbs and syntactic structures when dealing with textual data.

Secondly, some studies indiscriminately equate the annotation, retrieval, statistics, and analysis of performative verbs with the study of speech acts. Although many researchers incorporated various linguistic forms and the contexts of the utterances into the investigation of speech acts, what should be noted here is that the annotation, retrieval, and analysis of linguistic forms (such as performative verbs and established linguistic forms) should be recognized as a priority in the study of speech acts. In particular, early corpus-based research focused exclusively on speech acts with fixed forms or established usage (e.g. Aijmer, 1996). For example, when examining the speech act of 'apology,' researchers usually search for terms like 'sorry' or 'pardon' in English corpora, and the same practice is adopted in Chinese corpora.

Thirdly, in most studies, every single utterance is treated equally as an independent speech act. On account of that, methods and software contributing to the automatic annotation of speech acts were developed by Stolcke et al. (2000), Leech and Weisser (2003), Stockley (2006), and Weisser (2003, 2015), etc. However, the fact is that in natural conversation, speech acts are entirely functional. More specifically, one speech act may reside in a variety of utterances, and there is no strict one-to-one correspondence between the two.

In general, as the automation of annotation improves as well as research continuing, the study of speech acts based on text corpora will soon earn an important place in the international pragmatic community, despite difficulties and challenges.

1.1.3.2 Emergence of multimodal corpora and speech act study

In the most up-to-date corpus-based research, utilizing text corpora to study speech acts has become dominant, but the traditional text-based approach can only partially reflect the copious information in spontaneous conversation. Many studies particularly highlight the investigation of performative verbs and syntactic structures while under-evaluating the importance of illocutionary devices such as prosodic features and gestures.

As the corpora evolved from single text corpora to audio corpora, researchers began to be aware that there are more cues that make a difference in language use, e.g. voice and prosody. In this sense, the research scope has been further enriched. Importantly, prosody's pragmatic function (especially intonation) is valued (Wichmann & Blakemore, 2006). Some scholars studied in depth the important cues embedded in voice and prosody in the analysis of speech acts. For example, the relation between French intonation and the performance of illocutionary force was detected by Fónagy, Bérard, and Fónagy (1983); Wennerstrom (2001) discussed the potential influence that English prosody exerts on illocutionary force, while Meyer and Mleinek (2006) examined how prosody in Russian affects the delivery of the speaker's intended meaning; and by bringing the dimension of emotions into the picture, de Moraes and Rilliard (2014) explored how illocutionary forces interact with affective prosody and facial expressions in Brazilian Portuguese. It was found that, in general, social attitudes and prosodic features seem to have a sound correlation in various types of illocutionary forces, while propositional attitudes are not related to any specific prosodic features; and by observing facial features, it is easy to distinguish social attitudes and propositional attitudes. As far as studies of illocutionary force are concerned, the above is some of the few studies from many different perspectives (prosody, facial expressions, etc.).

With the advancement of modern computer multimedia technology and storage technology, and the improvement of people's understanding of the essence of language activities, the linguistic data that researchers concerned themselves with changed from monomodal to multimodal, which gave birth to multimodal corpora and facilitated the rise of multimodal corpus linguistics[7] accordingly. The scope of pragmatic research (including the study of speech acts) was extended, and its methodologies were enriched by using multimodal approaches, making pragmatic research with multimodal corpus approaches a new concern (Romero-Trillo, 2008) and placing this study at the frontier of pragmatics.

Multimodal corpora reflect much information that was previously unrecorded and display the relationship between language, context, or other factors, which is of high reference value for us in more comprehensively and accurately interpreting the meaning of language in use. The adoption of the multimodal corpus in pragmatic research provides the classic theories originating from logical philosophy with fresh perspectives and new practices (Knight & Adolphs, 2008), and the scope of classic pragmatics theories will be further enriched. Some researchers examined the affordances of prosody, facial expressions, and bodily movements in meaning-making, resorting to multimodal corpora (either based on partial linguistic data or small corpora). For example, Kendon (1995) devoted analytical focus to the connotations of four frequently used gestures like sign- and meaning-makers in authentic conversation in southern Italy via video documents, with the assumption that they can be the apparent indicators of various speech acts in mind. Likewise, building on filmic documents, the similarities and differences between the translation of English and French speech acts were compared by Mubenga (2008, 2009), who proposed the multimodal pragmatic analysis based on Halliday's theories of systematic grammar and semiotics. In China, under the guideline of semiotic multimodality, Chen and Qian (2011) brought up a multimodal framework for pragmatic analysis to delve into the intricate interactions of the multidimensional information embedded in live speech.

So far, few approaches based on multimodal corpus linguistics have been utilized in the study of speech acts. Nevertheless, Gu (2013a) made a breakthrough in this regard. Based on the analysis of individual discourse in multimodal corpora, he constructed a conceptual model for analysing performative acts and discussed the relationships between illocutionary forces, emotions, and prosody in situated discourse. Daily verbal communication experience tells us that the same utterance accompanied by different prosodic features, expressions, or physical movements often conveys different interpretations intended by the speaker, i.e. different illocutionary forces. For example, the use of 'This is your book' is open to two interpretations: (1) if the speaker says it in ascending tones, it can be intended as a sign of scepticism, that is, the speaker doubts the attribution of this book; (2) if he/she uses descending tones, it can be intended as a sign of affirmative or, to put it differently, he/she is making the statement that the book belongs to you. This shows how many prosodic features influence speech communication. Sometimes, the speaker's facial expressions and bodily movements often give away the occurrent emotions and, at the same time, affect the prosodic features, causing the same utterance to yield different illocutionary forces. For example, the utterance 'You are so smart' can be perceived in different ways when the speaker gives various non-verbal cues. More specifically, if the speaker articulates it in a rising tone or even plays with dancing eyebrows, it shows that he/she is delighted, and the utterance can be construed as a genuine compliment by the hearer; on the other hand, if the speaker gives it a falling intonation in an indifferent manner, frowning and squinting, the utterance may be perceived and interpreted as dissatisfied and biting by the hearer. These are what we usually mean by 'listening to the intended meaning behind someone's words' and 'reading faces.' This indicates that speakers are all involved in multimodal interactions with outsiders (including the physical surroundings) in natural conversations where they perform various performative acts that correspondingly form particular authentic illocutionary forces. In this respect, researchers need to take syntactic structures, prosodic features, bodily movements, and other cues into consideration.

Meanwhile, basic approaches in multimodal corpus linguistics, steps for Simulative Modelling, and research perspectives are elaborated by Gu (2013b), providing a theoretical framework and analytical tactics for the investigation of speech acts based on multimodal corpora. Although it is outside the mainstream to explore pragmatics questions such as live speech acts in natural discourse with multimodal corpus-based methods from 'the whole person' perspective, it still places such research at the frontier of pragmatics.

Gu (2013a) carried out research that treats speech acts as collective human behaviour. However, previous research prefers the concept that language use is an activity in which people exchange information with the help of language; hence, the core issue of pragmatics is to reveal how language serves communication (Gu, 2010: xiv). In this way, expressions, gestures, and bodily movements are only considered as para-/extra-linguistic information whose research value lies in how they contribute to information exchange. However, when facing real-life situations of language use, this mindset tends to be out of place, because establishing the research foothold and theoretical foundation in information communication theory is insufficient for us to get the whole picture of language use. Instead, it is necessary to borrow concepts from sociology in connection to actions to restore the soul of speech act theory. Moreover, we have to bear in mind that due to the fact that language use is intersected in social activities, the matter of language should be analysed from the perspective of behaviour, and language use should be understood under the framework of general behaviour theories (Gu, 2010: xvi). Other cues, including gestures and prosody, are indispensable components of the felicitousness of illocutionary acts performed by 'the whole person,' and therefore must be taken into account when studying the illocutionary forces located in situated discourse. Detailed strategies are presented in Chapter 3.

Generally, the research carried out by Gu (2013a) provides inspiration for further investigation into speech acts in situated discourse by providing analytical ideas and theoretical frameworks. Additionally, the STFE-match assumption (what is said, what is thought of, what is felt, and what is embodied) proposed by Gu (2013b), on the other hand, provides solid research principles and analytic perspectives for this research. As noted already, there is no doubt that these studies are quite cutting-edge and full of inspiration. However, importantly, they still require refinements, which constitutes the very reason for the present study. The relationship between this pilot study, Gu (2013a), and Gu (2013b) can roughly be described in two aspects:

1 The theoretical basis and conceptual model of real-life illocutionary force have already been established by Gu (2013a, 2013b), where only a few sampling corpora are exemplified for analytical purposes, and a reasonable scale of multimodal corpora for statistics, analysis, and verification are absent. To fill this gap, more types of illocutionary forces are included in this research, in addition to a multimodal corpus in Chinese situated discourse being constructed for verification.
2 On the whole, Gu (2013a) shed light on the interaction between illocutionary force, prosody, and emotion, while such non-verbal cues as

facial expressions, gestures, and bodily movements are absent from the procedure of analysis. In this study, the analytic framework is refined to some extent, and the non-verbal act is added to the analytic tiers.

In brief, following the basic ideas and theoretical framework of Gu (2013a, 2013b), this study fine-tunes the analytic framework in both researches when necessity arises, and self-constructed multimodal corpora of a proper size will be adopted for quantitative analysis in the hope of facilitating the development of speech act study toward multimodal corpus pragmatics.

In conclusion, although some studies have examined the functions of prosody and gestures arising in the course of speakers performing speech acts, in general, the use of multimodal corpora methods to study pragmatic questions including speech acts is still in its infancy. Especially in China, extant research is still at the theoretical discussion stage, which means that empirical research is needed for refinement and enhancement. Furthermore, many problems remain unsolved in the adoption of corpus methods for pragmatic research, and on account of that, more effort should be made in this field to promote the development of pragmatic theories (Rühlemann, 2010).

1.2 Research contents and objectives

1.2.1 Overview of the research

This research locates itself in the field of multimodal corpus pragmatics, which is also an emerging research field that the author advocates. By consulting the basic methods in multimodal corpus linguistics and taking multimodal data in live speech, this research is to discover how speakers employ multiple means of expression in situated discourse to perform illocutionary acts. More specifically, the research is also designed to explore how the triadic interaction of emotion, prosody, and non-verbal acts produces a variety of live illocutionary forces, that is, to describe the mechanism of expressions carrying various illocutionary forces adopted by speakers and their characteristics, displaying regularity, and possibly being influential factors of IFIDs.

In general, the research aims to unveil how illocutionary force is associated with emotional states, prosodic features, and gestures from both original and extended viewpoints, hence enriching the theories in pragmatics by applying a multimodal corpus approach.

1.2.2 Primary research objectives

1 To explore the relations between speakers' occurrent emotions and illocutionary acts in live speech and examine how emotions interact with prosody and gestures;

2 To describe speakers' multimodal means when expressing different illocutionary forces, and to define the forms and functions of IFIDs in situated discourse, in addition to analysing potential factors that would affect their presentation and interaction;

3 Inspired by the research on the multimodal study of speech acts, the author believes that multimodal corpus pragmatics is far more than a speech act study. Therefore, based on the previous multimodal study of illocutionary acts, the author attempts to enrich the emerging multimodal pragmatics and provide further inquiry into this field.

1.3 Research methodology and analysis approach

In this book, language communication is deemed an intricate system. This work deals with speech acts and their corresponding illocutionary forces produced by Chinese illiterate people in live speech, to discover how speech content, prosodic features, and non-verbal acts, which are impacted by speakers' occurrent emotions, interact with each other to produce a variety of live illocutionary forces, and to describe the mechanism of the emergence of different IFIDs. Inspired by mixed-method research, the author blended qualitative and quantitative approaches in launching a qualitative analysis on speech behaviour and types of illocutionary force, in addition to a quantitative corpus-based investigation capitalizing statistical methods.

1.3.1 Basic methodology: Simulative Modelling

Simulative Modelling is widely used in the scientific research and engineering field, to model an object or phenomenon that already exists at the time of modelling. It requires researchers to collect authentic information on what the research object is, then select a particular perspective of research, mine data, and carry out the research. In terms of language research, Simulative Modelling is the attempt to reproduce the saturated meaning conveyed by human multimodal discourse activities under modelling (Gu, 2016: 4). Simulative Modelling is also an elemental approach in the construction of multimodal corpus linguistics. In this study, the speaker's authentic illocutionary force in situated discourse is modelled through three stages: (1) concept modelling, (2) data modelling, and (3) implementation and verification. On this basis, analysis is then carried out on illocutionary force in situated discourse.

1.3.2 Data resource: multimodal corpus

Situated discourse is a multimodal process filled with total saturated significance with the participation of the interlocutors in a conversation (Gu, 2009: 435). Therefore, it would be rather inadequate not to investigate speech acts from a multimodal perspective. In this research, the author transcribed and processed situated discourse with the employment of multimodal corpus linguistics, modern audiovisual technology, and multimodal corpus analysis software (Elan). Besides, with the combination of qualitative and quantitative paradigms, this research presents a dynamic and multidimensional description of illocutionary force conveyed by speech acts in the multimodal corpus of Chinese situated discourse.

1.3.3 Research scope: interdisciplinary perspective

Multimodal language research is comprehensive and includes several sub-disciplines. In authentic interactions, the expression of speech acts and their corresponding episodes depend on both intonation or gesture in creating particular functions and the actual words or discourse structures being used. In this sense, this research's principal concern is to figure out how speakers perform illocutionary acts and form illocutionary force through their multimodal interactions in situated discourse. From the perspective of linguistics, the integration of several branches of linguistics, including phonetic analysis, pragmatics, and corpus linguistics, is adopted through the application of statistical techniques and qualitative methods. Additionally, it provides the analysis of emotional states conveyed in the speakers' prosodic features and gestures when performing illocutionary acts, and the research also integrates some proper methods and background knowledge from sociology, psychology, and non-verbal studies.

1.4 Framework of this book

Overall, this book is organized as follows:

Chapter 1 reviews the research phylogeny of speech act theory and illocutionary force, and then introduces the backdrop, content, objectives, and methodologies of this study.

Chapter 2 comprehensively discusses basic concepts such as situated discourse and multimodal corpus to reveal the theoretical issues behind them. Introductions of the construction of multimodal corpus and multimodal corpus linguistics are presented in this chapter as well.

Chapter 3 elaborates on the core methodology of this study in detail, namely, Simulative Modelling. Also, an analytic framework is developed here by considering STFE-match assumption and live illocutionary forces in situated discourse. Insights into the taxonomy of illocutionary forces and its rationale are provided in this chapter as they forge paths and perspectives for a well-organized and segmented research later.

Chapter 4 zooms in to look closely at the Discovery Procedure of live illocutionary forces in situated discourse, including its principle, basis, characteristics, framework, and connotations.

Chapter 5 explains how the raw multimodal data located in situated discourse is collected and processed in this study.

Chapter 6 displays the details of the self-built multimodal corpus of Chinese illiterate people's illocutionary force, including selecting and implementing the segmentation and annotation schemes of illocutionary forces. Additionally, the testing of the annotation's consistency, validity, and reliability will be available here.

In Chapter 7, a conceptual analysis of various types of illocutionary force in the multimodal corpus is carried out, and the analysis of specific situations sums up each token corresponding to the illocutionary force type.

In Chapter 8, based on the constructed corpus and collected data, the speakers' multimodal expressive devices when performing illocutionary forces are broken down into further examined items using statistical methods. After that, we conclude with the mechanism and rationale beneath the surface phenomenon and surmise what may affect speakers' multimodal expressive devices and IFIDs.

In Chapter 9, the author summarizes and presents the conclusion of our research. Meanwhile, the innovation, research values, and constraints of this study and the future agenda in the ongoing research are discussed here.

In Chapter 10, fresh insights into the topics of multimodal corpus linguistics are provided by introducing research paradigms and various approaches to multimodality. Finally, the basic logic behind the construction of multimodal pragmatics and some associated research problems is proposed.

1.5 Summary

In brief, this chapter mainly reviews the research phylogeny of speech act theory and illocutionary force, and then introduces the backdrop, content, objectives, and methodologies of this study.

The study of speech acts can be divided into three distinct stages: (1) Establishing speech act theory and early research by analytical philosophers, (2) then including linguists from a purely linguistic perspective, (3) and now expanding with the corpus-based approach.

By adopting basic multimodal corpus linguistics methods and taking multimodal data in live speech, the research is designed to discover how the speakers employ various devices in situated discourse to perform illocutionary acts. More specifically, the research is also designed to explore how the triadic interaction of emotion, prosody, and non-verbal acts produces a variety of live illocutionary forces, i.e. to describe the mechanism of expressions carrying various illocutionary forces adopted by the speakers and their characteristics, displaying regularity, and possible influential factors of the IFIDs.

Notes

[1] Situated discourse refers to the here-and-now utterances made by certain language users without preparation, which is the most primitive form of verbal communication. See Section 2.1.1 for more details.

[2] Austin and Searle did not clearly define ‘illocutionary force’ in their classic works. Leech (1983: 15) defined it as the meaning of an utterance. If we adopt Leech’s definition, ‘illocutionary force’ in this study can be defined as: the meaning of an utterance yielded in verbal communication through speakers’ multimodal interaction.

[3] Gu (1989: 32) pointed out that illocutionary acts are performed by speakers, while illocutionary forces are a contextual function of the discourse.

[4] Illocutionary point is correlated with illocutionary force, but they are not the same. The former is a component of the latter. For example, ‘require’ and ‘order’ share the same illocutionary point – asking somebody to do something – but they have different illocutionary forces. Searle and Vanderveken (1985) explained the corresponding illocutionary points of the five illocutionary forces respectively: To state something in a determined way (representative), to promise something (commissive), to make somebody to do something (directive), to impact objective things through utterances (declaration), and to express some feelings or attitudes (expressive).

[5] It should be noted that Grice’s cooperative principle is based on the assumption that both the speaker and the hearer are rational beings, differing from Gu’s STFE-match principle (based on the live, whole person). This study adopts Gu’s STFE-match principle, which is elaborated in the following sections.

[6] The ‘intentionality’ used by Searle does not mean intentions or aims, but the generalization of mental activity abilities, e.g. beliefs, wishes, and the five senses.

[7] Different scholars interpret the philosophy of multimodal corpus linguistics differently. See Baldry and Thibault (2006) and Gu (2013b).

2 Situated discourse and multimodal corpus

2.1 Situated discourse and multimodal linguistic research

This study examines the illocutionary forces in the situated discourse that bear a close relation to multimodality. In this section, the ties between situated discourse and multimodal linguistic research are introduced.

2.1.1 Situated discourse

In its formation, situated discourse is the utterances (or writings) made by user(s) of a particular language at a specific time and place (Gu, 1999a: 3) with little pre-preparation. In essence, such discourse is situated in the sense that it is situated to (Gu, 2002b: 490):

1 An actual social situation;
2 Actual users;
3 An inter-subjective world of discourse;
4 Actual goals;
5 A spatial and temporal setting;
6 The cognitive capacity of actual users;
7 The performance contingencies of actual users who are engaged in spontaneous talking with little pre-planning.

In this study, the linguistic data is confined to land-borne situated discourse (LBSD) (see Gu, 2012b), i.e. interlocutors' face-to-face interaction in a given time and place. Such face-to-face discourse is the oldest linguistic form that existed before writing was invented. We need to look into the following facts: Some ethnic groups only communicate in spoken languages without a writing system, children normally acquire their mother tongue from LSBD before knowing how to read, and most of the verbal activities of an illiterate who has never received education in a written form are composed of situated discourse, so it is reasonable to conclude that human language develops from situated discourse, which serves as the most fundamental form of human verbal activities. Thus, situated discourse deserves intensive inquiry.

DOI: 10.4324/9781003251774-2

It should be noted that the study of situated discourse in this book is not wholly equivalent to the research of spoken language in traditional linguistics. Spoken language has a broader scope than situated discourse. In terms of speakers, spoken language can be a single person's self-talk, while situated discourse usually involves two or more speakers in conversation. In terms of preparing speech content, if the speaker prepares the speech content in advance (in written/oral form), such as oral presentation of a written text, drama, sketch or crosstalk, it falls into the scope of spoken language instead of situated discourse.

From the history of linguistic theories, situated discourse is treated with different importance and handled in different ways in theories developed by Chomsky, Saussure, Hymes, Halliday, Bloomfield, Austin, and Searle, and in various linguistic research paradigms, including discursive analysis and conversational analysis (see Gu, 1999b). However, the attention of some related research in conversation analysis of situated conversations is quite inspiring. Conversation analysis was first carried out by sociologists, among which the representative scholars include Sack, Sehegloff, and Heritage. Early conversation analysis research mainly described the structure of daily social activities (Sacks, Schegloff, & Jefferson, 1974), and later on, its scope was extended to social interactions in institutional contexts (González-Lloret, 2010: 59). Today, conversation analysis has widened its research scope to investigate speech acts (Drew, 2013: 4). Researchers of conversation analysis set store on many phenomena that are not of interest to previous linguistic studies, such as turn-taking rules, turn transition relevance places, situated repair, emergency response, interruptions, and pauses. Indeed, their studies are mainly aimed to expatiate on the social phenomenon behind conversations. Ethnomethodologists, Harold Garfinkel as a representative, probed into 'social interactions' in daily life (see, e.g., Garfinkel, 1967; Sacks, 1984; Schegloff, 1984, 2007). If we use one sentence to summarize ethnomethodology, then it would be 'back to the things themselves,' which means studying the related issues of sociology by returning to the facts of everyday activities in the sociological aspect. Later on, with the emergence of interactional linguistics, a discipline influenced by conversation analysis in its methodology, scholars began to pay more attention to language structure, turn-taking orders, prosodic features, and the syntactic–semantic functions in everyday verbal activities (Lindström, 2009: 96–97).

From a methodological point of view, all research paradigms, including ethnomethodology and conversation analysis in sociological research, and the up-and-coming interactional linguistics, stress the importance of authentic information in people's daily practices, which serves as the warp and woof for analysing various social and linguistic issues. Their revelation to linguistics is 'back to language use itself,' i.e. linguistic research should focus on the real look of everyday language use and generalize the linguistic patterns from the most straightforward facts. However, taking situated discourse as an object of linguistics research involves the classification of linguistic branches and the construction of related theories. In this respect, Gu (1994, 1997, 1999a, 1999b, 2002b, 2002c, 2006a, 2006b, 2009) includes a series of theoretical constructions and practical explorations.

Compared with previous research, the investigation of speech acts grounded in situated discourse can be deemed an extension in terms of investigating perspectives and objects, which encourages researchers to make breakthroughs in methodology. For more explorations on the speech act and its recommended method (simulative modelling), detailed elaborations can be found in the author's relevant works (see Huang, 2014a), and we will not go in detail here.

2.1.2 Definition of multimodality

One concept closely related to situated discourse is 'multimodality.' The term 'multimodality' has become a 'buzzword' across research in natural science, engineering technology, humanities, and social science, involving linguistics, cognitive science, philosophy, brain science, clinical medicine, computer science, and many other fields (Gu, 2015a: 448). Multimodality can be interpreted in different ways and examined by different approaches in different domains. More information about this is provided in Chapter 10.

The term 'multimodality' in this book originates from brain nerve science. Modern brain science research prefers to use the term 'modality' to refer to general sensory organs and their interconnected neural networks (see Kolb & Whishaw, 2005: 135). Many significant results have been achieved in modern neurology and embryology research on multimodal sensory systems. For example, the visual modality shapes if the interconnected neural system dealing with visual signal processing is connected to the eyes. Vision, hearing, touch, smell, and taste are the most conventional multimodal sensory systems that are often investigated. In this regard, 'modality' is defined as how humans interact with external environments (e.g. people, animals, machines, objects) through sensory systems (e.g. vision, hearing, touch). Kolb and Whishaw (2005: 135) pointed out that sensory modality can have one or more functions. For example, body sense includes tactility, pressure, joint sense, pain and thermoception. Only when perception and production modality[1] work together can a complete verbal communication proceed. For example, the speaker gives utterances (including prosodic features) through vocal modalities, and the hearer receives this information through auditory modalities. If the interactants employ gestures like facial expressions, hand movements, or postures, other channels, such as vision, touch, and body sense, will also be included. Overall, the interaction between physically and mentally healthy people and the outside world (including interpersonal interactions) is multimodal or at least bimodal, particularly face-to-face situated discourse. For example, experienced doctors of traditional Chinese medicine (TCM) usually adopt the method of 'inspection, auscultation and olfaction, inquiry, pulse-taking, and palpation' during their diagnostic interaction with patients, involving multiple modes ranging from vision, hearing, taste, and vocalization to touch.

The situated conversation is a kind of multimodal interaction where people use their multimodal sensory systems to encode and decode the outside world's meanings. Evidence of this can be found in neurology. With the help of functional magnetic resonance imaging (fMRI), event-related potential (ERP), positron emission tomography (PET), electroencephalography (EEG), and other techniques, researchers have proved that different functional areas of the brain control the processing of different information in discourse activities. For example, the process of emotional prosody at and above sentence level is typically lateralized to the left hemisphere (Robin, Tranel, & Damasio, 1990), mainly located in the inferior frontal gyrus, superior temporal gyrus, and parieto-occipital region. There are also studies in brain imaging that indicate that both the left and right hemispheres of the brain are activated when processing speech and prosody, i.e. the right hemisphere in charge of the production and understanding of emotional cues in verbal communications is dominant in the knowledge of emotional prosody (Ireland & Tenenbaum, 2008: 213; Whitaker, 2010: 225), while the left hemisphere mainly deals with phonological processing (some scholars are sceptical about such distinctions; see Wharton, 2009: 166). Besides, the motor area involved in the bodily movements is found to be localized in the front half of the brain (Ireland & Tenenbaum, 2008: 211), and the front area of the left hemisphere handles positive emotions, positive intervention, or forward movements, while the front area of the right hemisphere is relatively more involved in negative emotions and avoidance behaviours. Research using modern brain imaging technology indicates that when a speaker is engaged in a situated conversation, many regions and the corresponding information processing systems in his/her brain operate simultaneously. These interconnected brain processing regions complement and support each other to carry out multimodal processing in people's interactions in situated conversations or with the outside world.

Fruitful research achievements in modern neurology and embryology have encouraged many scholars to develop 'multimodal senses-based' linguistic theories (or "multimodal conception" in Chapter 10). The so-called 'multimodal senses-based' implies a default premise – the occurrence and development of language require the support of the multimodal sensory system to construct linguistic theories. Studies based on multimodal sensory systems include Piaget's empirical constructivism, Karmiloff-Smith's cognitive development, Johnson's embodied philosophy, and Gu's modelling of 'saturated experience' and 'saturated significance' (Gu, 2015a). Guided by such theoretical beliefs, language research should take multimodal interaction between people and the outside world as an essential data source. We use situated conversation, which provides saturated data as a multimodal interaction between participants and the outside world, known as 'saturated significance from saturated experience' in Gu (2009). Essentially, such multimodal research belongs to behavioural research.

It should be noted that although the definition of 'modality' adopted in our study comes from brain science or physiology, in actual practice, our attentions are centred on the process where the hearer perceives the speaker's multimodal output through diverse clues, while the neural mechanism of multimodal production and perception is not included. Therefore, particular emphasis should be laid on speakers' multimodal cues when they perform illocutionary forces in situated discourse, for they serve as the basis of our investigation.

2.2 Situated discourse, situated experiencing, situated cognition, and multimodal data[2]

Now let us put situated discourse into a process to look at its multimodal nature. The subject of a situated conversation is a person with vivid sounds, emotions, and gestures, which Gu (2013) regarded as a 'live, whole person.'[3] This 'live, whole person' is the experiencer in a situated conversation, whose experiencing is closely related to his/her situated cognition. The so-called situated cognition refers to the cognitive activities in which the cognitive subject interacts with the outside world through a multimodal sensory system in a real spatial and temporal setting, concrete and specific (Gu, 2016). In this view, language is formed as the speaker makes utterances, and fades away as he/she stops talking. Among peoples whose cultures are passed on by word of mouth, children who have not learned writing, or illiterate people, their language used in verbal communications resides in their authentic experiencing, which manifests itself most commonly in situated conversation. So it is easy to understand that the language experience of a 'live, whole person' in a situated conversation is a kind of total saturated experience. The process of verbal communication involves the joint participation of multiple senses through which total saturated signification is constructed (Gu, 2009: 435–437). Gu (2016) referred to this as the 'principle of multimodality and saturatedness.' From the perspective of everyday behaviours, situated discourse, generated in people's saturated acts and the conceptualized environment, is the most fundamental form of meaning expression that existed before any other medium, including writing (Giddens, 1987: 91–92).

If we regard situated cognition as the process of continuously enriching the speaker's experience with information, then the experiencer would be a big integrated database, the experiencing process would continually add data to it, and the data in his/her possession would turn into experiences. Based on this, Gu (2016) proposed a '3E model' (Experiencer, Experiencing, Experiences) for the research of language experience. This model claims that 'the experiencer experiences the internal and external environments to give birth to personalized experiences.' Therefore, 'the experiencer-experiencing-experience is lifelong and non-stop from womb to tomb. Both experiencing and cognition are framed and enabled by multiple sensory organs, and the ever-experiencing experiencer is engaged in situated cognition' (Gu, 2016: 475–476), and his or her situated discourse is conceived of as dynamic complex systems.

Certainly, humans can break the constraints of time and space. Humans process information through multimodal sensory organs that produce multimodal memories. This kind of memory is stored in images (by image, Damasio (1999: 9) means 'a mental pattern in any of the sensory modalities'), with little reliance on linguistic symbols or other signals. Afterward, they can tell others their experiences by re-encoding them in language. Language makes people communicate with peers, taking modality as medium, and can express any modal data in its content. But 'the language faculty is not tied to specific sensory modalities'(Chomsky, 2000:121) and therefore it is 'amodal' in nature. Fresh insights in human language, cognition, and communication relationships can be gleaned from revelations in human multimodal sensory organs, multimodal memories, coding systems, the total saturated signification of situated discourse, and the 'amodality' of language faculty (see Gu (2016) for detailed analysis).

Language experiencing plays a different role in different linguistic theories. Here are some examples. Personal experience is filtered out in Saussure's structural linguistics, for he clearly stated that linguistic attention should be paid to *langue* (an abstract system of signs) instead of *parole* (the discourse interlocutors are experiencing). Halliday and other systemic functional linguists believed that language encodes language experiencing. Lakoff-Johnson from the cognitive linguistic school regarded language experiencing as 'embodiment' and 'metaphorization.' Chomsky's formal linguistics declared that language experiencing plays a triggering role in children's language development.

When studying situated discourse, we should record the speaker's multimodal saturated information to the utmost. At present, situated discourse can be converted into computable audio-visual digital streams through modern image technology, which forms 'multimodal linguistic data' for research. Strictly speaking, such records cannot preserve the saturatedness and richness of the original data in its fullest form. Once the situated discourse is 'experienced' by the interlocutor, it may tuck away into the interlocutor's memory. Even though it can be recorded in various forms, such as video, audio, or written forms, the recorded data has lost its saturated signification in real-life communication. Since 'multimodal linguistic data' recorded by video means belongs to the 'situated discourse' that has been 'experienced' by the 'live, whole person' (namely, the experiencer), it is no longer entirely in accordance with the original situated discourse but turns into experiences.

In brief, the research of situated discourse should be based on experiencer's multimodal sensory organs, shored up by the theoretical framework of human cognition, and delved into from total saturated experience and total saturated information. It should fully represent people's saturated experience of the multimodal interactions in situated discourse, and multimodal data should be collected for comprehensive and pooled analysis. Of course, multimodal data has broad coverage. In addition to audio and video data, data of assorted types generated in the speaker's verbal interactions can be collected from various dimensions, e.g. ERP, eye trackers, fMRI and other high-speed digital imaging systems can be used to collect the real-time acoustic and physiological data presented by the speakers in the situated discourse. Such data can be built into an extensive data set prepared for later research on eye movements, tongue positions, vocal cords, vocal tracts, breathing, emotions, activated brain regions, etc. From the perspective of big data, all information can be counted as the data source of language research as long as it meets the research aims and is amenable to being processed by the techniques.

What is the proper approach for studying illocutionary forces in situated discourse? In earlier passages here, the multimodal corpus-based approach is suggested. The following sections introduce what guidance and methodology multimodal corpus linguistics can provide to researchers.

2.3 Introduction to the multimodal corpus

For space reasons, this section cannot enlarge on all the relevant theoretical and technical issues of the multimodal corpus but introduces some common underlying problems. Elaborations on the application issues of the multimodal corpus can be found in some literature, and readers interested in that may check the references in this section for more information.

2.3.1 Some basics of the multimodal corpus

The last 60 years have witnessed corpus linguistics contributing much to many fields in linguistic studies. Meanwhile, corpus linguistics itself has developed successfully. Born in the 'pre-electronic' era, the corpus has speedily developed into an indispensable paradigm due to computer technology sophistication and researchers' more comprehensive outlook on language.

If the corpus manually built by researchers in the 'pre-electronic age' is called 'corpus 1.0,' then those corpora compiled with primitive computerized data would be called 'corpus 2.0.' In the 'corpus 2.0' version, researchers' hands are set free. Researchers can conduct large-scale data collection and processing and build large-scale computerized corpora with improved computer-processing capabilities. By so doing, the corpus data has expanded from texts to audios and videos, and 'corpus 3.0' was born (Knight, 2011a). At this stage, corpora are prevalent in linguistic research, even in many fields of humanities and social sciences study, raising their theoretical status to 'corpus linguistics.'

In the past decade or two, besides the large-scale studies of lexicogrammar on the basis of written corpora, 'there has been a consistent effort in the exploration of spoken discourse' (Knight & Adolphs, 2008: 175). Linguists began to pay increasing attention to the fact that verbal and non-verbal cues fit together into integrated messages in daily interaction. Additionally, with the technical development of modern computer multimedia and storage, and people's deepening understanding of the nature of speech activities, monomodal corpora are gradually evolving into multimodal corpora (Knight, 2011a: 9–13).

If the researcher collects audio data through a voice recorder and transcribes it into text, then two modalities are involved in data sampling. In this regard, the multimodal corpora's sample data is collected by recording, videoing, and text transcribing, involving three modalities. If researchers go on to use other techniques such as fMRI, ERP, and PET to collect brain activity data, more modalities are involved. For transcribed multimodal data, we can divide the interlocutor's activities into the 'content layer' and 'medium layer' (see Gu, 2006b). The so-called content layer refers to the video images streams of the camera's interpersonal interactions, while the media layer carries its content. Multiple media are indispensable to the presentation of multimodal content. For example, content linking to auditive modality is carried by soundwaves, to visual modality by lightwaves, and to gustatory modality by chemical molecular, etc. (Allwood, 2008: 208–209). These media need to be processed by various computer-processing tools (or synthesis tools that integrate diversified processing media). Thus, the camera's multimodal data carries the multimodal content and embodies the multimedia and is integrated to form multimodal texts. A multimodal corpus may cover various types of data, among which situated discourse plays a vital role. Examining the situated discourse with a multimodal corpus-based method, researchers have to record its relevant information to the utmost by resorting to various technologies, i.e. to sample as much authentic data from the situated discourse as possible. They also have to make full use of the multimedia to record the interlocutor's multimodal content to present the researchers with an intimate picture of how the multimodal interaction proceeds and simulate the saturated signification of the situated discourse. On the current technical level, such multimedia generally refers to cameras and voice recorders that can capture audio-visual information, yet none of the current techniques can completely reproduce the saturated multimodal information. General research may choose video and audio devices that can objectively and continuously record various information, including sounds, gestures, space, and settings.

Therefore, the scope of the 'corpus' has been greatly extended. Print-based texts combined with different layouts of pictures constitute a 'multimodal text corpus,' and audios, videos, texts, and other forms of data can be developed into new forms of corpora. Generally, a multimodal corpus is defined as 'an annotated collection of coordinated content on communication channels, including speech, gaze, hand gesture and body language, and is generally based on recorded human behavior' (Foster & Oberlander, 2007: 307–308). Multimodal corpora are an integration of multiple information, including audios, videos, and texts, whereby researchers can process, retrieve, and conduct statistical analysis of linguistic data in multimodal ways. In this sense, multimodal corpora can provide rich and valuable information for both quantitative and qualitative investigation of language-in-use in different situations. An increasing number of researchers believe that 'multimodal corpora are an important resource for studying and analyzing the principles of human communication' (Fanelli et al., 2010: 70). With such a background, the multimodal corpus comes into being as 'corpus 4.0' (Knight, 2009: 16–28), which provides a richer representation of different aspects of discourse, including participants' utterance content, prosody, and gestures, and also the context or co-text in which interaction takes place.

Over the years, academic conferences, special collections, and other academic resources overseas have grown relatively mature in the basic and applied research of multimodal corpus. Judging from the themes and content of these academic conferences, journals, or collections, overseas research on multimodal corpus has transcended by far the traditional level of data collection, transcription, processing, annotation, construction, and verification, and has dug deep into the correlations between various resources such as speech, prosody, facial expressions, gestures, and posture; progress has been made in the development of tools for qualitative analysis or quantitative data extraction. In addition to linguistic experts, scholars in education, psychology, anthropology, sociology, physiology, AI, and other fields are involved in its research. Many research institutions strongly support multimodal corpus application in interdisciplinary research where abundant innovative results have been produced. With state-of-the-art research techniques of high applicability, the basic and applied research of the multimodal corpus are organically combined. In contrast, Chinese multimodal corpus research and related studies in the linguistic domain are quite young, and interdisciplinary cooperation is not commonly seen.

2.3.2 Building the multimodal corpus

Though multimodal corpora are still in their infancy and few large-scale corpora have been published or are commercially available due to the high cost of construction and copyright restrictions, a range of multimodal corpora have been developed using different scales, registers, and languages (see Knight, 2011a: 5–8, 2011b: 392–393). These corpora have been used for a variety of research purposes, including language perception and production and HCI. Examples include the UCLA Library Broadcast NewsScape for news programmes from the USA and around the world, AMI (Augmented Multi-party Interaction) Meeting Corpus for developing meeting browsing technology, Multimodal Instruction Based Learning Corpus for teaching and learning discourse, and SmartKom Multimodal Corpus that is HCI-based. Many European countries have achieved plenty of research results in theoretical basis, data collection, data processing, data annotation, and analytic framework by developing and launching research on multimodal corpora extensively (see Bernsen & Dybkjær, 2007; Kipp, Martin, Paggio, & Heylen, 2009).

The construction and research pertinent to multimodal corpora in China also thrive. Currently, the largest-scale Chinese multimodal corpus is the multimodal corpus affiliated to the spoken Chinese corpus of situated discourse in the Beijing area (SCCSD BJ-500) (Gu, 2002b), which now contains several subordinated branch corpora, including the Multimodal Corpus of Gerontic Discourse (currently being developed by the author and Professor Yueguo Gu), the Multimodal Corpus of Children's Language Development, and the Multimodal Corpus of Court and Criminal Investigation Discourse. Other research projects or doctoral dissertations have also created some multimodal corpora, such as the 'Corpora of English Education in China (CEEC)' built by Professor Anping He of South China Normal University, the multimodal corpus of oral English for Chinese science and engineering majors constructed by Liu and Pan (2010), and Multimodal Corpus of Classroom Teaching (MCCT) by Yuxiang Li of Tongji University. But in general, since the multimodal corpus development is still in its infancy in China and the construction of multimodal corpora is more time-consuming and labour-intensive than that of text corpora, the current multimodal corpora still trail behind the text corpora either in terms of the total data amount or in the size of a single corpus. Also, completely open or commercially released multimodal corpora are rare.

The construction of the multimodal corpus involves theoretical and technical issues. In comparison, on the one hand, the multimodal corpus shares similarities with the traditional text corpus, while on the other hand, it features a distinctiveness from it. This section mainly introduces some basic knowledge about the construction of the multimodal corpus.

Wittenburg (2008) and Allwood (2008) expatiated on the working definition, construction methods, and related theoretical issues of multimodal corpora, including discussions on data collection, data processing, and data storage. Knight (2011a: 20) provided valuable suggestions for multimodal corpus construction by analysing its status quo from the perspectives of the corpora's size, coverage, naturalness, shareability, and reusability. Adolphs and Carter (2013) introduced the issues related to monomodal corpora, multimodal corpora, and their transition from monomodality toward multimodality. More updated discussion on multimodal data and multimodal corpus can be found in Bateman, Wildfeuer and Hiippala (2017) and Norris (2019).

A series of Gu's studies (1996, 1997, 1999a, 1999b, 2002b, 2002c, 2006a, 2006b, 2009) have built from the nature of situated discourse a theoretical framework for processing multimodal texts (video stream data), analysing multimodal texts, and modelling multimodal data, which set a crucial theoretical basis for the collection, processing, annotation, and analysis of multimodal data in situated discourse.

2.3.2.1 Data collection

1. Recording devices. Both video cassette recorder (VCR) and digital video camera can be utilized for video recording. The former collects analogue signals, but part of the information will be missing during the process of digitization, where analogue signals are converted to digit value by connecting the VCR to a computer through a data cable. Nowadays, digital video recorders usually have a high-sampling rate that can meet general research requirements. Therefore, it is recommended that all digital high-sampling rate video recorders be used through which digital data can be directly imported into the computer, avoiding the loss of digital information in the middle stages. Certainly, technical standards for collecting, storing, and processing data in different multimodal corpora can be formulated according to actual research needs. For further common issues related to the techniques and standards of collecting and processing multimodal data, see Wittenburg (2008: 664–684).
2. Filming angles. Researchers must adjust the filming angles properly according to different data sources and research purposes in data collection. Generally speaking, the interlocutor's facial expressions and body movements are essential information in multimodal language research. When collecting the data, if the speaker is performing some tasks, in principle, continuous wide-angle filming should be adopted; that is, to capture the interactional process where all participants and the speaker's whole body are involved to make sure all the speech act activities and multimodal clues given by the speaker are fully covered (Allwood, 2008: 215). If the speaker is sitting while talking, and the lower body shows no apparent body movements, the upper body can be photographed in close-up. Should it be a multi-party conversation, multiple filming angles should be selected according to the situation. Conditions permitting, multiple cameras can be used to film simultaneously, some for panoramic scenes, and some for close shots. Figure 2.1 shows some examples of the common filming angles:

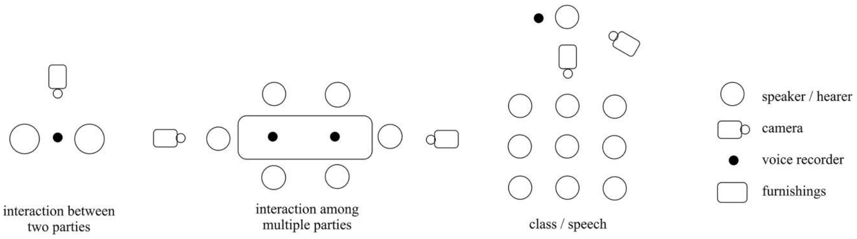

interaction between two parties
interaction among multiple parties
class / speech
speaker / hearer
camera
voice recorder
furnishings

Figure 2.1 Filming angles.

3 Naturalness. Naturalness refers to the data's authenticity, whether the speaker talks with a plan or script beforehand, or is completely spontaneous. To some extent, many factors affect the naturalness of the data, including recording conditions, settings selection, the use of disruptive equipment, the speaker's psychological tension, and the degree of interest correlation. Of course, this does not mean all data in multimodal corpora must be totally natural. Its degree of naturalness is contingent on the corpus-based research content and aims and is often closely related to the given situations. Knight (2011a: 26–27) plotted four scenarios for the collection of multimodal data, and each reflects different degrees of naturalness: (1) in highly conditioned situations, in studios, for example, when the speaker reads the pre-prepared script during data collection; (2) in partially conditioned situations, such as when the speaker has to complete a specific task or is involved in conversations in academic settings (such as lectures and tutorials); (3) in informal contexts, when massive audio and video data is collected without any preparation or instructions, such as in a family where the speaker can move about freely; and (4) in an unrestricted live recording, for example, in scenes from the speaker's life or work, when real-life data can be collected. Besides, the overuse of equipment would make the speaker nervous no less than the presence of researchers or data gatherers, which would further affect the data's naturalness. Despite small recording equipment as a result of technological development that can significantly reduce the unnaturalness caused by equipment, there still exists an inherent contradiction between the all-round collection of natural data and the minimization of intervention in conversation, yet the two vary in magnitude.

4 Collection records and metadata. During the collection of corpus data, we should fill in the corpus collection information card with such information as the recorder, interlocutor (e.g. name, gender, age, occupation, hometown, or accent), the time, place, reason, and purpose of the conversation, setting, background activities, and recording equipment. This information is known as metadata, which is a structured description of data or information resources. Metadata is applied in the corpus construction, aiming to make an objective, complete, and standardized description of language resources. It fulfils three specific tasks (Fu & Song, 2005: 574–575): Firstly, it gives an overall account of the corpus, including its language, date of construction, data sources, the gatherer's information, data textural features, and coding format. Secondly, it describes the language resources target and how these resources are managed, e.g. ways of expression, media type, and storage forms. Thirdly, it annotates the data or information to facilitate data retrieval, which is related to the research aims, corpus application aims, data display forms, and understanding of the information contained in the data. Recording such information and metadata is of considerable significance, because not only does multimodal research need to take into account

the background information (e.g. discourse production environment, speaker), but building the corpus needs to consider the sampling and representativeness of the corpus data (Gu, 2002b).

5 Ethical issues on the collection and use of corpus data. In light of the particularity of the multimodal corpus, researchers need to ponder relevant ethical problems when building the corpus (Adolphs & Carter, 2013: 149; Knight, 2011a: 50–54), including whether the speakers agree to have their images recorded by the researchers, and whether they agree to make them public for the needs of academic research. Generally speaking, researchers should ask for permission from the speaker through an oral or written form before data collection. The speaker's facial expressions, actions, and postures are a part of the multimodal corpus data, so researchers have to make a trade-off between protecting the speaker's privacy while fully displaying the relevant information. When publishing certain research results, researchers may resort to technologies to protect the speaker's privacy by blurring his/her face image while sticking to the premise on which the presentation of information would not be affected.

2.3.2.2 Data transcription

The benchmark and methods for natural discourse transcription have been a heated topic for many scholars. In the 1970s, the conversation analysis (CA) school has discussed issues concerning the transcription of natural discourse, which enriched the research methodology of ethnography. In the 1990s, with computer technology development, discussions extended from transcription issues to linguistics, sociology, anthropology, ethnography, psychology, etc. (O'Connell & Kowal, 2009).

The transcription of multimodal data, on the one hand, shares similarities with that of traditional monomodal data (usually for speech corpus), but on the other hand, a multitude of differences exist between the two. Edwards (1992) pointed out that data transcription should take into account the conditions of occurrence (e.g. settings, activities, interrelationships between participants), and manner of speaking (e.g. cadence, pauses, sound quality). He added that the hearers, discourse content, times (e.g. length of pauses, sequence of events, turn-taking between different speakers, utterances, and gestures overlapping timing), certain meta comments, and explanatory 'annotations' should be recorded. Garside, Leech, and McEnery (1997), Allwood et al. (2003), and Gu (2006b) emphasize the differences between transcription and annotation. Data transcription deals with the presentation of information that can be perceived through sensory organs from direct observation. At the same time, data annotation refers to the selective processing and presentation of information according to the field the researcher engages in and the theory he/she follows, which can be defined as the process where information is transformed into data (Gu, 2006b). That means that data transcription does not just record the discourse content in words, but records all kinds of information that can be directly observed through human senses, so that researchers can understand the conversation objectively and accurately. However, the information carried by multimodal data is much more vibrant than that of monomodal data (e.g. speech corpus or text corpus). Therefore, researchers generally transcribe it selectively according to the actual research needs (Cook, 1990) rather than transcribing everything. In theory, no such transcription system can fully express real-life conversation, particularly the multimodal data of situated discourse. In other words, any forms of data transcription lose some of the messages carried by the original information, more or less, and hence the total signification of the situated discourse cannot be fully presented through a monomodal carrier.

There are multiple transcription systems for natural spoken information, including those proposed by John W. Du Bois, Konrad Ehlich, John Gumperz, Norine Berenz, Gail Jefferson, and Brian MacWhinney. Scholars agree on what information to transcribe but have different opinions on how to perform it (see O'Connell & Kowal, 2009). In general, currently, there has not formed a so-called 'standard' method or system for data transcription in the development of the multimodal corpus (Cameron, 2001: 43), and transcription methods vary from one method to another. In fact, the forms and methods of transcription are directly related to the researcher's theoretical postulates and the research objects. How much data to transcribe and how to transcribe it depend on different purposes and natures of the research. In this regard, Meyer (2002: 72) pointed out a compromise approach, in which researchers may first accurately take down what people say in the conversation, and then add other information using the resources available.

2.3.2.3 Data segmentation and annotation

Here, segmentation does not mean partitioning the data physically according to the length of time but refers to the definition of boundaries between analysis units in a specific study. Given the particularity of multimodal corpus, data segmentation is a sticking-point when handling multimodal data. For the multimodal corpus in situated discourse, corpus data is often located in an interaction where speech and actions are intertwined. Thus, the segmentation of corpus data involves structural issues. The segmentation of situated multimodal data can be carried out according to the structural layers of the speaker's discourse and different research needs.[4]

Gu (2006b: 214–222) proposed several crucial clues for data segmentation, including spatial-temporal settings, social roles/functions/purposes realization pattern, consistency of situational purposes, asymmetry in purpose realization, participant gestures, and sound quality. Human interaction is deemed a set of various behaviours that one conducts with 'social actors' in a social context, including talking, doing, and/or talking and doing (Gu, 2006b: 138). Consequently, when labelling the data, except for demonstrating the hierarchical structure of multimodal texts on the labelling interface, a thorough presentation of the sequence, starting and ending points, and connotations of talking and doing, as well as other related information (e.g. emotions, intentions) should also be provided, in that such annotations will have a direct bearing on the subsequent analysis of that episode. Content labelling should be selected within a certain predetermined data range to build a systematic and standardized labelling system. In the practice of multimodal data segmentation, due to different research purposes, researchers may select different layers of multimodal texts to set up segmentation units.

Leech (1993) proposed seven principles for data annotation, serving as broad guidelines for general corpus construction. In addition to complying with the general corpus common rules, the annotation of multimodal data also has its own particularities. Some annotation frameworks with universal relevance have been developed, such as the linguistic annotation framework developed by the International Organization for Standardization (ISO), which attempts to provide a possible solution that is referable, comparable, and adjustable according to specific research needs for the annotation of various corpus data. Additionally, some other research project groups have also developed a variety of distinctive annotation standard systems (see Wittenburg, 2008: 673–677). However, any annotation scheme is determined to serve the analysis of linguistic facts in various research objectives, handling the corpus data from the view of pragmatism. Compared with text corpora, the annotation of multimodal corpora is quite special. Generally, it refers to the information annotation of the multimodal content carried by the audios and videos. It not only contains the annotation of traditional text corpora, such as vocabulary, syntax, semantics, and prosody, but also includes other non-verbal information, such as hand/head movements, facial expressions, gaze movement, and postures or bodily movement.

Some scholars such as Blache, Bertrand, and Ferre (2009) launched the Tools for Multimodal Annotation (ToMA) project in an effort to explore a general annotation framework for the annotation of multimodal corpora. Previous research results proposed a relatively comprehensive multimodal annotation framework, which involves metadata, morphology and syntax, phonetics and prosody, gestures, and text/discourse analysis. Gestural analysis includes facial movements, gaze directions, expressions, and hand movements. Besides, attention should be centred on the gestural annotation system established by Mcneill (1992), as it further stratifies the bodily movements from multiple dimensions.

The author believes that methodologically, the annotation of multimodal corpora matches the stage of data modelling in the simulative modelling[5] approach. When observing a complex matter, researchers cannot present all the information in an all-around way, so they usually process and present a certain amount of information each time. With the help of modern imaging technologies, massive amounts of information can be captured throughout multimodal interaction (information sources), which has various uses. After the research questions are framed, and the researchers construct the relevant hypotheses, they need to dig into and organize the targeted information, which will serve as the foundation for question analysis. Then, such information is transformed into data. Considering the simulative modelling approach behind multimodal corpus linguistics, the creation of annotation layers in data modelling depends on concept modelling's investigating perspectives, whose diversity plays a decisive role in deciding from what view and to what extent the data is processed and information is extracted in terms of corpus data.

Generally speaking, it is impossible and unnecessary for researchers to convert all the multimodal corpus information into data. Instead, they should design a reasonable and adequate data mining scheme (namely, an annotation scheme) that adapts to the research object and concept modelling needs. Multimodal text research follows *the principle of multiordinality* (see Gu, 2006b: 129), which correspondingly manifests itself in the multiple-layered approach in multimodal text annotation, i.e. creating various annotation layers on the annotation software, and annotating the corpus data from different perspectives and aspects.[6]

Meanwhile, the annotation of the multimodal corpus data also corresponds to the implementation and evaluation[7] stage in simulative modelling. Before annotating a large scale of linguistic data and building a corpus, the validity, reliability, and consistency of multimodal corpus data annotation should be verified (Cavicchio & Poesio, 2009). Here, validity refers to content validity that assesses whether such annotation can fairly represent all the information required in the research, or whether the intended information is labelled. There are no statistical procedures or verification formulas available for checking the content validity. Still, annotation schemes' content validity can be verified according to experts' grading or comments on how much they are satisfied with it (through interviews or questionnaire surveys) (Huang, 2012: 59). This study also adopts this approach. Reliability is used to measure the authenticity of the annotation process and results, i.e. how accurate the results are. Consistency entails intra-annotator consistency and inter-annotator consistency. The former refers to the degree to which the annotation results on the corpus data done by the same annotator at different times are consistent, while the latter refers to the degree to which the annotation results on the same corpus data done by two or more annotators are consistent.

After completing large-scale annotation and construction of the corpus, researchers can arrange and display various categories of multiple modalities on a window interface or on an operation platform under the corpora's data integration function, which is convenient for observing and analysing the interactions between modalities (Gu, 2006b).

Besides, the markup language of the corpus should comply with international standards for subsequent utilization and extension. The most common generic markup language is Extensible Markup Language (XML), a metalanguage used to create markup languages. Most of the corpus data (for both text and multimodal corpora) is annotated by XML, a global universal language. The 'multimodal corpus of Chinese situated discourse' in this study is annotated by Elan, a multimedia annotation tool with an XML-based data format.

2.3.2.4 Corpus building and sharing

The last step in building a corpus is to consider how to present annotated information to serve as a reusable resource for related research and subsequent research for other purposes. Shareability and reusability define the extent to which researchers can capitalize on the compiled multimodal corpus. Gu (2002b: 489) put forward archive and corpus concepts, distinguishing the selected data for research from the collected corpus data. The so-called archive refers to data collection consisting of all the collected corpus data that fits into the definition of situated discourse. In contrast, the corpus is a research-oriented library compiled from linguistic data extracted from the archive according to certain standards. One may follow these three principles when sieving corpus data from the archives and building the corpus. Firstly, take into account the representativeness of the corpus data. The selected corpus data should include the interlocutors and cover different representative settings according to research needs. Secondly, take into account the representativeness of the research content. Try to cover the research content in an all-round aspect and select the representative corpus data. Thirdly, take into account the quality of the corpus data. Since the recording takes place in situated discourse rather than in a noiseless studio, researchers should try to select high recording quality data.

Different from traditional text corpus, up to now, there is no unified or standardized representation form for multimodal corpus. Adolphs and Carter (2013: 170–173) discussed a variety of methods for corpus representation, among which some representation forms seem easier for later retrieval and research. For raw corpus data that has only been transcribed but not marked, transcribed content can be synchronized in a certain time frame (e.g. three to five minutes) with audios and videos available to the researchers via retrieval. For corpus data that has been annotated, the transcribed content, audios, and videos can be represented synchronously through annotation software. An advantage of doing so is that researchers can get a clear picture of the annotated content and conduct uniform retrieval of marked files' plurality. Both the original text transcriptions and the documents processed by multimodal annotation software are included in corpora. For example, the author used Elan for annotation, and five file types were generated in the self-built multimodal corpus, namely: (1) Video files in MPG format; (2) audio files in WAV format; (3) Elan-annotated files in

EAF format; (4) Praat-annotated file in TEXTGRID format; and (5) text transcriptions in DOC format or subtitle files synchronized with the video.

Researchers can store the annotated multimodal corpus data by grouping it into independent folders according to certain classification standards and entitling it with names fitting the research needs and the principle of easy retrieval. In this way, a small multimodal corpus dedicated to a specific research purpose is constructed. For easy retrieval in the future, researchers can also restore the basic information about the small annotated corpus in Excel or Access, including the speaker, setting, and corpus name, and connect it to the annotated file by a hyperlink through which they can have access to the relevant files by clicking on that link.

Currently, multimodal annotation tools allow for a certain extent of retrieval. However, it should be noted that this kind of retrieval is based on transcriptions, not on images or videos, which is known as semantic retrieval. When researchers search for a certain linguistic form in the transcriptions using tools like Elan, that linguistic form will be shown in the search result's contexts, i.e. the text corpus's concordance. Alternatively, when researchers retrieve a specific annotation item, e.g. prosodic features (e.g. extended tone, rising tone, and falling tone) or gestures (e.g. nodding, waving, and smiling), that annotation item would be positioned in a certain range of the targeted corpus data.

Generally speaking, after the multimodal corpus has been built, it can be published in some form (e.g. storage device or internet) to share relevant corpus data with other researchers and improve its efficiency. However, at present, due to the time-consuming and laborious construction of multimodal corpora, compounded by issues related to speaker privacy and copyright, most multimodal corpora have not been fully open and shared on the internet, though a few corpora have been partially thrown open to the public.

2.3.2.5 Toolkit introduction to multimodal corpus tools

With the development of computer and video technology, research into multimodal annotation tools is piling up. Various annotation tools have been well received in different research projects and fields, such as Anvil, Elan, Semiomix, MMAV, EXMARaLDA, TASX, and MacVisTA. Here we do not go into detail, but introduce two of the most frequently used annotation tools, Elan, and Praat, and one metadata management software, Arbil.

Elan was developed by the Max Planck Institute for Psycholinguistics in Nijmegen, the Netherlands.[8] It is a visual, linguistic annotation tool to simultaneously annotate, analyse, summarize, and retrieve video and audio recordings before building them into a corpus. Elan supports video formats, including Media format, MPEG format, WAV format, MP4 format, and QuickTime format. In addition to the annotation function, Elan is also a search tool for information retrieval in the multimodal corpus basing on annotated texts.

Praat[9] is an annotation tool for phonetic research developed by the University of Amsterdam in the Netherlands, with which one can segment, annotate, process, and synthesize digitized speech signals. It also has statistical analysis functions and can generate various sonograms. Currently, many Chinese phonetics training courses introduce the application of Praat. An introduction to Praat and its use also can be found in Wang and Peng (2006); a publicly published book called *Language, Speech, and Technology, Manual of PRAAT (voice soft)*, authored by Xiong at the Chinese Academy of Social Sciences, is a primer with an exhaustive explanation of how to use it[10] in China.

Channels exist for data transmission between Elan and Praat.

Arbil[11] (Archive Builder) was developed in the programming language Java to create and manage IMDI format metadata. It is an auxiliary tool for multimodal corpus construction. This software supports the export of created metadata to Elan, linking it with the annotated corpus data, which improves the efficiency of metadata management. Before the multimodal corpus data is segmented and annotated by software such as Elan, researchers can use Arbil to create its metadata, such as data attributes, recording settings, and speakers. This data should be stored in IMDI format before being imported into Elan. Arbil also supports downloading metadata templates from the remote corpus, which after modification can be saved directly and imported into the local corpus. Also, with Language Archive Management and Upload System (LAMUS), the metadata can be uploaded from the local corpus to the remote corpus.

2.3.2.6 Multimodal corpus construction: existing problems

Despite multimodal corpus research being a burgeoning field in recent years, many problems still occurred in the construction process: (1) Most multimodal annotation tools cannot conduct complicated statistical analysis between data (especially between different annotation layers) according to the researcher's needs, and real full-scale data integration functions and relatively powerful retrieval functions have not yet been implemented. (2) There are few widely accepted annotation schemes, and therefore further exploration is needed on the annotation of the multimodal corpus. (3) The enormous workload of processing usually limits the multimodal corpus size, so improving processing efficiency and taping into the valid data fully still requires further research.

In the current development of the multimodal corpus, the multimodal corpus of situated discourse complied by Gu (2006a, 2006b, 2009) and the agent-oriented modelling proposed show the outstanding progress in recent years (Feng, Zhang, & O'Halloran, 2014). Based on specific operational technologies such as XML language, RDF framework, and TML language, agent-oriented modelling language can help build models for social activities and agentive relationships in multimodal texts and build a corpus. Based on Holland's research on genetic algorithms in biology, agent-oriented modelling and simulation is a practical approach for studying intricate systems. After that, it rapidly spread to other fields, including economics, sociology, and military science. Nowadays, the philosophy of the agent is very active in various research and application fields. It evolved in the first place from a technical solution to a way of thinking (Liao, Wang, & Zhang, 2015: 9–10). Agent-oriented modelling (AOM) and simulation have been well received and play an active role in sociological research. 'People' in the social system and the 'agent' are similar in nature (Liao, Wang, & Zhang, 2015: 11).[12] As mentioned above, constructing linguistic theories based on multimodal senses regards the interaction between humans and the outside world as a kind of human behaviour. Such multimodal research belongs to behaviour research. With this knowledge, it is feasible for researchers to adopt agent-oriented modelling and the simulation approach to model, simulate, and extract the features of verbal communication. Thus, AOM has become an essential approach to the construction of a multimodal corpus, which serves as an important data source for multimodal interaction studies.

It can be observed that the multimodal corpus is a novel type of corpus upon which researchers can base the investigation of verbal communication and provide traditional linguistics with fresh research paradigms and improved theories. Based on the multimodal data, multimodal corpus linguistics is a subset of corpus linguistics. Baldary and Thibault (2006) first used the term multimodal corpus linguistics; Gu (2013b: 4) set the ultimate goal for multimodal corpus linguistics as using all kinds of state-of-the-art technologies to collect all the multimodal data generated in the interaction between humans and the outside world, to adequately represent real-life human multimodal activities and to study the process of human interaction. In general, the basic ideas, basic features, and pursuit of corpus linguistics all come down to testing and developing linguistic theories based on linguistic facts. At present, the basic ideas and techniques about the multimodal corpus have been well received in many fields. For more information about the multimodal corpus, multimodal corpus linguistics, and its research methods and application fields, see Huang (2015c); here the author does not go into details.

2.4 Summary

Constructed in the experiencing process of the 'live, whole person,' situated discourse appeared before the invention of written language. It is multimodal in nature and has a close relation to human situated cognition. The investigation of situated discourse should be based on multimodal sensory organs, shored up by a theoretical framework of human cognition, examined from the starting point of total saturated experience and total saturated significance, while following the clues of simulative modelling. Its goal is to represent the total saturated experience in human authentic multimodal interaction to the fullest. In this sense, the multimodal corpus linguistics approach has been recognized as a concrete practice for exploring situated discourse.

The construction and related research of the multimodal corpus, known as the corpus version 4.0, has grown into a promising research field in corpus linguistics. Its construction fits into the general corpus overall patterns, but it also has its distinctive features. It brings fresh perspectives and new practices in linguistic research, extending linguistic research scope and unearthing new rules. Furthermore, the study of the multimodal corpus has risen to the level of multimodal corpus linguistics.

Notes

[1] McBurney (2002: 351) divided modality in interaction into perception and production.

[2] It is important to distinguish between 'information' and 'data.' Information is objective and saturated, which is impossible to be handled once and for all. If a certain amount of information is sampled from a certain angle and treated as the research object, it becomes research data. Therefore, data is the constructed information that serves the research question.

[3] The idea of the 'live, whole person' originated with Firth (1957: 19). Firth claimed that we should stop the binary opposition between mind and body, as well as thought and word, and treat the interlocutor as a whole human whose mind and acts are integrated as a whole and are socially connected with others. The linguistic study should be carried out based on the whole human pattern of living. Gu's STFE-match principle is also based on the above philosophy.

[4] See Section 3.2.1 for details.

[5] See Section 3.1 for more details about simulative modelling.

[6] More and more processing and annotation tools for multimodal data came out, e.g. ANVIL, Elan, MacVisSTA, DRS, and Exmeralda. See Allwood et al. (2003).

[7] See Section 3.1 for details about implementation and evaluation.

[8] Elan's homepage and download link: http://tla.mpi.nl/tools/tla-tools/elan/

[9] Praat's homepage and download link: http://www.fon.hum.uva.nl/praat/

[10] Download link: http://ling.cass.cn/yuyin/staff/praat_manual.pdf

[11] Arbil's homepage and download link: https://tla.mpi.nl/tools/tla-tools/arbil/

[12] However, Gu (2009: 445–446) pointed out that his comments on the 'agent' mainly come from Gibson's (1986) ecological approach on perception study, varying from the 'agent' definition in other fields, such as AI.

3 Illocutionary force study

Basic methodology and theory

3.1 Applications of Simulative Modelling

Since verbal communication is essentially a complex system, it should be examined under complex system paradigms, and researchers should pay attention to its elemental composition and the interactive relationships between each element. Faced with such a complex system, researchers may borrow the core methodology from Simulative Modelling, a common approach in the realm of multimodal corpus linguistics (Gu, 2013b: 4).

3.1.1 A complex systems view of verbal communication

A 'system' is a fundamental concept with diverse connotations, which is commonly used in multiple disciplines. The biologist Ludwig von Bertalanffy, the founder of general systems theory, claimed that a system is a complex of interacting elements (Deng et al., 2009: 2). The well-known Chinese cyberneticist Qian Xuesen defines a system as a whole that is composed of interrelated, interacting, and mutually influenced components, and has specific functions (Yu, 2014: 5).

According to the system's complexity, Qian divided systems into simple systems, simple giant systems, complex giant systems, and open complex giant systems. From the late 1980s to the early 1990s, Qian, Yu, and Dai successively proposed a 'meta-synthesis from the qualitative to the quantitative' and 'Hall for Workshop of Metasynthetic Engineering from the qualitative to the quantitative' to deal with complex giant systems in China. These methods focus on organically uniting qualitative and quantitative approaches throughout the entire research process, translating qualitative comprehension into quantitative comprehension. Also, they united different knowledge (scientific theory and human experience) and integrated various disciplines into meta-synthesis (Deng et al., 2009: 43–44). The philosophy of meta-synthesis has been applied in the field of social sciences.

From the perspective of diachronic evolution, language is a complex adaptive system. This indicates that some systems evolved from an elementary starting condition can lead to exceedingly sophisticated and mysterious

DOI: 10.4324/9781003251774-3

outcomes in certain environments. Concepts related to complex adaptive systems have expanded to many disciplines. Scholars draw on such concepts as phase transition, emergence, stability, and equilibrium to model, analyse, and predict various phenomena. From the perspective of language evolution, human communication originated from primitive and simple manual gestures or vocal calls, which gradually switched to a verbal communication system with diverse resources. This is called the phase transition of language. Of course, such a phase transition is not completed at one single time. In this systematic and dynamic evolution process, language continuously adjusts itself and adapts to various environments (Wang, 2006a: 5–6). Researchers cannot go back to tens of thousands of years ago to study language evolution to observe its emergence and evolution. Nevertheless, with computer technology, they can explore the origin and acquisition of language utilizing Simulative Modelling (Wang, 2006b). Therefore, the theories and tools used in studying complex systems provide help and valuable references for language origin (Wang & Ke, 2001: 197).

After a long process of complex adaptation and evolution, verbal communication became a complex system from a synchronic perspective, called a 'dynamic complex system.' The dynamic complex system theory was born in natural science research but has already been widely applied in physics, chemistry, mathematics, biology, and other fields. Researchers may launch their research on a dynamic complex system from 'qualitative' and 'quantitative' aspects. In qualitative analysis, dynamic complex systems have six core characteristics: Elements, interactions between elements, formation and operation, diversity and variability, environment, and activities (Bar-Yam, 1997: 5). In quantitative analysis, related research includes mathematical modelling, quantitative calculation, and big data analysis.

The author went out of his way to introduce dynamic complex systems for believing that human verbal communication is essentially a complex system of dynamics. One's experience can be regarded as a dynamic complex system, and one constantly fills up the system with data that arise from situated cognitions (Gu, 2016: 485–486).

Accordingly, the study of illocutionary forces in situated discourse is theoretically based on situated experience and situated cognition, and it can be carried out by adopting 'qualitative' and 'quantitative' approaches.

The 'qualitative' approach to illocutionary forces in situated discourse includes analysing their ontological properties and constituent elements – perspectives and aspects. After analysing the fundamental elements of illocutionary forces, we adopted modelling, a well-received approach in the research fields of dynamic complex systems, physics, chemistry, and mathematics for the methodology in this study.

The 'quantitative' approach includes using quantitative methods such as statistics to analyse each element and the interactive relationships between elements, mainly employed in the basic practice of multimodal corpus linguistics.

Whether complex adaptive systems or dynamic complex systems, their underlying logic is a 'complex systems paradigm,' containing a set of cognitive patterns and methodological standards developed by mainstream theorists in the research of various complex systems (Yin & Wang, 2016: 63). As complex systems deal with complex issues, researchers must have complex thinking to renew research paradigms. Important enlightenment that the complex system paradigm brings to scientific research is that attention should be focused on the systems' elemental and relational constitutions and the interaction between individuals and the environment. The formalized representation of such elements and the dynamic relations between them are quite crucial in the face of a complex system. The complex systems model says that 'those seemingly random and intricate system behaviours may be formed by interactions among a few simple variables' (Yin & Wang, 2016: 72). Therefore, figuring out the core elements and unearthing the interactive mechanism between them is vital for understanding complex systems' behaviour. That is why we represent the Conceptual Model of illocutionary force and examine the relations between IFIDs in the form of 'sets.' Researchers should also employ various research methods when dealing with complex systems, including qualitative and quantitative approaches. It is called 'mixed-methods research' in social sciences, whereby researchers collect and analyse data from both quantitative and qualitative dimensions within the same study to understand and interpret the research object more comprehensively (Creswell, 2014: 3–4). Some scholars have already applied mixed methods to linguistic studies.

In brief, verbal communication is a complex system where lexis, semantics, prosodic features, gestures, emotions, and intentions are interconnected to create a multidimensional interaction with those elements depending on and impacting one another. With this in mind, if someone were to study only verbal communication and ignore other interrelated dimensions, he/she would probably be caught up in a dilemma whereof not being able to 'see the wood for the trees.' An essential principle in the study of complex systems is that its structure and environment and relations determine the overall interconnected system and its functions. In other words, the overall system and its functions are the results of the integration of its internal structure and the external environment, echoing the phenomenon of emergence in complexity sciences (Liao, Wang, & Zhang, 2015: 55). The same is true in human verbal communication. In different social situations, people employ multimodal resources to express emotions, intentions, and meanings. In the following introduction, readers may notice that the illocutionary force performance in situated discourse is a linguistic phenomenon with complex system features. In this case, the coexisting multimodal resources (IFIDs) are interrelated and used by the speaker to perform illocutionary forces collaboratively under the influence of factors like social situations.

Noting that language is a complex system involving many elements, linguists should not be confined to isolated elements when exploring linguistic phenomena. It is understandable why researchers must focus on aspects at a particular stage due to research conditions and technology limitations. However, problems will arise if they treat the combination of some interacting elements as the composition of the entire system. How can we launch our study when facing complex systems? The core methodology of Simulative Modelling is introduced in the next section.

3.1.2 Definition of Simulative Modelling

Understanding things is a process of modelling. Modelling is a common and important method in understanding the objective world, especially in natural sciences and engineering technologies. Modelling is used mainly because the research objects are too complicated, and humans can only deal with limited information at a time. Therefore, people resort to modelling to make things less complicated for handling the essential information at once (Blaha & Rumbaugh, 2006: 15–16). Modelling is a process that builds a particular model in the end. A model is a theoretical description that can help people understand a particular phenomenon or product in advance.

Simulative Modelling is one of the commonly used methods in scientific research, which requires researchers to collect information as close as possible to what the object is. On this basis, they can select a particular investigating perspective, mine the data, and conduct research. When it comes to selecting investigating perspectives and the amount of data for modelling, it depends on the research aims and how researchers look at the object (Gu, 2009: 444). Theoretically, the angle of observation is unlimited. Another modelling method in contrast to Simulative Modelling is Product Modelling. The fundamental difference between the two is the former models on objective things or phenomena, while the latter models on something new that is defined, designed, and constructed by its developers. In this regard, Simulative Modelling should fully represent all traits of the objective thing or phenomena.

Situated interaction between an alive, whole person and the outside world is multimodal, producing saturated signification. There is no method available that can thoroughly and objectively reproduce such total saturated signification. As previously mentioned, Gu (2016: 475–476) points out that

> the experiencer-experiencing-experience is lifelong and non-stop from womb to tomb, which gives rise to big data that are developmentally and integratively accumulated…the situated discourse produced by the ever-experiencing experiencer, conceived of as dynamic complex systems, is approached to by way of simulative modelling.

Therefore, we can only borrow Simulative Modelling to simulate the saturated state of the signification of multimodal discourse and do our utmost to represent the whole picture of the research object. Situated discourse is very complicated, for it involves many phenomena and problems. It is difficult for researchers to investigate all these phenomena in one study. Generally, they only select several vital parts of the discourse for each study. As the speaker also considers a complicated object when he/she performs illocutionary acts/forces in situated discourse, issues like how and from what investigating perspectives to study live illocutionary forces need careful consideration. In this regard, Simulative Modelling can provide a valuable reference.

Simulative Modelling, the methodology of multimodal corpus linguistics, is applied to simulate various saturated significations produced in multimodal situated discourse in linguistic studies (Gu, 2013b: 4). Gu (2013b: 6) used 'breaking up the whole into parts' to explain the process of Simulative Modelling upon a 'live, whole person.' In this case, he referred 'the whole' to the 'live, whole person' and 'parts' to the modelling perspectives. From a single perspective, we can only observe one aspect of the whole person. In contrast, if we put all perspectives together, we will get a whole picture of that person. This is called data integration in modelling (Gu, 2009). We can model a live illocutionary force from a multiplicity of perspectives. First, we can break it up into parts to examine them, and then treat it as a whole by taking all perspectives into account. Understandably, limited by technologies, methodologies, and even necessities in actual practice, it is impossible for researchers to restore all original saturated data in the situated discourse completely.

In the multimodal exploration of situated discourse, the situated discourse produced by the 'live, whole person' is what researchers from various perspectives model upon, as shown in Figure 3.1.

After building a multi-perspective model of situated discourse, researchers have to extract various data from live speech. At present, the most common practice is to record the situated discourse on video and audio, whereby to build a multimodal corpus.

3.1.3 Procedure and application of Simulative Modelling

Simulative Modelling contains three stages: concept modelling, data modelling, and implementation and evaluation (Gu, 2013b: 4–5).

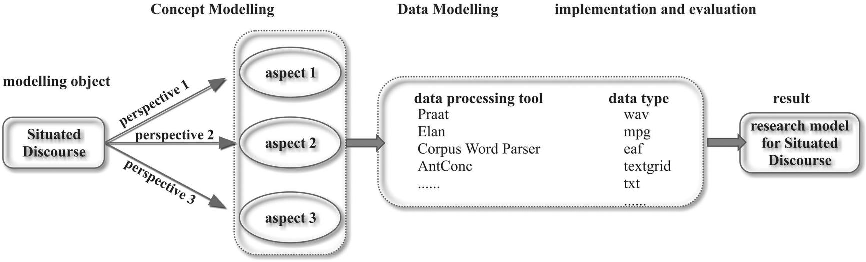


Figure 3.1 Simulative Modelling process of situated discourse.

3.1.3.1 Concept modelling

Concept modelling is a crucial method for software engineering (Gu & Zhang, 2013: 225). The conceptual model may refer to models formed after a conceptualization or generalization process of a system. It could be a mathematical, logical, or linguistic description of the object developed for specific research (Sargent, 2009: 164). Conceptual models can be described by the knowledge engineering approach and formal method, among which a formal method is a hotspot in line with the trend of current research. The so-called formal method is a technique that describes the target system based on mathematical concepts, methods, and tools (Wang & Wang, 2007: 783). Following Gu and Zhang (2013: 225), this study borrowed the mathematical concept 'set' as the form of concept modelling. Theoretically, all the individual and relational values are included in the subsets. However, limited by research techniques, perspectives, and capacity, researchers usually adopt a few modelling perspectives to establish the corresponding subsets, which further constitute a set.

Here we would like to distinguish between two philosophical concepts related to constructing the illocutionary force's conceptual model, i.e. types and tokens. The American logician-philosopher Peirce first made the critical distinctions in 1931. 'Types' embody the patterns of similar individual instances (tokens), or a type is a class of similar tokens. Nicholas Bunnin, an expert in contemporary Chinese philosophy from the University of Oxford, pointed out,

> A token is a particular, individual sign, a single object or event, while a type is the embodiment of the patterns of similar individual instances (tokens), or is a class of similar tokens. Type is not a single thing or event, and it can only exist through the representation of its tokens.
>
> (Bunin & Yu, 2001: 1030)

Goldman (1970) took this pair of notions to distinguish between human act-types and act-tokens.[1] An act-type is a kind of action, such as weeping, running, writing letters, or giving a speech. An act-token is a particular act or action performed by a particular person (agent) in a particular circumstance. An act-type is an action property, while an act-token is an exemplification of such a type. In other ways, people's actions vary greatly in actuality. Even in the same act-type class, different action instances may occur subject to the influence of agent, time, place, or environment. Likewise, we can apply Goldman's view to distinguish between illocutionary act-types and illocutionary act-tokens (Gu, 2013a). An illocutionary act-type is an action property of an illocutionary act in a speech community, while an illocutionary act-token is an exemplification of such a type in various situated discourse. Different speakers exemplify certain illocutionary act-types at different times and places in situated discourse, forming diverse illocutionary act-tokens. These illocutionary act-tokens retain many features of their corresponding illocutionary act-type, activity type, and social situation, while having their own characteristics.

As discussed above, the author believes that researchers cannot examine illocutionary act-types in isolation from illocutionary act-tokens. To understand illocutionary act-types, first of all, we should analyse illocutionary act-tokens in situated discourse. This can be achieved by concept modelling. Illocutionary act-tokens in situated discourse are abstracted and generalized from different angles through concept modelling. They form a number of sets that reflect the attributes or features of the corresponding illocutionary act-type, setting it apart from other illocutionary act-types. In other words, the object of concept modelling is illocutionary act-tokens, and the product of concept modelling is the description or depiction of the illocutionary act-types.

Concept modelling is the most fundamental step, which sets the basis for corpus-based multimodal data mining and analysis. The perspective of concept modelling is in accordance with researchers' investigating perspective of illocutionary forces. By describing a particular illocutionary act-type property in several ways, concept modelling defines and analyses it in the form of 'sets.' The concrete forms, establishment, and analysis of the illocutionary force's conceptual model is elaborated in Section 4.2.1.

3.1.3.2 Data modelling

The second stage of Simulative Modelling is data modelling, and its core task is to prepare the corresponding data types for the research perspectives provided by the conceptual model. Different research perspectives may require different data types. For example, in Chinese mythology, Shennong (Divine Farmer), a Chinese deity, identified hundreds of medical (and poisonous) herbs by personally testing their properties. The interaction between Shennong's eyes and the herbs constitute an investigating perspective, which needs to be supported by biological data. When light rays entered his eyes, they were converted into electrical signals. This process also involves other data types like biological and chemical. Besides, data modelling has to deal with a series of technical issues, such as metadata, data integration, and data exchange (Gu, 2013b: 5).

In this study, data modelling aims to prepare the required data types on Elan and Praat for the study of live illocutionary forces from the perspective of the STFE-match principle, i.e. what is said, what is thought of, what is felt, and what is embodied (see Gu, 2013a, 2013b). It should be noted that Elan and Praat were involved in the development of part of the data modelling process, such as metalanguage and data integration. To some extent, this determines which data types can be represented by such software, and which cannot, due to technological limitations. For consistency, our 'segmentation and annotation scheme of illocutionary forces in situated

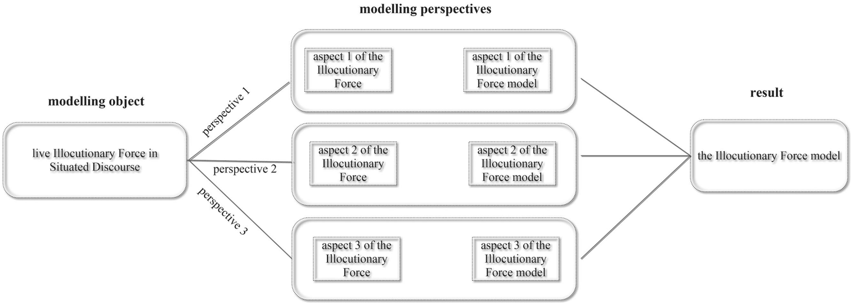

modelling perspectives
modelling object
live Illocutionary Force in Situated Discourse
perspective 1
perspective 2
perspective 3
aspect 1 of the Illocutionary Force
aspect 1 of the Illocutionary Force model
aspect 2 of the Illocutionary Force
aspect 2 of the Illocutionary Force model
aspect 3 of the Illocutionary Force
aspect 3 of the Illocutionary Force model
result
the Illocutionary Force model

Figure 3.2 Procedure of modelling live illocutionary forces from multiple perspectives.

discourse' plotted out which data type each research perspective needs. For example, for emotional states, we established 'emotional descriptors' as its data type. However, the identification of the data type is only a small part of data modelling. When annotators annotate the data on multiple levels with Elan and Praat, they are entering the stage of implementation and evaluation, i.e. using advanced digital technologies to handle modelling objects based on the data model. The segmentation and annotation scheme of live illocutionary forces based on concept modelling and the STFE-match principle is introduced in Section 6.1.

3.1.3.3 Implementation and evaluation

The third stage of Simulative Modelling is implementation and evaluation. In this stage, the previous data model is applied to the research object by digital technologies to determine whether the results are in accordance with the conceptual model. If any conflicts with the conceptual model or the data model arise from the testing, the established model should be modified. This is called the model-modifying process. In this study, this stage involves researchers' trial annotation of some corpus data, and experts' testing and assessment of the data samples marked under the multimodal segmentation and annotation scheme. After experts' evaluation, researchers can select the most typical data sample as the later annotation template. Discussions on the verification of corpus data annotation can be found in Section 6.4.

In this study, as the live illocutionary forces are modelled from multiple angles in concept modelling and data modelling, the modelling processes construct a multiple of simple models (see Figure 3.2). Specific perspectives and strategies of modelling are discussed in the following sections.

3.2 Discourse activities: methodology and modelling objects

The first chapter reviewed previous research on speech acts, which contains three distinctive stages. It also pointed out that such research failed to pay enough attention to cues like facial expressions, gestures, and postures by believing that language use is an activity in which people exchange information with languages and that the central issue in pragmatics is to reveal how language served the purpose of communication (Gu, 2010: xiv). In this way, facial expressions, gestures, postures, etc. are only regarded as paralinguistic or non-linguistic information. This hindered researchers from conducting speech act research from the perspective of general behaviour theory.

This section mainly discusses the relevant ideas and theories Gu put forward on situated discourse, which have been adopted as the theoretical framework and methodology in our study.

3.2.1 Social situation: research status and approach

Context has always been a critical factor in the realm of linguistic research, especially in pragmatic research. Traditional pragmatic research treats the social situation as the context of linguistic activities. In contemporary pragmatics, contexts cover a wide range of factors related to the expression of meaning in language, including the environment where a language is spoken, identity, social status, and gender of the speaker, linguistic and cultural backgrounds, ideology, and customs. Those factors are dynamic (see He, 2011: 2; Wang, 1983: 64). As pragmatics develops, the study of contexts presents diversified perspectives. Hu's (2002) analysis indicated that the connotation of 'context' has evolved from unitary (language environment and context) to binary (language environment and non-language environment), ternary (language environment, the physical environment, and shared knowledge), and diversification. The context does not solely refer to the 'co-text' in the past but has a broader connotation. It can be defined as 'the situation in which discourse is produced' (McCarthy, 1991: 64). However, many crucial factors in contexts were not fully explored in previous investigations for short corpus linguistic methods, owing to past technological and conditional limitations. Statistically, context was previously treated as a static concept in applied linguistics and corpus analysis, but the context of authentic speech activities is indeed dynamic, multidimensional, and multilayered. Many scholars have expressed their agreement on this (Knight, 2011a: 185–187).

It can be observed from the above that although the connotation of context has been expanded, much previous research still separates it from discourse activities. The rationale is that issues of language use are oversimplified down to sheer information exchanges. Gu (2010: xvi) proposed that language-use issues should be handled within the basic framework of general behaviour theory because language use and social activities are intertwined. In this sense, the social situation should not be regarded as a context in traditional linguistic studies.

Human interactions are 'social actions over space and time' (Gu, 2006b: 133). Logically, situated discourse is also based on a certain time and space as a kind of human interaction. Therefore, situated discourse recorded in

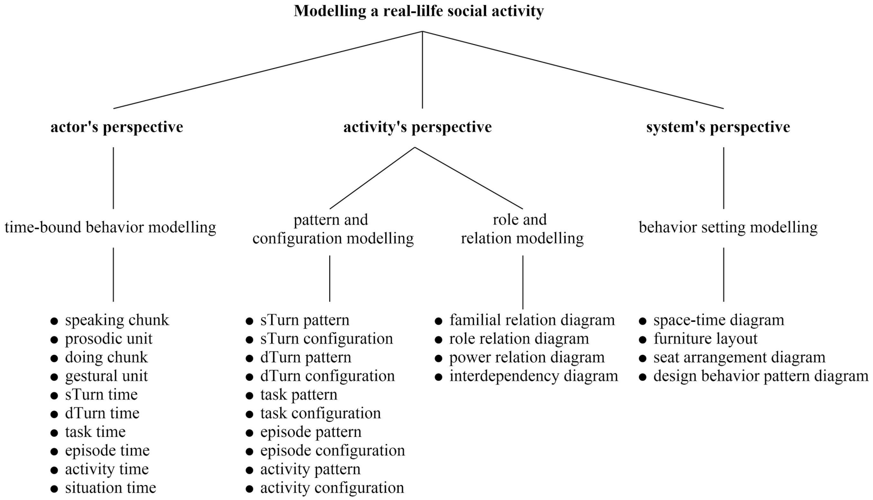


Figure 3.3 Three modelling perspectives.

multimodal texts could be analysed with situations-based structural analysis methods (Argyle, Furnham, & Graham, 1981).

Daily life is composed of various social situations where situated discourse is produced. The social situation is an intersection where the individual interacts with society (Gu, 2006b). When recording discourse activities in audio and video streams, we are taking segments that produce meaningful chunks out of continuous speech streams. For research purposes, multimodal texts recorded in audio and video streams can be seen as various social situations (Gu, 2006b). For example, a multimodal text that entirely recorded the 80th anniversary of the founding of Beijing Foreign Studies University in audio-visual streams can be treated as a complete social situation and a meaningful chunk.

Certainly, chunks can be structured. It depends on how researchers view the situated discourse, or from what perspective it is modelled. Theoretically, there are infinite modelling perspectives, and their selection depends on the researcher's research needs. Gu (2009) provided three possible modelling perspectives on discourse activities, namely the perspectives of the actor, the activity, and the system (see Figure 3.3). They are allowed to choose from them the perspectives that meet their research needs.

Researchers can conduct time-bound behaviour modelling from the actor's perspective, pattern and configuration modelling, and roles and relation modelling from the activity's perspective, and behaviour setting modelling from the system's perspective. When studying live illocutionary forces in situated discourse, Gu (2013b) adopted the actor's perspective, which was further broken down into four sub-perspectives, i.e. what is said, what is thought of, what is felt, and what is embodied to get a whole picture of the modelling object. This study also describes and analyses illocutionary forces in the adoption of the actor's perspective. Aside from that, the study also refers to the activity's perspective and system's perspective when analysing the influential factors of multimodal illocutionary force's expressive devices and IFIDs (see Section 8.4 for details).

If we look at social situations from the activity's perspective, each of these perspectives can be examined through structural analysis.

'Activity type,' proposed by Levinson (1979: 69), is a concept related to social situations. There are both similarities and differences between Levinson's 'activity type' and the 'social situation' proposed by Argyle, Furnham, and Graham (1981). While Gu (2006b: 127–167) believes that these two concepts are essentially similar, he pointed out that the distinction between them is that social situation is the general environment where discourse occurs, which may contain one or more activity types at the same time. He once gave an example that in the social situation of a real-life archaeological excavation site, four activity types might co-occur: Field personnel may be excavating the cultural relics, journalists carrying out interviews with the experts, photographers recording the scenes, and security guards keeping order at the spot.

An activity type is comprised of a series of tasks and episodes. Whether a specific action is a task or an episode depends on the objective of the activity type. A task is an action aimed to achieve the core objective of an activity type, while an episode is a minor action parallel to it (Gu, 2006b: 124). For example, in a conference, the host's introduction to the guests is a task while the waiters serving tea is an episode. Overall, the social situation and its subordinate activity types are composed of tasks and episodes and are objective-oriented.

From the perspective of human social interaction, stratification, social situation, activity type, tasks, and episodes mentioned above all constitute the sociopsychological layer. The next layer of the tasks or episodes is individual behaviour, which belongs to the individual behaviour layer, and it includes a series of talking, doing, speech acts, the prosodic units of illocutionary force, etc. (Gu, 2006b: 132). These layers are shown in Figure 3.4:

Gu (2006b: 201) believes that a social situation is not a context or background in the analysis of situated discourse as it was in traditional pragmatics. Instead, it should be treated as the highest level of the analysis units. Similarly, various units of lower social situations can also be analysed. For example, the illocutionary force/act at the individual behaviour layer is seen as an object in this study.

3.2.2 Gestures: research status and approach

As mentioned above, the agent of real-life discourse activities is a 'live, whole person' with dynamic sounds, emotions, and gestures. Likewise, the

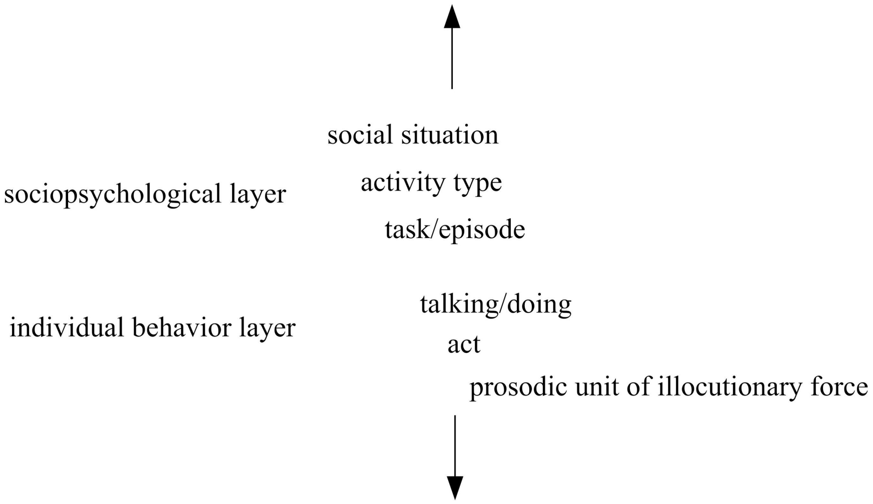


Figure 3.4 An analytic scheme of a multimodal text.

performers of a speech act that our study deals with are also live, whole persons, rather than those idealized talking persons (or 'talking heads' in Thomsen's (2010) words) who can only exchange information through language. If situated discourse is examined by Simulative Modelling, the whole person in the discourse activities will become the modelling object (Gu, 2013b: 6). Likewise, if the live illocutionary forces in the situated discourse are examined by Simulative Modelling, the live illocutionary forces produced by the 'live, whole person' would become the modelling object.

As prosody, gestures, movements, etc. are parts of the live, whole person's illocutionary force performance, they should be incorporated into the research scope and treated as significant as other analytic units such as discourse content and prosodic features. Theoretically, this is fundamentally different from facial expressions, movements, and postures as paralinguistic or non-linguistic information in past pragmatic studies.

So here comes the question: What perspective should we adopt for modelling live illocutionary forces?

3.3 STFE-match principle and analytic perspectives for live illocutionary forces

3.3.1 STFE-match principle and Simulative Modelling

According to the methodology of 'breaking up the whole into parts,' Gu (2013b) proposed four perspectives that represent the 'parts' in modelling a 'live, whole person' in situated discourse, i.e. what is said, what is thought of, what is felt, and what is embodied.

These four perspectives provide a fertile source for observation in this study, each holding a wealth of meanings (see Gu, 2013b: 6–8). What the speaker says involves utterance content and prosodic features, as well as various live illocutionary forces. What the speaker thinks of involves his/her thinking, attitudes, and beliefs about someone or something. What the speaker feels involves his/her emotions that fall into many categories. Moreover, what the speaker embodies involves postures, bodily movements, facial expressions, looking into the eyes, etc. In actual practice, the subdivision of these four perspectives depends on the research needs (this is to determine the 'fine granularity' of modelling in linguistic terminology) (see Gu, 2013b: 8).

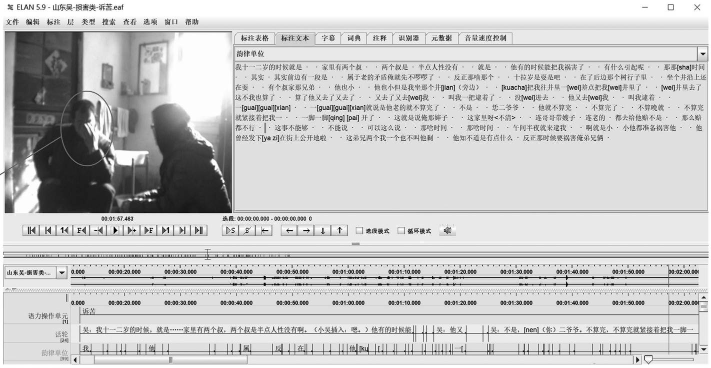
The live illocutionary forces are produced through illocutionary acts by the live, whole person with dynamic sounds, emotions, and gestures in situated discourse. In this study, an illocutionary act is regarded as a performance unit of the live, whole person's discourse activities. Likewise, we can also conduct concept modelling upon the live illocutionary forces by consulting the four perspectives of modelling a 'live, whole person' while considering other factors.

Here is an introduction to the STFE-match principle proposed by Gu (2013b).

Suppose the speaker shows consistency in what is said, what is thought, what is felt, and what is embodied when performing illocutionary forces. In that case, it can be concluded that the speaker follows the STFE-match principle (Gu, 2013b: 2). Generally speaking, what a child says, thinks of, feels, and embodies are consistent (of course, there are exceptions such as lying). For example, a child who cries for a teddy bear would show consistency in his/her words (expressing his/her wants through words and crying), thoughts (wanting something), feelings (frustrated because of not getting it), and appearance (frowning, crying, waving hands). When a topic directly relates to their interests, adult speakers usually display their true feelings with cues like prosodic features and gestures. A typical example can be found in our corpus, which falls into the illocutionary force type of 'complaint.'

In that example, the speaker was sitting at home, smoking, and recalling the experience of being bullied by a relative who almost pushed him into a well when he was little. Since the speech content is directly and negatively related to the speaker's interests, this example falls into the group of harmful illocutionary acts (the classification standard of illocutionary acts is discussed later).

In terms of prosodic patterns, the whole discourse process is marked with low pitch, and most of its discourse is in slow speech rate. Verbal and non-verbal cues such as long pauses (the most prolonged pause lasted 10.8 seconds, and the average 7.6 seconds), and sobs also appeared in the instance.



我十一二岁的时候就是 · · 家里有两个叔 · · 两个叔是 · 半点人性没有 · · 就是 · · 他有的时候能把我祸害了 · · 有什么引起呢 · · 那那[sha]时间
· · 其实 · 其实前边有一段是 · · 属于老的矛盾俺就先不啰啰了 · · 反正那啥那个 · · 十拉岁是耍是吧 · · 在了后边那个树行子里 · · 坐个井沿上还
在耍 · · 有个叔家那兄弟 · · 他也小 · · 他也小但是我坐那个井[jian]（旁边） · · [kuacha]把我往井里一[wei]差点把我[wei]井里了 · · [wei]井里去了
这不我也算了 · · 算了他又去了又去了 · · 又去了又去[wei]我 · · 叫我一把逮着了 · · 没[wei]进去 · · 他又去[wei]我 · · 叫我逮着 · ·
一[guai][guai][xian] · · 一[guai][guai][xian]就说是他老的就不算完了 · · 不是 · · 恁二爷爷 · · 他就不算完 · · 不算完了 · · 不算晚就 · · 不算完
就紧接着把我一 · · 一脚一脚[qing] [pai] 开了 · · 这就是说俺那婶子 · · 这家里呀<不清> · · 连哥哥带嫂子 · 连老的 · 都去给他赔不是 · · 那么赔
都不行 · | · 这事不能够 · · 不能说 · · 可以这么说 · · 那啥时间 · · 那啥时间 · · 午间半夜就来逮我 · · 啊就是小 · 小他都准备祸害他 · · 他
曾经发下[ya zi]在街上公开地啦 · · 这弟兄两个我一个也不叫他剩 · · 他知不道是有点什么 · 反正那时候要祸害俺弟兄俩 ·

诉苦

吴：我十一二岁的时候，就是……家里有两个叔，两个叔是半点人性没有啊。（小吴插入：嗯。）他有的时候能

吴：他又

吴：不是，[nen]（你）二爷爷。不算完，不算完就紧接着把我一脚一

我 他 属 反 在 他 [ku [一[

looking down (eyes droping), weeping, and wiping the face long pause for 10.8 seconds

Figure 3.5 Scheme of gestures.

According to the statistics of the annotated gestures, non-linguistic acts such as looking down (eyes dropping), weeping, and touching the face are frequently seen in this case, indicating that the speaker's emotional state is sad and angry, and what he thought about was the personal tragedy of being bullied.

This is a stark example illustrating that the speaker was talking about a past misfortune with his prosodic features and gestures showing a negative emotional state. These overt cues provide crucial clues and basis for the judgement of emotional states. As they are always consistent and can genuinely reflect the speaker's inner thoughts and emotions, we can say that they are in line with the STFE-match principle.

Undeniably, everyday experience tells us that in situated discourse among adults, some situations follow the STFE-match principle and some do not. Gu (2013b: 8–10) analysed seven situations that violated the principle, such as political stunts pulled by politicians and 'insincere' illocutionary acts performed by those trying to conceal their real thoughts and emotions. For example, when expressing the illocutionary force of 'sympathy,' the speaker may frown and say, 'I do feel for you' with low pitch and slow speech rate, but what he/she thinks of may be 'Who cares?' and he/she does not feel at all sympathetic. Additionally, noting that actors usually act with pre-prepared scripts, the emotions they express in the drama are very likely inconsistent with their true feelings, but are acted out for the stage. For example, an actor who is complaining over the misfortunes he/she has gone through in a film dialogue may show signs of misery in words or gestures, but he/she may not be pained at heart. Certainly, there are actors whose outward displays of emotions conform to what they think as they are so much in the characters that they attune themselves to the characters' fates. From this point of view, the STFE-match principle does not fit all situations. Here are two more exceptions that violate the principle. One is that the speaker performs infelicitous illocutionary acts by saying things without meaning them, and the other is that the speaker deliberately violates this principle to imply certain pragmatic meanings. New meanings can be interpreted by consulting relevant contexts or other background information. This also shows that the investigation of illocutionary forces is very complicated. Researchers are required to take into account the discourse content, prosodic features, emotional and gestural cues, and the relevant contexts of the situated discourse.

Furthermore, as some adults become more educated and gain more social experience, they tend to be more conservative in displaying emotions, with fewer prosodic and gestural cues presented. Speakers from certain cultural backgrounds show a propensity for an outward display of prosody, gestures, facial expressions, etc. while people from some cultural backgrounds do the opposite. For example, as an old Chinese proverb goes, 'One should not betray his or her emotions.'[2]

This section has generally introduced the relations between the STFE-match principle and Simulative Modelling. The abstraction of a simulation model, namely concept modelling and data modelling, and their relations to each perspective in modelling a 'live, whole person' is discussed in Section 6.2.

3.3.2 Significance of the four perspectives

In essence, the STFE-match principle is part of the 'felicity condition' of the illocutionary force instances in situated discourse. If the illocutionary force is 'felicitous,' it must follow the principle, showing consistency in what is said, what is thought, what is felt, and what is embodied. Conversely, even if an illocutionary force shows consistency in the four perspectives in question, it does not mean such illocutionary force is felicitous for many other factors determining the felicity. For example, the host of a sports meeting announces its opening through specific procedures in the right place and time, with a qualified identity. Therefore, the STFE-match principle is necessary for judging whether an illocutionary force is felicitous, not a necessary and sufficient condition.

For this study, the significance of the four perspectives and the STFE-match principle lies in two aspects:

1 The four perspectives proposed by Gu (2013b: 6) are the modelling perspectives of a 'live, whole person' in Simulative Modelling, yet it also provides important insights on modelling the illocutionary forces in situated discourse. Of course, there are far more modelling perspectives of a live, whole person than these four aspects. The live illocutionary forces generated by the 'live, whole person' in situated discourse can also be modelled from several other perspectives. However, these four perspectives are vital at least for exploring verbal communication (concept modelling is discussed in Section 4.2.1, and the relations between

each perspective, concept modelling, and data modelling are elaborated in Section 6.2). Accordingly, these four perspectives could guide researchers on discovering various clues left by the illocutionary force instances in multimodal corpus data via video or audio forms in the stage of data modelling.

2 The STFE-match principle also provides a research model and analysis sequence to investigate what the 'live, whole person' says, thinks, feels, and embodies. Words and gestures are 'demonstrative cues' that can be observed visually, while emotions and thoughts belong to 'inner states' that cannot be seen. For speakers, their demonstrative cues are driven by their inner states, which indicate that the order of these four perspectives is 'thoughts → feelings → words → embodiment.' However, the order of these four perspectives for hearers is just the opposite since they perceive the speakers' inner states through demonstrative cues, i.e. 'words → embodiment → feelings → thoughts.' In this study, the author infers the speaker's inner states by considering various demonstrative cues and other factors, such as the illocutionary forces' social situation. This is a process of inferring the speaker's inner states (including emotions and thoughts) from the demonstrative cues (including speech and gestures). Without question, some demonstrative cues cannot be observed visually even though they are also driven by inner states. In actuality, researchers usually conduct such analysis time and again, and integrate different factors into consideration.

3.4 Illocutionary force instances: classification, basis, and significance

3.4.1 Classification perspective

The classification of speech acts (illocutionary forces) has been the focus of relevant research since Austin put forward the speech act theory. Since then, many scholars established many classification standards by preserving, criticizing, revising, or reconstructing previous classifications of illocutionary force, which have been discussed in the first chapter.

In this study, to bring to the fore the close relations between illocutionary force tokens, the speaker's factors, and the specific discursive environment, illocutionary forces are examined in the context of situated discourse. Faced with miscellaneous illocutionary force tokens in situated discourse, Gu (2013a) suggested that they should be classified from the perspective of the relations between what is talked about and the speaker/hearer's interests. In this regard, speech acts are classified according to discourse activities which carry illocutionary forces/acts, following the method of examining speech acts at the level of behaviour activities (discourse activities). Classifying Speech act from a discourse perspective, oriented toward the speech acts in natural discourse, has built a bridge between speech acts and the discourse in which they are located (Adolphs, 2008: 47).

This study's illocutionary act-tokens fall into four classes: neutral, beneficial, harmful, and counterproductive.

Neutral illocutionary forces refer to the illocutionary forces in which the speaker or hearer's interests are not directly related to the discourse content, which in turn has no positive or negative impact on their interests.

Beneficial illocutionary forces refer to the illocutionary forces in which the discourse content favours the speaker/hearer's interests or has a positive impact on it.

Harmful illocutionary forces refer to the illocutionary forces in which the discourse content is unfavourable to the speaker/hearer's interests or has a negative impact on it.

Counterproductive illocutionary forces refer to the illocutionary forces in which the discourse content runs contrary to the primary expectation on the speaker/hearer's interests or leads to a contradiction between the expectation and the reality. Its utterance content first favours the speaker/hearer's interests but ends up as unfavourable due to the real world's changes.

3.4.2 Results and significance

Emotional changes are usually directly related to the speaker/hearer's interests, and the latter often affects or even defines the former. For example, in performing interest-neutral illocutionary acts, the speaker and hearer can emotionally distance themselves from the acts, while in interest-beneficial/harmful illocutionary acts, the speaker and hearer can be emotionally positive/negative (Gu, 2013a). Again, the speaker's emotional states are closely related to the prosody of his/her utterances, facial expressions, and gestures. This also happens in our daily communications – an interlocutor whose emotional state is angry would perform the illocutionary forces with clenched teeth and dilated pupils, one who is feeling happy would perform them with dancing eyebrows and beaming eyes, and one who is proud would hold his/her head high. Today, some scholars believe that facial expressions hold the key to the study of emotions (Ekman & Rosenberg, 2005; Keltner & Ekman, 2000: 236). As *Xici* from *Zhouyi* (*The book of changes)*, a Confucian classic, says,

> Those who are about to rebel may speak with a look of shame; those who have doubts in their hearts are ambiguous and unorganized; those with virtues give a concise and incisive speech; those with impetuous temperament tend to would speak redundantly and illogically; those who slander the good-hearted stammer; and those who fail in their duties must be weak-minded and swim with the tide.

In this way, words, thoughts, emotions, and gestures are connected.

Classifying illocutionary forces from the perspective of how the discourse content is related to the interlocutor's interests helps examine the interactions between emotional states, prosodic features, and gestures in different illocutionary force instances. By so doing, the interlocutors' interests become a tie between the illocutionary acts and the perlocutionary acts. Due to the difference in stakeholder relations, the speaker and the hearer usually present different emotions that are reflected in various data clues (such as phonology and gestures) in the multimodal corpus, not only allowing researchers to analyse the illocutionary acts/forces through the speaker's IFIDs but also to examine the perlocutionary acts through the hearer's facial expressions, gestures, utterances, etc.

It should be noted that the grouping of illocutionary acts/forces discussed in this section and the traditional illocutionary force type are two different concepts that should not be confused. The former classifies all illocutionary forces from a certain perspective (not toward the illocutionary force types), while the latter categorizes the abstract form (the ontological properties) of specific illocutionary act-tokens, namely the illocutionary force types. According to our classification standards, one illocutionary act-type may translate into different illocutionary act-tokens in situated discourse that fall into different groups of illocutionary force. It means that the categorizations of illocutionary act-types in previous traditional speech act studies could overlap with the four illocutionary force classifications discussed in this section, as they are viewed differently and applied to different standards. For example, if a speaker states his/her opinions in three different situated discourses, these statements would all belong to the illocutionary force type of 'statement' according to the categorization of traditional speech act theory. However, if these three statements vary in content, they would fall into different illocutionary force groups in this study. Let us say that the first statement talks about the speaker being praised by the leader, which is favourable to the speaker's interests, the second statement involves topics harming the speaker's interests, such as being criticized or blamed by the leader, and the third statement mentions that someone is ill (not any hearers present), which has nothing to do with the interests of any hearers present. According to our classification based on stakeholder relations (directly related to emotions), these three instances would fall into the beneficial, harmful, and neutral groups.

Even though the speaker's or the hearer's emotions are mostly dependent on stakeholder relations, it does not mean that they are the only factors that influence the speaker's emotional state when performing illocutionary forces. In other words, stakeholder relations are not the sole condition that induces emotional state. Other conditions include situational factors, personal styles, and cultural conventions, and the like. For example, although the neutral illocutionary forces' utterance content does not directly relate to the speaker or hearer's interests, it does not mean that they will not have a positive or negative emotional state. From a social psychological point of view, as humans are social animals with emotions, even if something is not directly related to their own interests, they will show empathy for the experiencer. For instance, the death of a stray cat often makes old ladies or the soft-hearted sad, though its life and death have nothing to do with any passers-by. In this case, the interlocutors' emotions are no longer neutral but sympathetic. Likewise, when someone wins awards or something good happens to them, their friends may empathize with them and feel truly happy for them. One of the differences between empathy and sympathy is that sympathy results from and echoes others' emotions, containing a causal relationship, i.e. if there is no such person feeling sorrow, none will feel sorry for them. However, empathy is similar to emotions based on the same facts and experienced by several people, which does not involve such a causal relationship (Huang, 2007: 36). Whether it is empathy or sympathy, the speaker is likely to have a corresponding emotional state serving other stakeholders' interests. This deserves special attention in the analysis of neutral illocutionary forces. Ken Binmore, Emeritus Professor at University College London and a well-known economist, believes that to sympathize means the subject would be to feel concern for others' welfare—to incorporate the object's welfare into subject's own with a feeling of sympathy (Binmore, 2005: 101–115). In other words, in a sympathetic mentality, changes in the object's welfare will lead to changes in the subject's welfare. Currently, modern neuroscience reveals that the psychological phenomenon of human sympathy is related to the mirror neuron system (Wang, 2014: 53). Binmore's understanding backs up our classification standard that distinguishes illocutionary forces on the basis of the speaker's/hearer's relevant interests. When the core content of the illocutionary force's performance is not directly relevant to the speaker's interests but involves other parties' direct interests, the speaker may include the relevant interests of the parties discussed into his/her own and produce a corresponding emotional state. This is called moral emotion (see Gu, 2015b). Different people may express sympathy or empathy differently in real life as they are also related to a series of factors, such as personal experience, morals, and ethics. Accordingly, to judge the speaker's emotional state in an illocutionary force instance, we have to consider the utterance content as well as other factors, including contexts and individual factors.

Nevertheless, not all types of illocutionary act-tokens fall into or overlap with the four groups in question. For example, the 'complaint' speech act only belongs to the harmful class because this type of illocutionary act is bound to be unfavourable to the speaker/hearer's interests. When Austin first proposed the theory of 'doing things with words,' he classified illocutionary acts according to performative verbs. Although equating the two is unscientific, it does not deny that many performative verbs have certain semantic properties that define the illocutionary act-token to a large extent. The 'complaint' speech act mentioned above is an illustrative example. But it should also be noted that in more cases, the semantic properties of performative verbs alone cannot determine to which group illocutionary force belongs to (because it can be proved that there is no such one-to-one correspondence between performative verbs and illocutionary acts (see Gu, 2002a: F33)). Instead, we should take discourse content, performative verbs, prosodic features, gestures, and other aspects into account. That means we should judge the classification of illocutionary force from the perspective of discourse activity.

3.4.3 Analysis of the stakeholder relations of illocutionary force instances

This study annotated the speakers' emotional states in all 134 instances in the multimodal corpus of illocutionary force (see Chapter 6) and analysed their stakeholder relations involved in the core contents, in an effort to unravel the correlation between stakeholder relations.[3] The results are shown as follows.

In the neutral group, the discourse content is irrelevant to the speaker/hearer's interests, but it may involve a third party's interests. The primary/universal emotions are mostly positive and neutral. The social emotions are mainly positive, followed by negative and neutral. In the neutral group, the situations of emotional state are more complicated.

In the beneficial group, the discourse content shows a positive correlation between the interests of the speaker/hearer as well as that of the third party involved. The primary/universal emotions are mostly positive, and only a few are neutral. The social emotions are also mainly positive, with a few isolated cases shown as negative or neutral.

In the harmful group, the discourse content shows a negative correlation with the interests of the speaker/hearer as well as the third party involved. The primary/universal emotions are mostly negative, while some are neutral. The social emotions are also mainly negative.

In the counterproductive group, the discourse content shows a negative correlation between the interests of the speaker/hearer as well as that of the third party involved. The primary/universal emotions include negativity and neutrality. The social emotions are roughly negative.

It can be observed how much of the discourse content is relevant to the interests of the speaker/hearer and significantly informs the speakers' primary/universal emotions and social emotions when performing the illocutionary force, especially for beneficial, harmful, and counterproductive groups where stakeholder relations are highly correlated with emotional states.

3.5 Summary

Consistent philosophy, methodology, and specific techniques can be found in the way we view verbal communication via complex system paradigms, Simulative Modelling, the mining of linguistic data with saturated signification from the real world, and the multimodal corpus-based approach.

Simulative Modelling is a research method adopted by multimodal corpus-based linguistics, and also serves as the core methodology of this study. Simulative Modelling requires researchers to collect information as close as possible to what the object is. On this basis, they can select a particular investigating perspective, mine the data, and conduct research. Simulative Modelling contains three stages: concept modelling, data modelling, and implementation and evaluation. In this study, the author chose four perspectives to model the live illocutionary forces in situated discourse, i.e. what is said, what is thought, what is felt, and what is embodied, each holding a wealth of clues or manifestations.

This research examines speech acts in light of general human behaviour. The social situation should not be regarded as a context or background as it was in traditional pragmatics in the analysis of situated discourse. Instead, it should be treated as the highest level of the analysis units. Similarly, various units lower than social situations can also be analysed. In this study, the illocutionary force/act at the individual behaviour layer is seen as a research object.

This study examines illocutionary forces in the context of situated discourse. So, an unparalleled classification method is adopted. According to how the utterance content is relevant to the interlocutors' interests (speakers/hearers), all illocutionary forces are divided into four classes: neutral, beneficial, harmful, and counterproductive. By distinguishing illocutionary forces in this way, researchers are allowed to explore the illocutionary act-tokens with varied emotions, prosodic features, and gestures through comparison and contrast, and to discover their interactive relationships between different tokens. Moreover, this classification, resorting to the interlocutors' interests, also links illocutionary acts with perlocutionary acts.

Notes

[1] Goldman (1970: 10–11) listed four decisive factors for an action instance, including its agent, property, time, and way. Goldman (1970: 16–19) expatiated on the role of purposes, intentions, and reasons in determining the properties of actions. He also believed that agents' purposes and intentions are the core of human actions, which is consistent with the findings in speech act studies.

[2] Source: *Book of Shu: Biography of the Former Lord* in *Records from the Three Kingdoms* (vol. 32.): 'His face never gives nothing away. He loves making friends with heroes, and many young talents are eager to assist him in achieving his aspirations.'

[3] For the definition and connotation of the core content, interested third parties, and emotional state, see Section 4.2.

4 Discovery Procedure of live illocutionary force

4.1 Principle, rationale, and characteristics of the Discovery Procedure

Logically, the principle, rationale, and characteristics of the Discovery Procedure depend on our object of investigation and fundamental methodology, namely the illocutionary force in situated discourse and Simulative Modelling.

4.1.1 Fundamental principle

Fundamentally, the Discovery Procedure is developed according to the ‘live, whole person’ principle proposed by Gu (2013b), which advocates that cues like prosody, bodily movement, and postures are all indispensable components of the felicitous performance of illocutionary acts in situated discourse. Therefore, the eloquent ‘live, whole person’ in discourse activities constitute the object of Simulative Modelling. Accordingly, such cues as prosody, facial expressions, gestures, postures, and bodily movements should be considered. Additionally, since speech acts are considered ordinary behaviours in this study, social situations are viewed as an independent analytic unit rather than the background or context speech acts. Social situations, activity types, and speech acts are only systemized results from different perspectives. In this sense, situational analysis is also an important object in this study.

4.1.2 Basic rationale

Under the fundamental methodology of Simulative Modelling, illocutionary forces in situated discourse are modelled and investigated through the four perspectives, i.e. what is said, what is thought, what is felt, and what is embodied. The Discovery Procedure is developed according to such methodology and perspectives, providing a series of concrete and detailed steps for further analysis. By so doing, the abstract methodology and perspectives are put into concrete analytic steps and multiple tiers practically. The first

DOI: 10.4324/9781003251774-4

step for Simulative Modelling is concept modelling, whereby the specific instances of speech acts in situated discourse are given abstract definitions, classified into different groups, summarized, and analysed, so it should be placed first in the analytic procedure. Subsequent issues like data modelling, implementation and evaluation, and multimodal linguistic data mining, as well as the evaluation and fine-tuning of both the analytic procedure and the annotation schemes, are preparations for the Discovery Procedure to carry out a detailed analysis. Therefore, all of them are excluded in the procedure. However, the four perspectives of what is said, what is thought, what is felt, and what is embodied should be stratified into different layers in the analytic framework. After annotation by data modelling, we further analyse the illocutionary force in situated discourse layer by layer and step by step based on the data mined and displayed.

4.1.3 Overall characteristics

The approach of Simulative Modelling determines the characterized openness inherent in the analytic procedure. Specifically speaking, modelling can be defined as the process that allows researchers to reproduce the object's structure according to the research objective, the primary concern of the investigation, and the analytical needs. Different studies can be modelled from different perspectives, but the chosen perspectives will determine the structure and the framework of the model, and in turn, the model will accordingly shape the frame of the analytic procedure. In this study, analyses are carried out from four perspectives: situations, emotional states, prosodic features, and gestures. However, other perspectives can also be taken into account when examining the illocutionary force in situated discourse; thereby, the analytic procedures should be adjusted accordingly. Furthermore, with the improvement of research skills and cognitive levels, researchers will be increasingly capable of capturing multimodal data, and their understanding of the relationships between different data will continue to deepen. In this sense, the perspectives of modelling and corresponding Discovery Procedures can be enriched and adjusted in future studies. For instance, current research techniques cannot sample the haptic data of communicators in authentic communication, but in some specific situations, different haptic sensations may convey different pragmatic meanings, say, a weak or a firm handshake indicates different illocutionary forces.

4.2 Framework and connotation of the Discovery Procedure

The so-called Discovery Procedure refers to the technical process whereby a certain dimension of language is analysed. As a set of techniques applicable to the language samples analysis, it was primarily widely adopted by structuralist linguists and used by linguists to discover the patterns and structures of linguistic data. This method is rooted in empiricism, viz. language description should be backed up by observed phenomena or data. It strikes a similar chord with the fundamental methodology of Simulative Modelling adopted in this study since Simulative Modelling is also based on the close observation of the object. Besides, our basic approach is to observe, describe, and summarize the object (live illocutionary force). Therefore, by borrowing the basic concept of Discovery Procedure in structuralism, this pilot study observed and described the live illocutionary force in situated discourse based on the construction of a multimodal corpus, which conforms to the tradition of descriptive linguistics.

To analyse the interactive relationships between illocutionary force, emotions, and prosody, Gu (2013a) proposed a 'Discovery Procedure for the tripartite interaction' in his research, which has also been the basis for our Discovery Procedure presented in this chapter. After aiming to improve such a tripartite procedure of analysis, we established our procedure for illocutionary forces in situated discourse. The most distinctive improvements are the addition of gestures (including facial expressions, physical attributes, and postures) and situational factors. Moreover, since this study focuses on the speaker's illocutionary acts, the hearer's perlocutionary acts will be left out here. Thus, we cancel the respective programme segments analysed from the hearer's side.

The Discovery Procedure contains five programme segments. Here, programme segments refer to the sequence of illocutionary force analysis from different dimensions, as well as the corresponding substances, including the conceptual model of illocutionary force, situational factors, emotional states, prosodic features, and gestures, as shown in Figure 4.1.

In the following sections, elaboration is given on the basic framework, connotations, and analytic methods of the Discovery Procedure developed for the illocutionary force in situated discourse.

4.2.1 Conceptual Model

Concept modelling is treated as the first programme segment in this study's Discovery Procedure of illocutionary force. The underlying reasons are illustrated below:

On the one hand, the Conceptual Model, echoing concept modelling, the first stage of Simulative Modelling, defines the way we investigate live illocutionary force and paves the way for other subsequent steps like data modelling. The logic behind it lies in the methodology of Simulative Modelling. In light of illocutionary act-types varying from illocutionary act-tokens, this procedure segment emphasizes the overall characteristics of a certain type of illocutionary force that are extracted and summarized from specific illocutionary act-tokens. To put it differently, the Conceptual Model aims to explain the illocutionary act-types that reside in the collective consciousness in a community from some proper perspectives.

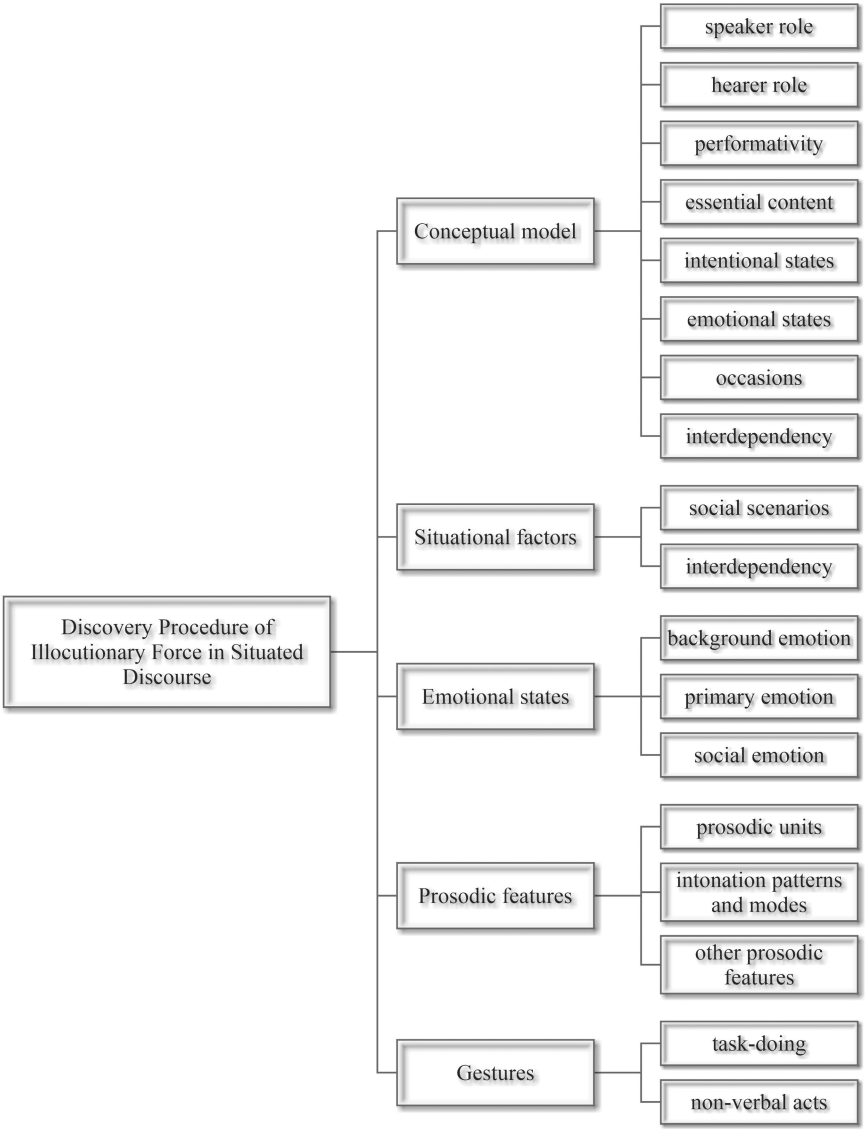


Figure 4.1 Analytic framework of illocutionary force in situated discourse.

On the other hand, the Conceptual Model's adopted perspectives also allow us to describe the internal properties, feasible conditions, and rules of the illocutionary act-types corresponding to different live illocutionary forces in situated discourse. Searle (1971: 42) pointed out that the implementation of an illocutionary act is based on a series of constitutive rules,[1] including preparatory conditions, sincerity conditions, essential conditions, and so forth (Searle, 1971: 53), which are the conditions where some type of illocutionary force is produced. If these constitutive rules are violated and the implementation conditions are not met, then the corresponding illocutionary act will not fail to perform. For example, the preparatory condition of 'order' is that the speaker is in a more authoritative position compared with the hearer, the sincerity condition means that the speaker truly wants something to be done, and the core condition here means that the main idea of the discourse is imbued with the speaker's intention of calling on the hearer to do something. Establishing a conceptual model for a certain type of illocutionary force enables us to examine the corresponding implementation conditions, constitutive rules, or even the properties of illocutionary force itself, which gives objective definition and analysis of the illocutionary force type in question.

4.2.1.1 Format of the Conceptual Model

It should be noted that here we adopt Gu's (2013a) 'octet scheme' to shape our Conceptual Model. Having considered the basic forms of illocutionary force and the relevant characteristics and influential factors of the discourse it situated, the 'octet scheme' is represented in the formula borrowed from set theory to draw a conclusion in an open and formalized way.[2] The Conceptual Model established by this formalized method describes and defines the characteristics and properties of a certain type of illocutionary force and provides perspectives for investigating the felicity/infelicity of certain illocutionary forces type in situated discourse.

The Conceptual Model of illocutionary force adopts the 'octet scheme,' as shown in Figure 4.2.

Here, we use 'subset'[3] to refer to the grouping of some specific values. The values assigned to the subsets are varied in type and correspond to different perspectives in concept modelling. The term 'subset' is used because every aspect of the sets in the Conceptual Model of illocutionary force is

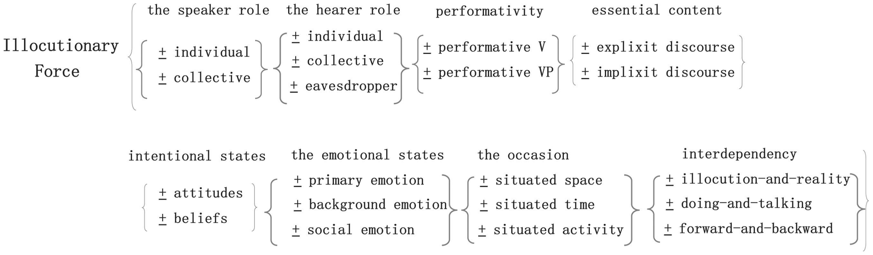


Figure 4.2 Conceptual model of illocutionary force.

essentially a configuration of several specific values, e.g. at least three kinds of values are contained in the set of 'emotional states': primary/universal emotion, background emotions, and social/secondary emotions. And all together, they constitute the subset of 'emotional states' in the whole set of illocutionary forces in the Conceptual Model.

Let us look at the illocutionary force's conceptual model from the philosophical relationships between 'type' and 'token.' The aforementioned subset is a generalization of types encompassing the properties of a certain type of illocutionary act/illocutionary force (also called 'attributes' in technical terms, cf. Section 6.2). Those subsets will take on different values according to the given contexts in situated discourse where they create various tokens of illocutionary force. Certainly, such values should be confined to a certain scope, which conforms to the nature of a certain type of illocutionary force itself. In this sense, a specific token of illocutionary force can be defined as the configuration of an array of specific values given by the speaker/actor when he/she is performing speech acts (Gu, 2013a).

4.2.1.2 Connotation analysis of the 'octet scheme'

1 Speaker role and hearer role

Both Searle (1976: 5) and Mey (2001: 119) believe that the speaker's and the hearer's roles and statuses should be taken into account in the classification of speech acts. From the perspective of verbal communication, since the speech act theory covers the whole communicative process, apparently at least the speaker role and the hearer role should be involved. These two subsets vary according to the contexts that the live illocutionary force is located in.

In terms of speech act agents, Hu (2009) believes that a third party (he/she/they), in addition to the speaker and the hearer, should be taken into account. Speech acts are basically organized by the first person 'I/we,' who are the performers of speech acts and the basic frame of reference. In view of the present Conceptual Model of illocutionary force, the first person 'I/we' refers to the 'speaker role,' while the 'hearer role' is more problematic since it can be subdivided into the addressee (you) and the audience:

a The hearer role only includes the addressee. There is no other audience or hearer. Precisely, only the speaker (I/we) and the addressee (you) are involved in the situated discourse's social situations. This time, the third party (he/she/they) is absent;
b In addition to the addressee, the hearer role also includes a third party (he/she/they), such as the audience and any other hearer;
c Moreover, it sometimes happens that the speaker role and the hearer role coincide, viz. the speaker talks to him/herself.

Speech act agents like the hearer role and the speaker role are directly related to the interests of the corresponding instances of live illocutionary force in our analysis.

It should be noted here that the third party involved in the essential content does not equal the third party of participants in a conversation. In this study, 'I/we,' 'you,' and 'he/she/they' refer to the agents of speech acts involved in the social situations where the situated discourse is located. The interested third party and the third party of the agents of speech act sometimes coincide (e.g. 'he/she/they' observe a discussion concerning their own interests), but sometimes do not (e.g. 'he/she/they' observe a conversation between the speaker and the hearer that concerns another's interest).

Here are more cases:

Case 1: The speech act's essential content does not directly involve the interests of the third party of the speech act agent ('he/she/they'), nor does it involve any interested third parties.

Case 2: The speech act's essential content does not directly involve the interests of the third party of the speech act agent ('he/she/they'), but does involve an interested third party.

Case 3: The speech act's essential content involves the interests of the third party of the speech act agent ('he/she/they'), but does not involve the interests of others.

Case 4: The speech act's essential content involves the interests of both the third party of the speech act agent ('he/she/they') and that of others.

2 Performativity

Performativity refers to whether an illocutionary act is performed explicitly through performative verbs/phrases, or implicitly in other ways. Austin states that the difference between explicit performatives and implicit performatives lies in whether the performative verbs/phrases arise in the discourse structure or not. If performative verbs or phrases are used in the utterances, then it is an explicit performative. For instance in the utterance, 'I bet it will rain tomorrow.' In this utterance, 'bet' is a performative verb. However, sometimes in situated discourse, it is unnecessary to use specific performative verbs/phrases to perform illocutionary force explicitly (Mey, 2001: 118) in that the performative function of an utterance can be made through its content, contexts, and so forth. For example, the utterance 'Now, I announce the opening of the University's first Sports Day!' is an explicit performative classified as the illocutionary force type of 'announcement,' since the performative verb 'announce' is used in the utterance. The same utterance above can also be put in this way: 'The University's first Sports Day is now open!' Although no performative verbs or phrases are used in the second utterance, we can still conclude that it is an implicit performative by judging from cues like the discourse content, tone of voice, prosodic features, accompanying gestures, and the settings. In this study, there is a very small number of explicit performatives with performative verbs or phrases in the constructed corpus based on situated discourse; as a contrast, most of them are implicit performatives.

3 Essential content

Essential content means that every instance of illocutionary force is bound to have a certain substance. The most basic thing for any speech act is 'what has been said (the discourse content),' echoing Austin's 'locutionary act.' Of course, the substantive content can be expressed either through explicit discourse or through implicit means. Under certain circumstances, some gestures also play a vital role in producing illocutionary force.

4 Intentional state

Two cases[4] were mentioned in Austin's discussion on the infelicity of the illocutionary act: Misfire and abuse. The former term refers to those acts that fail to go through an accepted conventional procedure that includes the uttering of certain words by certain persons in certain circumstances, while the latter term indicates that the speaker's feelings, attitudes, beliefs, and intentions are improperly matched with the acts, i.e. the performed action is infelicitous and insincere. The above cases are responsible for the infelicities of illocutionary acts. In Austin's (1962: 14–15) discussion of six rules that make felicitous performative utterances, he said that 'the programme is designed for use by persons having certain thoughts or feelings … must in fact have those thoughts or feelings.'

On the other hand, to perform a certain illocutionary force felicitously, the above rule must be followed. At the same time, the thoughts and feelings in question also become elements of a conceptual model of illocutionary force, which correspond to the 'intentional state' and 'emotional state' respectively in Gu's (2013a) octet analytic framework of illocution.

Here we first look into the intentional state. A 'felicitous' illocutionary act must contain a certain intentional state of the speaker. This study's 'intentional state' derives from Grice's meaning and intention theory and Strawson's relative theories on intentions. Grice proposed the concept of non-natural meaning (meaning embedded in verbal communication), which is generated in the mutual interaction between the speaker and the hearer. His main contribution to speech acts is recognized by the view that daily communication should be established because the speaker shows certain intentions, which the hearer can perceive through the utterances. Assuming both the speaker and the hearer are rational, the cooperative principle he proposed is designed to analyse how the hearer perceives the speaker's intention under certain circumstances. The intention of illocution proposed by Strawson (1964) is closely related to Grice's meaning and intention theory.

Austin (1962: 116) pointed out that an illocutionary act cannot be performed without securing uptake, which contains the understanding of the speaker's intentional state. In performing an illocutionary act, if the intention of the speaker (S) cannot be understood by the hearer (H), the latter, who fails to grasp the force, will consequently make no perlocutionary acts. In the end, no speech act is literally produced between S and H. In terms of the speaker, when S performs an illocutionary act/force, there must be some appropriate intentional state accompanying the performance: S expects a certain result; S hopes his/her intention to be understood by H; S presumes that his/her intention should be part of or the exact reason for H to make a perlocutionary act. Such intentions can be summarized as primary intentions (Strawson, 1964), which constitute the basic elements of various types of illocutionary force, e.g. request and entreaty. However, when performing an illocutionary act/force, except for the aforementioned three constituents of primary intentions, other factors, such as attitudes, beliefs, hopes, and wants, which are also closely related to the social situations where illocutionary forces are located, may become key defining features between some types of illocutionary force. Let us say, in the example of 'request' and 'entreat,' when making a request, S may show a firm and tough attitude, but if S is entreating, the request may be expressed in a mild way; after all, it is the other party that has the say. Primary intentions, combined with attitudes, beliefs, hopes, and wants, which are affected by many other factors, constitute the speaker's intentional state when performing an illocutionary act/force and distinguish different types of illocutionary force. As a result, when analysing the speaker's intentional state, we have to set the common primary intentions among various illocutionary forces apart from those attitudes, beliefs, hopes, or wants exclusive to a particular type of illocutionary force.

5 Emotional state

As it is undeniable that a speech act agent would never be an emotionless individual ruled by rationality, emotional factors should be covered in pragmatic studies.[5] To examine speech acts from the perspective of the whole framework of common behaviour theory, we must acknowledge the fact that a normal 'live, whole person' always endows some feelings to his/her uttering. This also becomes a crucial part in the Discovery Procedure in this study. Searle (1969: 64) gave an example of saying 'Hello,' believing that there was no propositional content and no sincerity condition with the utterance. However, Gu (2013a) argued that some illocutionary forces may have no sincerity condition, though they always have some emotions. Austin (1962: 40) also pointed out that feelings are necessary in performing an illocutionary force. The emotions behind an illocutionary force directly affect the felicity of its performance. That is to say, corresponding emotions must back up some illocutionary force types. Note that in the conceptual model, the corresponding emotional state of an illocutionary force type, presented in the formula of a subset, is 'justified' for the speaker, based on the author's communal experience. The absence of a corresponding emotion may render the illocutionary force instance in situated discourse infelicitous, e.g. a speaker in a default state of ‘happiness’ when congratulating someone. Therefore, in our research, it is necessary to make a distinction between the speaker’s justified emotional state of a certain illocutionary force type and his/her real emotional state.

6 Occasion

As mentioned earlier, occasion is one of the analytical objects in our study. Austin (1962) pointed out that the circumstances of the utterance are also an exceedingly important aid for the discrimination of illocutionary forces. When discussing the six rules that render performative utterances felicitous, he mentions the importance of circumstances (Austin, 1962: 14–15):

A1 That procedure [is] to include the uttering of certain words by certain persons in certain circumstances;

A2 The particular persons and circumstances in a given case must be appropriate for the invocation of the particular procedure invoked.

Searle and Vanderveken (1985: 27) referred to the context in which an illocutionary act is located as ‘the context of utterance,’ a crucial defining factor for an illocutionary act, in that an utterance can produce diverse illocutionary acts in different contexts. Mey (2001: 110) also clearly stated that speech acts and speech act verbs only make sense in appropriate contexts, and occasionally a speech act verb/phrase is not needed to perform an illocutionary act (Mey, 2001: 118). In Searle’s (2001: 30) later research, he also proposed that in real-life speech, it is unnecessary to exploit an explicit performative verb to perform an illocutionary act, because its context will foreground the performative function of the utterance. After an adequate corpus-based study, McAllister (2015: 45) also confirmed that situational types play a role in determining the type of a speech behaviour.

In short, a speech act is an intentional act performed by the speaker for some specific purpose in a real-life situation. The occasion is quite significant for a speaker to perform an illocutionary act/force, and for the hearer to perceive and interpret the utterance (Gu, 2015). If we are going to study live speech acts in situated discourse thoroughly, we must introduce occasion as a subset into our study.

7 Interdependency

In this study, the interdependency in the conceptual model mainly manifests itself in the following three aspects:

i Relationship between a speech act and former/later utterances, known as the *forward-and-backward interdependency*. Plus, this is one of the 12 dimensions of Searle’s classifying system for illocutionary acts (1976: 5).

The occurrence of some illocutionary force types is based on their relations with previous discourse. Let us look at ‘proposing’ as an example. The speaker could make a proposal based on a suggestion offered by someone earlier, some previously stated opinion, or a projection from some aforementioned premise (Searle, 1976: 5). By the same token, the occurrence of some speech acts may affect the speaker’s later utterance or the relative illocutionary force, e.g. in the case of ‘warning,’ the speaker may give the hearer a warning first before intimidating him/her if the warning is in vain. In terms of discriminating illocutionary force types, the hearer and the third party (such as the researcher), more often than not, can only make an inference from the relation or function of the former/latter utterance since it is strongly associated with the speaker’s intentional state.

ii The generation of a certain type of illocutionary force bears a causal relationship with what is happening in a situated discourse at a given time and place or what is happening (or has happened or will happen) at a random time and place, known as the *illocution-and-reality interdependency*. For example, in the case of ‘accusation,’ if the speaker is to accuse someone, a fact of damaging the speaker’s interests done by the accused must be found in the situated or non-situated discourse, which is how the illocutionary act/force of ‘accusation’ is formed. In this sense, illocution-and-reality interdependency is similar to Searle’s preparatory conditions (1971: 53).

iii As the speaker is performing an illocutionary act, his/her doing and talking are interconnected and interdependent, which is known as the *doing-and-talking interdependency*. Since our study examines live speech in situated discourse, cues like gestures and utterances are inseparable players for the speaker to perform an illocutionary act and thus produce an illocutionary force. For this reason, analysis of the relationship between the speaker’s doing and talking directly involves the interpretation of his/her implication, which is considered to be the ground of circumstanced illocutionary act/force studies.

4.2.2 Situational factors

In the Discovery Procedure of situational factors, two aspects are involved: (1) analysis of the social situations of the situated discourse where a specific illocutionary force locates, and (2) analysis of the interdependency of a specific illocutionary force in a given context in situated discourse.

4.2.2.1 Social situation analysis

As discussed in Chapter 3, Gu (2006b) referred to social situation as the maximum unit in the analysis of multimodal texts, that is, the social situation itself is also an analytical object. As the superordinate concept (maximum unit) in the analysis of situated discourse, social situations contain various types of activity that further involve our analytical objects, namely illocutionary acts and their corresponding illocutionary forces. Below the level of every analytical object, prosodic units and discrete gestural units are included. Note the relevant units involved in the specific analysis are adapted to the practical research needs.

So for every illocutionary force, there is a time, space, and scenario where it is located. And the analysis of social situation will provide crucial clues to our study of illocutionary acts.

Moreover, in this programme segment, what is to be analysed is the social situation where an illocutionary force instance is embedded, as well as its subordinate activity types. Explanations are required if the social situation in question includes more than one activity type.

4.2.2.2 Interdependency analysis

As mentioned above, our investigation of interdependency contains three main aspects:

1 The relationship between a speech act and former/later utterances, known as one of the 12 dimensions of Searle's classifying system for illocutionary acts (1976: 5). The occurrence of some illocutionary force types is based on a certain relationship with previous discourse. At the same time, the relationships and functions of the former/latter utterances are essential cues in determining illocutionary force types.

2 The generation of a certain type of illocutionary force is correlated to what is happening in the situated discourse at a given time and place or what is happening (or has happened or will happen) at a random time and place (illocution-and-reality interdependency). In this aspect, conceptual analysis and the relevant situations are taken into account. For instance, in the performance unit of 'complaint,' there must be a painful or tragic experience that has occurred before the speaker performs the act; and likewise, in the case of 'naming,' the act should be preceded by the existence of a thing or person.

3 The interdependent relationships between the speaker's doing and talking when he/she is performing an illocutionary act (doing-and-talking interdependency). In this aspect, it is necessary to combine the utterances and gestures given by the speaker under a specific situation in our analysis. According to the experience of data analysis in the situated discourse corpus, Gu (2002c) proposed eight relationships between doing and talking:

a Talking is doing, such as meetings, discussions, interviews, etc.;

b Talking is the main constitutive part of the task, e.g. a teacher giving a lecture in the classroom, and a doctor attending to a patient in the consulting room;

c Talking is a constitutive part of the task. However, compared with the previous relationship, talking accounts for a smaller proportion of the task; that is, talking is only incidental to the primary task of doing. For example, an engineer enquires and chats while debugging an electrical appliance;

d Talking and doing run in a conflicting parallel, for example, talking while eating;

e Talk is an embedded social part of the task, e.g. a couple chatting at the table; this time, talking is indispensable in maintaining their intimate relationship;

f Talking is a decorative part of the task, e.g. introducing oolong while making tea;

g Talking is a hindrance to the task, e.g. whispering in the exam;

h Talking and doing are independent of each other, e.g. picking out the edible parts of vegetables while making small talk.

Let us make a summary of the eight situations listed above. An array of descriptors can be used, whether singly or in combination, to describe the diverse relationships between talking and doing: overlapping, conflicted, parallel, independent, and related. By using those words, those eight situations above can be labelled as follows:

(a) & (e) Overlapping.

(b), (c), & (f) Parallel and related. (The status of doing and taking varies according to different situations and illocutionary force types. Sometimes, talking is subordinate to doing, sometimes the reverse is true).

(d) Conflict. (Talking and doing simultaneously are conventionally forbidden).

(g) Parallel and conflicted. (Doing and talking can happen at the same time, but they are not allowed to do so according to some conventions).

(h) Parallel and independent.

By observing those behaviours discussed by Gu (2002c), a conclusion can be drawn that all of them are tasking behaviours with explicit purposes. If viewed from a broader perspective, there are also random, subconscious, and non-verbal acts accompanying a speaker's utterance, such as frowning, touching one's head, or waving. Therefore, in this study, the author divided doing into task-doing and non-verbal doing (ibid. Section 4.2.5); the doing proposed by Gu (2002c) mainly refers to the task-doing in our study.

4.2.3 Emotional state

As suggested above, emotional state[6] is a crucial criterion for judging the 'felicity' of an illocutionary force, and it is also an important factor that influences such IFIDs as speech content, prosodic features, and gestures.

Emotion is an old subject. Some scholars believe that it even appeared before human consciousness in the process of human evolution (Damasio, 1999: 37). Since ancient times, emotions have been recognized by philosophers as occupying an important theoretical position in ancient Chinese Confucianism. In modern research literature, Western scholars mainly studied emotion in the fields of psychology, sociology, physiology, and health research. Also, many Chinese scholars carried out investigations into human emotions and feelings (see Meng, 1989, 1998, 2002). Since the late 19th century, scholars including Charles Darwin, William James, and Sigmund Freud have delved into emotions from various aspects. For example, from the perspective of evolution, Darwin investigated emotions from culture to culture, species to species, drawing the conclusion that emotional expression is a product of evolution. *The Expression of the Emotions in Man and Animals,* published in 1872, is Darwin's representative work. In the 20th century, researchers began to pay attention to the relationship between emotions and languages. People gradually realized that the direct or indirect impact that emotion made on language is reflected in all forms of language, including phonetics, syntax, semantics, and pragmatics. Jackson did pioneering research on emotions from neuroanatomy and discovered that the right cerebral hemisphere mainly handles emotion while the left handles language. In psycholinguistics, research on the relationships between language and emotion also concentrated on how emotions affect vocabulary storage and retrieval (van Lancker Sidtis, 2008: 199). Some scholars advocated that human language has a tradition of emotional expression, and that verbal expressions like crying, shouting, and singing play a crucial role in the early development of our language (Code, 2005). From that, the judgement can be made that human language is handled by a dual-process model in collaboration (van Lancker Sidtis, 2004). When neuropsychological disorders of emotion arise, our daily communicative competence will be affected. Right-hemisphere damage can affect the perception of emotional states, having consequences for linguistic expression (van Lancker Sidtis, 2008: 203–205). Similarly, speech and language disorders interfere with efficient communication of emotional and attitudinal information (van Lancker Sidtis, 2008: 200). Everyday experience tells us that in communicative situations, such cues as facial expressions, bodily manifestations, and extra speech sounds can all reveal a speaker's emotional state; conversely, language with emotional information may cause emotional changes in interlocutors, e.g. 'talking therapy' in clinical psychology is a typical example (van Lancker Sidtis, 2008: 202). A comprehensive, interdisciplinary, and multiple-perspective introduction to the present studies on domains like emotions and language can be found in Fussell's (2002) and Scherer's (2003) research, where a variety of research paradigms and models are analysed, and suggestions for future research are made.

From the perspective of bio-anatomy, the limbic system handles the processing of emotions (Ireland & Tenenbaum, 2008: 209), though some scholars such as Ledoux (1992) believe that it is the amygdala that really makes a difference (cf. Tang, 2006: 122). Researchers realize that the brain's emotional circuits have a certain processing area and conduction path just like the senses of seeing, hearing, and feeling. Interestingly, Kolb and Whishaw (2005: 411) pointed out that many amygdala neurons are multimodal, which are able to react to more than one sensory modality, such as photic stimulation, acoustic stimulation, haptic stimulation, taste stimulation, olfactory stimulation, etc. In this sense, emotions are the reactions produced by multiple neural systems, based on the evaluation of stimuli (Tang, 2006: 121). Stimuli that cause this kind of response are diverse; it could be just a sensory interaction with the outside world, such as 'the sight of something reminds someone of somebody' (visual); or the interaction of two or more senses, such as feeling timid in the presence of a tiger while hearing its roar (visual and auditory), or weeping for joy when bumping into a long-lost relative or friend by seeing his/her face (visual), talking with him/her (auditory), or touching him (haptic). That is to say, emotions, processed by a specialized neural system, are generated in the interaction of multiple senses with the outside world. The speaker obtains information and produces emotional reactions to it through multimodal interactions between sense modality and the outside world. Meanwhile, he/she can express feelings via a variety of sensory modalities.

Based on a review of previous studies, Damasio (1999: 51) states that all emotions use a person's body as their theatre and that emotions affect brain operations. Emotion is a complex responsive combination of chemistry and nerves, forming an array of patterns (Damasio, 1999: 51). The psychologist Plutchik (2001: 345–346) pointed out that emotion is not simply a feeling state but is a series of loose but complex consecutive events. Starting from stimulus, it includes sensations, psychological changes, action stimulations, and specific, goal-oriented behaviours. In other words, researchers may gain some clues about the speaker's inner emotions from his/her external physical performance, which is a direct embodiment of emotions. This also provide the author with theoretical foundations to infer and figure out the interlocutor's emotions through his/her prosodies, gestures, or other relative performance. In brief, this is a research process that traces from external forms like speech and gestures to internal emotions (Gu, 2013b: 2).

4.2.3.1 Classification and stratification of emotions

In terms of duration, emotions can be roughly divided into two subsets: (1) 'Occurrent emotions' (see Gu, 2013a: 327) or 'situated emotions,' which occur in a given context, time, and place in the situated discourse, and (2) 'dispositional emotions,' which stay with the speaker for quite a long time. For example, some people are as pessimistic as Lin Daiyu (an emotional girl in *Dream of the red chamber*[7]), who continually feels depressed and unhappy; other people are naturally happy-go-lucky, wearing a smile all day. Situated emotions can be stratified into different layers due to a variety of influential factors, e.g. situated atmospheric emotions, the interlocutor's well-being, and his/her usual mood in recent days. Under the sway of situated atmospheric emotions, there emerges a subset emotion that echoes with what Hatfield, Cacioppo, and Rapson (1994) referred to as 'contagious' emotions. Occasionally, collisions may appear between occurrent emotions and atmospheric emotions due to the interlocutor's factors. For example, an interlocutor may feel down in spirits and in poor physical health but attends a celebration that is supposed to be emotionally uplifting and cheerful. Alternatively, the interlocutor has been in high spirits in recent days but attends a grave and sorrowful occasion with positive and jubilant feelings.

Previous research on emotions reveals that, on the one hand, emotions are a product of human evolution, generated by the physiological parts in our body as the response to external stimuli; on the other hand, most emotions are gradually formed in the process of social acquisition (see Elster, 1999). Therefore, human emotions are characteristically complicated and multilayered. Damasio (1999) mentioned background emotions, universal/primary emotions, and social/secondary emotions in the stratification of emotions; Ekman (1973, 2003) conducted in-depth research on primary/universal emotions; and Gu (2013a) further layered social emotions. It should be pointed out that researchers stratified emotions into three tiers from three perspectives, which can co-occur, but different emotional tiers are under different degrees of conscious control. For example, social emotions are the most likely to be consciously controlled, while background emotions are the least likely (Gu, 2013b).

1 Background emotions

Background emotions refer to emotions directly associated with the speaker's physical conditions, such as fitness, listlessness, tranquillity, nervousness, vibrancy, lassitude, passion, and fatigue. Therefore, Gu (2013b: 7) also named these emotional types of 'constitutional emotions.' For example, some older adults are vigorous and lively despite their age, while some youths look listless and ill on account of their physical weakness. Damasio (1999: 52) believed that background emotions originate from intrinsic causes. Strictly speaking, the tripartite interaction between body functions, human organisms, and the outside environment induces physical changes.

In a situated discourse, cues like prosodic features and physical attributes to a large extent depend on the speaker's background emotions, which directly impact the fundamental prosodies. For example, on a given occasion, a speaker who is in good health will speak loudly in a high pitch and make some corresponding gestures, showing that he/she is full of energy; on the other hand, if the speaker is in poor health, his/ her voice will be low and he/she will look listless. Traditional Chinese Medicine even takes the tone of voice and the speaker's gestures as one of the diagnostic aids (which respectively echoes with inspection, auscultation, and olfaction in the four diagnostic methods). For example, congratulations given by a dispirited, frail young patient vary vastly from those given by an elderly person who is in fine fettle in terms of prosody and gestures. Even though the former is in a primary-emotional state of 'joyfulness' and the social-emotional state of 'respect,' his/her prosodic features are usually marked with low pitch, slow speed, long pauses, and frequent sighs, accompanied by gestures like frowning, lounging in a chair sluggishly, and so forth.

2 Primary emotions/universal emotions

Primary emotions are the basis of human emotions and are the main concentration of the previous studies in this domain. In the 19th century, with the development of evolutionary biology, the impact that emotions had on human facial expressions and verbal expressions received increasing attention from scholars. From the perspective of evolution theory, both humans and primates experience primary emotions, also known as primitive emotions, which often trigger physiological changes, e.g. accelerated heartbeat and increased blood pressure. Darwin (1872) was the first person to propose the universality of emotions, i.e. primary emotions. He believed that the expression of certain emotions was a legacy from the survival tendency in the course of human evolution. From an anatomical view, the amygdala plays a prominent role in deciphering the affective meaning of emotional events. It is a major part of the brain that involves the acquisition and expression of fear conditioning and the activation of self-reward (a form of motivational state) upon external stimuli. Moreover, the amygdala is also involved in the expression of emotional states, ranging from joy to sadness and fear to disgust in response to stimuli in the autonomic neural system and behaviours (Tang, 2006: 122). As a universal subset of emotions, primary emotions span cultures and races and occupy a central position in other emotions.

For centuries, various ways to classify primary emotions can be found in ancient Chinese Confucian literature and in ancient Western rhetoric. Philosophers and psychologists later also made significant contributions to this. For instance, *The Conveyance of Rites* (one chapter in the *Book of Rites*), an ancient Chinese Confucian classic, records that 'It is natural for us to experience happiness, anger, sorrow, fear, love, disgust, and desire'; and there exist emotional types in traditional Chinese medicine, namely happiness, anger, worry, anxious, sorrow, fear, and horror. All the above emotions refer to primary emotions. In the West, the Stoics in ancient Greece mainly divided emotions into four types, i.e. pleasure/delight, distress, appetite, and fear (Cicero, around 45 BC, Tusculan Disputations, iv: 14–15), while Descartes listed six kinds of emotions: wonder, love, hatred, desire, joy, and sadness (1649, The Passions of the Soul, 353). Today, rigorous research has been conducted into the classification of primary emotions, e.g. Turner (2000: 68–69) spelled out the typical viewpoints of at least 20 emotional types under primary emotions.

Despite researchers differing on how many subsets are there of primary emotions and which emotion can be counted as a primary emotion, their conclusions about their scope are virtually the same, and some of their classifications overlap, in that they all agree that happiness, fear, anger, and sadness are quite common. The basic classifications include happiness, sadness, fear, anger, surprise, and disgust. Paul Ekman, a renowned scholar who studied emotions through facial expressions, confirmed that a series of facial expressions are universal to human species, an inherent part of humans throughout the process of evolution. Also, he proposed six types of primary emotion: anger, fear, disgust, surprise, happiness, and sadness (Ekman, 1973: 2003). Turner (2000: 73) even made a primary emotion paradigm to illustrate primary emotions as well as their shifting forms, holding the view that primary emotions can be divided into three levels (high, medium, and low). Although every language has an intensive vocabulary to describe emotions, various public studies have shown that hundreds of words, due to their similarities, usually belong to the same family so that they can be summarized into several subsets of the primary emotions.

Primary emotions can be identified either by facial expressions (Ekman & Friesen, 1975) or by other cues like posture, voice, etc. (Turner, 2000: 13).

3 Social emotions/secondary emotions

In addition to the aforementioned universal and fundamental emotional forms in humankind, some emotional forms also exist beyond the scope of primary emotions, due to the influence of various factors like cultural norms and social situations. Such emotional forms are known as social emotions or secondary emotions. Stemming from primary emotions, social emotions are highly dependent on social settings. Their development is closely related to social cultures, and their generation often based on social cognition (Tang, 2006: 122). Social emotions are acquired through the speaker's experience and movement towards socialization (Culpeper, 2011: 59; Turner & Stets, 2005: 15–19). Scholars believe that social emotions originate from a mixture of primary emotions (Kemper, 1987; Plutchik & Kellerman, 1980; Turner, 2000: 15), and such emotions are more socially constructive (Kemper, 1987).

Their classification varies from culture to culture as it relates to cultural traditions. Gu (2013a: 324) divided social emotions into three categories: positive, negative, and neutral. Each category can be further divided into two major directions: Other-directed and self-directed. Social emotions often occur in a conversation when interlocutors exert influences on each other, e.g. interlocutors show respect, affection, or admiration while talking, or make comments on something/someone. This study borrows Gu's (2013a) classification and relevant descriptive framework. This is shown in Figure 4.3.

In different cultures, social emotions may feed into our judgements on the felicity of some illocutionary force instances. For example, in traditional Chinese culture, many students may embrace a social emotion that is other-directed and positive, with awe towards their teachers, while such emotion

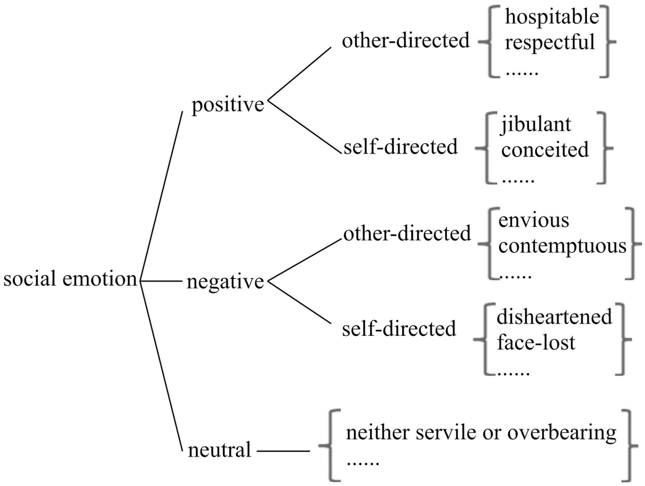


Figure 4.3 Social emotion: a classification.

is absent in the West. In Western countries, teachers and students are equal, and students' social emotions are usually neutral, neither overbearing nor servile. When Chinese students present a report to teachers, professors, or leaders, some may feel the social emotion that is other-directed and positive, with awe as well, which has become an accompanying emotion in Chinese culture and in some social situations for the act of 'reporting.'

4.2.3.2 Three-tier emotions: judgement, correlation, and subdivision

1 Judgement of three-tier emotions

In this study, we mainly rely on the speaker's extrinsic gestures and the situated contexts to speculate the speaker's emotional state, which is, as suggested above, intrinsic to the speaker.

In terms of judging emotional states, here, three analytic methods are introduced. The first is called the 'three-stage model of emotion representation,' proposed by Feng and Qi (2014) based on cognitive evaluation theories. This model provided some valuable experience in the judgement of emotions and attitudes, though it was designed to serve the multimodal construction of attitudinal meaning in the evaluation system, that is, how to express attitudinal meaning through such multimodal resources as emotions, judgement, and appreciation (Feng & Qi, 2014: 586). According to the research on cognitive evaluation theories in Frijda (1986) and Lazarus (1991), the subject's cognition of the outside world can stimulate certain emotions and attitudes in himself/herself, and further induce the subject to express them. Therefore, this model includes three stages: eliciting condition, feeling state, and expression (Feng & O'Halloran, 2013: 82; Feng & Qi, 2014: 587). Ortony, Clore, and Collins (1988) further divide the eliciting condition into three aspects: consequence of event[8] (pleased or displeased), action of agent (approving or disapproving), and aspect of object (liking or disliking). The expression of emotions incorporates a broad spectrum of cues displayed by the speaker such as utterance, prosody, facial expression, and gesture. Generally speaking, the speaker, driven by the same emotional state, usually shows consistent and coordinated emotional cues (see Section 3.3 for an introduction to the STFE-match principle). Otherwise, the inconsistent cues may be related to other implications, or the speaker's occurrent emotions are complex and multilayered.

This model is a valuable reference for judging the speaker's situated emotions when he/she is performing an illocutionary force. Still, in general, it is too simple – it only involves three steps from the elicitation of emotions to the expression of emotions, ignoring the speaker as a 'live, whole person,' who can carry out multimodal perception, rational evaluation, and behavioural motivation, etc.

In Scherer (1987, 2009), emotion was conceptualized as an emergent, dynamic process based on an individual's subjective appraisal of significant events, and the 'component process model' focusing on the dynamic aspect of emotion processes was proposed. In this model, emotions contain a process in which five elements are synchronized, including cognitive appraisal, motivation and action tendencies, physiological responses, motor expression, and subjective experience. From there, Scherer (2014: 218–220) proposed the 'tripartite emotion expression and perception model' to explain how the speaker continuously conveys emotions through a variety of resources, and how the observer speculates on what kind of emotions the speaker expresses through various clues.

The model clearly shows that after evaluating actual events, the speaker can employ multiple resources in such forms as face, voice, and body to express emotions. The observer can also speculate and perceive the speaker under the influence of sociocultural contexts and psychobiological architecture through visual modalities, auditory modalities, etc.

Scherer's theoretical model is valuable for the judgement of emotional states in our study based on the following considerations. Firstly, the model regards emotion as a dynamic process instead of a static state, consistent with our view, so emphasis is given to situated emotions in this study. Secondly, the model involves a process where the speaker expresses emotions, and the hearer interprets emotions, which together formed the theoretical basis for researchers to analyse various clues and speculate on emotions from the observer's perspective by taking the vital role of multimodal cues into account. Third, Scherer (2014: 210) pointed out that the speaker's emotions

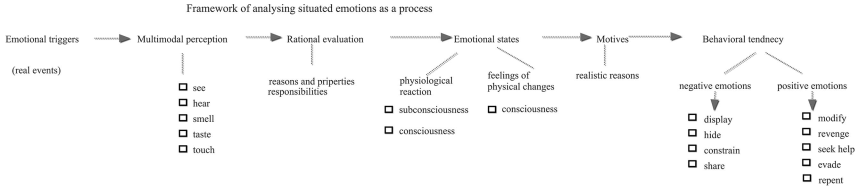


Figure 4.4 Framework of analysing situated emotions as a process.

affect pragmatic forces, which conforms to the view that ‘emotional states influence the performance of illocutionary forces’ in this study.

Scherer’s theoretical model set the theoretical basis for examining emotional expressions in multimodal behaviours. Here we introduce the approach of the ‘six-step procedure of emotional analysis’ proposed by Gu (2015b). This emotional research model also laid emphasis on analysing the emotional process and took elements involving emotional expression into account, such as multimodal perceptions, rational appraisal, motivation, and action tendencies. Such an analytic approach borrowed the plot of ‘Song Jiang Slays Po Xi in a Fit of Anger’ from Chapter 21 of *The outlaws of the marsh* to analyse Song and Po’s respective emotional states, as well as the readers or audience’s possible moral emotions. Figure 4.4 illustrates the process analysis approach.

Here, we take the analysis of the illocutionary force of ‘accusation’ as an example. First of all, after analysing the illocutionary act ‘accusation’ in the conceptual model, we discover that the occurrence of negative events against or that harm the interest of parties concerned generally constitute the precondition of ‘accusation’ (i.e. illocution-and-reality interdependency is in this book’s terminology). In the situated discourse, a real event that has occurred or may occur in the future that harms the interests of the speaker turns out to be a realistic starting point that triggers the accompanying emotions of ‘accusation,’ such as a real event where someone damages the speaker’s belongings but is unwilling to make compensation. Then, the speaker perceives the event through a multimodal sensory system (including seeing, hearing, touching, etc.) before carrying out an appraisal of the event and analysing its causes, nature, responsibility attribution, and the physical environment. Making an appraisal is an important step, because emotional states may vary according to different results of the appraisal. For example, if the speaker sees the event as a deliberate injury, he/she will generally produce negative emotions such as anger. Nevertheless, if the event is caused by a force majeure, the speaker may just be sad and helpless. By evaluating the event, the speaker feels the corresponding positive, negative, or neutral emotions, in addition to a series of physiological reactions. After that, certain motives are generated with reference to the concluded reasons before the

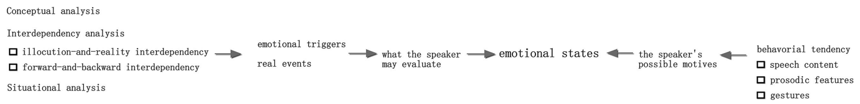


Figure 4.5 Bidirectional estimation process of emotional states.

speaker actually acts on them. According to different accompanying emotions, i.e. positive, negative, or neutral emotions, the speaker may choose to display such emotions explicitly, hide them inwardly, seek revenge, avoid, repent, or ask for help.

This study adopts a bidirectional approach in exploring the emotional states behind live illocutionary forces in the situated discourse.

One direction is to identify the 'usual, justified' emotions of the speaker in terms of a certain type of illocutionary force based on communal experience, e.g. 'accusation' is usually accompanied by negative emotions like anger, while 'congratulations' generally provoke positive emotions. Additionally, we also consider the illocutionary force's contexts and analyse the triggering event through 'interdependency analysis' to figure out illocution-and-reality interdependency, forward-and-backward interdependency and so forth.

The other direction is to infer and determine the speaker's motives through observing his/her actual behavioural performance (speech and gestures) in the multimodal corpus, including utterances, prosodic features, gestures, etc.

From both of the above directions, we are able to judge the situated 'emotional state'[9] of the speaker when performing an illocutionary force. Figure 4.5 illustrates such an inference process:

2 Correlations between three-tier emotions

As the speaker's emotions are complicated in situated discourse, they affect the illocutionary acts/forces differently and complicatedly, so researchers need to analyse them tier by tier. This also explains why different tiers of emotions are occasionally not consistent with each other in the multimodal annotation of some illocutionary act-tokens (example later).

Background emotions (constitutional emotions) are closely related to the speaker's physical condition for a while, and are therefore relatively stable and often remain the same in the situated discourse. Background emotions are the basis of primary emotions and social emotions, which always occur in the situated discourse.

Primary emotions and social emotions change directly as events or activities change, and do not coexist with background emotions. In previous sections, we have discussed the types of emotions that directly affect the felicity/infelicity of illocutionary acts/forces. According to the aforementioned concept modelling of illocutionary forces, the occurrence of some emotions constitutes the crucial factor of a felicitous illocutionary force, which refers to primary emotion and social emotion in this study. Specifically, if certain emotions are absent from a given illocutionary act in the situated discourse, it would be infelicitous. However, if they appear correctly, they become a necessary condition for the illocutionary act-token felicity. In reality, primary emotion and social emotion can show up synchronously or asynchronously. If only one of them occurs, the absent emotion is considered to be neutral or insignificant.

3 Subdivision of three-tier emotions

In this study, primary emotions and social emotions are further divided into occurrent emotions/situated emotions and reported emotions.

Emphasis is laid on the interlocutor's situated emotions, which coexist with the illocutionary forces (Gu, 2013b: 7) and directly affect the speaker's illocutionary force.

We also focus on non-present/non-situated emotions, known as reported emotions, which refer to emotions accompanying a past event that happened in 'other time' and is described by the speaker through ways like narration. 'Other time' (non-present/non-situated time) is compared with the moment when the illocutionary act/force occurs (i.e. when the speaker is performing the locutionary act), and refers to the timing of something that happened before or after the locutionary act. If it is necessary to distinguish between situated emotions and reported emotions as the speaker is performing an illocutionary act/force, and how to analyse and characterize both emotional tiers, we have to consider 'interdependency,' a subset in concept modelling, as well as the specific situation of the illocutionary force instance; in other words, whether it is necessary to make a distinction between situated emotions or reported emotions, and how to annotate them not only relating to the illocutionary act/force's properties (by analysing the conceptual model) but also relating to the specific circumstances of the corresponding instance in a given situation (by analysing situational factors).

Here we take the illocutionary act 'fear' as an illustrative example.

There are several illocutionary act-tokens of 'fear' in our constructed multimodal corpus. In adopting the analytic methods from concept modelling, analysis of the illocution-and-reality interdependency of 'fear' cases involves a frightening event that has occurred/occurring/will occur.

a A particular case would be when the event in question occurred before the illocutionary act of 'fear' where the further classification of emotions is required. Taking the illocutionary act-token 'Zhang – talking about children's education + mentioning past experiences – fear 2' as an example, the speaker said that in her childhood, she was scared by her father who used to intimidate her with a knife. Recalling an experience from the past is called a reported illocutionary act, producing a corresponding reported illocutionary force. At this time, primary emotions and social emotions should be subdivided into two tiers: situated emotions and reported emotions. Situated emotions refer to the emotional state the speaker is experiencing in performing the locutionary act of 'fear,' while reported emotions refer to the emotional state that the speaker was experiencing during the frightening event. The occurrence of both emotions involves two aspects: (i) When both situated and reported emotions under primary emotions reflect the state of 'fear,' the illocutionary act is felicitous. That is, the speaker is not only frightened in performing the locutionary act but also in recalling the event; (ii) when the chances are that the situated emotions are inconsistent with the reported emotions, i.e. the reported emotion of the speaker is fear, which is default or weakened in the situated emotion, then the illocutionary act is infelicitous.

b When frightening events occur synchronously with the speech acts, there exist only situated emotions.

c When frightening events occur after the locutionary act, this also involves only situated emotions. In brief, whether it is necessary to subdivide situated and reported emotions under both primary emotions and social emotions depends on the interdependency analysis of different kinds of illocutionary acts, i.e. whether it involves an event provoking certain justified emotions in the speaker, which occurred before the speech act. Such differentiation explains why sometimes the speaker embodies distinctive features in his/her prosodies and gestures when describing personal experiences and things that scare him/her, and sometimes appears calm without relevant features. The reason is that the reported emotions and the situated emotions do not seamlessly match with each other.[10]

4.2.4 Prosody features

Prosody features[11] are also essential aspects of the analysis of illocutionary acts/forces. Although Austin did not explicitly include prosodic features in the research scope of speech acts and elaborate on them, he once pointed out that some features like tone of voice, cadence, and emphasis in colloquialisms also serve as aids in performing speech acts (Austin, 1962: 73–76). Obviously, these are prosodic features. Prosodic changes arise from the changes in pitch, intensity, and duration of sound at the suprasegmental level. These three elements collectively produce various effects and pragmatic functions at different perceptive layers, generating diverse illocutionary forces.

Gu (2002a: F32) once gave an example:

(At a train station)

A: Go! Get out of the way!
B: What's wrong with you? No way!

B was upset with A because A's remarks did not correspond to B's expected perlocutionary acts. Obviously, in this example, A's tone, embodied through some prosodic features, is to be blamed. Judging from A's tone, B believes A was issuing an 'order' that failed to meet the felicity condition of communication. Thus, it led to an unexpected result. It would be the right thing for A to resort to justified prosodic features to show a tone of negotiation and apology to convey the speech act of 'request.' It is observed that prosodic features, to some degree, determine the types of illocutionary force.

Research into prosody enjoys a long history. The relevant arguments on speech prosody found in Ancient Greek rhetoric can be regarded as the origin of prosodic studies in the West. Prosodic research in China stems from the study of rhyming and rhythms, which belong to the aesthetic category of literary language. After the birth of phonology, our forefathers carried out relatively systematic research of the beautiful rhythms of Chinese. Moreover, by integrating the findings of rhyming and rhythms in poems, the science of phonology, centring on the study of sounds, rhyming, and rhythms, began to take shape. After the May Fourth Movement (May 4, 1919), the Chinese linguistics community began to systematically accept the Western theories pertinent to phonology while embracing traditional phonetic theories, which paved the way for the modernization of Chinese phonetics (Wu & Zhu, 2001: 11). Famous Chinese linguists such as Fu Liu, Yuen-ren Chao, Changpei Luo, Wangdao Chen, Shengshu Ding, Zongji Wu, and Jun Wang all made outstanding contributions to different aspects of the study of Chinese prosody. From the perspective of disciplinary development, Liu (2007) gave a detailed discussion on the study of prosody in Chinese by probing *The study of the Peking Dialect Intonation*, authored by Yuen-ren Chao and published in 1929. Generally speaking, the study of prosody in Chinese evolved from philology studies to phonology and prosodic grammar. Modern linguists began to pay attention to the study of the relationships between semantics and pragmatics at the prosodic and syntactic levels. With the establishment of Chinese phonetic databases, prosodic studies reached the discourse level.

The basic function of prosody is to express discourse meanings. Intensive research has been conducted on prosody's expressive function (especially intonation), regarding intonation as a defining feature in the discrimination of particular syntactic structures (declarative sentences, interrogative sentences, imperative sentences, and exclamatory sentences). Overseas scholars such as O'Connor and Arnold (1961), Bolinger (1986, 1989), and Ladd (1980, 1996) explored diverse intonational meanings in English, including speech acts, attitudes, and beliefs or feelings of the speaker, in addition to syntactic modality. Among them, Halliday (1970) studied the functions of intonation in standard British English and distinguished five types of intonation, which he investigated by taking into account contexts, the speaker's intentions, etc. Under the significant influence of Halliday's research, scholars later conducted in-depth explorations of the communicative function of intonation in English, such as Bing (1985), Gussenhoven (1984), and Tench (1996). Brazil (1975, 1978, 1985, 1997) attached particular attention to pitch and analysed units' intonational meanings below the pitch curve. Hirschberg (2002) explored the relationships between speech prosodic features and Grice's 'theory of conversational implicature,' and concluded that prosodic features convey various pragmatic meanings in different contexts. These studies were all carried out from the functionalist intonational view. Many scholars delved into the types and communicative function of intonation in Chinese (that is, the study of functional intonation); for example, Shen (1998) made a classification of functional intonation in Chinese.

Another example is Wang (2008) who pointed out that prosodic cues like sentence stress and intonation depend on pragmatic factors. The stress of the sentence is mainly contingent on its pragmatic focus. To be more specific, the speaker underlines with emphasis such information that he/she believes to be unfamiliar or to have been misunderstood by the hearer to draw the hearer's attention. Intonation indicates whether the speaker is sure about something and the role he/she plays in the communication. Some scholars analysed the acoustic features of various tones with communicative functions based on phonetic experiments. Others studied how intonations in Chinese convey various tones by adopting mathematical modelling methods.

As the research scope expanded from sentences to larger units such as discourse, prosody's pragmatic function (especially intonation) was valued again (Couper-Kuhlen, 2009: 177). Researchers explored the logic behind prosody by examining the prosodic phenomenon in natural conversations, treating prosody as a crucial clue of natural speech analysis. Halliday (1967, 1970) was the first to realize that intonational units can be used as a way for the speaker to divide chunk information as well as linking old and new information in the discourse; Bolinger (1972a, 1972b) also discussed the pragmatic functions of intonation. However, the majority of studies delved into how prosody conveys information primarily based on monologic discourse instead of talk-in-interaction data. With audiovisual technology development, researchers could collect much prosodic information appearing in natural conversations, such as pauses, inhalations, laughs, stress, and intonation. Gumperz (1982) and Gumperz and Berenz (1992) took prosody as a vital clue to social interaction. In the relevant studies, prosodic features became the focus of interactional sociolinguistics. Such features include intonation, loudness, stress, vowel length variations, registers, and meaning groups (to separate and highlight chunk information through pausing, speeding up, or slowing down speech rate). Couper-Kuhlen and Selting (1996: 11–56) suggested a way to deal with prosody in natural speech in an interactional perspective and believed that prosody emerged in social interactions as a strategy deployed by interactants in the management of turn-taking and floor-holding; some scholars such as Tao (1996) and Tseng (2009) examined the important role of rhythm in expressing meaning in naturally occurring verbal interaction; still other scholars delved into the vital role of rhythm in analysing multilingual conversation (Couper-Kuhlen & Ford, 2004); Selting (2010: 14–16) summarized four paradigms for prosodic research from the talk-in-interaction perspective; Couper-Kuhlen (2009: 178–182) reviewed prosodic studies in natural interactions from structural and interactional perspectives; Barth-Weingarte, Reber and Selting (2010) reviewed the origin, scope, and method of prosodic research in verbal communications, and presented the relevant results achieved in prosodic studies in natural multilingual conversations.

4.2.4.1 Prosodic hierarchy

The Western linguistic community has not reached an agreement on the definition and scope of prosody. Firth (1957) thought prosody to be the syntagmatic relationships between syllables that are not determined by words or the structure of discourse. The definition does not strictly distinguish between linguistic and paralinguistic characteristics in the traditional sense and categorizes suprasegmental features resulting from changes in pitch, loudness, duration, and voice of quality as prosodic features. Such a general definition is more applicable to conversational prosodic research (Selting, 2010: 5).

Following Firth, the principal prosody features examined in this study include speed rate, pause, stress, pitch, tone of voice, and other prosody-related paralinguistic information. Sagisaka, Campbell, and Higuchi (1997) summarized the effects of prosodic features into three aspects: linguistic (lexical, syntactic, and semantic), hyperlinguistic (purposes, attitudes, and rhetoric), and non-linguistic (e.g. emotions). Linguists have investigated the prosodic phenomenon in languages from many different theoretical perspectives. For example, Feng (1995, 1997, 2000) delved into the relationships between vocabulary, syntax, and prosody. Explorations on prosody in this study are conducted at phonetic and pragmatic levels to discover how prosodic features embody the speaker's emotions in performing illocutionary forces and how they facilitate the performance of illocutionary forces, and to analyse the changes in speech prosody caused by emotions via perceptual judgement. The interactive relations between prosody and syntax, and intensive statistical analysis of prosody's acoustic features are not covered.

In our analysis, prosody plays two major roles in the performance of the illocutionary act/force. The first is its differentiation role: It helps differentiate the performance of one act/force from another when we classify illocutionary acts/forces or set up corpora. The second is its effects on the felicity/infelicity states of performance as an important subset in the Discovery

Procedure. It should be noted that since different languages own different prosody features, their roles are also different. It is necessary to distinguish between a 'behaviour-selection prosodic system' and a 'latent prosodic system.' The former is a prosodic resource that potentially exists but is not presently evident or realized for all languages, while the latter, the major concern in this study, is the prosodic performance of a specific language in situated discourse (Gu, 2013a).

To study prosodic features in illocutionary forces, it is necessary to stratify prosody, that is, to limit the prosodic layers in our research. There exist many theories about the layering of prosody in phonetics and phonology. Vogel (2009: 62) argued that, despite the differences between languages, prosodic constituents at all layers exist in all languages, and some of them are universal, including prosodic words, clitic groups, phonological phrases, intonational phrases, and phonological utterances; Lin (2000, 2001, 2002), Cao (2001), Li (2002), and many other scholars explored the layering and segmentation of prosody in Chinese. It should be noted that prosodic hierarchy and syntactic prosodic hierarchy are two different concepts. The former is characterized as 'pure prosody units,' involving hierarchical units entirely defined by the superficial prosody of the utterances, while the latter takes into account the syntax (Wang, 2008: 248). Among many prosodic hierarchy models, the most famous and representative one for the aforementioned prosodic hierarchy and syntactic prosodic hierarchy can be found in Seilkirk (1984) and Nespor and Vogel (1986), respectively. Wang (2008: 249–250) points out that these two models are independent layered schemes that vary in nature and meet different research purposes. Although expressions like prosodic words and prosodic phrases are seen in both schemes, they refer to different prosodic layers (i.e. whether the syntactic layer is covered).

Next, let us look into the prosodic hierarchy through the information units. Sinclair (1991) believed that everyday language is composed of a series of phrases or multi-word units. Accordingly, the speaker combines a series of 'semi-preconstructed phrases' when making utterances. In conversational analysis and corpus linguistics, phrases or multi-word phrases are the main analytic units (Adolphs & Carter, 2013: 111–141). Prosodic segmentation can divide the flow of speech into basic units, and the occurrence of a pause (including pauses in dragging tones like 'um/uh') is the boundary dividing these basic units that some scholars treat as the units of information/thought (Chafe, 1980; Halliday, 1967; Kjellmer, 2003; Rühlemann, Bagoutdinov, & O'Donnell, 2013).

In this study, we refer to Chafe's (1979, 1980, 1988) research in intonational phrases, the fundamental unit of discourse production, to segment the basic prosodic units for analysis. Levelt (1989) associated intonational phrases with the information the brain processes before utterances are produced. Chafe (1994) pointed out that intonational phrases are related to discourse information units processed by the working memory in our brains, and that their occurrence also allows hearers to process the discourse information unit as a whole. Chafe associated intonational phrases with the focus of consciousness – the amount of attention the speaker pays to certain information in a certain period. Lin and Adolphs (2009) attempted to prove that phraseological units constitute a whole from the viewpoint of psycholinguistics and are consistent with the information chunk, by examining whether they are phonologically coherent. The relevant cognitive findings indicate that in long-term memory, mental vocabulary is stored, extracted, and processed as a whole in the form of phrases (Dahlmann & Adolphs, 2007: 49–56; Erman, 2007: 47–48; Schmitt, 2004). When intonational phrases are matched with the chunks in our brains, it involves the issue of brain representation. Are the chunks psychologically realistic? Based on preliminary research, chunks, imprinted with psychological reality, are units in the mental lexicon stored and processed as a whole. Currently, many relevant theories have been developed and proved through experiments, and the relevant reviews can be found in Huang and Zhan (2011) and Yi and Lu (2013).

In summary, since our principal purpose in researching prosody features is to discover how they serve the transmission of illocutionary forces, they are segmented according to the prosodic properties of the surface speech flow, and syntactic analysis is excluded. For the same reason, we adopted Seilkirk's (1984) prosodic hierarchy model that only deals with the final prosodic forms without involving syntactic factors. Therefore, the smallest prosodic units in our analysis are intonational phrases that can be raised to a higher prosodic level when the need arises.

4.2.4.2 Categorization and analytic method of prosody features

In previous sections, the prosody features examined in this study include speech rate, pause, stress, intonation, and pitch as well as other prosody-related paralinguistic information, because those cues play a significant role in shaping illocutionary forces.

Speech rate, pause, stress, intonation, and pitch can be observed through acoustic parameters. Of course, the corresponding parameters only make sense when compared with illocutionary forces in various contexts. As a channel of information exchanges in interpersonal communication, paralinguistic information has two distinctive features: One is that it carries sound elements, and the second is that it bears no fixed semantic meanings (Li, 2002). Traditionally, voice information generated by phonetic acts, including intonation, loudness, duration, speech rate, and the quality of voice, is categorized as paralanguage. Such voice information was covered in the previous analysis of prosodic features. Therefore, paralanguage here is narrowed to functional utterances, including laughing, crying, sighing, grumbling, coughing, booing, etc.[12] Speech act theory defines illocutionary acts as giving idiomatic and meaningful utterances. In the traditional sense, paralanguage cannot be called a speech act because it carries no fixed semantic meaning, though it is phonic. That is, it is a phonetic act, but not a phatic act. Nevertheless, from the linguistics perspective, although paralinguistic information relies heavily on the context or individual factors of the speaker, it does convey a certain pragmatic purpose in connection with the discourse and plays a supporting role in the transmission of illocutionary forces.

Based on the above discussions, prosody features are divided into three layers: prosodic units, tonal patterns and modes, and other prosodic features.

In terms of acoustic parameters, the author borrowed the matrix and parameters of the Jiugong mode of basic Chinese intonation proposed by Wu and Zhu (2001) (specific standards are introduced in Chapter 6).

4.2.5 Gestures[13]

Scholars have valued the relationship between gestures and meaning expression for quite a long time. Early researches can be found in Goodwin (1981), Heath (1986), and Kendon (1990). Kendon (2004) introduced the classification of these gestures through a review of studies on how gestures express meanings, and explained the channels of meaning interaction between languages and gestures from the perspective of human behaviour.

In this section, we first spell out the rationales behind gestural analysis, which is followed by an introduction to approaches to the classification of gestures.

4.2.5.1 Gestures and speech acts of a 'live, whole person'

We have mentioned in previous sections that pragmatic research in the past often held the belief that language use is an activity in which people use language to communicate, and the crux of pragmatics is to clarify how language is used for information exchange (Gu, 2010: xiv). In this respect, facial expressions, gestures, and postures are only regarded as paralinguistic information or non-verbal information. In other words, the language subject in the theories of previous mainstream pragmatics is not a 'whole person' but a talking head who can only communicate with people. However, real language use tells us that it is insufficient to use only information communication theory as the starting point and theoretical foundation of our research, which has been common practice in the past pragmatic studies and turns out to be idealized 'language usage.'

Researchers, faced with this problem, need to view language issues from a behavioural perspective by borrowing general behaviour theories in sociology and philosophy and deal with language-use issues within the analytical framework built on the fundamental structures of ordinary act theories, since language use is intertwined with social activities (Gu, 2010: xvi). Gu (2013b) proposed a 'whole person' model through modelling a subject engaged in situated discourse activities with dynamic sounds, emotions, and appearance, known as the ‘live, whole person’ in concept modelling. STFE involves four perspectives for exploring a ‘live, whole person,’ i.e. what is said, what is thought, what is felt, and what is embodied. Accordingly, gestures are an indispensable part of a ‘live, whole person’ in such a model. In situated discourse, a ‘live, whole person’ is extremely rich in gestures, including postures, facial expressions, the look in one’s eyes, hand/head/leg movements, and so on. Still, gestures are correlated to speech, thoughts, and feelings, e.g. staring at others as a gesture gives away a person’s thoughts, and hands shaking usually shows excitement, anger, or anxiety, but sometimes it is related to the individual’s physical condition (background emotions), such as being old, frail, or ill (Gu, 2013b: 8).

So our theoretical framework follows the idea proposed by Gu (2013a, 2013b) to examine speech acts from the perspective of the ‘live, whole person’ in live speech in the adoption of the multimodal corpus-based approach. In this way, gestures are justified to be included in the research scope.

However, it should be pointed out that our investigation focuses on those gestures that accompany utterances or are based on former/latter discourse. Austin (1962: 19) stated that many well-established acts are performed in a non-verbal way, such as betting or transferring property, but such behaviours directly expressing certain pragmatic purposes through single actions are not counted as our research objects. The underlying reason is that there must exist a locution act in the first place because the original meaning of speech act theory is ‘to do things with words.’ Locutionary act involves three kinds of acts at an abstract level: phonetic acts, phatic acts, and rhetic acts. Acts performed in non-verbal ways are excluded from our investigation due to their lack of locutionary acts.

4.2.5.2 Classification of gestures

A speaker’s gestures often show in diversity and complexity in the situated discourse. Kendon (1992: 328) pointed out that if the speaker performs such acts as clearing the throat, stretching out the legs, smoking, or taking a sip of coffee during verbal communications, the interlocutors will not take them as a meaning-loaded part of the conversation except in some particular occasions where pragmatic meanings can be conveyed through such acts (Wharton, 2009: 152). Occasionally the same act carries different meanings in different situations, e.g. shaking one’s head can be interpreted as objection, suspicion, confusion, and so forth (Carter & Adolphs, 2008: 277). This means we should distinguish between those acts that are performed in the talk to fulfil a certain task (e.g. the interlocutor hopes to speak in a clearer voice in the act of throat clearing) and those that are done in a specific condition in the hope of sending extra information through gestures (e.g. the interlocutor attempts to draw the other’s attention or make a complaint about something in the act of throat clearing).

Therefore, in speech act theory, acts are divided into those with communicative function and those without. In the framework of speech acts, the speaker's particular intention is also embedded in those acts with communicative functions (like some particular gestures) as speech does. Grice (1957) gave an example to illustrate this. If you want to drive someone out of the room as quickly as possible, you can either throw a coin out of the window or point in the direction outside the door while pushing the person gently. Obviously, although both acts indicate that the speaker/actor hopes the person to leave, the latter is considered to be an act with communicative function, because the person who is being urged to leave can recognize the speaker/actor's intention.

In this study, our annotation and analysis of gestures are not confined to the movements created by a certain part of the body (e.g. face), but cover all movements that the speaker displays. In our annotation scheme, the author examined the speaker's gestures as a whole from the perspective of behavioural theories and annotated all the notable bodily movements that would be divided into movements with overt communicative functions and those without – only incidental to speech act.

Based on our discussions above, it is necessary to classify and stratify the speaker's gestures in detail. By so doing, we are able to distinguish between a specific task-doing and a non-verbal emotional act. The former, examining speech acts under the framework of behavioural theories, is a part of the background analysis, e.g. doing something with one's hands; while the latter, interacting with other cues (e.g. prosody), is a kind of gesture in connection with an emotional state, e.g. throwing back one's head laughing. The chances are that the same gesture that might belong to different groups under different circumstances, like the previous example, clearing one's throat. Another example is the act of rubbing one's hands, which can be a task-doing under a specific circumstance having nothing to do with speech acts, e.g. the speaker rubs his/her hands to keep warm in a cold environment. Alternatively, it can be a non-verbal doing that is an external manifestation of a specific emotion, e.g. the speaker rubs his/her hands unconsciously due to tension or shyness. Besides, it should be pointed out that such classification is not meant to utterly separate task-doings from non-verbal doings or to sever the connection between task-doings and the pragmatic meanings of the discourse, since they are united and intertwined in many cases where the speaker's task-doings should be taken into account in the analysis of illocutionary acts/forces.

In short, this study divides the gestures into non-verbal doings[14] and task-doings due to their differences in implications and purposes. Non-verbal doings can be further divided into facial expressions, head movements, hand movements, posture, etc. Some non-verbal doings co-occur with speech, like facial expressions or pointing. At the same time, others usually have nothing to do with the situated discourse, such as scratching

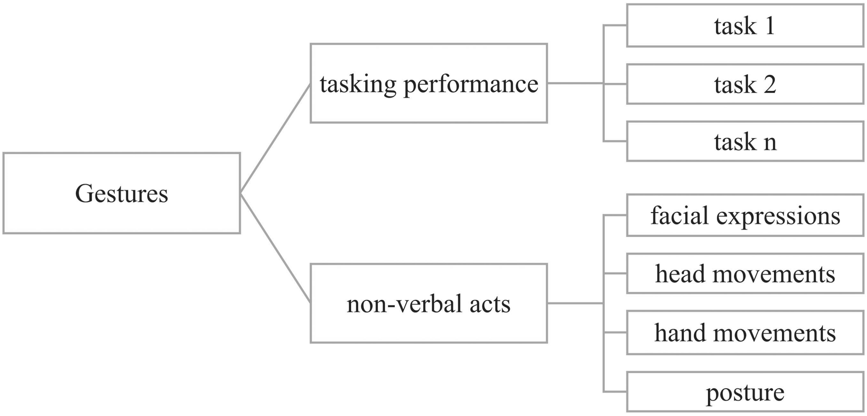


Figure 4.6 Subsets of gestures.

one's head, crossing one's hands, and tapping one's feet. However, at times they also reflect a specific emotional state (e.g. one might nervously scratch one's head). Figure 4.6 shows our classification of gestures.[15]

In this study, part of the gestures and prosodic features are defined as paralinguistic information because they are embodiments of the speaker's emotional states and play an essential role in the level of intensity in emotional expressions as well as prosodic structures, such as breathing, extended tones, pauses in the middle of a sentence, laughing, and crying. This paralinguistic information is also included in our annotation and analysis scope, respectively presented in the multilayered annotation scheme after being further classified into the category of phonology or gestures. For instance, wheezing, extended tones, and pausing in the middle of a sentence obviously belong to the phonological category, which is analysed in the layer of prosodic features, while coughing, laughing, and crying are obviously accompanied by bodily movement, which are annotated in the layer of gestures.

4.3 Summary

This chapter has introduced the Discovery Procedure of situated discourse, including its principles, rationale, characteristics, framework, and connotation.

The Discovery Procedure in this study refers to the analytic steps in describing and analysing live illocutionary forces in situated discourse from various aspects according to concept modelling and the four perspectives provided by STFE-match assumption, i.e. what is said, what is thought, what is felt, and what is embodied. Five programme segments are involved in the Discovery Procedure, i.e. conceptual model, situational factors, emotional states, prosodic features, and gestures.

The conceptual model borrows the concept of 'set' as a formalized tool that contains eight subsets: speaker role, hearer role, performativity, essential content, intentional state, emotional state, situations, and interdependency.

Situational factors involve the analysis of the social situations and contextual interdependency of the situated discourse, where a particular illocutionary force is located.

Emotional states are divided into three subsets: background emotions, primary emotions, and social emotions. Background emotions refer to emotions directly associated with the speaker's physical condition, such as fitness, listlessness, tranquillity, nervousness, vibrancy, lassitude, passion, and fatigue. Primary emotions are the basis of human emotions, which can be recognized by facial expressions or detected from other clues like posture and voice. In addition to these two universal forms of human emotion, some emotions go beyond basic human emotions and are influenced by cultural norms, social situations, and other factors, known as social emotions.

Prosodic features mainly include speech rate, pause, stress, pitch, pitch, and tone of voice, and some other paralinguistic information related to prosody. Our prosodic analysis focuses on three aspects: prosodic unit, tonal patterns and modes, and other prosodic features.

Gestures are divided into purposeful task-doings under specific situations and non-verbal doings that are incidental to the discourse. Non-verbal doings can be further divided into facial expressions, head movements, hand movements, and postures.

Notes

1 For the meaning of constitutive rules and corresponding regulative rules, see Searle (1971: 41).

2 Gu and Zhang (2013: 225–226) explained the reasons for using set theory as a concept modelling method.

3 Sets are the basic concepts in a relational database. A relational database is a database whose values are created on the basis of a relational model. The relational model is used to express various entities in the real world and various connections between entities. This collective representation is closely related to Description Logics, which are not described here. For details, see Baader et al. (2003), quoted in Gu and Zhang (2013: 222–224). This kind of connected thought can also be applied to the relationships between the various subsets in the conceptual model of force, and because they are also related to each other, such as the intent state (attitude, belief) and emotional state, are closely connected.

4 In fact, Austin used a variety of other words to name inappropriate behaviours, such as non-plays, misplays, miscarriages, misexecutions, non-fulfilments, and disloyalties (Austin, 1962); misfire and abuse are just two of them.

5 Regarding pragmatics research, the role of affective factors in language communication cannot be ignored. For related discussions, see Jiang (2014).

6 In psychology, feelings, emotions, and attitudes are three concepts with different meanings; linguists tend to separate these three concepts but do not distinguish

between feelings and emotions. Moreover, this research accepts this view and believes that attitude is the external manifestation of people's reflections and that the experience of people on whether objective things meet their own needs is different from the internal state of feelings and emotions.

[7] Also called *The Story of the Stone*, or *Hongloumeng*, composed by Cao Xueqin, this is one of China's Four Great Classical Novels. It was written sometime in the middle of the 18th century during the Qing dynasty. Long considered a masterpiece of Chinese literature, the novel is generally acknowledged to be one of the pinnacles of Chinese fiction – Wikipedia https://en.wikipedia.org/wiki/Dream_of_the_Red_Chamber

[8] The author believes that the 'event' here should be broad, including the concept of 'state' in a narrow sense. For example, the speaker is in a state of illness or old age, which also becomes a factor stimulating the corresponding inner state (including emotions and intentions) and subsequently affects the speaker's speech (e.g. frustrating utterances), prosodic features (e.g. low voice), and gestures (e.g. frowning and lying down from weakness).

[9] In fact, this process is also applicable to the judgement of the speaker's 'intentional state' (i.e. what is thought of) in the performance of illocutionary forces.

[10] Background emotions (constitutional emotions) do not distinguish between reported emotions and situational emotions.

[11] The definition of prosody varies in different research fields, and people have understood prosody differently in different historical stages, but rhythm, supersegment characteristics, and nonlinear characteristics are always related.

[12] 'Gestures' used in this study are an important type of paralanguage in English, yet disagreements centring on whether interjections belong to non-linguistic or paralinguistic discourse exist in the history of Western linguistics, e.g. conceptualist view vs. procedural view (see Wharton, 2009: 70–106). Anyway, the important role of interjections in verbal communication has been increasingly valued.

[13] 'Gestures' used in this study refer to a broad category, including hand movements, posture, facial expressions, and a series of body-related movements. Kendon (2004) believed a gesture to be a visible action of any body part of a speaker.

[14] Scholars have not come to an agreement on the category of non-verbal doings (see Siegman & Feldstein, 1987: 351–352). In this section, the category of non-verbal doings is narrowed to information at the gestural level, excluding such paralinguistic information that obviously belongs to phonological analysis, e.g. prosody.

[15] The reason that 'act' is used to refer to gestures rather than 'behaviour' is because they are subtly different (see Gu, 2010: 208). In this study, 'behaviour' refers to the speaker's external movements that can be observed directly through sensory organs or recorded by modern imaging technology, while 'act' is an interpreted behaviour. Therefore, 'act' is used in the term 'speech/illocutionary act,' which is an observable behaviour that has implicatures after interpretation.

5 Collecting and processing multimodal data

5.1 Data collection

5.1.1 Objects and methods

This study takes speech acts and their corresponding illocutionary forces produced by Chinese illiterate people in live speech as our data source for the following reasons:

For the discourse's producer, 'illiterate people' usually refers to those who do not know how to read or write. As this group of people have hardly been affected by the written system, the discourse they produce is more like authentic spoken language, and most of their interaction with the outside world uses situated discourse, which echoes 'doing things with words,' which is the original intention of developing speech act theory. After that, some researchers investigated speech acts in written forms, though they were not the initial focus of speech act theory. Additionally, carrying out explorations in illiterate people's language-processing mechanisms, characteristics of their daily language, and the mental lexicon and phonetic strategies they employed in daily communications will help further study their language processing characteristics. It will also benefit human society by creating a favourable ground for understanding the mechanism of human language processing and brain function and providing explanations for numerous linguistic phenomena. Hence, the study of illiterate people's language use involves a series of issues related to language theories and social values (see Huang, 2014b). According to UNESCO's definition, illiteracy refers to the total inability to read or write simple sentences in any language (UNESCO, 1953: 20). The ability to read or write has become a universal criterion for the international community to judge whether someone is illiterate. For example, encyclopedias such as the World Encyclopedia (2005) and the Columbia Encyclopedia (2015) both define illiteracy as 'an inability to read or write.' After that, people's focal concern shifted from reading and writing to broader aspects, such as written language functions in various social contexts, which introduced the concept of 'functional illiteracy.' However, UNESCO's concept of functional illiteracy (1953: 20) is not discussed in this

DOI: 10.4324/9781003251774-5

study. We only defined illiteracy from the perspective of reading and writing competencies.

The Encyclopedia of China (2009), with reference to the specific definition of the 'illiterate' in China, defines 'illiterate people' as follows:

> people who have not completed the fourth grade of primary school, unable to read, or can only understand fewer than 1,500 words. People who lack the competency to apply words to participate in social activities, even though they can understand 1,500 words or more, are also illiterate. This is so-called functional illiteracy.

The Chinese government's official illiterate standard is applied in this study – urban residents who know fewer than 2,000 words and rural residents who have a vocabulary of fewer than 1,500 words are illiterate people.[1] Before collecting and recording the linguistic data, we undertook a comprehensive survey of the interlocutors' backgrounds. According to China's current literacy standards in elementary schools, a vast majority of the interlocutors had not been educated or studied in literacy classes, i.e. most of them were illiterate. Very few interlocutors were educated in the first grade of elementary school, even if they had had a good grasp of all the vocabulary taught fully in class (according to the Chinese textbook of ten years of schooling compiled in 1961, students should learn to read 1,044 words in the first grade[2]) after one year of study, they would have still fallen short of the literacy level set by China. Moreover, those students said they had little opportunity to write or read during most of their lives, so they had almost forgotten the Chinese characters learned in the past. Since the interviews did not involve the interests of the relevant party and admitting that one is illiterate implies adverse effects in the conventional sense, the interviewees usually told their educational backgrounds and literacy levels. Therefore, generally, there was no situation in which interviewees concealed the truth and claimed to be illiterate. Technically speaking, all interviewees should have taken a literacy test, and quantitative comparisons should have been made between their test results and the illiteracy standards. This study failed to conduct the test due to limited research time and energy, which is a drawback of this study.

Seeking illiterate people who are willing to accept corpus data collection is a tricky task. The author asked relatives, friends, and acquaintances to help to look for illiterate people, with the search scope expanded to other areas outside Shanghai (including Zhejiang Province and Shandong Province) to avoid illiterate people coming from a single place.

Corpus data collection information card was filled in when all corpus data was collected, including information about the speakers and interlocutors (e.g. name, gender, age, occupation, hometown, or accent), time and place, reasons and purposes of the conversation, settings, and background activities.

The collection started in August 2011 and ended in February 2012, lasting seven months.

Before the collection, recorders explained to all speakers that the collected corpus data would serve as academic research and would not be provided to third parties for profit. They also stated that internal research would be carried out while protecting the collection objects' relevant rights and interests, and the collection would be done openly after obtaining their consent to eliminate their psychological concerns. In the end, they orally allowed us to publish the linguistic data for academic exchanges in the future.

There are two ways to collect corpus data: pure observation and participant observation (Tao, 2004: 51–52). In the former case, the recorder is not at the scene; only electronic devices are set up to record the situated discourse. In the latter case, the recorder presents him/herself at the recording scene to observe the conversation as an outsider or directly engage in the conversation as a participant. To avoid the interviewee getting nervous in the face of the camera and recording devices and militate against the relevant corpus data, the recorder should first have natural communication with the interlocutor for a while, and the camera and recording devices should be placed somewhere for the required corpus data to be recorded while not making the interlocutors nervous. The study adopted the approach of participant observation.

5.1.2 Recording devices

Multimodal linguistic research requires corpus data of good quality. With current information technology, multimodal data can be carried by videos and audios. Devices adopted in this study included the Sony cassette video recorder (video data is transferred to the computer through 1394 interface cards), Sony digital high-sampling rate recorder (up to 48 kHz), and Sony high-sampling rate voice recorder (up to 44.1 kHz). We used a Sony cassette video recorder for recording in the early stage, which collected analogue signals. Nevertheless, later, we found that losing some signals was inevitable when transferring those analogue signals to the computer through a data cable, lowering the data quality. The current digital video recorders usually have a relatively high sampling rate, so they should be sufficient for this study. Compared with cassette video recorders, digital video recorders are superior in definition and convenience as their files can be directly imported into the computer, avoiding the loss of digital information in the middle. In this view, Sony digital high-sampling rate video recorders were used in the later stage of our study.

5.1.3 Recoding personnel

Recorders include the author (about 20 hours of collection time), the son of a speaker (about one hour of collection time), and two registered students in the master's programme in linguistics from the author's university (about 0.5 and 5 hours in data collection). One of the graduate students comes from the region with the same dialect as Wu, one of our speakers from Ju County, Shandong Province. That graduate student is also a relative of the speaker Wu. Since they come from the same dialect region and speak the same dialect, that student was also invited to collect and record the speaker's data and transcribe part of it. Before the recording, the author introduced the research and the recording requirements to the recorders. They were given instructions on how to use the cameras and recorders and how to fill in the information card.

Sometimes in our corpus, the recorder participated in the situated discourse and communicated with the speaker, while in other cases, the recorder was a bystander who only recorded the process where the speaker spoke with other persons.

5.1.4 Two explanations about corpus data collection

In an attempt to avoid samples from a single source, we included discourses from different dialects. Some have worried that variables could impact the research results, but this will not become a problem when looking into our research purpose and methods. There are three main reasons. First of all, this research mainly explores the speaker's emotional intonations in terms of prosodic features, i.e. the correlation between different emotions and their prosodic features. We also probe into the co-occurrence relationship between facial expressions, gestures, and prosodic features caused by a particular emotional state. A neutral emotional tone serves as a reference system in this study. Chao Yuen-ren, a Chinese linguist, once pointed out in *The Study of the Peking Dialect Intonation* (in Chinese, 1929) that 'neutral intonation' is distinctive, which varies from one place to another, while 'emotional intonation' is almost the same across countries, even sharing similarities with some foreign countries. This is firstly because in addition to tone groups made up of basic tones and regular linking tones, there are also many pitch movements that express the speaker's moods and attitudes (*Tone and Intonation in Chinese*, Chao Yuen-ren, 1933). Secondly, from the perspective of specific research procedures, this study first examined the prosodic features (including intonation) of neutral emotion shown by speakers from various dialect regions, which serve as a reference system for later research on illocutionary forces. Then, linguistic data concerning other emotions was selected from the same speaker's discourse and compared with it (comparisons were made with the same speaker). In this sense, collecting linguistic data given by speakers from different dialect regions can work out. Of course, this issue has yet to be further verified in the future.

Pijper and Sanderman (1994) pointed out that the corpus data used for prosodic research should meet at least the following two requirements: (1) The speaker should be unaware of the fact that his/her speech will be used for prosodic study, so that he/she will not deliberately emphasize a specific prosodic phenomenon (such as stresses, pauses, and speech rates) when giving utterances. In this case, the discourse will be quite natural with rich prosodic cues. (2) Major sentence patterns in the targeted language should be involved to promise the appearance of as many prosodic phrases/words of different types and different degrees of intensity as possible. Since all data collected in our corpus is from situated discourse, it should contain diverse sentence patterns that usually appear in natural discourse. Moreover, the speaker had no idea about what research the corpus data was to be used for. Thus, our collected data meets these two requirements. It is credible to explore the interactive relationship between prosody and illocutionary forces on this basis. As this study collected and recorded situated discourse, the surrounding noise was inevitably mixed into the record due to technical limitations and environmental factors, making it impossible for us to analyse phonetic features at layers lower than intonational phrases (though this is not vital according to our research purposes).

In short, the collected corpus data is qualified for the purpose of this study.

5.2 Original data and the composition of the corpus

The corpus data was collected from 12 Chinese speakers (six male and six female), with a wide age-span. There are various social situations in the corpus recorded in the regions of Shanghai, Shandong, and Zhejiang in China. To ensure the diversity of recording settings, homes, markets, construction sites, factories, office buildings, offices, canteens, etc. were taken into account as places where situated discourse occurs.

Table 5.1 shows the statistics of all conversation participators whose discourses have been collected into our corpus.

Table 5.1 Information on the participants

Recorded subject	*Gender*	*Age*	*Hometown*	*Recording location*
Guo	male	40+	Lu’an, Anhui	Yangpu, Shanghai
Zhang	male	40+	Lu’an, Anhui	Yangpu, Shanghai
Tang	male	70+	Chuzhou, Anhui	Zhabei, Shanghai
Fu	male	50+	Tonglu, Zhejiang	Tonglu, Zhejiang
Wu	male	40+	Ju County, Shandong	Ju County, Shandong
Pan	male	50+	Fengxian, Shanghai	Fengxian, Shanghai
Shi	female	40+	Lu’an, Anhui	Yangpu, Shanghai
Zhang	female	40+	Lu’an, Anhui	Yangpu, Shanghai
Wu	female	50+	Tonglu, Zhejiang	Tonglu, Zhejiang
Bo	female	50+	Tonglu, Zhejiang	Tonglu, Zhejiang
Zhou	female	70+	Tonglu, Zhejiang	Tonglu, Zhejiang
He	female	50+	Tonglu, Zhejiang	Tonglu, Zhejiang

Data obtained from fieldwork made up the original resources for this research. Videos were stored in an AVI format and audios in a WAV format when imported into the computer. Later, videos were converted into an MPG format. Although the AVI format has a higher sampling rate than the MPG format video display, it usually occupies more space and is not convenient for large-scale annotation and processing. The MPG format resolution equates to that of Super-VHS video recorders, satisfying corpus data-processing requirements. Furthermore, an MPG file occupies little computer space, convenient for data transfer and annotations carried out by Elan or other annotation tools in the future. Audios are stored in a wav format for two reasons. Firstly, a WAV file, with a sampling rate of 44.1 kHz and a bit depth of 16 bits, is lossless. With a high resolution, it can reflect various sounds in situated discourse with high fidelity. Secondly, WAV is the supported audio format of Elan in processing corpus data.

This corpus has 39 videos and 45 audios altogether, with a total duration of about 28.98 hours (equivalent to 1,739 minutes). The numbers of audio files and video files are not equivalent in the original recording data, and their file names do not match each other. There are two main reasons. On the one hand, our recording is conditioned by features of the Sony digital recorder's operation and the maximum recording time. For example, our recording devices cannot keep on recording after a pause. It will automatically save the previous record and start recording a new file, while the camera can continue to record as the same file. On the other hand, when conducting video recording of some situated discourses, their auditory information was not captured.

Subsequently, through the first filter, the author removed the corpus data marked with low quality and in which the speakers did not talk or talked little.

The filtered data resources last approximately 24.43 hours (1,466 minutes in total), given by 10 speakers. Videos were divided into 35 files and stored in MPG format according to the dates they were collected. Each file may involve more than one speaker. The files were named by recording dates, recording places, and speakers; the corresponding audio files stored in the WAV format were also named by recording dates, recording places, and speakers. They stored computer hard disks, and mobile hard disks, with a total size exceeding 83.1 GB. The filtered data resources amounted to 364,498 Chinese characters after transcription, forming our 'multimodal corpus of situated discourse.'

Information on the filtered multimodal corpus of situated discourse is shown in Table 5.2.

For the convenience of storage, the author physically split the chosen videos and audios into units about one hour long. Because the collected original multimodal corpus data was unprocessed, it belonged to the raw corpus data. Our first segmentation targeted raw corpus data. Such segmentation was not endowed with specific meanings and did not follow a specific rule.

Table 5.2 Information on the filtered multimodal corpus of situated discourse

Recorded subject	*Native place*	*Dialect quarter*	*File name of video*	*Duration (minutes)*	*Transcription (Chinese characters)*
Zhang	Lu’an, Anhui	Jianghuai Mandarin – Hongchao area – Luzhou zone	11y08m15d talking about work 1 11y08m15d talking about work 2 11y08m15d talking about children’s education + past experience (start from 00’22’’10) 11y08m18d pitching tents 11y08m26 talking about theft and campus freak 1 11y08m26 talking about theft and campus freak 2 110816 Zhang talked in the food market and the dormitory 110818 Guo Hui’s birthday 110922 Shi – drank pesticide 110927 colleague collected trash 110929 assembling children’s bicycle 111006 cleaning; Zhang talked about his son working as an automotive technician 111006 having a meal at Guo’s 1 111006 having a meal at Guo’s 2 111006 chatting with Guo; Tongtong 111006 Shi bought food in the supermarket 20110905 cleaning 20110912 mid-autumn festival 1 20110912 mid-autumn festival 2 20110914 woodcarving; talking about the son 1 20110914 woodcarving; talking about the son 2 20110926 chatting with Guo 20120229 – 10:30 Guo putting up a script on the wall of a new building at Tongji	909	249,202
Shi	Lu’an, Anhui	Jianghuai Mandarin– Hongchao area – Luzhou zone			
Guo	Lu’an, Anhui	Jianghuai Mandarin – Hongchao area – Luzhou zone			

(*Continued*)

Recorded subject	*Native place*	*Dialect quarter*	*File name of video*	*Duration (minutes)*	*Transcription (Chinese characters)*
Tang	Chuzhou, Anhui	Jianghuai Mandarin – Hongchao area – Nanjing zone	Li Shuangfu (before lunch) 1 Li Shuangfu (before lunch) 2 Lunch-afternoon	147	28,096
Bo	Tonglu, Zhejiang	Wu dialect –Taihu area – Linshao zone	1st January, talking with the shopkeeper in Tonglu 1 1st January, talking with the shopkeeper in Tonglu 2	113	27,564
Wu	Tonglu, Zhejiang	Wu dialect – Taihu area – Linshao zone			
He	Tonglu, Zhejiang	Wu dialect – Taihu area – Linshao zone	120113 talking with He Xiaoyi's aunt	27	7,406
Zhou	Tonglu, Zhejiang	Wu dialect – Taihu area – Linshao zone	Talking with Zhou	41	12,971
Pan	Fengxian, Shanghai	Wu dialect – Taihu area-Su – Hu – Jia zone (Suzhou – Shanghai – Jiaxing)	20120125 a talk (after lunch)	62	15,842
Wu	Ju County, Shandong	Jiao-Liao Mandarin – Qianglai area – Juzhao zone	20120206 – 17:00 (chatting + coring the vegetables) 20120209 – 11:15 (working outdoors) 20120209 – 12.38 (working in an empty house)	167	23,417
				1466	364,498

It was done to trim the files that occupied a larger digital space into smaller ones for the convenience of computer processing and physical media storage. The segment of videos and audios and their format conversions was carried out through Corel VideoStudio Pro X4, a multimedia processing software.

The storage media for corpus data are mobile hard disks and CD-ROMs. Simultaneously, to avoid data loss, the author stored copies of the corpus data on multiple media.

5.3 Notes on corpus data transcription

In this study, the multimodal corpus of situated discourse (24.43 hours) was transcribed into Chinese characters with Pinyin (the official romanization system for Standard Chinese). Since most of the speakers spoke in dialects or Mandarin with regional accents, there were usually no corresponding Chinese characters that conveyed the intended meanings. At this time, we could either transcribe them into Chinese characters with similar pronunciations or annotate them with Pinyin. Besides, a proper form of text transcription should be adopted according to research purposes (see Ochs, 1979). Considering that the study of speech acts should take into account the structure and content of former–latter utterances to represent the interdependence between daily conversations, we adopted a top-down transcription. The transcribed corpus data was stored in the computer in Word documents using Microsoft Office.

It should be noted that corpus data was transcribed into Chinese characters for the convenience of reading, annotation, and retrieval. In actuality, the interlocutors only participated in the conversation with sound and prosodic cues, without Chinese characters. Concerning the illiterate people examined in our study, there was no concept of Chinese characters in their minds any time they were conversing. Their conversation skipped over the interpretation of meaning to the production of speech (currently, it remains to be confirmed whether non-illiterate people experience the stage of encoding meanings into Chinese characters).

For the multimodal corpus of Chinese illiterate people's situated discourse with a scope of 24.43 hours (1,466 minutes), we only transcribed its speakers' discourse content into texts. However, the multimodal corpus, which contains 134 illocutionary force instances, was transcribed more in detail – the speakers' prosodic features, gestures, and other information were included. The whole process of transcription was completed through Elan.

In this study, the transcription of all multimodal corpus data was done manually. There were four transcribers in total: the author, one of the speaker's relatives, and two persons familiar with the dialect. Before transcribing, other transcribers were trained by the author and given instructions on how to conduct transcriptions. Also, transcribers needed to proofread the transcripts.

5.4 Summary

This chapter introduces the acquisition and processing of the original multimodal corpus of situated speech in this research.

This study targeted illiterate people. Twelve illiterate people were involved in the original database, six males and six females, ranging from middle-aged people to older people. Our recording took place at various locations, including urban areas, outskirts, and rural areas in Shanghai, Shandong Province, and Zhejiang Province. The original database encompasses 39 videos and 45 audios altogether, with a total duration of about 28.98 hours (1,739 minutes). After the first filter, corpus data marked with low quality, and the data indicates that the speakers who were not talking or those who spoke very little were removed. In the end, valid data that lasts approximately 24.43 hours (1,466 minutes) was collected. When all valid data was transcribed, our multimodal corpus of situated discourse began to take shape with the transcription of 364,498 Chinese characters.

Notes

[1] Under Article 7 of the 'Regulations on Elimination of Illiteracy in China' (promulgated by No. 8 Document of State Council in 1988, and amended on August 1, 1993), the standard for individual to eliminate illiteracy is that farmers can understand 1,500 Chinese characters, employees of enterprises and institutions, and urban residents can understand 2,000 Chinese characters. They should also be able to read newspapers and articles written in simple terms, keep simple accounts, and do simple practical writing.'

[2] Judging from the speakers' age distribution in our corpus, the youngest speaker is in his forties. As a reasonable guess, his first grade of elementary school should have started before 1980. According to the statistics of Chinese literacy requirements in primary schools in China's history given by Li (1985), Sun (2009), and others, the highest literacy requirement before reform and opening-up was set in the Chinese textbook of ten years of schooling compiled in 1961, asking first-year students to grasp 1,044 Chinese characters.

6 Developing a multimodal corpus of speech acts in situated discourse

6.1 Segmentation and annotation scheme and its implementation

Under the guidance of concept modelling and the STFE-match principle, the scheme of segmentation and annotation was designed according to the Discovery Procedure of live illocutionary forces. The working definition, segmentation standard, and annotation method of all tiers in the multimodal annotation scheme were introduced according to the exemplification of a specific instance. The formulation of the scheme was at the stage of data modelling. At the implementation stage of segmentation and annotation, the author carried out the pilot annotation with Elan to label the audios and videos in layers according to the scheme. In this way, we could produce processable, storable, readable, retrievable, and computable data for the computer, which prepared itself for further statistical analysis in the framework of Discovery Procedure.

Here is a comprehensive introduction to the segmentation and annotation scheme of illocutionary force with an example.

This illocutionary force instance is labelled 'Mrs. Zhou – harmful – grumble.' Set in the social situation where the 70-odd-year-old elderly lady Zhou from Tonglu, Zhejiang, was talking to the author in her living room; next to her was her husband listening to the conversation. This illocutionary force falls into the group of 'grumble,' in which the speaker recalled the hardship she experienced in childhood. The data lasts 77 seconds.

Transcription (// = pause):

Alas//really suffering//at that time//they all said//it would be more consummate //if your mother is still alive//your mother//didn't enjoy her life//but only suffered a lot//we suffered from hunger

I told my father when returned//told my father//two [error] three persons [repair] took only about a pound of flour for two days//stir with water

how miserable//we were then.

DOI: 10.4324/9781003251774-6

6.1.1 Performance unit of illocutionary force: working definition, segmentation standard, and annotation method

WORKING DEFINITION

We combined speech act verbs search and line-by-line analysis of the transcription (McAllister, 2015: 31) in searching and determining the illocutionary force instances in the 24.43 hours of multimodal corpus of situated discourse.[1]

Speech act verb search is suitable for illocutionary acts that contain explicit performative verbs, such as searching the keyword of 'satisfy' to find illocutionary force instances of 'satisfaction,' or searching for 'thanks' to look for instances of 'gratitude.' In contrast, line-by-line analysis aims to determine those illocutionary acts without an explicit performative verb.

Basic methods of identifying the illocutionary act-type of a specific illocutionary force.

When any explicit performative verbs appeared, the instance was judged according to the traditional classification method based on the performative verbs. For example, as there is an apparent performative verb in the sentence 'I declare the sports meeting open!' it falls into the 'declaration' group.

Meanwhile, we considered the following factors when explicit performative verbs do not show up:

1 Since context plays an important role in determining illocutionary force types, we had to analyse the speaker's discourse in the context of social situations and take into account its interdependencies, i.e. we had to conduct the 'situational analysis' mentioned earlier.
2 We also took the speaker's speech content, prosodic features, gestures, emotional states, and the like as essential clues in judging illocutionary types.

Given that natural discourse was used as the corpus data in analysing and identifying the types of illocutionary forces (Adolphs, 2008: 9), we also resorted to the Conceptual Model as the primary validation tool to reexamine our results.

Figure 6.1 presents the identification flow of instances of speech acts.

In situated discourse, the length of the performance unit varied from one illocutionary act to another in different situations, and other types of illocutionary force or multiple turns of speaking might also show up. According to our work definition, as long as that illocutionary force fitted into the conceptual model of a particular type of illocutionary force, we would still categorize it into that group.

SEGMENTATION STANDARD

When segmenting the corpus data, the researchers dealt with various performance units of a live illocutionary act located in situated discourse (the

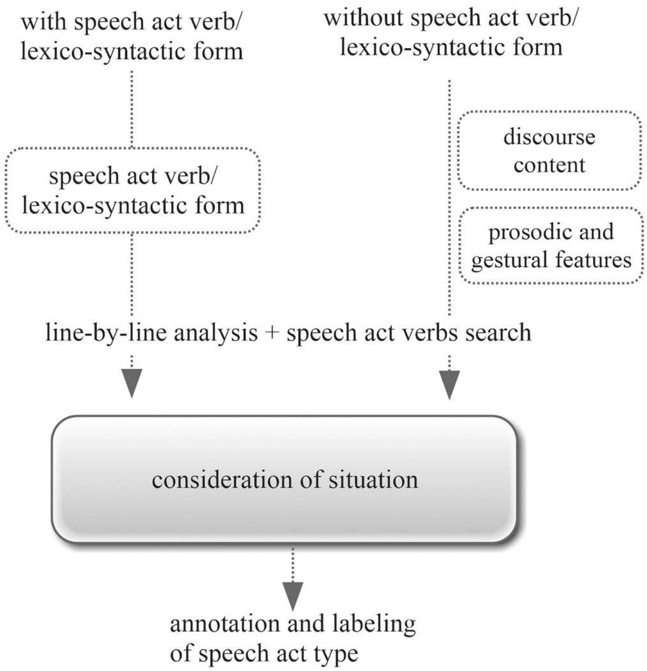

with speech act verb/ lexico-syntactic form
without speech act verb/ lexico-syntactic form
speech act verb/ lexico-syntactic form
discourse content
prosodic and gestural features
line-by-line analysis + speech act verbs search
consideration of situation
annotation and labeling of speech act type

Figure 6.1 Identification flow of speech act-type.

performance of an illocutionary act falls into a token). A single performance of an illocutionary act could vary in the length of time in different situations. Sometimes an illocutionary act-token had only a single turn of speaking, and its shortest length could be a single syllable such as ‘go!’ that represents the ‘urging’ speech act. Sometimes a single performance of an illocutionary act might last a long time, containing more than one turn of speaking, with utterances and gestures alternatively appearing in it. Regardless of the length of time, we used ‘each illocutionary act/force’ to define the segmentation units. Such segmentation and analysis units are called ‘performance units’ (Gu, 2013a).

In this study, a performance unit is equivalent to an instance of a speech act. The former is a token corresponding to the speech act-type in live speech, while the latter is a working unit or the object of study containing all the properties and information held by a live speech act, which is artificially defined for research convenience.

Following Gu (2013a), we used intonation group as the basic unit for analysing and segmenting the performance unit. A performance unit can be either a long one or a short one. It can be as short as an intonation unit that only contains a monosyllable or as long as multiple intonation units. For example, when introducing a few people, repeated utterances and gestures might occur in the intervals, such as applause and thanks.

We defined the segmentation boundary of the performance of an illocutionary act as follows:

In this study, in principle, the beginning and end of an intonation group on the time axis marked the segmentation boundary of a performance unit or the performance of an illocutionary act.[2]

After such segmentation and definition, we determined the core analytic unit for the study of illocutionary forces. In the corpus data multimodal analysis framework, the 'performance unit' of illocutionary force is an umbrella-term referring to the working unit of a live speech act in situated discourse. Its hyponym includes the segmentation and annotation of prosodic features, emotional states, and gestures. In terms of Elan, it can be explained in this way: The performance unit is the core working unit, which belongs to the parent layer, while the rest become its child layers. This study did not set the parent layer mainly due to technical reasons and its demanding workload.

ANNOTATION METHOD

As we segmented the performance unit according to its functions in a speech act, there is no such fixed boundary of a performance unit (single or multiple turns of speaking). More specifically, it is identified from the ontological attributes of a speech act's performance, without using any formal judgement standards. Therefore, the annotation tags of our illocutionary force instances are equivalent to their corresponding types, e.g. 'explanation,' 'complaint,' 'boast,' and 'compliment.'

CASE STUDY

Identification of the types of illocutionary force:

'Alas, really suffering. Suffering at that time…' In such utterances, we could conclude that the speaker was 'grumbling' if we combined it with the setting and other multimodal information, including non-verbal information such as frowning, shaking the head, and dropping the eyes.

The speech act 'grumble' can be via a formalized method called 'octet scheme,' as we mentioned before, enabling us to summarize and generalize this illocutionary force type (see Figure 6.2).[3]

Next, we analyse this instance according to the octet scheme, which includes eight subsets, while placing it in the specific contexts.

<Speaker's role>

Here we only have one speaker, i.e. an old lady whose surname is Zhou.

<Hearer's role>

Here are two hearers – the recorder (the author) and Zhou's husband. The speaker was talking to the author while her husband was sitting on her right side, but he was outside the frame.

<Performativity>

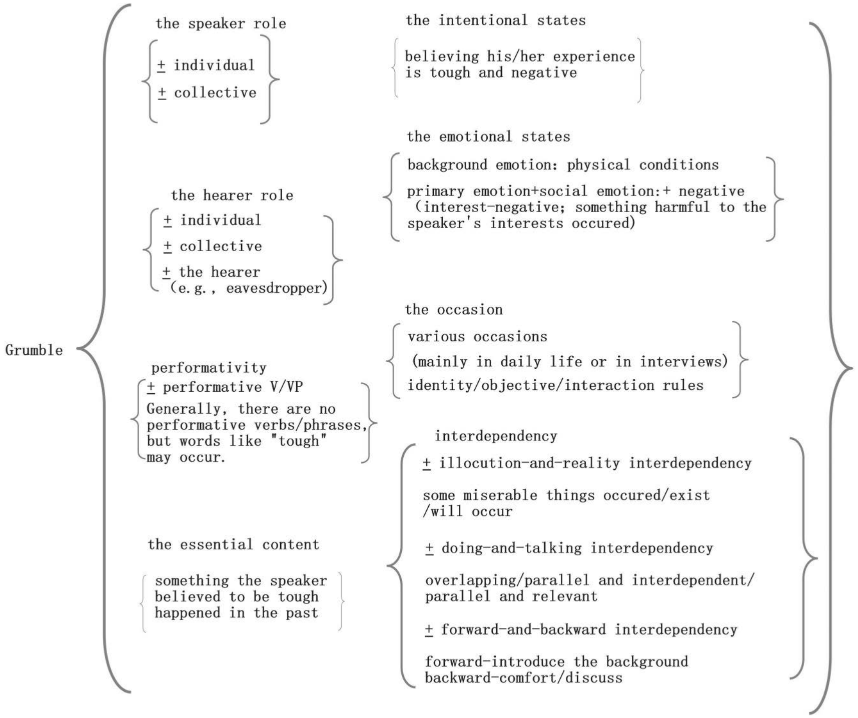

Grumble
the speaker role
± individual
± collective
the hearer role
± individual
± collective
± the hearer (e.g., eavesdropper)
performativity
± performative V/VP
Generally, there are no performative verbs/phrases, but words like "tough" may occur.
the essential content
something the speaker believed to be tough happened in the past
the intentional states
believing his/her experience is tough and negative
the emotional states
background emotion: physical conditions
primary emotion+social emotion:+ negative (interest-negative; something harmful to the speaker's interests occured)
the occasion
various occasions
(mainly in daily life or in interviews)
identity/objective/interaction rules
interdependency
± illocution-and-reality interdependency
some miserable things occured/exist /will occur
± doing-and-talking interdependency
overlapping/parallel and interdependent/ parallel and relevant
± forward-and-backward interdependency
forward-introduce the background
backward-comfort/discuss

Figure 6.2 Analytic scheme of illocutionary forces.

Although there are no explicit performative verbs, we can still decipher its performativity through other discourse content. For example, the speaker said, 'Alas, really suffering. Suffering at that time....' To some extent, the term 'suffering' serves as a discourse marker suggesting this instance's performativity.

<Essential content>

This instance's essential content is conveyed through overt discourse carrying the propositional content that directly shows the speaker's 'grumbling,' namely, the speaker's mother died when she was a child and suffered from starvation.

<Intentional state>

In this instance, through the speaker's prosodic features, emotional state, and gestures, we can observe that she felt pained about her past life experience. In other words, she believed that the previous experience was harsh and miserable.

<Interdependency>

The interdependency of the speech act 'grumble' covers three aspects. Forward-and-backward interdependency: forward interdependency – this family lived by picking out herbs from weeds. Backward interdependency – the hearer asked if any of her family members were martyrs. Illocution-and-reality interdependency – something tough and tragic did happen in the speaker's experience in the past. In principle, this instance was based on the past since the speaker recalled her hardships in the past. The doing-and-talking interdependency was overlapping (talking is doing).

<Occasion>

In this instance, the old lady Zhou was talking to the recorder, recalling her memory of living a harsh life in her childhood, with her husband sitting nearby.

<Emotional state>

According to the aforementioned analytic methods of emotional state, three subsets of emotional state all occurred in this instance, i.e. primary/universal emotion, social/secondary emotion, and background emotion. Altogether they reflected the speaker's overall sadness. Details are provided below in the annotation notes elaborating how we layered the emotional state.

After the speech act was determined as 'grumble,' we analysed its segmentation and labelling:

According to our previous definition, the beginning and end of an intonation group on the time axis mark the segmentation boundary of a performance unit or the performance of an illocutionary act. The starting point and ending point of a performance unit, respectively, equate to that of the corresponding intonation group.

In the instance above, the speaker's emotional state, prosodic features, and gestures co-occurred. However, in many cases, gestures may occur before any utterances. This is referred to as follows: 'anyone who gives away his anger before saying anything is full of rage; anyone who gives utterances tinged with anger is feigned' (*The Classified Characters and Political Abilities*, written by Liu Shao, quoted in Gu, 2013b: 11). This case involved the segmentation of 'pre-gesture performance unit' and 'post-gesture performance unit.'

In this study, we did not directly take facial expressions, postures, and other gestures as boundaries to segment the performance unit of speech acts. If any closely related gestures that lasted longer than the performance unit arose before or after it, we segmented them into 'pre-gesture performance unit' and 'post-gesture performance unit' and annotated them accordingly.

Here is an example from our self-constructed corpus in which prosody, emotions, and gestures did not co-occur – this required us to segment the 'pre-gesture performance unit' and 'post-gesture performance unit.'

This instance belongs to the speech act 'comment,' which lasts 60 seconds (an interview taken after lunch in Fengxian – neutral class – talking about the factory director's son). The speaker talked about the factory director's son, including his work in the factory. Its task-doings were sitting, chatting, and smoking. Nevertheless, the beginning and end of the speech did not overlap the instance's scope. This was because he made some non-verbal gestures before and after his comment. Before the comment, the speaker shook his head for about 1.5 seconds, and after the comment, he smiled for 2.5 seconds. This is more commonly the scenario of natural conversation – generally, the speaker's gestures and the stream of speech do not co-occur. Instead, the former may show up before but disappear later than the latter. These gestures sometimes speak louder than the speech itself (Bolinger, 1986: 213–214). Therefore, they become very vital clues in analysing the speech act. In this instance, we still used the boundary of the intonation group to define the boundary of the 'performance unit' and marked those non-verbal behaviours and task-doings that appear before or after the performance unit as 'pre-gesture performance units' or 'post-gesture performance units' (see below). Such gestures were meaningful when analysing the speech act 'comment' since they demonstrated the speaker's attitude towards the factory director's job placement and the job his son was doing.

Among the 134 instances in our multimodal corpus of speech acts, a total of six instances were marked with 'pre-gesture performance unit' or 'post-gesture performance unit' (4.48%). Those six instances are listed as follows:

1 The interview was carried out in Fengxian (after lunch) – neutral class – talking about the factory director's son;
2 The interview was carried out in Fengxian (after lunch) – beneficial class – expressing satisfaction over staff meals in the factory;
3 110914 Discussing the character of the factory director's son – harmful class – complaining;
4 20110926 Guo's talk – neutral class – explaining how to spray pesticide;
5 Wu Yulin from Shandong (fieldwork) – neutral class – explaining;
6 20120215 Interview with Pan (after lunch) – judging.

6.1.2 Activity types: working definition, segmentation standards, and annotation methods

WORKING DEFINITION

Previous sections discuss the connections and difference between social situation and activity type: An activity type is essentially similar to a social situation (Gu, 2006b: 127–167), but a social situation refers to a general environment where discourse takes place, containing one or more activity types at the same time. In our corpus, generally, a social situation only involves one activity type, but there are also situations where multiple activity types co-occur in a single social situation. In our self-constructed multimodal corpus, there are also examples in which several activity types are involved in one social situation. Example 1: The speaker was 'bargaining' over the price of vegetables in the food market while some people discussed its quality, and other people were buying fish or fruit. In this case, the speaker, vendor, and hearer formed an activity type. Aside from that, other independent activity types also existed in this social setting – the food market. Example 2: In a broad social situation of 'assembling a toy car,' the speaker was assembling a toy car, the mother was holding the child back from creating trouble, and some colleagues around were talking about the toy car. In this case, at least three activity types were involved.

Besides, activity type has a close relation with task-doing, which belongs to the tier of gestures.

SEGMENTATION STANDARD

In this study, the segmentation of activity types is carried out according to the scope of performance units. If multiple turns of speaking arise, e.g. the conversation is interrupted by others, interrupters should also be included in the annotation scope of the discourse type in question because all these verbal activities belong to the same activity type. For instance, chatting, discussing, and setting up a tent together can fall into the same activity type.

ANNOTATION METHOD

Making use of Chinese markers to describe the activity types, such as pitching a tent, sitting and chatting, and having a haircut.

CASE STUDY

This performance of the illocutionary act 'grumble' took place in the speaker's home, where she was talking with the recorder about her past hard life experience, with a hearer sitting nearby. Therefore, this instance was labelled as 'conversation.'

6.1.3 Turn-taking: working definition, segmentation standard, and annotation method

WORKING DEFINITION

Turn-taking refers to every continuous stream of speech given by each speaker in the situated discourse. In this study, turn-taking is only a form of unit, a method of organization in conversation and discourse where participants speak one at a time in alternating turns. It should be noted that turn-taking is only a unit, while illocutionary forces are functional, so the two do not always seamlessly match. Sometimes one turn corresponds to multiple illocutionary forces; sometimes, multiple turns occurring before or after the same speaker's turn correspond to one illocutionary force.

Segmentation standard

Generally speaking, the scope of turn-taking is very apparent, and there are few disputes about its segmentation. Typically, we would take the switch of speech sounds to define the scope of turn-taking. If the performance unit only corresponds to one turn-taking, its scope will overlap the turn-takings. This is true in this instance of 'grumbling.' However, some exceptional cases also exist:

1 Interruption

Interruption by one or more speakers in the same speaker's turn-taking can be further divided into two situations: one is that the interrupter's discourse overlaps the original speaker's, and the other is that the original speaker stops speaking until the interrupter finishes his/her talk. In this study, interruptions are annotated differently according to different discourse content and property. If the interruptions are relatively short without any discourse with explicit analytical meanings but only contain terms expressing agreement, confusion, or exclamation, such as 'well,' 'oh,' 'huh,' generally we did not annotate them individually. We surrounded the interrupter's discourse content with parentheses and insert it into the original speaker's discourse. Such interruptions, either overlapping the speaker's turn-taking or occurring alone, are known as backchannel (see Knight, 2011a; Knight & Adolphs, 2008). If the interruption lasts a long time and is worthy of analysis, we segmented and annotated it in particular.

2 Interval

Occasionally in the situated discourse, talking while dealing with other things, talking while thinking, or talking while experiencing some specific emotions (e.g. too sad to talk) can prolong the intervals between the same speaker's turn-takings or of different speakers. In this study, the scope of turn-taking was defined by speakers, i.e. it started with and ended with the same speaker. The scope of the next turn-taking began with the initial syllable uttered by the next speaker.

3 Relationship between the segmentation of turn-taking and gestures

Two tiers concerning non-verbal behaviour and task-doing were established (both belong to gestures). Occasionally, turn-takings did not co-occur with gestures, which resulted in 'pre-gesture performance unit' or 'post-gesture performance unit.' At this time, these non-verbal behaviours (such as laughing, crying, and coughing) and task-doings (such as trimming vegetables for cooking and slicing up food) were individually processed in their respective tiers of annotation. Their scopes started at the moment they appeared and ended when they vanished. Turn-takings were entirely divided by the switches of voices.

ANNOTATION METHOD

Annotations were added to the speaker and the discourse content. The discourse content was annotated with previous transcriptions.

CASE STUDY

As only the old lady Zhou was involved in the turn-taking of the illocutionary act 'grumble,' and there was no noticeable sustained interruption or delay, we took the start and end of the utterances to define the scope of a complete turn of speaking. Accordingly, this instance lasted 77 seconds.

6.1.4 Emotional state: working definition, segmentation standard, and annotation method

There are three subsets in this study's tier of emotional state: primary/universal emotion, social/secondary emotion, and background emotion. In situated discourse, these subsets sometimes co-occurred, sometimes not. Therefore, they were annotated respectively.

This study refers to the classification of emotions developed by previous researchers and treats it as an essential parameter in our multimodal annotation system. Sophisticated instruments for psychological measurement were applied to detect and record the speaker's emotions. The annotator or investigators mainly judged the speaker's emotions according to their analysis and induction. Based on the STFE-match principle, they observed the speaker's demonstrative cues and made judgements. The discourse content, prosody features, and gestures, including facial expression, hand gestures, and postures, were worthy of observation.

The steps and methods of judging emotions are elaborated in Section 4.2.3.2. Other scholars' research results on emotional expression forms and judgement of emotional cues set the basis for defining, segmenting, and annotating the three subsets of emotional state.

Sometimes, the speakers would directly encode their emotions in the verbal language in situated discourse. There are two possibilities. First, the encoded emotions are equivalent to the speaker's illocutionary acts/forces, e.g. in the sentence 'I am happy,' 'happy' is both an emotion and an illocutionary act/force. Second, the encoded emotions only serve as an emotional state of the illocutionary acts/forces, e.g. in the sentence 'I have to invite them to dinner politely,' 'invite them to dinner' is an illocutionary act/force, and the speaker's emotional state was respectful, pleasant, etc.

Since prosody features serve as crucial clues for expressing emotions, scholars have conducted intensive research on how they express emotions (see Scherer, 2003). In terms of acoustic parameters, the fundamental frequency is one of the core clues. The maximum, minimum, and average fundamental frequency and domains directly reflect the ways of emotional expression. For example, the fundamental frequency can effectively distinguish between neutral emotions and angry emotions and even identify the intensity of anger (Mathon & de Abreu, 2007: 81). Duration is another important acoustic feature. The lengthening and shortening of a sentence or stressed syllables in duration are necessary forms of emotional communication. Some studies even conducted a quantitative analysis of Chinese emotional intonations based on phonetics experiments (e.g. Cowie & Cornelius, 2003; Hirose et al., 2000; Liu, 2009). Liu (2011) provided a comparative summary of the distribution and performance of Chinese neutral emotions and the prosodic features (e.g. stress, fundamental frequency, duration, and pauses) of some emotional intonations (e.g. happiness, anger, sadness, and astonishment). Other scholars explored the relationship between the grammatical mood of a sentence (feelings, attitudes, and emotions in one's speech), mood particles, and intonations (e.g. Jin, 1992; Shi, 1980; Zhang, 2008). Liu (2009: 30) claimed that pitch-level changes with emotional development in imperative sentences, serving as the leading indicator of the speaker's excitement intensity. Whether the emotion is positive or negative, fierce excitement matches high pitch level and vice versa. This also applies in exclamatory sentences – low pitch level usually represents a state of excitement with low intensity, while high pitch level correlates with excitement with high intensity (Liu, 2009: 23).

People also express emotions through gestures, including facial expressions, hand movements, and postures. Sometimes, these gestures are more effective than other ways (such as utterance content and prosodic features) in expressing the speaker's emotions (Allwood, 2002). Since Darwin (1872) claimed that facial expressions are a universal form of human emotions and that humans are born with the ability to display their emotions through facial emotions, many scholars held in-depth discussions about the universality of facial expressions and cultural specificity. It has been proved that gestures, including facial expressions, physical attributes, postures, and the like, can reflect the speaker's emotional state. Proven facts can be found in Ekman (1972), Scherer and Ekman (1984), Ekman and Davidson (1994), Russell and Fernandez-Dols (1997), and Kipp and Martin (2009). Tao and Tan (2004), from the Institute of Automation of the Chinese Academy of Sciences, found that the relationship between individuals' postures and emotions carried certain messages and generally changed with the progression of interaction. For example, if someone makes more extensive use of gestures, it usually emphasizes his/her attitude. If tremors occur in a certain part of one's body, it usually implies that they are nervous.

All in all, previous studies have shown that emotional state would manifest itself in a variety of external cues. Speech, prosody, bodily movements, facial expressions are important ways to identify the speaker's emotions (Planalp, 1998). Therefore, in the situated discourse, we mainly judge the speaker's occurrent emotions by various interconnected clues, such as the speech content, prosodic features, and gestures (Gunes, Piccardi, & Pantic, 2008). Reported emotions are mainly inferred based on the abovementioned clues and the conceptual analysis of illocutionary act-types. In the following sections, we enlarge the working definition, segmentation standard, and annotation method of background emotion, primary/universal emotion, and social/secondary emotion.

6.1.4.1 Background emotion

WORKING DEFINITION

The background emotion hinges on the speaker's physical condition and directly impacts the prosody. For example, a healthy speaker speaks loudly in a higher pitch on a specific occasion, indicating that the speaker is full of energy, while a speaker in poor health tends to speak in a low tone of voice, which sounds listless.

SEGMENTATION STANDARD

Identifying background emotion can be based on analysing clues about the speaker's prosodic features, postures, and body movement's speed and trajectory (Damasio, 1999). In this study, we also made use of clues like prosodic features (e.g. voice quality, the strength of voice, intonational changes) and gestures (e.g. facial expressions, gestures, postures, and bodily movements) to identify background emotions, and segmented them according to performance units. Part of the prosody features was positively correlated with the corresponding emotions. Although studies in the past treated prosody features including fundamental frequency, duration, intensity, and voice quality as important clues (Devillers & Vidrascu, 2007: 35), many scholars still believe that prosodic features alone are not adequate to accurately identify and categorize emotions (e.g. Batliner et al., 2003). Therefore, we must combine multiple clues when making judgements.

In the previous illocutionary act-token 'Mrs. Zhou – harmful class – grumble,' although Zhou was grumbling about her previous suffering, she spoke energetically. Most of the time, she kept an erect posture while sitting and sometimes leaned back. She made frequent use of her gestures and facial expressions, such as pointing, clapping, shaking her head, and frowning. All of these showed that the speaker was still relatively healthy and energetic despite old age. Besides, in light of the prosodic units' segmentation, the frequency and duration of her pauses were moderate, and she could

Table 6.1 Analytic scheme of tags

Attribute name	*Tag type*	*Tag value*	*Remark*
emotional state	background emotion	fitness listlessness tranquillity nervousness vibrancy lassitude passion fatigue	emotional markers that mainly reflect the speaker's physical conditions

talk in a smooth voice flow. Based on the above clues, it can be concluded that when performing the illocutionary force instance of 'grumbling,' the speaker's background emotion was 'energetic and healthy.'

ANNOTATION METHOD

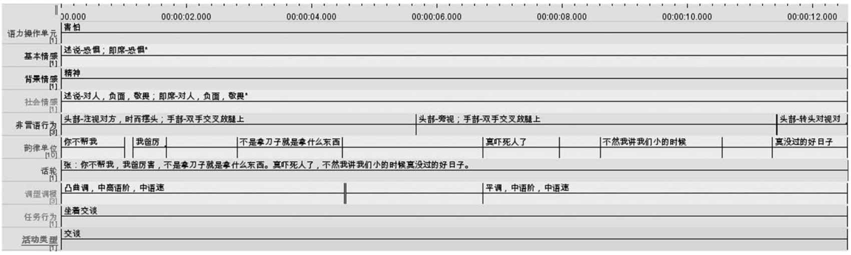
Tag value refers to emotional markers that mainly reflect the speaker's physical conditions as shown in Table 6.1.

6.1.4.2 Primary/universal emotion

WORKING DEFINITION

As a universal subset of emotions, primary emotion spans cultures and races, occupying a central position, among other emotions. Although researchers have discrepancies on the number of primary emotions and the definition of primary emotions, some emotions have been widely recognized as primary emotions. This study analysed the seven most fundamental emotions: happiness, sadness, fear, anger, surprise, disgust, and worry. Occasionally, one primary emotion can further produce a stronger or weaker subset, which still belongs to primary emotion, such as delight being a weaker form of happiness, and ecstasy a more vital form of it. If that were the case, the subset adjectives were used to label the corresponding emotions more appropriately.

In modern emotional research, researchers have found that many emotional expressions are universal. Body languages including facial expressions, gestures, and postures all provide evidence for annotating primary emotions. In recent years, some researchers have built a multimodal corpora of emotions to examine the relationship between various acts and emotions (such as Bänziger, Pirker, & Scherer, 2006). Ekman (2003) analysed the correspondence between various clues of facial expressions and each subset of primary emotions. For example, surprise can be expressed in two areas of the face, and the meaning of two different combinations is different: A



语力操作单元 [1]	害怕
基本情感 [1]	述说-恐惧；即席-恐惧*
背景情感 [1]	精神
社会情感 [1]	述说-对人，负面，敬畏；即席-对人，负面，敬畏*
非言语行为 [3]	头部-注视对方，时而摆头；手部-双手交叉放腿上 头部-旁视；手部-双手交叉放腿上 头部-转头对视对
韵律单位 [10]	你不帮我 我爸厉 不是拿刀子就是拿什么东西 真吓死人了 不然我讲我们小的时候 真没过的好日子
话轮 [1]	张：你不帮我，我爸厉害，不是拿刀子就是拿什么东西。真吓死人了，不然我讲我们小的时候真没过的好日子。
调型调模 [3]	凸曲调，中高语阶，中语速 平调，中语阶，中语速
任务行为 [1]	坐着交谈
活动类型 [1]	交谈

Figure 6.3 Annotation of emotion in three tiers.

combination of eyes and eyebrows expresses questioning surprise, a combination of eyes and mouth expresses astonished surprise, and a combination of eyebrows and mouth expresses dazed or less interested surprise (Martin & Devillers, 2009: 271). In many cases, the speaker may have two or more primary emotions, such as feeling angry and sad. Ekman and Friesen (1975) also described in detail these facial expressions with mixed emotions.

As mentioned before (Section 4.2.3), based on interdependency analysis, primary emotions can be divided into situated emotions and reported emotions according to research needs. Examples are given in this study to illustrate the definition of situated emotions and reported emotions. We used the instance 'Zhang – talking about children's education and past experiences – Fear 2' for illustration. Here, we explain it again with its annotations.

In this illocutionary act-token, the speaker said that her father used to threaten her with a knife when she was little, which scared her greatly. Since this incident occurred before she expressed 'fear,' two tiers should be established under primary emotion and social emotion, namely situated emotion and reported emotion. Situated emotions refer to the emotional state the speaker is experiencing when performing the locutionary act of 'fear', while reported emotions refer to the emotional state that the speaker was experiencing in the frightening event.

For instance, we could speculate that the speaker's primary emotion was fear from her utterance 'It freaked me out.' Again, judging from her prosodic features and gestures in situated discourse, we could infer that the speaker's situated emotion was also fear, even though its degree might be relatively weakened. Similarly, her reported and situated social emotions were 'other-directive, negative, and awed,' but the latter was relatively weaker in intensity. The background emotion here was 'fitness,' and there was no need to divide it into situated emotion and reported emotion.

The annotation is shown in Figure 6.3.

SEGMENTATION STANDARD

In this study, the speaker's primary emotion was identified according to various clues, including his/her corresponding utterance content, prosodic features (e.g. intonational changes), and body language (e.g. facial expressions, gestures, postures). Based on this, annotations were added to it, and then it was segmented based on the scope of the performance units. There were two experts to verify the reliability of every illocutionary act-token's annotation of emotional state.

Here are some examples of typical primary emotions:

1 Happiness

In the performance of the illocutionary act 'Bo – interview with a shopkeeper in Tonglu on January 14th – satisfaction', the speaker Bo always wore a smile when performing the illocutionary force. Utterance content with distinctive features includes the following:

> xiang qi lai ye shi hen shu fu de o
> xiang wo men zhe yang // yi qian ku de xian zai hao le
> wo de xin hen ping de
> feels very comfortable at the thought of it
> like us//we do not suffer anymore
> I am feeling peaceful

In terms of prosodic features, the pitch reached a high and middle level, and the corresponding utterance was 'like us//we do not suffer any more.' In conclusion, Bo's situated primary emotion was 'happiness.' See Figure 6.4.

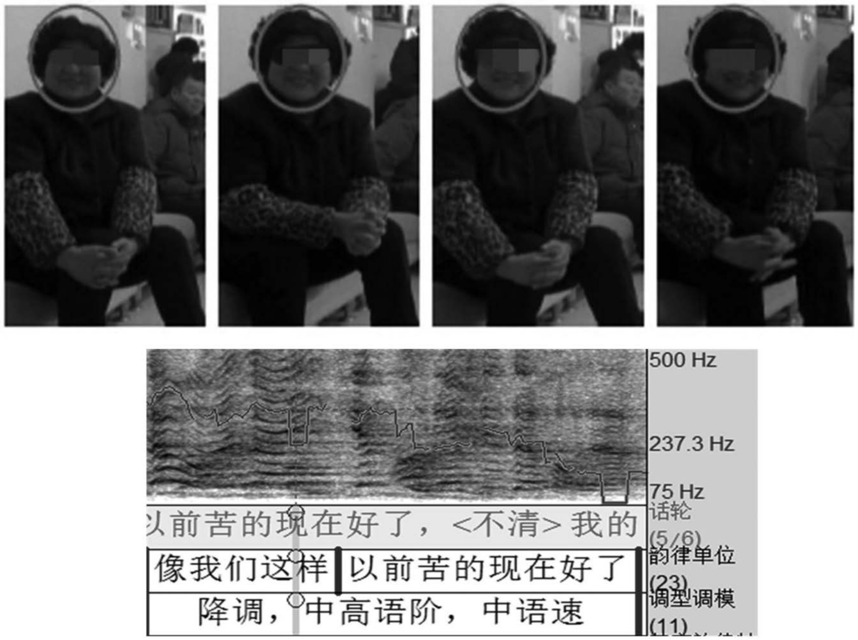


Figure 6.4 Instance of 'happiness.'

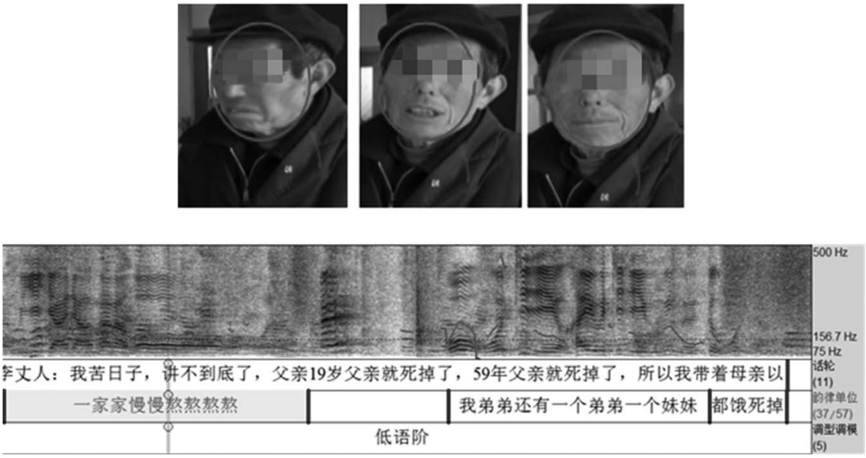

500 Hz
156.7 Hz
75 Hz
李丈人：我苦日子，讲不到底了，父亲19岁父亲就死掉了，59年父亲就死掉了，所以我带着母亲以
话轮
(11)
一家家慢慢熬熬熬熬
我弟弟还有一个弟弟一个妹妹
都饿死掉
韵律单位
(37/57)
低语阶
调型调模
(5)

Figure 6.5 Instance of ‘sadness.’

2 Sadness

In the instance ‘Li Shuangfu’s father-in-law (before lunch) 2 – harmful class – complaint,’ the speaker recalled some tragic experiences – his father died early, he married late due to poverty, and his younger brothers and sisters starved to death in a famine. The speaker looked serious and grave throughout the conversation, frequently shook his head, and waved his hands. From prosodic features, the speaker spoke in a low and slow voice with many pauses. These clues all suggested that the speaker’s situated emotion was ‘sadness.’ See Figure 6.5.

3 Anger

In the instance ‘20110905 cleaning – harmful class – complaint & curse: angry 2,’ the speaker Shi complained that someone always soaked the mop in water in the sink and closed the doors and windows of the toilet upstairs tightly, resulting in an unpleasant odour. One characteristic of complaining is that it often co-occurs with such speech acts as threatening, fulfilling one’s duties, and cursing. In this case, cursing appeared before and after complaining. The speaker’s voice reached a very high level of pitch. Regarding the fundamental frequency, the average prosody of anger is in the high-frequency band (Liu, 2011: 61) and rapid speech rate, studded with stresses. Let us look at the gestures. Frowning was an ever-present facial expression that lasted for a long time (six seconds), directly reflecting the speaker’s corresponding primary emotion (anger). See Figure 6.6.

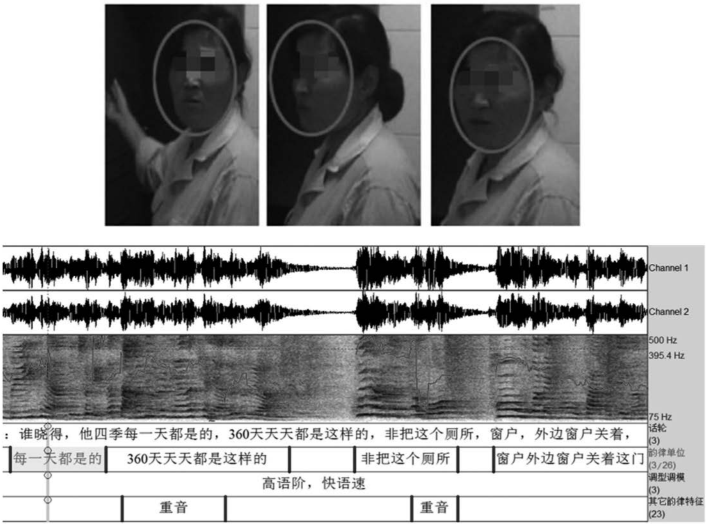


Figure 6.6 Instance of 'anger.'

Table 6.2 Emotional tag code analysis

Attribute name	*Tag type*	*Tag value*	*Remark*
emotional state	primary emotion	anger fear sadness disgust surprise happiness worry nothing	other emotional markers produced by varied emotional intensity

ANNOTATION METHOD

According to the discussion in Section 4.2.3, seven interculturally and interracially accepted emotional tags and 'nothing' (no obvious emotional cues) are adopted as tag values here. The tag codes are shown in Table 6.2.

6.1.4.3 Social emotions

WORKING DEFINITION

Social emotion is closely related to the setting where situated discourse locates and directly impacts illocutionary forces. Their classification varies from culture to culture as it relates to cultural traditions. Social emotions fall into three groups: positive, negative, and neutral (Gu, 2013a). There are two subsets in each group, i.e. other-directive and self-directive. These two subsets are to analyse whether social emotion serves oneself or is stimulated by others.

As mentioned in Section 4.2.3, based on the interdependency analysis among different illocutionary forces, social emotions can be divided into situated emotions and reported emotions, depending on research needs. Since this has been elaborated in previous sections, here we will not repeat it.

SEGMENTATION STANDARD

Social emotions are tucked away in many physical cues, such as often smiling shows friendliness (Burgoon, Buller, Hale, & de Turck, 1984), while holding one's head up may imply harshness (Carney, Hall, & LeBeau, 2005). Studies have shown that the speaker's overall behavioural cues are of great significance for analysing his/her attitudes and stances[4] (Escalera et al., 2010). Thus, our identification of social emotions starts from the speaker's overall cues, which mainly include prosodic features (e.g. voice quality, intonational changes, pitch), gestures (e.g. facial expressions, hand gestures, postures), and utterance content. Social emotions are segmented by the same demarcation of performance units.

Here is an instance of a neutral illocutionary act – 'Bo – interview with the owner of a small shop in Tonglu on January 14 – explanation.' In this instance, the speaker Bo explained the archway's connotation and construction to the hearer. Bo's voice reached a medium and high pitch with a moderate speech rate, studded with stresses during the explanation. In terms of gestures, Bo kept looking into the hearer's eyes, as pointing and making hand gestures were in between, showing great interest and confidence in the introduction.

ANNOTATION METHOD

As discussed in Section 4.2.3, concerning the classification in Gu (2013a), social emotions can be divided as follows: First of all, social emotions can be divided into 'other-directive' and 'self-directive' regarding directedness. The former refers to feelings towards others, such as respect, jealousy, and contempt, while the latter refers to feelings towards the speaker himself/herself, such as inferiority, complacency, and shyness. Indeed, the same emotion can be both other-directive and self-directive, say, satisfaction. One can be satisfied with another's speech, behaviours, and performance, or just with oneself. Next, they fall into three groups: positive, negative, and neutral.

Social emotions are tagged with relevant Chinese expressions, such as jealousy, envy, contempt, and respect. The form of annotation is shown in Table 6.3.

Table 6.3 Analytic scheme of illocutionary force instance

Attribute name	*Tag type*	*Tag value*		
emotional state	social emotion	positive	other-directive	markers of social emotions (e.g. respect, jealousy, shyness)
			self-directive	
		negative	other-directive	
			self-directive	
		neutral	other-directive	
			self-directive	

In the following sections, we comprehensively analyse the three emotional tiers in the example of 'Mrs. Zhou – harmful class – grumble.'

This case is about the speaker's recalling of the hardship in her past life. From her utterance content, intonation, facial expressions, and other cues, inferences are made below.

Background emotion: This related to the speaker's physical condition and directly impacted the basic intonation. Even though the speaker was over 70 years old, she was quite emotional and energetic during the talk, so her basic emotions were judged as vibrancy and fitness.

Primary/universal emotions: As this instance was about complaining, there should be some suffering before the speech act, according to the interdependency analysis of the Conceptual Model. In this sense, basic emotions should be divided into situated emotions and reported emotions. From the speaker's utterance content, we learn that she was very heartbroken at that time and experienced some misfortunes, such as starvation caused by poverty and the premature death of her mother. On this basis, this instance was annotated as 'reporting – sad and angry.' Judging from the speaker's prosodic and physical cues, we believe that she also felt a little sad in the situated discourse, so it was labelled as: 'situated – sad*' (* means the situated emotion may be weaker in intensity than the reported emotion).

Social emotions: Since this was a recollection of her sufferings in the past, it was self-directive and negative. Three relevant episodes showcased anguish, helplessness, and self-pity, respectively. Like basic emotions, here, social emotions should also be divided into situated emotions and reported emotions.

6.1.5 Prosodic features: working definition, segmentation standard, and annotation method

We have discussed the scope of analysis of prosodic features and how they would be stratified in Section 4.2.4. This study's corpus data is segmented into intonational phrases, which are the smallest units in prosodic units. Prosodic words and syllables or other smaller phonetic units are not segmented or annotated.

Naturally, a single performance unit can vary in length, containing single or multiple intonational phrases. In other words, the segmentation boundary of a performance unit is the boundary of a corresponding intonational phrase or the outermost boundary of several intonational phrases.

Pause-extension is the interruption and prolongation of voice in the speech stream when speaking or reading, mainly manifesting as a blank and extended voice in acoustics (Wu & Zhu, 2001: 35–83). We can say that pause-extension is a prosodic feature of the discourse and an external acoustic cue conducive to the distinction of prosodic levels and prosodic units segmentation. The fact that pause-extension can distinguish prosodic levels has been proved by Wu and Zhu (2001: 41). They developed a structure consisting of inner and outer three levels of the extension and pause. Wang (2008) believed that pauses generally appear at the boundary of intonational segments or larger prosodic units, while prolongation mainly appears at the boundary of prosodic phrases.

Pause-extension in Chinese can generally be divided into three groups: physiological pause-extension (the most fundamental way of pausing, occurring out of physiological needs), logical pause-extension (the commonest way of pausing, occurring for better expression or to facilitate the hearer's understanding), and the emphasized pause-extension (differing from the regular logical pause-extension to highlight something in the speech stream). Daily verbal communication experience tells us that unusual pause-extension that contradicts logical pause-extension always carries a particular emotion or pragmatic intention in the situated discourse. It often conveys different illocutionary forces with varied durations or positioned differently or made in a different manner. Li Yu was a Chinese playwright and novelist who lived during the late-Ming and early-Qing dynasties. He wrote in *Xianqing ouji* (*Pleasant diversions*, written in ca 1667): 'In playwriting, pauses are often avoided on purpose in sentences when they are needed, and appear when coherence is needed. Likewise, chunks are often intentionally split when they should be linked and linked when they should be separate. One can only perceive its wisdom, but not explain it to others, at least not in words.' Wu and Zhu (2001: 71) summarized it thus: 'Logical pause-extension serves the reason, while emphasized pause-extension serves the emotion.'

Cruttenden (2002: 29–30) pointed out that the boundaries of intonational units (i.e. intonation phrases) ideally should be defined based on 'external standards,' namely, phonetic cues around the boundaries. Nevertheless, those cues (e.g. pauses) are sometimes very ambiguous. Accordingly, 'internal criteria' should be taken into account, for example, combining the phonetic cues with those discursive chunks that fit into the entire intonation pattern. When the necessity arises, we should also consider grammatical and semantic factors. Wang (2008: 252–254) suggested that several common concepts related to the classification of prosodic units include (intonation) pitch range, pitch line, bassline, pitch declination, downstep, pitch reset, and pause-extension. These concepts are conducive to the stratification of prosodic units. However, this study examined situated discourse, mostly colloquialisms, which may have failed to observe grammar in written language (e.g. syntactic structures). Besides, pause-extension in spoken language is often associated with the number of speech information units that the speaker's working memory can process, and therefore it would be more appropriate for us to utilize it as a crucial sign in segmenting the intonational groups while taking into account other factors (including 'internal criteria') when the necessity arises.

What should be noted is that researchers used to adopt a uniform annotation system in corpus-based phonetic research. For example, they utilized the International Phonetic Alphabet (IPA) to annotate the pronunciation of every single prosodic unit in detail. Meanwhile, they would resort to tones and break indices (ToBI) to annotate prosodic features. The Laboratory of Phonetics and Speech Science, Institute of Linguistics, Chinese Academy of Social Sciences, also developed a labelling system, C-ToBI for Chinese prosody.[5] Since we mainly investigated how prosodic features interact with emotions, gestures, and illocutionary forces, we did not pay much attention to analysing the details of phonetic features. Instead of adopting the aforementioned common annotation scheme, we describe those prosodic features in words and label them.

Our labelling scheme divides the prosodic features into three sub-tiers, namely prosodic units, intonation patterns and modes, and other features. All the labelling of the tier of prosodic features was carried out by Praat's voice labelling system, creating the above three sub-tiers established on the working interface. Automatically generated Praat TEXTGRID files were saved after labelling, which was finally imported in the corresponding Elan files through its data channel with Praat.

6.1.5.1 Tier of prosodic unit

WORKING DEFINITION

As mentioned earlier, occasionally, a performance unit only contains a single turn of speaking, a single intonational phrase, or even a single prosodic word, such as 'go,' when urging someone to leave. Sometimes it can also contain multiple turns of speaking or multiple intonational phrases within a turn; for example, a speaker introduces a guest with many words. This reminds us to combine performance units in the analysis of prosodic units. When processing the corpus data, researchers can analyse a discourse by segmenting it into several integrated performance units (sometimes only one); within each, it can further incorporate several intonational phrases (sometimes only one). In this way, the performance unit's boundary will become the outermost boundary of its corresponding intonational phrase(s), while the former also determine the latter's segmentation.

SEGMENTATION STANDARD

In this tier, the segmentation and annotation of prosodic units was equal to that of intonational phrases. We mainly used pause-extension (an important prosodic performance) to segment intonational phrases. Undeniably, this segmentation standard is not accurate enough. Many phonetic studies segment intonational phrases according to their phonetic characteristics, including intonational coherence, the boundary tones on the ending edges of an intonational phrase, the extension of the last syllable of a phrase, and the occurrence of pauses at the ends (Wang, 2003: 16). This study did not follow this standard, mainly because it does not examine the phonetic features of the prosodic units within illocutionary act-tokens. It only observed some prominent features of the prosodic units as a whole. Therefore, phonetic features such as boundary tones and the final syllable did not qualify as standards for segmentation or annotation, let alone our including them in our investigation. Nevertheless, this approach did not impede any investigation of other prosodic features, such as intonation patterns and modes, pause-extension, and stresses.

ANNOTATION METHOD

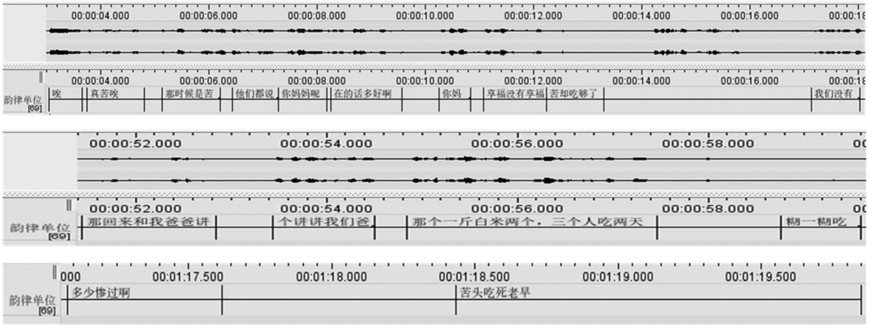
This tier was annotated with the Chinese characters corresponding to the intonational phrases. For the dialectical pronunciations without matching Chinese characters, the corresponding phonetic symbols were used. Parameters including sound wave, spectrum, and pitch were established on the workspace after the audio data of the speech act was imported into Praat, enabling us to parse the prosodic units by viewing the spectrum sonogram while listening to the audios.

CASE ANALYSIS

In this case, we still used external standards, i.e. pause-extensions, to segment prosodic units whose boundaries are defined by their duration. Blanks were left during the pause-extensions. The segmentation, in this case, was carried out as follows (// represents the segmentation boundary of the prosody unit, which is also a sign of the pause-extension). Here is the segmentation of prosodic unit (Figure 6.7).

ai // zhen ku ai // na shi hou shi ku a // ta men dou shuo de // ni ma ma ne // zai de hua duo hao a // ni ma ma // xiang fu mei you xiang fu // ku que chi gou le // wo men mei you de chi'

na hui lai he wo ba jiang // ge jiang wo men ba // yi jin bai mi liang ge [error] san ge ren [repair] chi liang tian // hu yi hu chi



韵律单位 [69]
唉 | 真苦唉 | 那时候是苦 | 他们都说 | 你妈妈呢 | 在的话多好啊 | 你妈 | 享福没有享福 | 苦却吃够了 | 我们没有

韵律单位 [69]
那回来和我爸爸讲 | 个讲讲我们爸 | 那个一斤白米两个，三个人吃两天 | 糊一糊吃

韵律单位 [69]
多少惨过啊 | 苦头吃死老早

Figure 6.7 Instance of the segmentation of prosodic units.

duo shao can guo a // ku tou chi si lao zao

[sigh] alas//really suffering//at that time//they all said//it would be more consummate//if your mother is still alive//your mother//didn't enjoy her life// but only suffered a lot//we have suffered from hunger

I told my father when returned//told my father//two [error] three persons [repair] took only about a pound of flour for two days//stir with water

how miserable//we were then

6.1.5.2 Tier of intonation patterns and modes

WORKING DEFINITION

Annotations of intonation patterns and modes of intonational phrases and higher-level prosodic units were carried out in this tier. From the acoustic perspective, pitch and sound length (the tempo of speech) were highlighted. Here are the classes of the annotation: intonation (flat tone, rising tone, falling tone, fall–rise tone, and rise–fall tone), pitch (high, medium, and low levels), and the tempo of speech (fast, medium, and slow).

SEGMENTATION STANDARD

Segmentation was done according to the intonation patterns and modes' starting time, and left the breaks without distinctive characteristics blank.

ANNOTATION METHOD

After importing the speech act audio files into Praat, a new workspace including format, spectrum, pitch, and other parameters was created.

Definition of intonation: Wu (2004: 267) pointed out that the term 'intonation' mainly referred to changes in pitch in sentences in recent years.

Table 6.4 Sound parameters of the pitch levels

Pitch level	*Upper and lower limits of basic intonations*	
high pitch level (C)	female: 450–270 (Hz)	male: 370–190 (Hz)
medium pitch level (B)	female: 330–150 (Hz)	male: 250–100 (Hz)
low pitch level (A)	female: 250–110 (Hz)	male: 170–90 (Hz)

Table 6.5 Sound parameters of the tempo of speech

Tempo of speech	*Parameters of the average time of each syllable*
fast tempo (a)	135–300 (ms/syllable)
medium tempo (b)	250–450 (ms/syllable)
slow tempo (c)	400–650–1000 (ms/syllable)

Therefore, we mainly observed the sonogram's pitch curve to identify the intonation modes of prosodic units (flat, rising, falling, fall–rise, and rise–fall).

Definition of pitch: By observing the overall fundamental frequency range of the curve in a period on the sonogram, the pitch (high, medium, and low pitch level) of the prosodic unit was defined by consulting the sound pitch parameters of three pitch levels developed by Wu and Zhu (2001: 396) based on phonetic experiments[6] (shown in Table 6.4).

Definition of the tempo of speech: We referred to the sound pitch parameters of three pitch levels developed by Wu and Zhu (2001: 398), shown in Table 6.5.

Noting that prosodic features (such as pitch and the tempo of speech) vary from person to person, comparisons should be made between different illocutionary acts performed by the same speaker when analysing specific instances.

After the parameters were determined according to the ranges mentioned above, we could mark them in each tier in TEXTGRID, a Praat workspace. We only labelled the prosodic units with distinctive prosodic features, and those insignificant were not labelled in detail.

The tagging codes are shown in Table 6.6.

CASE ANALYSIS

In this case, the tier of intonation patterns and modes were marked with 'low pitch level' five times, 'falling tone' five times, and 'flat tone' twice.

In these two prosodic units, 'really suffering//suffering at that time,' they both displayed more of a falling tone for intonation patterns. In general, when people mention something sad or helpless, their tones often appear weak or low, which manifests as a falling tone in the intonation patterns.

Table 6.6 Annotation of prosodic features

Attribute name	*Tag type*	*Tag sub-type*	*Tag value*
prosodic features	intonation patterns and modes	intonation	flat tone rising tone falling tone fall–rising tone rising–falling tone
		pitch level	high pitch level medium pitch level low pitch level
		tempo of speech	fast tempo medium tempo slow tempo

6.1.5.3 Tier of other features

WORKING DEFINITION

In addition to the aforementioned intonation patterns and modes, we still focused on other prosodic features, including pause-extension, stresses, sound quality, and some other prosody-related paralinguistic information, because in situated discourse, these prosodic features played a role in the formation of illocutionary forces in a specific context. Paralanguage here mainly refers to functional vocalization, including laughing, crying, sighing, groaning, and coughing. During annotation, we only considered the characteristics of a prosodic unit, such as ‘soft,’ ‘rough,’ and ‘vigorous’ in terms of sound quality, rather than adding functional explanations to them, such as whether a soft tone represents the speaker being friendly or just pretending to be adorable.

Here are some notes on the annotation of pause-extensions. Pause-extension serves as signs of boundary segmentation in the tier of the prosodic unit in our annotation scheme, but its scope and connotation are different from that in this tier. In this tier, pause-extensions are annotated to explore the reasons behind their formation and their possible impacts on the formation of the illocutionary force. Usually, they can be divided into multiple categories, such as choking, extended tones, and thinking time. These pause-extensions often carry special meanings, and some are directly related to the property of the illocutionary force. For example, in the complaint speech act, the speaker was too sad to talk. Here, the pause-extensions served as the signs of boundary segmentation and as a crucial role in conveying the speaker’s feelings.

In this case, the speaker was weeping while recalling in a choked voice the experience of being bullied by others in childhood. Such non-verbal gestures lasted for a long time. Since these prosodic features and non-verbal gestures are essential to our analysis of illocutionary forces, they are marked with 'pause-extensions' in the tier of 'other prosodic features.' Other insignificant pause-extensions are ignored in this tier.

Therefore, pause-extensions that serve as signs of boundary segmentation have a broader scope than those that carry special meanings, because all pause-extensions can serve as separators if they cause any stops in the acoustic wave. However, only those with special meanings are annotated in this tier as pause-extensions; others (pauses caused by coughing, throat clearing, taking a breath, etc.) are labelled with other markers according to the context.

Here are some notes on the annotation of stresses. Up to now, researchers have not reached a consensus on an accurate definition of stress. There are two forms of stress: word stress and sentence stress (Wu & Zhu, 2001: 284). This study only focuses on sentence stress, which can be further divided into grammatical stress, emphasized stress, and rhythmic stress. Grammatical stress is the pattern of stressed and unstressed syntactic structure and semantic expressions across a sentence; emphasized stress (pragmatic stress) is used to highlight certain significant words or syllables within a sentence, and rhythmic stress is used to emphasize the prosody or to compare and contrast different semantic meanings (Wu & Zhu, 2001: 289–298). We mainly focus on emphasized stress (pragmatic stress). It primarily functions at the sentence level to highlight a specific meaning or message, so they are particularly annotated in this tier. Speakers usually resort to them to draw hearers' attention to what they have no idea about, what they might have misunderstood, or what they would like to emphasize. It can be said that the speakers are subjectively using prosody to underline their intentions or feelings by enhancing the pitch, intensity, and length of the sound of the relevant words in a sentence. In this sense, the stressed words can be explained by interpreting the speaker's intention or feelings in a text with a complete semantic structure (Zhang, 2014: 121–122). Here, they were defined by observing the shadowed spectrum in the Praat sonogram and the annotator's feelings on the pitch, length of the sound, etc.[7]

SEGMENTATION STANDARD

Segmentation is done according to the prosodic features' corresponding to beginning and end times on the time bar, and the period without distinctive features are left blank.

ANNOTATION METHOD

Pause-extension is judged and marked by the breaks between the boundaries of the aforementioned prosodic units, while stress, quality of sound, and other paralinguistic information are marked at the corresponding point in time.

Table 6.7 Sample of tags

Attribute name	*Tag type*	*Tag sub-type*	*Tag value*	*Remark*
prosodic features	other prosodic features	phonetic message	pause-extension stress the descriptor for the quality of sound	other descriptors for phonetic message
		paralinguistic information	coughing clearing one's throat taking a breath	other expressions for paralinguistic information

Tags include pause-extension, stress, and descriptors related to sound quality. Paralinguistic information is annotated by the corresponding descriptors that depict the phenomenon. Table 6.7 shows some examples.

CASE ANALYSIS

In this case, pause-extensions and stress occurred four and five times respectively.

From the sonogram of the 'complaining' speech act performed by old lady Zhou in previous sections, it was observed that before the complaint, there was an extended sound (also known as an extended tone, belonging to the scope of pause-extension) which formed the prosodic unit of stress, indicating that the speaker was sad and helpless about her misfortunes. The second stress fell on such prosodic units as 'they all said//your mother.' It revealed that others sympathized with the speaker over her mother's premature death. The speaker wanted to stress this to express her grief over her mother's death.

Figure 6.8 is a screenshot of the Praat workspace presenting the aforementioned three prosodic features in the tier.

6.1.6 Gestures: working definition, segmentation standard, and labelling method

As mentioned earlier, our annotation scheme divides gestures into non-verbal acts and task-doing acts to distinguish two gestures with different meanings and purposes. Non-verbal acts can be further divided into acts accompanying speech (such as facial expressions and hand movements) and acts having nothing to do with the speech (such as scratching one's head, crossing hands, and shaking feet, but sometimes these gestures also show a specific emotion).[8]

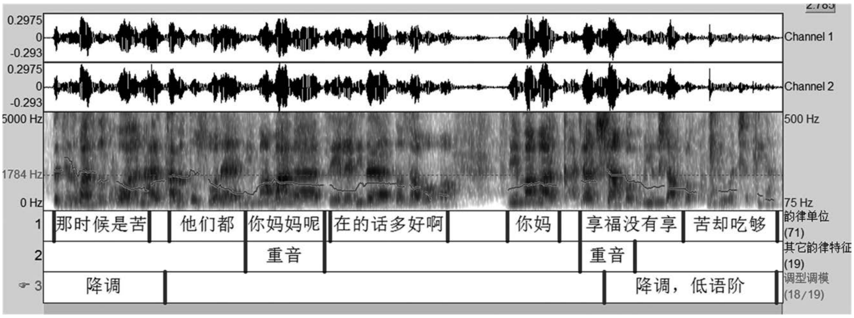


Figure 6.8 Sample of the prosodic analysis.

Faced with the multimodal corpus, if researchers transcribe and annotate all the gestures indiscriminately, they will be overwhelmed by the heavy workload. As a matter of fact, this is time-consuming, laborious, and unnecessary. Most researchers only pay attention to the gestural information that has distinctive communicative functions (Allwood, 2008: 217). Therefore, such a method was adopted in our annotation of gestures.

6.1.6.1 Non-verbal acts

WORKING DEFINITION

Annotation in this tier deals with gestural information with intimate interaction with other modalities, such as gestures closely related to emotions and prosody (e.g. rocking backwards while laughing). Those non-verbal acts fall into four groups with a physical aspect. See Table 6.8.

SEGMENTATION STANDARD

The beginning and end of the movement serve as the segmentation boundary of non-verbal acts, and the time without distinctive features is left blank.

ANNOTATION METHOD

Considering that not all gestures mentioned above occur in the situated discourse, we just annotated such information in one tier succinctly rather than categorizing them into different body parts. Only when overlap arose, then the overlapping gestures were annotated as a whole but described respectively.

In this annotation scheme, we only focused on the functions and properties of the gestural cues and annotated them as a whole, instead of looking

Table 6.8 Classification of non-verbal acts

Attribute name	*Tag type*	*Tag sub-type*	*Tag value*	*Remark*
gestures	non-verbal acts	head movement, 'head'	including nodding, shaking, or turning one's head	no grouping of the gestures by intensity
		facial expressions 'expression'	including smiling, crying, frowning, staring, mouth opening	describe facial expressions in different categories respectively
		hand movement 'hand'	including palm gestures (e.g. palms up), finger gestures (e.g. pointing), shoulder gestures (shrugging)	mark the direction, and mark whether it is a single-handed or two-handed gesture
		Other postures 'posture'	including the movement of the belly, back, leg, and the whole body	mark the direction, degree, and other information

into their forms, ranges of motion, or internal structures (e.g. the beginning, climax, and end of a movement). For example, leaning back was not annotated with its extent of movement but only its duration marked with the movement name. When the movement showed some distinctive features in its extent or range, it was annotated with the corresponding descriptors, such as 'leaning back slightly' and 'shaking vigorously.' Such a method has also been applied in some multimodal corpus annotation schemes, e.g. Allwood et al. (2004). However, it differs from the annotation scheme proposed by Li et al. (2008) because the latter resorted to statistical methods to conduct a quantitative study on the relationship between gestures and speech. Nevertheless, the selection of annotation schemes was based on the actual needs of the research.

Tags include body parts + descriptors for the aforementioned non-verbal acts, such as head + shaking head.

CASE ANALYSIS

A total of four types of non-verbal acts related to the head, hand, posture, and expressions emerged in this case. Compared with other speakers in this corpus, the speaker displayed more cues with respect to non-verbal acts, which might be relevant to the types of speech act. The speaker recalled a miserable experience in the past, so he was quite emotional. Undeniably, this might also depend on one's style.

6.1.6.2 Task-doing acts

WORKING DEFINITION

Part of the corpus data indicates that the speaker performed a specific task-doing act while speaking. As task-doing acts are closely related to the analysis of the speaker's discourse content and intentions, this tier mainly annotated what tasks the speakers were doing when performing the illocutionary force.

SEGMENTATION STANDARD

The beginning and end of the movement served as the segmentation boundary of the non-verbal acts, and the time without distinctive features was left blank.

ANNOTATION METHOD

Describe task-doing acts, such as fiddling with a laundry hanging rod and peeling pomegranates. See Table 6.9.

CASE ANALYSIS

In this example, the speaker and the recorder were sitting in the living room during the free talk. The speaker did not involve in any other activities at that time, so talking became the speaker's task-doing act. From the beginning to the end of illocutionary act-token, there only existed one task-doing act – talking.

6.1.7 Intention state: working definition, segmentation standard, and annotation method

WORKING DEFINITION

In Section 4.2.1, we discussed the rationale for setting {intentional state} in the form of sets in concept modelling. When performing a speech act,

Table 6.9 Annotation of task-doing acts

Attribute name	*Tag type*	*Tag value*	*Remark*
gestures	task-doing act	standing coring vegetables smoking pitching a tent	tags used to describe task-doing act

the intentional states a speaker might experience included intending a particular outcome, intending understanding from the hearer, intending understandings from the hearer entirely or partly result in his/her performance of perlocutionary acts, and attitudes to, beliefs about, hopes regarding, or requirements from others.

SEGMENTATION STANDARD

The corresponding linguistic forms, prosodic features (e.g. intonation), gestures (e.g. facial expressions, hand movements, and postures), and speech content all provided cues for intentional state identification and that of a performance unit mostly defined the segmentation boundary of an intentional state.

ANNOTATION METHOD

Intentional states could be divided into basic intentions with some common features of a particular illocutionary force type and other intentions with distinctive features of an illocutionary force type. This study mainly focused on the latter, so its information is presented in detail in the annotation.

Tags used to describe the speaker's intention state are shown in Table 6.10. Here, {…} indicates what the speaker is thinking about.

CASE ANALYSIS

In this case, an old person recalled some past misfortunes, which belonged to the illocutionary act of complaining. Based on the prosodic and gestural cues, it can be concluded that the speaker's intentional state was as follows:

intending that he/she was complaining about some sufferings in the past

intending the hearer to feel that he/she really suffered

intending the hearer to understand that such sufferings accounted for his/her complaint

attitude: believing that it was a miserable experience

Table 6.10 Annotation of Intentional States

Attribute name	*Tag type*	*Tag value*	*Remark*

Intentional states	basic intention
	attitude
	belief
	hopes
		believe{…}
		think{…}
		want {…}
		doubt {…}
			tags or descriptors used to describe the intentional states

6.1.8 Interdependency: working definition, segmentation standards, and annotation methods

WORKING DEFINITION

As discussed in Section 4.2.2.2, interdependency relation includes three aspects: (1) forward-and-backward interdependency, i.e. before-and-after relation of the utterance; (2) illocution-and-reality interdependency, i.e. the relation between speech act and what is happening at the here-and-now behaviour setting/beyond the here-and-now/both; and (3) doing-and-talking interdependency, i.e. the relation between doing and talking. Therefore, these three aspects of interdependency should be presented in the annotation.

SEGMENTATION STANDARD

Context of the illocutionary force, prosodic features of the speaker, and speech content all provided cues for identifying interdependency and that of a performance unit mostly defined the segmentation boundary of interdependency.

ANNOTATION METHOD

Descriptors used to describe interdependency are shown in Table 6.11, where {…} indicates the actual content.

Table 6.11 Interdependency annotation

Attribute name	*Tag type*	*Tag value*	*Remark*

interdependency	forward-and-backward interdependency	before {…}
		after {…}
			judgements are made according to the utterances before and after the illocutionary force

	illocution-and-reality interdependency	happened/happening/will happen	describe real-life situations
	doing-and-talking interdependency	overlapping conflictive parallel and relevant (primary/secondary) parallel and conflictive parallel and independent	tags in tag value can be combined to describe the actual situation

CASE ANALYSIS

In this case, an old person recalled some past misfortunes, which belonged to the illocutionary act of complaining. The speaker's task was to sit and talk, telling the hearer about his own miserable experience in the past. Before complaining, the speaker explained how her family lived by sorting herbs from weeds. After the complaint, the hearer asked if any of her family members were martyrs. The relation between the situated discourse and the relevant facts or things is that 'the speaker underwent something miserable in the past,' while the relation between doing and talking was overlapping. That is, talking was completing the task.

6.1.9 Metadata and annotation language

The metadata of the corpus of Chinese situated discourse and multimodal corpus of Chinese situated discourse have been presented on information cards for raw corpus data collection. These cards keep a record of the source of the corpus data, details of its collection, its social situation, and background activities, as well as the speaker and collector's information. We expatiate on the forms, media types, and storage forms of the multimodal corpus in Section 6.3.2.

6.2 Correspondences between concept modelling, data modelling, and modelling from a 'live, whole person' perspective

What is said, what is thought, what is felt, and what is embodied represent four perspectives for the simulative modelling of a 'live, whole person' in situated discourse, for which the simulative modelling in this study targets all sorts of live illocutionary forces produced by the 'live, whole person.' These four perspectives for the modelling of a 'live, whole person' also provide essential clues for the live illocutionary forces. Of course, there are more perspectives for the modelling of the 'live, whole person.' Still, investigations into the live illocutionary forces can be carried out from other angles. This section mainly discusses the correspondences between concept modelling, data modelling, and modelling from a 'live, whole person' perspective.

Discussions are presented in cohesion with Figure 6.9 in the following sections.

6.2.1 Correspondence between a 'live, whole person' modelling perspective and a data modelling perspective

Let us start our discussion with the four modelling perspectives in concept modelling, namely, what is said, what is thought, what is felt, and what is embodied. As mentioned above, this research uses sets as a tool and form

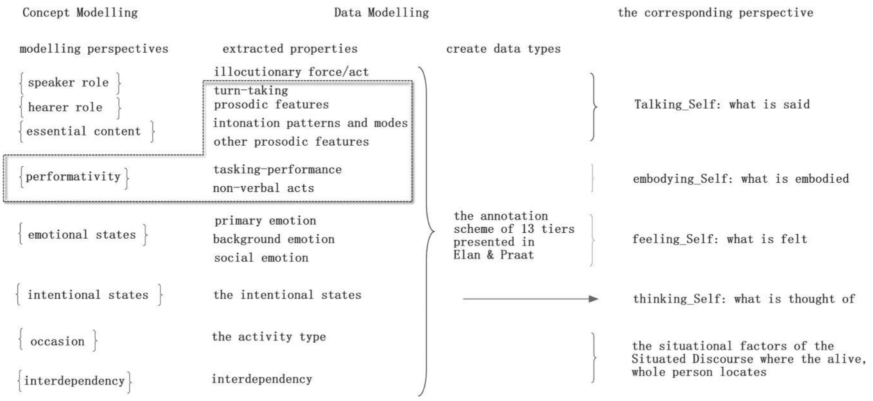


Figure 6.9 The corresponding perspectives of concept and data modelling.

for the concept modelling of research objects. In this sense, a 'live, whole person' can be modelled in these four perspectives in this way:

Talking_self {…}
Thinking_self {…}
Feeling_self {…}
Embodying_self {…}

These four sets represent four broad, overall perspectives, and their subsets can also encompass a few more aspects. In actual research, each perspective can also be further detailed and given different numbers of subsets, according to the research and modelling needs. For the four sets (perspectives) mentioned above, i.e. talking_self {…}, thinking_self {…}, feeling_self {…}, and embodying_self {…}, there are multiple subsets, that is, detailed aspects of a particular perspective. Here are some examples:

Talking_self {…} can be clarified as talking_self – illocutionary force {…}; talking_self – turn-taking {…}; talking_self – prosodic features {…}.

Among them, talking_self {…} – prosodic features {…} can be further subdivided into: talking_self – prosodic features – prosodic units {…}; talking_self – prosodic features – intonation patterns and modes {…}; talking_self – prosodic features –other prosodic features {…}.

Likewise, feeling_self {…} encompasses three subsets, namely: feeling_self – background emotion {…}; feeling_self – primary/universal emotion {…}; feeling_self – -social emotions {…}.

In this study, subsets concerning talking_self {…}, thinking_self {…}, feeling_self {…}, and embodying_self {…} can be seen in Figure 6.9. These subsets also constitute different attributes of the 'live illocutionary forces in the situated discourse' modelled in this study, with each matching to a different data type.

It should also be noted that since this study concentrates on illocutionary force, its context and interdependency (connotations of the two are discussed in Section 4.2.1) are crucial factors for investigation, and therefore are included in our research as one of the perspectives of concept modelling. Meanwhile, it also required us to establish their corresponding data types in data modelling. When analysing the context and interdependency of an illocutionary act's specific performance, it is also necessary to examine what the speaker said and what he/she embodied. Accordingly, a series of clues such as discourse content, prosodic features, and gestures are pivotal references when analysing a speech act's context and interdependency.

6.2.2 Correspondence between concept modelling and data modelling

Then we conducted concept modelling of the live illocutionary forces. Since the modelling is carried out from eight perspectives, the octet scheme was adopted (see Section 4.2.1). The octet scheme contains eight subsets that serve as eight perspectives for modelling the illocutionary forces in the situated discourse and eight attributes extracted from the illocutionary forces in the situated discourse ('attribute' belongs to the terminology of AOM, according to Gu (2009). For example, by subdividing the attribute {emotional state} (also a perspective), another three attributes can be extracted: Background emotion, primary/universal emotion, and social emotion. If presented in a set form, it would be:

> {Emotional state} = {{background emotion}, {primary emotion}, {social emotion}}

The number of attributes to be extracted depends on the research needs. In theory, infinite attributes can be extracted from a subject in concept modelling since the research needs to determine the research perspectives, yet the research perspectives may have a host of variations.

Subsequently, several data types are established for each attribute of each modelling perspective in the octet scheme (a conceptual model of the illocutionary force). In this research, the 'segmentation and annotation scheme for live illocutionary forces' was established to determine the annotation method and tagset for each attribute (representing different tiers in Elan and Praat). For example, the tagset of background emotion comprises a number of adjectives used to describe the speaker's feelings. However, limited by technology, not all attributes could be annotated. Therefore, the data types that could be presented in Elan and Praat were also limited. These steps were all part of the data modelling process.

It should be noted that each concept modelling perspective (or a subset from the perspective of set theory) can be supported by one or more data types determined by data modelling. In other words, there can be either one-to-one or one-to-many relationships between the attributes of concept modelling and the data types of data modelling. For example, in the above figure, the subset of concept modelling {performativity} can be further subdivided into another three attributes: illocution (represented as turn-taking in data modelling), prosody features (contain multiple data types, such as prosodic unit, intonation patterns and modes, and other prosodic features), and gestures (two data types: task-doing and non-verbal acts). These three attributes can also be subdivided into more attributes, as shown in the bracket. A specific data type supports each attribute, e.g. the data type of intonation patterns and modes includes pitch, speech rate, and the rise and fall of intonation.

In this sense, we can see the correspondence between the four perspectives of a 'live, whole person' in concept modelling, the eight perspectives of live illocutionary forces in concept modelling, and the data types of live illocutionary forces in data modelling. Such correspondence can be observed from the figure above.

6.3 Multimodal corpus of Chinese speech acts

Chapter 5 introduced some essential facts concerning the acquisition of original corpus data upon which the multimodal corpus of Chinese situated discourse was built. This section mainly introduces the development of the multimodal corpus of Chinese speech acts, which was created based on the multimodal corpus of Chinese situated discourse, according to the principle of 'stratified sampling.'

6.3.1 Principles for corpus establishment

Building a corpus for specific research purposes is a marriage of perfection and pragmatism (McEnery, Xiao, & Tono, 2006: 73). The size of a corpus should consider not only the representativeness of its samples but also its reality. Many studies are conducted by selecting a certain amount of corpus data from larger-scale corpus data, either through random sampling or frequency evaluation (Vaughan & Clancy, 2013: 59). In corpus-based pragmatic studies, in particular, researchers usually need to probe into the research object's contexts. However, due to limited technology, not all richly embedded information can be perceived and analysed. Therefore, large-scale corpora are not applicable here. It is difficult for the large-scale corpora to reflect all the original contextual information fully tucked away in the speech (Koester, 2010: 66–67), yet such information is particularly crucial for pragmatic research. Pragmatic research is not confined to observing the use frequency of specific terms (Vaughan & Clancy, 2013: 57).

In contrast, small-scale corpora are more in line with the actual needs of corpus-based pragmatic research (Vaughan & Clancy, 2013: 70). Mini-corpora for specialized purposes have the edge over larger corpora – they provide researchers with more information beyond the language, such as the speaker's information, contexts, and social situations, which is conducive to in-depth analysis (Cutting, 2001: 1208–1209). A corpus with a scale between 20,000 and 200,000 words can be called a small corpus (Aston, 1997); in corpus-based pragmatic research, the size of a small corpus is usually around 20,000–50,000 words. According to this standard and the amount of our transcription, the multimodal corpus of Chinese situated discourse belongs to a small corpus, and the multimodal corpus of Chinese speech acts (17,880 words) is almost equivalent to a small corpus, or we can call it a mini-corpus.

The research object of this book is the illocutionary forces in the situated discourse. Archer, Aijmer, and Wichmann (2012) pointed out that the difficulty of using corpora for speech act research lies in that they usually do not mark the types of speech acts in particular. As every speech act displays different linguistic manifestations (IFIDs) and shows no fixed lexical features or lexical hooks as Rühlemann (2010: 290) called them, researchers are hindered from directly searching and quantitatively analysing speech acts in the corpus (Vaughan & Clancy, 2013: 70). Here, Archer and Rühlemann were referring to text corpora. The same is even truer in multimodal corpora. It is because multimodal situated discourse involves prosody, gestures, facial expressions, etc. preventing the researchers from directly searching the illocutionary forces in the multimodal corpus of situated discourse. Therefore, a corpus of speech acts must be compiled based on the abovementioned corpus of situated discourse for this study.

Following the four principles below, we selected a certain number of illocutionary act-tokens in the multimodal corpus of Chinese situated discourse and segmented them using Corel VideoStudio Pro X4 (a multimedia processing software) to prepare for annotation. The four principles are:

1 Collection quality

Since what we collected was situated discourse, it was inevitable that we recorded some unexpected noise, which would interfere with the later analysis of prosodic features. To reduce the noise's impact, we tried to select corpus data as pure as possible. We also tried to choose that with high video quality that could provide resolvable detail and without shakes that significantly affect the picture quality during recording.

2 Social situations and activity types

Social situations and activity types are pivotal aspects of our investigation of illocutionary forces. The same discourse content may form different illocutionary acts/forces in different situations and activities and produce different pragmatic effects. Since this study collected speakers' speech in various contexts, we also tried to select a corpus of diverse activity types from multiple contexts.

3 Typicality of illocutionary force

To improve the coverage and representativeness of the corpus of speech acts, we tried to include as many illocutionary act-types as possible and select the corresponding episodes of typical illocutionary act-tokens in the same type.

4 Distribution of speakers

Influenced by multiple factors, the collected discourse given by different speakers was not equal in amount. Nevertheless, we still tried to cover all speakers into our corpus of speech acts, in the hope of expanding the source of illocutionary act-tokens. In the meantime, we also did our best to ensure the same type of illocutionary act-tokens from different speakers.

5 Stratified sampling

Stratified sampling is a method of sampling in which researchers partition a population into subpopulations (strata) based on members' shared characteristics or some other principle. Some random portion of the entire population is taken out from each stratum to represent the whole data set. Stratified sampling is widely used in sociological research because it ensures that the sample's structure is similar to that of the entire population, thereby improving the precision of the estimation. For example, in demographics, researchers first divide the sample units of the population into different categories based on different attributes possessed by the subjects, such as age, gender, and social class. Simple random sampling is then applied within each stratum (some call it stratified random sampling: McEnery, Xiao, & Tono, 2006: 20). Gu (2002b) adopted the same method and developed a corpus of Chinese situated discourse. Biber (1993) believed that stratified sampling provided a more representative sample data than simple random sampling. After checking the transcription line by line, we selected the illocutionary forces sample for this study and divided them into four classes (neutral, beneficial, harmful, and counterproductive).

Here are the specifics of how we got there:

1 According to the abovementioned classification standards for illocutionary act-tokens (how much the discourse content relates to the speaker/hearer's interests), they can be divided into four classes: neutral, beneficial, harmful, and counterproductive. A line-by-line analysis was

conducted to transcribe the multimodal corpus of situated discourse to determine their number and distribution;

2 No less than five typical illocutionary act-types were sampled from these four classes (strata);

3 In the multimodal corpus of situated discourse, we enumerated all the possible illocutionary act-tokens of 20 illocutionary act-types sampled from the 4 illocutionary force groups, and built a corpus of speech acts after stratified sampling;

4 According to the segmentation and annotation scheme, we segmented and added annotations to all illocutionary act-tokens in the corpus.

6.3.2 Composition and representation of the corpus of speech acts

With a high-quality collection, evenly distributed situated discourse settings, illocutionary act-types, and speakers, a mini-corpus named the 'multimodal corpus of Chinese speech acts' was built after the selection of samples. Subsequently, based on these principles and the previously formulated 'multimodal segmentation and annotation scheme for illocutionary force,' annotations were added to all the corpus data.

The basic information for the speakers in the corpus is shown in Table 6.12.

All the annotations were conducted in detail with Elan and Praat according to the requirements in the segmentation and annotation scheme of the illocutionary force in situated discourse. Each annotated performance unit contains typically four types of files:

1 Video file in MPG format;
2 Audio file in WAV format;
3 Elan-annotated file in EAF format;
4 Praat-annotated file in TEXTGRID format.

Unlike the text corpus, there is no standard or widely adopted form to represent multimodal corpus. For raw corpus data that had been transcribed but not annotated, we could synchronize it with its audio, video, and transcription for a particular duration (e.g. three to five minutes). Researchers could search in the transcription for what they needed. For cooked corpus data that had been annotated, it could be synchronized in transcription, audio, and video through annotation software. The advantage of doing this is that all the annotations could be seen quickly at a glance, and it also allowed us to search for multiple annotated files at a time. In the corpus, there were primitive texts, as well as software-annotated files.

In this study, the four files, as mentioned earlier to the same episode, were entitled with the same name, and those in the same types of speech acts were grouped into an independent folder. The files were named in this way:

Date/occasion-speaker's name-illocutionary force class – illocutionary act-token's name (sometimes we might switch the place of date/occasion

Table 6.12 Sample information of speakers

Speaker	*Native place*	*Dialect quarter*	*Class of illocutionary force*	*Number of illocutionary act-tokens*	*Proportion (%)*
Zhang	Lu'an, Anhui	Jiang – Huai Mandarin – Hongchao area –Luzhou zone	neutral beneficial harmful counterproductive	26	19.40
Shi	Lu'an, Anhui	Jiang – Huai Mandarin – Hongchao area – Luzhou zone	neutral beneficial harmful counterproductive	29	21.64
Guo	Lu'an, Anhui	Jiang–Huai Mandarin – Hongchao area – Luzhou zone	neutral beneficial harmful counterproductive	33	24.62
Zhang	Lu'an, Anhui	Jiang–Huai Mandarin – Hongchao area – Luzhou zone	harmful	1	0.75
Li Zhangren (Tang)	Chuzhou, Anhui	Jianghuai Mandarin – Hongchao area – Nanjing zone	neutral counterproductive	8	5.97
Bo	Tonglu, Zhejiang	Wu dialect – Taihu area – Linshao zone	neutral beneficial harmful	6	4.48
Shopkeeper Wu	Tonglu, Zhejiang	Wu dialect – Taihu area – Linshao zone	neutral beneficial harmful	6	4.48
Zhou	Tonglu, Zhejiang	Wu dialect – Taihu area – Linshao zone	neutral beneficial harmful counterproductive	9	6.72
Pan	Fengxian, Shanghai	Wu dialect-Taihu area-Su–Hu–Jia zone (Suzhou–Shanghai–Jiaxing)	neutral beneficial harmful	6	4.48
Wu	Ju County, Shandong	Jiao-Liao Mandarin – Qianglai area – Juzhao zone	neutral beneficial harmful counterproductive	10	7.46
Total				134	100

Table 6.13 Illocutionary force instances

Neutral		*Beneficial*		*Harmful*		*Counterproductive*	
explain	20	be satisfied	10	complain	14	be helpless	7
comment	16	praise	7	grumble	11	be disappointed	4
judge	8	urge	5	criticize	4	feel aggrieved	4
request	6	assume	3	worry	3	regret	2
promise	5	invite	1	fear	3	be surprised	1
subtotal	55	subtotal	26	subtotal	35	subtotal	18
						Total	134

with the speaker's name or omit the speaker's name, but the category of illocutionary force and the name of illocutionary act-tokens would never be omitted).

For example, 20120114 Interviewing a shopkeeper from Tonglu – Bo – neutral class – comment

20110905 Cleaning – Shi – harmful – complain

20110914-110922 Shi drank pesticide – Shi – express – feel aggrieved

According to the four illocutionary force classes, four folders consisting of the corresponding files were established. Then, a processed multimodal corpus of speech acts was built up.

Then we have 134 instances of speech acts (20 types) in four classes. See Table 6.13.

Researchers retrieve the annotated files corresponding to illocutionary act-tokens according to the speaker's name, the illocutionary act-token's name, etc. The storage media ranged in scope from hard disk drives and portable hard disk drives to CDs.

Two corpora mentioned above possessed the following attributes: They all belonged to synchronized corpora in terms of the duration of selected corpus data; there were annotated data (corpus of speech acts) and non-annotated data (corpus of situated discourse); they both belonged to monolingual (Chinese) corpora when looking into the language of the corpus data. The child corpus (corpus of speech acts) entirely differed from the parent corpus (corpus of situated discourse) in nature. For the parent corpus, its corpus data was situated discourse, but it was redeveloped for a specific research purpose for the child corpus.

6.4 Implementation and evaluation of annotation and tests of consistency, validity, and reliability

In this section, the implementation and evaluation of annotation and the tests of annotation consistency, reliability, and validity are introduced.

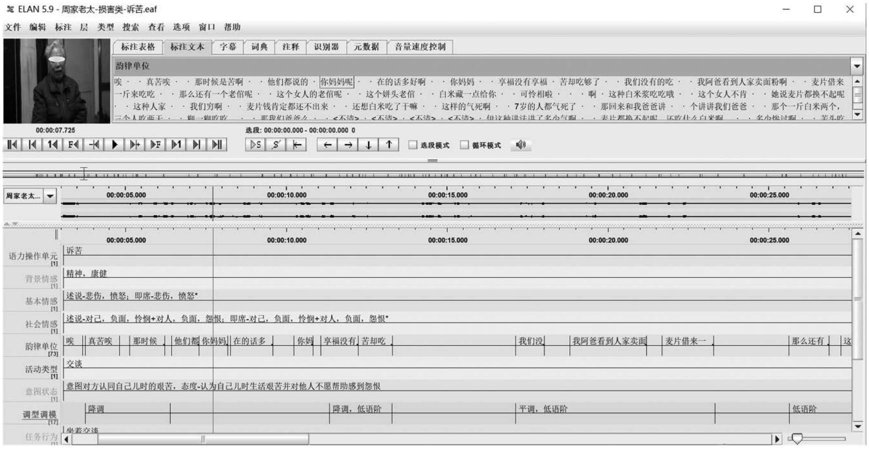
6.4.1 Annotation implementation and non-expert evaluation

After building a conceptual model and a data model for the live illocutionary forces in the situated discourse, we had to evaluate the rationality of such modelling and the implementation of annotation. In the implementation and evaluation stage of simulative modelling, implementation means adding annotations to data based on the modelling scheme. In this study, it means following the segmentation and annotation scheme of the illocutionary force in situated discourse. Evaluation means judging from the implementation of whether the adopted annotation scheme is rational. We invited linguistic experts and non-specialists to test and evaluate our pilot annotations for their reliability and validity to create an example for our annotation. Such pilot annotation was of great significance. It serves as a pivotal step in testing and adjusting the previous annotation scheme and provides the model with valuable insights on the upcoming large-scale annotation of the whole corpus. Once a typical example was set for the multimodal annotation of several illocutionary forces, there was a reference for annotating and analysing other corpus data, improving the accuracy and efficiency in analysing the speech acts in a multimodal manner.

All annotations in this study were conducted manually by the author. For research time and workload reasons, we did not adopt the method of offering annotations by multi-groups (multiple people in each group) and applying a statistical approach to find out the most reliable and valid way, which was frequently used in large-scale multimodal corpus research (e.g. Allwood et al., 2004). Instead, we graded the annotated samples through data validation by laypeople and resorted to a statistical approach to select the best-annotated sample universally acknowledged as an example for later annotation.

This study used two annotation tools: Elan (Version 4.8.1) and Praat (Version 5.4.12). The video streams in our multimodal corpus were segmented and annotated by the annotator with Elan. Annotations were mainly added to the activity type, performance unit, turn-taking, emotional state, gestures, and prosodic features of the instance. Also, simple statistical analysis was carried out leveraging the data statistics function of Elan. Praat was applied in segmenting and annotating the homogeneous audio data. As a scientific computer software package for the analysis of speech in phonetics, it can provide researchers with more detailed and specific parameters for phonetic analysis than Elan. Consequently, we mainly used Praat to conduct annotation and analysis of the prosodic features of speech acts. Once the annotation was completed, the annotation information was imported into the Elan file through the data transmission channel between Praat and Elan. Doing so enabled researchers to analyse the prosodic features of speech acts with the annotation information in Praat and explore the interaction between all multimodal cues by retrieving the data that was annotated in the different tiers of Elan files.

We invited several non-related linguistic experts who did not participate in the annotation to test and evaluate it. They were asked to anonymously assess, in scores, some samples after observing the multimodal corpus and the annotated files processed by Elan and Praat. If the experts' evaluations



唉 · · 真苦唉 · · 那时候是苦啊 · · 他们都说的 · 你妈妈呢 · · 在的话多好啊 · · 你妈妈 · · 享福没有享福 · 苦却吃够了 · · 我们没有的吃 · · 我阿爸看到人家卖面粉啊 · · 麦片借来
一斤来吃吃 · · 那么还有一个老倌呢 · · 这个女人的老倌呢 · · 这个姘头老倌 · · 白米藏一点给你 · · 可怜相啦 · · · 啊 · 这种白米浆吃吃哦 · · 这个女人不肯 · · 她说麦片都换不起呢
· · 这种人家 · · 我们穷啊 · · 麦片钱肯定都还不出来 · · 还想白米吃了干嘛 · · 这样的气死啊 · · 7岁的人都气死了 · · 那回来和我爸爸讲 · · 个讲讲我们爸爸 · · 那个一斤白米两个，
三个人吃两天 · · 糊一糊吃吃 · · 那我们爸爸么 · · <不清> · <不清> · <不清> · <不清> · <不清> · 但这种讲法讲了多少气啊 · · 麦片都换不起呢 还吃什么白米啊 · · 多少惨过啊 · · 苦头吃

选段: 00:00:00.000 - 00:00:00.000 0

语力操作单元: 诉苦
背景情感: 精神，康健
基本情感: 述说-悲伤，愤怒；即席-悲伤，愤怒*
社会情感: 述说-对己，负面，怜悯+对人，负面，怨恨；即席-对己，负面，怜悯+对人，负面，怨恨*
韵律单位: 唉 | 真苦唉 | 那时候 | 他们都 | 你妈妈 | 在的话多 | 你妈 | 享福没有 | 苦却吃 | 我们没 | 我阿爸看到人家卖面 | 麦片借来一 | 那么还有 | 这
活动类型: 交谈
意图状态: 意图对方认同自己儿时的艰苦，态度-认为自己儿时生活艰苦并对他人不愿帮助感到怨恨
调型调模: 降调 | 降调，低语阶 | 平调，低语阶 | 低语阶
任务行为: 坐着交谈

Figure 6.10 Illocutionary force instance.

were consistent with the author's, such as their judgements on the types of illocutionary act or emotions, high scores were given; if not, they were given low scores. After the evaluation, two tokens with the highest average scores were selected from each illocutionary force group to serve as the typical examples and reference standards for that group and other tokens in the same group. In the meantime, standard deviations among different evaluators in every token of the illocutionary act were also considered. Only those of low standard deviations were selected, i.e. those close to the mean (also called the set's expected value).

According to the segmentation and annotation scheme of the illocutionary force in situated discourse, we first selected 40 tokens of the illocutionary act from the corpus and added annotations to them according to the segmentation and annotation scheme of the illocutionary force in situated discourse. These 40 instances of eight speech act-types fell into four groups of illocutionary force in this mini-corpus. Figure 6.10 is an example of an annotated file by Elan and Praat, which is an instance of the speech act 'grumble.'

After the multimodal cues are annotated in the Elan file, researchers could record the annotated data in a Microsoft Excel table (illocutionary act-token annotation sample checklist) by using Elan's retrieval function. The table records information including data source, types of illocutionary act, evaluated value, standard deviation, and comments on the evaluation standard. Among them, we adopted the five-point Likert-type scale as our evaluation standard: (one: strongly disagree, two: disagree, three: neither agree nor disagree, four: agree, five: strongly agree).

Table 6.14 SPSS Calculation

Sample summary

		N	%
Instance	valid	40	100.0
	excluded[a]	0	.0
	Total	40	100.0

[a] Listwise deletion

Statistics of reliability

Cronbach's Alpha	*Number of items*
. 703	10

Hyperlinks between the serial numbers of the illocutionary act-tokens and their corresponding annotation files were established to open them quickly during the check.

Ten laypeople who were not involved in the annotation were invited to check and evaluate, in scores, the 40 pilot annotations above. Before the evaluation, the author first briefly introduced the study, data source, and evaluation tasks and methods to those evaluators. After making sure every evaluator understood their tasks by illustrating examples, they were shown relevant data and its annotation files and given scorecards to put down their scores for annotating the 40 tokens anonymously and independently.

After the check, all the scores given by ten evaluators were inputted into the checklist and were put in illocutionary force groups to calculate the average score and standard deviation for each group.

Meanwhile, those scores were also inputted into SPSS to calculate their reliability, and the Cronbach's alpha tuned out to be 0.703, which proves that those ten samples were quite reliable. See Table 6.14.

These forty tokens of the illocutionary act can be divided into eight types (each illocutionary act-type contained five illocutionary act-tokens) and four groups (each group included two illocutionary act-types). This study chose one standard sample from five tokens in every type, i.e. eight standard samples in total. In this way, we could ensure that those standard samples covering eight illocutionary act-types in the four mentioned groups of illocutionary force were reasonably representative.

We then computed the arithmetic mean and standard deviation of the 40 instances.

The arithmetic mean here reflected the overall evaluation of the annotation made by the ten evaluators. If the mean was higher than the median, it indicated that the evaluators mostly approved our pilot annotation. The standard deviation is the most critical and commonly used indicator in

Table 6.15 Annotation sample

Number	*Corpus data*	*Illocutionary act-type*	*Group of illocutionary force*
1	An interview with Pan from Fengxian (after lunch) – explain how to make corn dishes	explain	neutral
2	20110114 Shopkeer from Tonglu discusses house price – neutral – comment	comment	neutral
3	An elderly lady from Tonglu – beneficial – praise	praise	beneficial
4	Zhang talks in the dormitory – beneficial – be satisfied	be satisfied	beneficial
5	Wu from Shandong – harmful – grumble	grumble	harmful
6	20110905 Cleaning – harmful – complain & grumble; angry (2)	complain	harmful
7	Li (father-in-law) – counterproductive – be disappointed (not covered by the labour security)	be disappointed	counterproductive
8	20110914-110922 – Shi drank pesticide – express – feel aggrieved	feel aggrieved	counterproductive

measuring the amount of variation or dispersion of a set of values. The more spread out the data, the higher the standard deviation. Here, annotation samples with low standard deviations indicated that they had received relatively high evaluation from all raters and are quite reliable. Since no apparent extreme values were found in this evaluation, we set the arithmetic mean as our first choice for the reference standard. If the same average occurred, we would consult the standard deviation and select those with low standard deviation values (the standard deviations of all eight annotation chosen samples were less than one).

After all the computation and selection, the annotation information of the eight illocutionary act-tokens is presented in Table 6.15.

The eight annotation samples above respectively represent eight tokens from eight illocutionary act-types in four groups. They are the typical examples verified by laymen serving as a reference for annotating other tokens in later studies.

The arithmetic mean and standard deviation of the 40 tokens are 4 and 0.75, respectively, indicating that the ten raters' overall evaluations were satisfactory. The test also proved that our annotation samples were the true representation of the live illocutionary forces and that our multimodal segmentation and annotation scheme developed met the demands for the research of illocutionary force in situated discourse.

Additionally, it should be noted that the annotations of the 40 tokens mentioned above all served as pilot annotations for building a multimodal corpus of speech acts. In the later development of the corpus, adjustments were made to the annotation based on experts' comments and suggestions. Not all 40 tokens with pilot annotations were included in the corpus, and some of the included ones were renamed.

6.4.2 Consistency test

A consistency test is performed to check whether an annotator makes consistent and stable judgements on the data annotation in the first place and after some time. Hence, some studies also call it a stability test. Since all annotations in the mini-corpus were rendered by the author, and annotations within an interval were made at random times, it was necessary to perform consistency tests.

In total, 44 instances of 134 performance units of speech acts were chosen as the consistency test samples. The distribution of these 44 tokens in 20 illocutionary act-types and four illocutionary force groups was in proportion to that of the illocutionary act-types and the groups in the 134 tokens. The principle of rounding was applied in computing the percentage. In the end, a total of 44 illocutionary tokens covering all 20 types were selected. The 44 tokens were marked in detail for the first time following the annotation scheme mentioned above.

Table 6.16 shows its distribution.

We independently and separately annotated the 44 instances twice with a temporal gap of one month. In the second try, the annotator created new Elan and Praat documents for the same 44 illocutionary act-tokens and rendered annotations to them once again.

By comparing the annotation results of those tokens in the first and second try (each annotation file has 13 annotation tiers), their relations could be roughly divided into the following three kinds:

1 Almost consistent. There was no significant difference between the two. There were two possibilities: First, the judgements on the essential facts (including activity type, illocutionary act-type, emotional state, intentional state, and interdependency) of the data were the same, and second, the annotations of other tiers (including turn-taking, prosodic

Table 6.16 Illocutionary force instance

Neutral class		*Beneficial class*		*Harmful class*		*Counterproductive class*	
explain	6	be satisfied	3	complain	4	be helpless	2
comment	4	praise	2	grumble	4	be disappointed	1
judge	3	urge	2	criticize	2	feel aggrieved	1
request	2	assume	1	worry	1	regret	1
promise	2	invite	1	fear	1	surprise	1
subtotal	17	subtotal	9	subtotal	12	subtotal	6

features, and gestures) were consistent in both tries, but the use of the tag codes might be slightly different, though they meant the same thing in nature;

2 Subtly different. There was no radical difference between the two. Still, slight changes showed up in several tiers (including emotional state, intentional state, turn-taking, prosodic features, and gestures), such as adding or reducing some markers, yet the overall annotations remained the same;

3 Markedly different. Inconsistency appeared in the judgements of the types of illocutionary act, or the new markers distinct in kind were used in the annotation of other tiers.

After summarizing all the tiers in the 44 tokens mentioned above in both tries, we obtained the following results shown in Table 6.17.

A total of 84% of the annotations were proven to be consistent overall. Since the illocutionary act-type (the performance unit) was the critical factor in our judgement, it demanded a high degree of consistency. The annotation of the 44 tokens in both tries turned out to be 100% consistent. In general, this study proved that the same annotator's annotations two times showed good consistency in the test.[9]

6.4.3 Professional verification on annotation validity and reliability

After the consistency test and the lay-person validation of the pilot annotations, the researcher annotated all 134 illocutionary act-tokens in the multimodal corpus. Professionals were invited to evaluate and verify all annotated corpus data.

Expert verification was performed to see whether all annotated instances in the multimodal corpus were marked according to the earlier annotation scheme; more specifically, whether judgements on the properties of the illocutionary act-tokens were correct and whether their relevant features were accurately marked. Experts were required to propose suggestions on revision for inaccurately or wrongly annotated instances. The annotator would make corrections according to the expert opinions to ensure their validity and reliability, which prepared reliable data for further statistical analysis.

The word 'expert' has two meanings in this research:

Firstly, it refers to specialized teachers, researchers, or doctoral students who have been educated and trained in linguistics (especially in the fields

Table 6.17 Illocutionary force instance

Almost consistent	*Subtly different*	*Markedly different*
37 84%	7 16%	0 0%

of pragmatics or discourse analysis) or are engaged in learning, researching, or teaching in related fields. Linguistic experts familiar with speech act research, multimodal linguistic research, and corpus linguistics can quickly understand the substance, purpose, and annotation scheme of this research and bring up suggestions on annotation revision in a targeted manner.

Secondly, it requires the verifier to be a local born in one of the dialect regions or who has lived in one of those regions for a long time and speaks the dialect like a native speaker. Since most of the speakers in this study were illiterate people who had not been well-educated, they spoke Mandarin with strong dialectal accents (some speakers even spoke in dialects throughout the talk). To ensure objectivity and accuracy in our annotation (especially judgements on the emotional state and prosodic features), experts from those dialect regions or familiar with those dialects were invited to verify the annotation data and propose suggestions.

The researcher arranged the recruitment of experts. The instructions for the recruitment of annotation verifiers were compiled and published on online platforms or spread in sister colleges to recruit verifiers publicly. In total, six qualified candidates were recruited.[10] Their information is shown in Table 6.18.

Table 6.18 Information about the experts

Candidates	*Occupation*	*Work unit*	*Dialect region of the data to be verified*	*Region of the spoken dialect*
LM	Ph.D., professor, and doctoral adviser	Tongji University	Lu'an & Chuzhou, Anhui	Jianghuai Mandarin – Hongchao area
ZYY	Ph.D. student	Shanghai Jiao Tong University	Lu'an & Chuzhou, Anhui	Jianghuai Mandarin – Hongchao area
DZF	Ph.D. student and associate professor	Tongji University	Ju County, Shandong	Jiao – Liao Mandarin – Qianglai area
XXL	Ph.D. student	Beijing Foreign Studies University	Ju County, Shandong	Jiao – Liao Mandarin – Qianglai area
XMF	Ph.D., associate professor, and master's supervisor	East China Normal University	Tonglu, Zhejiang; Fengxian, Shanghai	Wu dialect– Taihu area
CXF	Ph.D. student and lecturer	Tongji University	Tonglu, Zhejiang; Fengxian, Shanghai	Wu dialect – Taihu area

The verification proceeded as follows:

Invited professionals opened the annotated files (files in EAF format marked by Elan) of the respective dialect group from region level to area level and then played the synchronized videos. In this way, they could first judge on the whole whether the types of the illocutionary act had been appropriately marked before detailed verification was performed into each tier. Emphasis was laid on the performance unit (to figure out whether the annotator had made a correct judgement on the types of illocutionary acts), background emotion, primary/universal emotion, social emotion, and intentional state. Subsequently, the experts would estimate, in scores, the annotated instances one by one. Still, the five-point Likert-type scale was applied as the estimation standard. Experts put down their comments in the remarks column for those instances with relatively low scores.

The verification was done in two ways: The researcher and the experts performed the verification face to face, or the experts accessed the corpus data sent by the researcher via the internet, performed the verification on their own, and made comments on it.

Among the 134 instances, 102 instances were evaluated as five points (76%), and the average value was 4.78.

First of all, SPSS was used to calculate Cronbach's alpha to verify the overall reliability of the 134 illocutionary force instances, which indicated that the reliability of the annotations of the 134 instances fell within a reliable range according to the professional's evaluation.[11] See Table 6.19.

Next, we used SPSS to perform a T-test between the two sets of values given by the two independent professionals in three dialectal groups to see whether the two evaluators held consistent views on the annotation.[12] The results indicated no obvious distinction among the evaluations made by different professionals on the same dialectal region, i.e. the professionals arrived at a strong agreement on the annotation.

Once again, we used SPSS to calculate the Pearson correlation coefficient between the two sets of values given by the two independent professionals in three dialectal groups. See Table 6.20.

Table 6.19 Sample summary

		N	*%*
Instance	Valid	134	100.0
	Exclude[a]	0	.0
	Total	134	100.0

[a] Listwise deletion.

Statistics of reliability

Cronbach's alpha	*Number of items*
.756	2

Table 6.20 Statistics of the correlation coefficient

Professional	*Number of instances*	*Dialect group*	*Pearson correlation coefficient*
Li Mei; Zhang Yuanyuan	97	Jianghuai Mandarin – Hongchao area (Lu'an & Chuzhou, Anhui)	0.653 (α = 0.01)
Cai Xiangfeng; Xu Mofan	27	Wu dialect – Taihu area (Fengxian, Shanghai; Tonglu, Zhejiang)	0.421 (α = 0.05)
Ding Zhaofen; Xu Xiuling	10	Jiao – Liao Mandarin – Qianglai area (Ju County, Shandong)	0.655 (α = 0.05)

The following conclusion can be made through all the tests above: (1) The 134 instances received high marks from the professionals; (2) the Pearson correlation coefficient reflected a moderate level of agreement between the two experts in each dialect group on the annotation of the respective performance units; and (3) the T-test performance indicated that these experts reached a consistent evaluation of all the performance unit annotations in each dialect group.

Overall, the segmentation and annotation in the multimodal corpus of speech acts in Chinese situated discourse were reliable and valid for further study.

6.4.4 Multimodal corpus of speech acts: publishing and sharing

Up to now, we had built up the multimodal corpus of Chinese speech acts. The process of building up such a corpus can be seen in Figure 6.11: data collection → the development of the multimodal corpus of situated discourse → deep-processing of the multimodal corpus of situated discourse → completion of the building of the multimodal corpus of speech acts.

After completion, the newly built corpus generally would be open to the public if the copyright issues were addressed and the speakers had given authorizations. To improve the reusability and application of the corpus, the researcher should share the metadata, annotation scheme, and processed data with other researchers in a proper form and manner. The multimodal

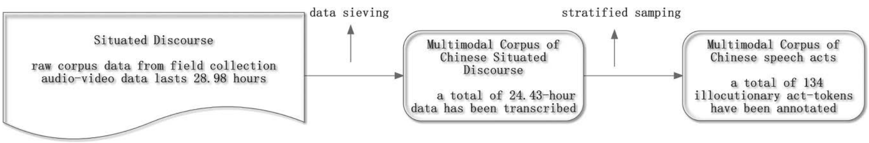

Situated Discourse
raw corpus data from field collection
audio-video data lasts 28.98 hours
data sieving
Multimodal Corpus of Chinese Situated Discourse
a total of 24.43-hour data has been transcribed
stratified samping
Multimodal Corpus of Chinese speech acts
a total of 134 illocutionary act-tokens have been annotated

Figure 6.11 Flow chart of building a multimodal corpus of speech acts.

corpus of Chinese speech acts built in this study will also be available to other researchers at a suitable time.

6.5 Summary

This chapter introduced issues related to the building of the multimodal corpus of Chinese speech acts, including the implementation and evaluation of all annotations and the tests of annotation consistency, reliability, and validity.

In our segmentation and annotation scheme, 13 tiers were established to annotate speech acts with the information from different perspectives conforming to our research target. These 13 tiers are performance units of illocutionary force, activity type, turn-taking, background emotion, primary emotion, social emotion, intonation group, prosodic pattern, other prosodic features, tasking-performance, non-verbal act, intentional state, and interdependency. In this chapter, the working definition, annotation standard, and annotation method of all 13 tiers have been elaborated. The annotator annotated all the speech acts in our multimodal corpus with Elan and Praat (for prosodic information). Lay validation was performed before standard samples are selected, which provides a reference for further annotation.

In this study, we invited linguistic experts to conduct tests of annotation consistency, reliability, and validity for all the data and annotation in this corpus. In general, the tests confirmed that the annotator offered consistent annotations in both tries. The result shows that all the data and annotation in this corpus were reliable and valid for further statistical analysis.

7 Types and tokens of illocutionary force in situated discourse

7.1 Ontological properties and case study of the neutral illocutionary force

The neutral illocutionary force refers to the illocutionary force in which the speaker or hearer's interests are not directly related to the discourse content that in return has no positive or negative impact on their interests. Exemplars include '*jieshi*' (解释, explaining), '*panduan*' (判断, judging), or '*yaoqiu*' (要求, requiring) others to complete a task irrelevant to the interests of the conversation interlocutors.

The five illocutionary act types of the neutral class in the multimodal corpus are '*jieshi*' (解释, explaining), '*pinglun*' (评论, commenting), '*panduan*' (判断, judging), '*yaoqiu*' (要求, requiring), and '*xunuo*' (许诺, promising). This section first analyses the ontological properties of different illocutionary act types based on the conceptual models in the Discovery Procedure of situated discourse and then conduct situational analysis. Only the analysis of the act of *jieshi* and *pinglun* is elaborated here.

7.1.1 The illocutionary act of jieshi (explaining in Chinese Pinyin)

(1) Conceptual analysis

<Speaker's role>

The speaker performing the illocutionary act of *jieshi* can be individual or collective. In the latter case, multiple speakers all together explain something and generate multiple illocutionary act tokens.

<Hearer's role>

The addressee perceiving the illocutionary act of *jieshi* can be individual or collective, e.g. explaining to one person or explaining a certain policy to all attendants at a convention. Under certain circumstances, the hearer is not necessarily the addressee. For instance, there are non-talking observers or hearers in the conversation.

<Performativity>

The illocutionary act of *jieshi* can be performed either explicitly or implicitly. The match between the act and force can be made explicit through

DOI: 10.4324/9781003251774-7

the performative utterance '我解释一下+解释内容'. Alternatively, it can be made implicitly when the speaker skips the performative verb and comes straight to the point. For instance, he/she uses utterances like '我说一下' or 'XX是这样的'.

<Essential content>

The relevant information of the object explained is the essential content of the illocutionary act of *jieshi*, due to the fact that if in the absence of such information being provided to the hearer, the act of *jieshi* will no longer exist. The essential content of *jieshi* can be delivered through utterances (i.e. what the speaker says) or through other channels such as actions (i.e. what the speaker does). In other words, the essential content of the act of *jieshi* can be either explicitly said (via explanatory speech) or implicitly indicated (via certain words or actions).

<Intentional states>

As mentioned earlier, when performing illocutionary acts, the speaker harbours various and complex intentions, including the basic intention and other intentions generated under the specific context, e.g. hope and attitude.

When performing the act of explaining, the speaker believes that the hearer is ignorant of a certain perspective, which will be understandable to him/her after the speaker's own interpretation and that the speaker can manage to deliver a clear interpretation. However, in real-life situations, some speakers do not genuinely have the conviction or attitude mentioned above and some may even go to another extreme to give intentionally misleading information. These cases are the mismatch between what is said, what is thought, what is felt, and what is embodied (i.e. the STFE-match assumption), which fail to meet the felicitous condition of the illocutionary force of *jieshi* but are endowed with other pragmatic intentions or meanings.

<Emotions>

The emotion of the act of *jieshi* can be neutral (when explaining an objective principle), positive (when explaining why you are happy, etc.), or negative (when explaining why you are upset, etc.). This study also analyses it via the three tiers of background emotion, primary emotion, and social emotion.

<Occasions>

The illocutionary act of *jieshi* can happen on various occasions. It is necessary to analyse in a combination of such conditions whether the speaker's illocutionary act of *jieshi* conforms to situational factors, the appropriateness of the speaker's identity, or the social goal of the activity under that context.

<Interdependency>

Interdependency includes three aspects, i.e. illocution-and-reality interdependency, doing-and-talking interdependency, and forward-and-backward interdependency.

Illocution-and-reality interdependency of the illocutionary act of *jieshi*: The objective content or object exists so that the speaker thinks it necessary to explain or the addressee is in need of having it explained. It can be something in the past, present, or future.

Doing-and-talking interdependency: The two most common cases of the illocutionary act of *jieshi* are: (1) The speaker performs *jieshi* as well as performing task-doing, such as explaining the mechanics of a car while repairing it or explaining the procedure when building a tent. In such a case, the performative act is intertwined with the task-doing, *jieshi* through performing the task-doing; (2) since the addressee is absent, the speaker has to resort to speech actions in a bid to explain or describe.

The forward-and-backward interdependency of the illocutionary act of *jieshi* should be analysed in view of the context. The previous discourse may be the hearer inquiring about something, so in this sense the discourse after is likely to be further inquiry on the same question.

To sum up, the analysis results of the ontological properties of the illocutionary act of *jieshi* can be presented in the formula of a set in Figure 7.1.

(2) Case study

There are 20 tokens of the illocutionary act of *jieshi* in the corpus, whose statistics are analysed and collated[1] in table format.

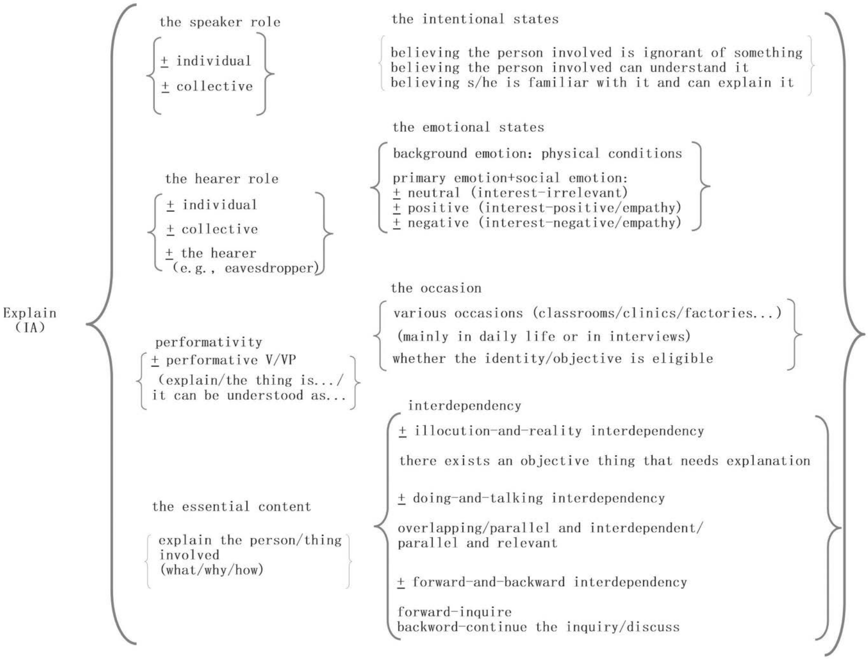

Explain
(IA)
the speaker role
± individual
± collective
the hearer role
± individual
± collective
± the hearer
(e.g., eavesdropper)
performativity
± performative V/VP
(explain/the thing is.../
it can be understood as...
the essential content
explain the person/thing
involved
(what/why/how)
the intentional states
believing the person involved is ignorant of something
believing the person involved can understand it
believing s/he is familiar with it and can explain it
the emotional states
background emotion: physical conditions
primary emotion+social emotion:
± neutral (interest-irrelevant)
± positive (interest-positive/empathy)
± negative (interest-negative/empathy)
the occasion
various occasions (classrooms/clinics/factories...)
(mainly in daily life or in interviews)
whether the identity/objective is eligible
interdependency
± illocution-and-reality interdependency
there exists an objective thing that needs explanation
± doing-and-talking interdependency
overlapping/parallel and interdependent/
parallel and relevant
± forward-and-backward interdependency
forward-inquire
backword-continue the inquiry/discuss

Figure 7.1 Analysis results of the ontological properties of the illocutionary act of jieshi.

Token 1: Interview with Mr Pan from Fengxian Town (in the afternoon) – jieshi – how to produce corn dumplings

zhen zhu mi xian zai //hua tou da lai //zuo //tang yuan //

zhen zhu mi zuo tang yuan ke yi //ai //

zhen zhu mi zha //zhen zhu mi yi zhi yi zhi shi fa //zha xia lai //qi bai le mi ji li //bao ke //ba ke bao te yi //bao te zi //man man de jin //jin shi //jin shi //shi er shi tian //dao yi ge hao tou //jin le jin le hao //nian tian zuo you //

nai nong ba yi liao qi lai //qi bai le //fen sui ji li fen //fen sui //fen sui nai nong qi //zuo tang yuan //zan e //bang nuo mi bi wu tan ban e //

shao dian //e //yao da mi de zhou e //xi zhou de fa shao le nian a //hao le // dao le yi dao nie //dao le yi dao nie //

bai e //zhen zhu mi o //you sha qu bie you hao ji hao ji zhong //nuo mi zuo le a chi zi shi nuo e //chi zi shi nuo e //chi le wu nuo e //jiu shi cu liang //jiu shi cu liang zuo e //wu nuo e //wo yuan lai tang yuan ji ben shang wu chi e nuo mi zuo e wu chi e //

huang e a you e ya //bai e nuo ya //huang e duo ai yi ban xing shi zhong huang yi ge de kuai xian zai //nao de ge er san nian a //nuo mi //di zheng de le me mei you le ma //hou lai zhen zhu mi //zhen zhu mi nong ai zhen zhu mi o //nuo mi wu tai ban e

Cornflour is now quite versatile //we can make dumplings with it//

Cornflour dumplings taste good //

Make corn into powder //put corns into the machine //start with peeling // and then soak them up //for 20 days or roughly one month //

After soaking all the corn shall be taken out and processed in the pulverizer //Such powder is just perfect for making dumplings //It even equals the taste of glutinous rice dumplings //Uh //Cook some rice porridge //and mix the porridge with the cornflour //

Cornflour is white //What's the difference? There are several kinds of flour. Dumplings with glutinous rice taste soft and glutinous. //If coarse grains, they would taste less glutinous //I don't eat ones made with glutinous rice. There are both white and yellow for cornflour //This region nowadays grows more yellow ones than white //There were two or three years then when farmland that we grow glutton rice was unavailable //Later we grow corn which can be made into cornflour that rivals glutinous flour // Cornflour is not bad//

The conceptual analysis of this illocutionary act token has been presented in the formula of a set. Here, we focus on the speaker's multimodal clues performing the illocutionary act of *jieshi* in its context, prioritizing the analysis of the speaker's emotion, prosody, and gestures (also applied to the analysis of all the following tokens).

This illocutionary act token's social situation is the factory office in a suburb of Shanghai where a factory employee in his 50s and the hearer (the author) had a casual conversation. The overall social situation is a factory where workers are on duty and products are being manufactured, with multiple relatively independent activity types. Due to the fact that there are only two interactants talking in the closed and quiet office, the activity type of this token is entirely independent. The discourse content is the speaker explaining to the hearer how to process self-grown corn into dumplings. Being rather proud of the food processing technology of the factory, he explained the process with enthusiasm.

1 Emotions

Background emotion: Since this tier hinges on the speaker's physical state and long-term mental condition, it exists in the emotion of each class (neutral, harmful, beneficial, and counterproductive). In this token, because all the speaker needs to do is talk with the author instead of doing heavy routine work in the workshop, his background emotion is **relaxed**.

Primary emotion: Although the utterance content does not involve the interests of either side in this token, what the speaker explains is his method of making corn dumplings. Therefore, he is delighted to share it with the hearer and to see his response. Judging from clues like his smile from beginning to end, the speaker's primary emotion then is **delighted**.

Social emotion: The speaker is confident and passionate when introducing the food processing method successfully implemented and explaining it to others. His smile throughout the entire conversation implies that he appears to be very enthusiastic about the explained subject and is quite confident and proud of it. In this token, social emotion can be divided into two layers: Pan is confident about his methods of making corn dumplings and introducing them to the hearer, and Pan is delighted to get the hearer's response of asking how to make corn dumplings. Therefore, the social emotion of the speaker in this illocutionary act token is: **self-directed positive, confident, proud, and other-directed positive and enthusiastic**.

2 Prosodic features

According to the segmentation and annotation scheme standards mentioned above, the external standard of the occurrence of pause is mainly adopted. Refer to the beginning part of this case analysis about how to segment this illocutionary act token into intonation groups.

One prominent prosodic feature in this token lies in many pauses in the utterance. The speaker has to pause to think over and then explain the complex steps and methods of making corn dumplings to the hearer. And meanwhile, the pause also marks the end of one step and the beginning of another, making it clearer and understandable for the hearer.

The speaker Pan has presented a relatively low overall pitch (about 100–150 Hz, and sometimes even 75 Hz) as well as mid-key.

As a male in his fifties, the speaker naturally has a relatively stable sound and lower pitch. However, compared with other parameters in his other discourses, his pitch here appears to be relatively lower. Generally, the illocutionary act of *jieshi* always adopts declarative sentences featuring low or medium pitch (Liu, 2009: 14). The figure above also verifies this. There are no other significant prosodic features in this discourse, indicating that the speaker is in a stable and positive mood. He is quite patient and willing to explain how to make dumplings with cornflour.

3 Gestures

Pan is sitting in a chair, facing the speaker, and has no other task-doing throughout the process.

He presents many non-verbal acts. In addition to the smile, his hand movements last for a long period and are particularly prominent.

His hand movements include demonstrating how to process corn (duration 22.98 seconds) and demonstrating with one hand (duration 7.17 seconds). These are typical non-verbal acts accompanying the illocutionary act of *jieshi*. To explain some parts or process more clearly, the speaker resorts to body gestures. This is a crucial feature that distinguishes this type from others.

Besides, due to his positive emotion, the speaker smiles for a long period throughout the conversation.

7.1.2 The illocutionary act of pinglun (commenting in Chinese Pinyin)

(1) Conceptual analysis

<Speaker's role>

Generally speaking, the speaker performing the illocutionary act of *pinglun* can be individual or collective, commenting on something and producing multiple illocutionary act tokens.

<Hearer's role>

The hearer can be individual or collective, listening to the speaker commenting on one person or one thing.

<Performativity>

The speaker comments to deliver his/her opinion. In this process, the speaker can either adopt the performative strategy of uttering '我来评论几句' or the non-performative strategy of *pinglun* without the performative verb.

<Essential content>

The essential content of the illocutionary act of *pinglun* is the speaker's opinion, perspective, or attitude towards a thing or a person. Without it, the illocutionary force of *pinglun* cannot be produced.

<Intentional states>

The intention of the speaker performing the illocutionary act of *pinglun* is that the speaker expresses his/her opinion on something that is believed to be valid and factual from his/her own perspective.

<Emotions>

The background emotion underlying the illocutionary act of *pinglun* is under the influence of the speaker's physical condition. The primary emotion and social emotion are more complicated, which can be positive, negative, or neutral, depending on the emotion triggering condition, i.e. whether the discourse content is related to the interlocutor's interests. If not relevant, then these tokens can be classified as neutral, and if relevant, the speaker could display either the primary emotion and social emotion of positive or negative due to influencing factors such as personal style (such as being highly sympathetic) or situational factors (like the atmosphere). Therefore, taking factors such as specific context and individual variation into consideration is imperative for analysing the emotions beneath each illocutionary act token.

<Occasions>

The illocutionary act of *pinglun* can occur in various occasions. It is necessary to examine whether the act of *pinglun* conforms to the situational factors, the appropriateness of the speaker's identity, and the social goal of the activity in that context.

<Interdependency>

Illocution-and-reality interdependency: the objective things, events, or phenomenon on which the speaker comments.

Doing-and-talking interdependency: This is basically consistent with the classification in Section 4.2.2.2. The speaker can solely make comments or do other things while commenting.

Forward-and-backward interdependency: In a specific context, the previous discourse may be the discussion on a real event and the discourse afterwards varies.

To sum up, the analysis results of the ontological properties of the illocutionary act of *pinglun* can be presented in the formula of a set, in Figure 7.2.

(2) Case study

There are 16 tokens of the illocutionary act of *pinglun* in the corpus.

Token (1): 11y08m26 talking about theft – neutral class – pinglun 2

xue xiao li mian si diao ren //wu suo wei //si diao jiu si diao le

pei shi me

pei //pei ta shi me qian a

shu xue xi na nian you ge xue sheng //en //si diao le //jia Zhang lai dao //dao ba lou ta jiang //en //xue xiao jiang //[ambiguous]//ni //zao jiu bu zai xue xiao xue xiao zhu le //ta ba dong xi quan bu reng diao le //xiu she li mian de dong xi quan bu reng diao //ta jiang bu zai xue xiao li mian zhu le //si diao bu jiu si diao le ma //ta jia ren lai nao

Someone died in the school.//The death doesn't matter.//

What the school owes to the dead student?

Compensation.//What compensation?

There was a student from the school of mathematics that year //Uh//He died//And his parents came to the eighth floor of the school. And the school gave them the response that their son had not been in the school for a long time[unclear]// Their son threw away everything in the dormitory and said he won't live here anymore.// In this sense, his death is totally irrelevant to the school// ever, his family came to find out what on earth happened.

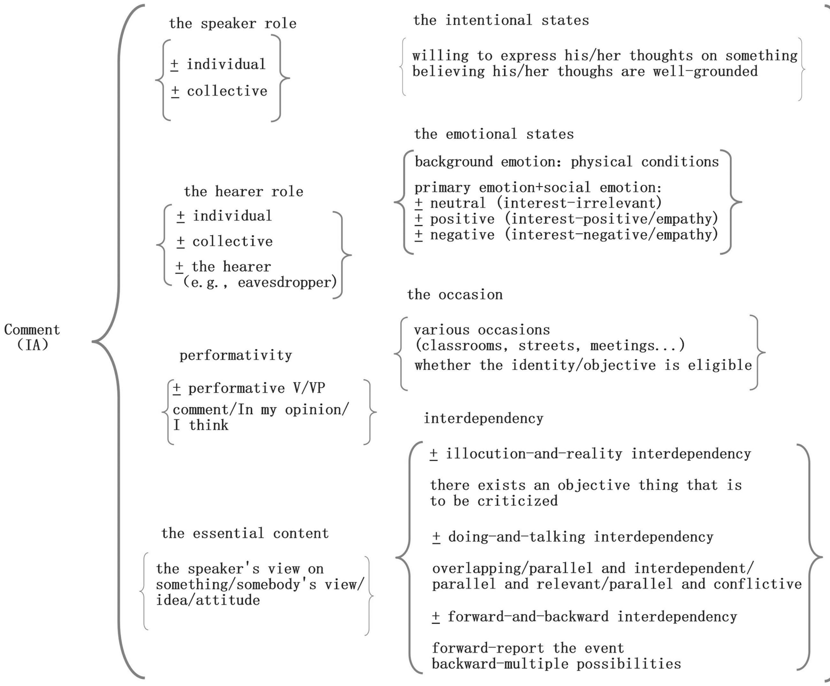

Comment (IA)
the speaker role
± individual
± collective
the hearer role
± individual
± collective
± the hearer (e.g., eavesdropper)
performativity
± performative V/VP
comment/In my opinion/ I think
the essential content
the speaker's view on something/somebody's view/ idea/attitude
the intentional states
willing to express his/her thoughts on something
believing his/her thoughs are well-grounded
the emotional states
background emotion: physical conditions
primary emotion+social emotion:
± neutral (interest-irrelevant)
± positive (interest-positive/empathy)
± negative (interest-negative/empathy)
the occasion
various occasions (classrooms, streets, meetings...)
whether the identity/objective is eligible
interdependency
± illocution-and-reality interdependency
there exists an objective thing that is to be criticized
± doing-and-talking interdependency
overlapping/parallel and interdependent/ parallel and relevant/parallel and conflictive
± forward-and-backward interdependency
forward-report the event
backward-multiple possibilities

Figure 7.2 Analysis results of the ontological properties of the illocutionary act of pinglun.

The social situation of this token is in a security room in the teaching building of one university where three people are chatting (the only activity type). In this conversation, the speaker is a security guard in his forties, the addressee is the author, and another hearer is the speaker's wife who occasionally interrupts the conversation. Due to his job responsibility, the speaker basically knows what is going on at the campus. And here he is commenting on how the department dealt with the accidental death of a student. Conventionally, accidents are something extremely upsetting and negative that stirs people's negative emotion. However, as a bystander without any interest involved in this event, the speaker has a neutral stance.

1 Emotions

Although the discourse content itself (the death of a student) is rather negative as mentioned above, the speaker shows neither sympathy nor negative emotions, for his own interest is not involved. This can be confirmed from the speaker's physical appearance and prosody.

The Elan annotation slot shows that the speaker's facial expression is relaxed with two slight smiles (duration 1.58 and 3.521 seconds, respectively) and that he appears to show a relaxed expression for a long time (roughly 8.19 seconds). Judging from this, his emotions are annotated as follows:

Primary emotion: **Situated – none** since there is no dominantly presented primary emotion.

Social emotion: **Neutral and indifferent**.

Background emotion: **Relaxed and calm**.

2 Prosodic features

Refer to the section above for the segmentation of the prosody unit of this token.

The prominent prosodic feature is the low pitch (below 150 Hz overall) with mostly mid-key. This reflects the speaker's emotion of calm and relaxation.

Due to the fact that the speaker and the author are acquaintances, the writer knows that the speaker does not always speak in a low pitch and mid-key and that he usually has perceptibly changing prosody in different moods.

A case in sharp contrast is the token entitled '20110912 the mid-autumn festival 2 – counterproductive class – weiqu.' Someone's bicycle that was parked near the security room was broken and he suspects that it was Guo who broke it on purpose and then argues with him. The speaker Guo is emotionally agitated and very aggrieved when saying:

wo jiang na zi xing che huai diao bu zheng chang ma (Isn't it normal for a bicycle to get broken?)

At this time, the speaker Guo was wronged so he is agitated with negative emotions (anger, etc.) and eventually bursts out, which are directly mirrored in the multimodal clues. It can be seen in Praat that he speaks at a high pitch (above 250Hz) at a fast tempo.

It can be verified from the token of the act of *pinglun* that the speaker's emotions and prosody are highly correlated.

3 Gestures

During the whole process of performing the illocutionary act, the speaker only talks to the author and has no other task-doing.

His gestures are mainly non-verbal acts such as smiling and showing a relaxed expression. This implies that despite the upsetting and negative discourse content (the accidental death of one student), the speaker displays no sympathy because his own interest has nothing to do with this event and because he and the hearer are acquaintances. His emotional state is relatively neutral without any dominantly presented negative emotion.

Token (2): conversation of the owner of Tonglu Inn about housing prices on January 14th – neutral class – pinglun

fang zi de shi qing //mei you jia qian ne //he //gen jia li de sheng huo yi yang de //you xie ren me //chi de hao chuan de hao //you xie ren chuan de ku yi dian //jiu shi zhe yang zi de

The house thing has no price (helpless laughter).// It's the same with the life at home.// Someone eat well and dress well while others have to tighten their belt to live. //This is life.

This token happens in the hall of a small inn where a woman in her 50s comments on housing prices to the hearer (the author), and during the conversation there is another interlocutor who is a friend of the speaker. The skyrocketing house prices bother everyone; however, the speaker does not get very bothered because she has already obtained an excellent relocation house. In this context, since her own interests are not involved with the housing prices, she has no negative attitude about it. This illocutionary act token is classified as neutral class.

1 Emotions

   Background emotion: Energetic, relaxed
   Social emotion: Other-directed neutral, calm, frank
   Primary emotion: No typical primary emotion

2 Prosodic features

   According to Praat's annotation, the overall pitch is relatively low, ranging from 150 to 260 Hz. High (or medium) pitch valuing about 350 Hz only appears when the topic begins. The pitch is descending and the speech rate is medium. This indicates that the speaker's attitude is generally positive. As this illocutionary act token is classified as neutral and the speaker presents few significant emotions, there is no special prosodic feature.

3 Gestures

   Sitting on the sofa and facing the writer, the speaker talks with her hands around her legs. Her non-verbal acts include smiling and staring to one side. The clues in Elan show the speaker feels rather relaxed and calm.

7.2 Ontological properties and case study of the beneficial illocutionary force

The beneficial illocutionary force refers to the illocutionary force that the discourse content is interest-positive to the speaker/hearer or will have a positive impact. Exemplars in the corpus include '*zanyang*' (赞扬, praising), '*manyi*' (满意, feeling satisfied) someone or something, '*yaoqing*' (邀请, inviting), or '*cuiqing*' (催请, urging) someone to do something in his/her interests.

It should be noted that there are two cases of the tokens in the beneficial class:

1 The corresponding illocutionary act type is positive and positively related to the interests of the interlocutors, such as the act of '*zanyang*' (赞扬, praising), '*zhuhe*' (祝贺, congratulating), and '*manyi*' (满意, being

satisfied). According to the conceptual analysis, ontological properties of these illocutionary acts such as essential content and intentions should be positive; otherwise, the pragmatic meaning of these performative verbs will be betrayed, thereby breeding unfit or unintended pragmatic meaning. Wierzbicka (1987: 18) believed that the meaning of speech acts themselves include different assumptions, emotions, thoughts, and intentions;

2 The corresponding illocutionary act type is neutral, such as '*jiashe*' (假设, hypothesizing). However, this token is still in the beneficial class since the discourse content is positively related to the interests of the interactants in the conversation. In the same vein, if the discourse content undermines the interests of the interlocutors, then these tokens should be classified as harmful.

The five types of the beneficial illocutionary force in the corpus are '*manyi*' (满意, feeling satisfied), '*zanyang*' (赞扬, praising), '*jiashe*' (假设, hypothesizing), '*cuiqing*' (催请, urging), and '*yaoqing*' (邀请, inviting). This section analyses the ontological properties of different illocutionary act types based on the conceptual models in the Discovery Procedure of situated discourse and then studies them under different contexts. Only the analysis of the acts of satisfaction and praise are presented here.

7.2.1 The illocutionary act of manyi (being satisfied in Chinese Pinyin)

(1) Concept analysis

<Speaker's role>

The speaker can be individual or collective. For example, one speaker is satisfied with a certain thing or state, or multiple speakers express their satisfaction together.

<Hearer's role>

The hearer can be individual or collective.

<Performativity>

When the speaker performing the act of *manyi* to the hearer, he/she can explicitly use the syntactic formula like '我对+具体内容+感到满意' or express it implicitly by talking about the content that satisfies him/her.

<Essential content>

The specific content or cause that meets the speaker's expectation is the essential content of the illocutionary act of *manyi*.

<Intentional states>

The speaker indeed thinks that a certain thing, experience, or state is agreeable to his/her desires and wishes.

<Emotions>

The background emotion hinges on the speaker's physical condition, while the primary emotion and social emotion are usually positive. This is because the discourse content is usually about the thing, experience, or state that conforms to the wishes or intentions of the speaker, therefore easily inducing positive emotions.

<Occasions>

The illocutionary act of *manyi* can occur on a variety of occasions. It is necessary to analyse in combination with such conditions whether the speaker's act of *manyi* conforms to situational factors, the appropriateness of the speaker's identity, and public order and moral behaviours.

<Interdependency>

Illocution-and-reality interdependency: the thing, experience, or state in line with the speaker's wishes or intentions and is logically 'satisfactory' does occur or exist. These objective facts can be a certain satisfactory thing or state in the past, present, or future (that is, its development or prospect is satisfying).

The doing-and-talking interdependency is basically consistent with the classification in Section 4.2.2.2; that is, one can express satisfaction solely when speaking, or express satisfaction while performing other task-doings.

The forward-and-backward interdependency should be analysed in its context. The previous discourse may be the statement of an objective state or event, and the discourse following may be further discussion on it.

To sum up, the analysis results of the ontological properties of the illocutionary act of *manyi* can be presented in the formula of a set in Figure 7.3.

(2) Case study

There are ten tokens of the illocutionary act of *manyi* in the corpus.

Token 1: Conversation with Ms Zhang in the dorm – beneficial class – manyi

xian zai ke yi le //jiu shi yi qian //na xian zai wo men ji hu tian tian dou you hun cai //ai dui le //xian zai sheng huo tiao jian you hao le yi hou //wo ye bu kou le //jiu yi ge xi fu qu shang lai le //zen me jiang a //fan zheng ye bu bei Zhang

ai //dou kuai you sun zi le //kuai yao you sun zi le //ai dui ya

en dui //nü er ne //fan zheng ta zi ji //yi hou zi ji pei jia ye gou le //fan zheng ta zi ji ye mang ge //qi ba wan //he he

Life now is fine. // Old days are gone. // Now we have meat dishes almost every day. // Oh. // Now the living conditions are better. // I don't have to pinch pennies and tighten my belt anymore. // And I have a daughter-in-law. // Another way to describe my current life is that // we don't take loan from the bank.//

I will have a grandson soon.//

Oh // about my daughter. // Anyway she // had already saved up 70,000 or 80,000. // That is enough to prepare for her wedding in the future. // Haha.

In this token, the speaker Zhang is a woman in her forties working as a school cleaner. After a day of work, she has this conversation with the writer in the evening at the school dormitory. Sitting on a stool and picking the vegetables meanwhile, the speaker talks with the author about her job and living conditions. On the right side is her nephew who does not

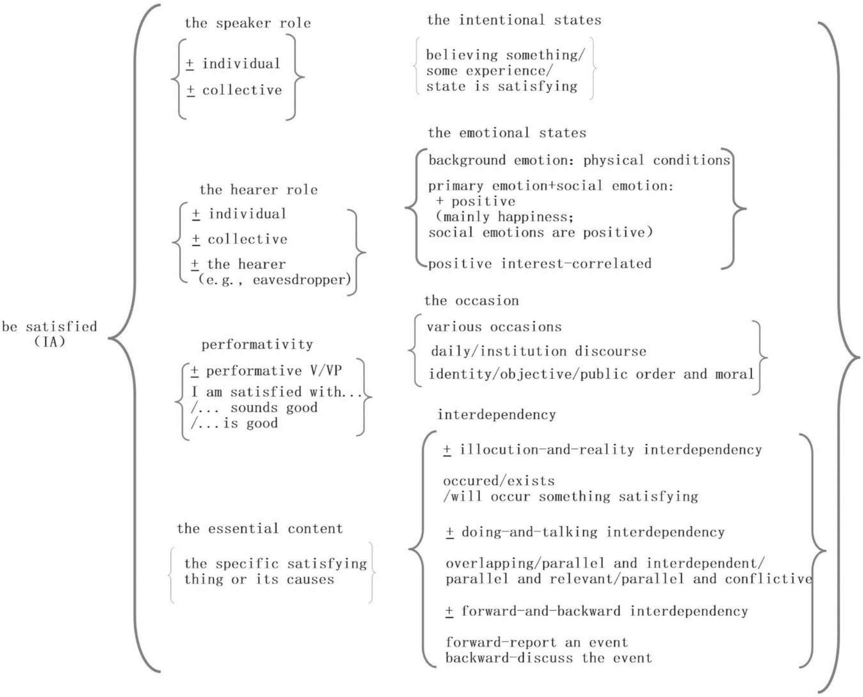


Figure 7.3 Analysis results of the ontological properties of the illocutionary act of manyi.

participate in the conversation and only records the conversation. So in this token, there is one speaker, i.e. the cleaner and two hearers, i.e. the author who participates in the conversation as well as the recorder who does not (the speaker's nephew).

Without the agent verb of *manyi*, other modal clues such as discourse content, prosodic features, and gestures are enough to confirm that the speaker is satisfied with her current state. This token's essential content is that the cleaner's work has improved her life quality and financial state. Her family has paid off all their debts and can now afford dishes that were too expensive for them. And her son had his own family, and her daughter who has prepared her own dowry is about to marry. Such content is worthy of the speaker's satisfaction, accompanied by her positive attitude, beliefs, and corresponding emotions.

1 Emotions

Background emotion: This is related to the speaker's physical condition and has a direct impact on the rhythm. Despite the fact that the speaker is still in good health in her forties, she cannot hide her exhaustion during the conversation after a day of high-intensity labour (outdoor sweeping). As far as the author knows, the speaker has presented fatigue expressions during this conversation, which is unusual when he talked with her another time. So here the background emotion is annotated as **fatigue**.

Primary emotion: **Delighted.** The speaker is delighted that her living conditions and economic situation are much improved, that her son has married and is going to have a baby, and that her daughter is now financially independent.

Social emotion: As the discourse content is about evaluating her own living conditions, the emotion is **self-directed positive** as well as **proud**. This is evinced by several clues when the speaker Zhang performs the act of *manyi*, such as 'xian zai ke yi le (my life now is pretty good)' and 'xian zai wo men ji hu tian tian dou you hun cai (now we can have meat dishes almost every day).'

2 Prosodic features

Statistics show that the main prosodic feature is the medium pitch, accompanied by several high pitches with fall-rise and mid-key.

Although the speaker is exhausted straight after a day of outdoor high-intensity work, she still presents several middle and high intonations, indicating that she is in high spirits. Furthermore, in view of her discourse content (including her attitudes on her current work and life), she is indeed very satisfied and therefore presents obvious joyfulness. Comparing with her intonation patterns and modes when performing the neutral illocutionary act (the pitch contour ranges from 180 to 190 Hz), the overall intonation here is at least 200 Hz with the highest even exceeding 300 Hz. During the entire process, the speaker has five sentential stresses and laughs out loud once.

To further look at the location of these stresses, they generally appear when the speaker is emphasizing the current good conditions.

The first accented word '*xianzai*' (now) and the second '*tiantian*' (everyday) are both accompanied by high intonation and falling tone, showing the speaker's firm satisfaction. The locations of these two stresses are at the beginning of a sentence, which is consistent with Liu's (2011: 61) research on the probability of the occurrence of stresses on the emotion of happiness.

3 Gestures

Task-doing: The speaker is picking vegetables while beginning to talk, and later when she becomes obviously enthusiastic, she stops picking vegetables to talk attentively with the author.

Non-verbal acts: There are three obvious smiles, relating to when the speaker talks about having a grandson, when the topic turns to her daughter, and when she talks about her daughter's deposit reaching about 80,000 Chinese RMB or so. The smiles then reveal her happiness and satisfaction.

7.2.2 The illocutionary act of zanyang (praising in Chinese Pinyin)

(1) Conceptual analysis

<Speaker's role>

The speaker can be individual or collective (such as collectively *zanyang* the motherland or leader).

<Hearer's role>

The hearer can be individual or collective.

<Performativity>

When performing the act of *zanyang*, the speaker can either adopt a performative or non-performative strategy.

<Essential content>

The essential content is about the merits or positive aspect of someone or something, and the speaker generally expresses this content when performing the act of *zanyang*.

<Intentional states>

The speaker thinks that the object praised enjoys an advantage or positive aspect that deserves complimenting. Indeed, such a belief or attitude is extremely subjective, that is, the speaker personally believes that a certain aspect is commendable, while the hearer/others may not literally share the same opinion. Sometimes the speaker may hold beliefs and attitudes that are not objective, but as long as he/she presents emotions in accordance with such beliefs and attitudes and the corresponding discourse content, the prosodic features and gestures fit, and at least one of the necessary felicity conditions of the performance of the illocutionary act is met.

<Emotions>

The background emotion depends on the speaker's physical condition, while the primary emotion and social emotion are usually positive. This is because the discourse content is usually the speaker praising advantages or positive aspects, therefore easily inducing positive emotions.

<Occasions>

The illocutionary act of *zanyang* can occur on a variety of occasions. It is necessary to analyse whether the speaker's illocutionary act conforms to situational factors, the appropriateness of the speaker's identity, and public order and moral behaviours.

<Interdependency>

Illocution-and-reality interdependency: The merit or positive aspect praised by the speaker occurs or exists and is found by the speaker. These objective facts can point to the merit or praiseworthy aspect of one thing or person in the past, or point to the present or future.

The doing-and-talking interdependency is basically consistent with the classification in Section 4.2.2.2; one can solely express the force of *zanyang* or express it while performing other task-doings.

The forward-and-backward interdependency should be analysed in its context. The previous discourse may be the objective statement of the state or behaviour, and the discourse following may be further discussion on it.

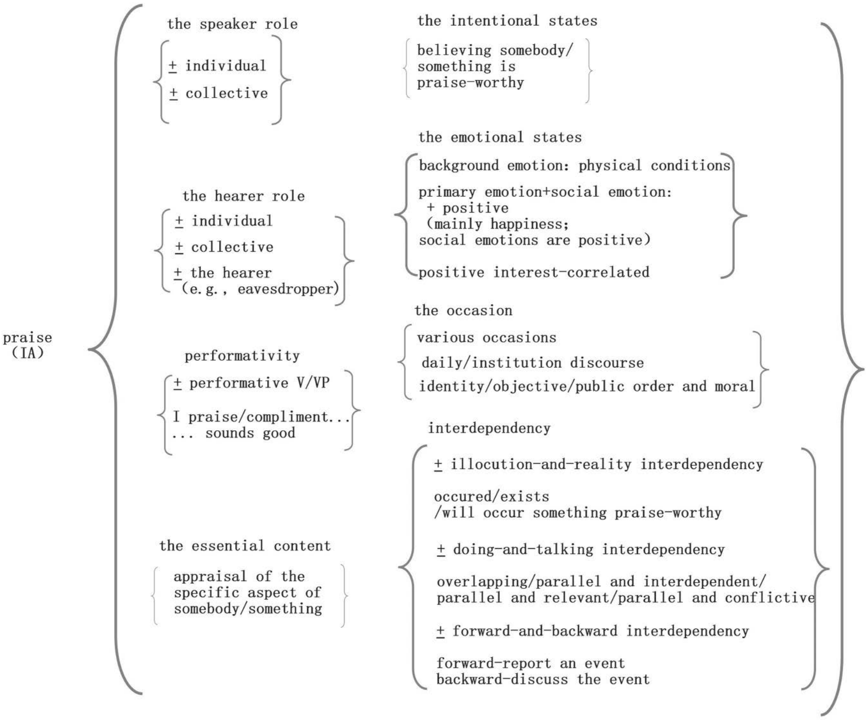


Figure 7.4 Analysis results of the ontological properties of the illocutionary act of zanyang.

To sum up, the analysis results of the ontological properties of the illocutionary act of *zanyang* are presented in the formula of a set, as shown in Figure 7.4.

(2) Case study

There are seven tokens of the illocutionary act of *zanyang* in the corpus.

Token (1): Mrs Bo in Tonglu – beneficial class – zanyang

xian zai wo men tong lu a //bi hang zhou na ge //fang zi hai hao //wo men yi jian //wo men na tian ji hao a //wo shi //jiu hao //jiu hao dao hang zhou qu wan a //hang zhou na ge fang zi //yao wo men tong lu yao bu lai de //ta dou shi lao fang zi a //wo men zhe li dou shi xin fang zi a

na ge fang zi zhe yang de fang zi //wo men zhe li yi ge di fang dou zhao bu dao le //na ge fang zi dou shi di di de //xiao xiao de //jiu jiu de //wo men zhe li mei you de

wo na tian qu wo kan kan wo zhe bian hai shi wo tong lu hao le o //shi wo men tong lu hao //wo men tong lu fang zi dou shi gao fang zi la

dou shi xin fang zi //fa zhan hai shi wo men tong lu kuai yi dian //hang zhou hai shi lao yang zi //he he he he

Houses in Tonglu are nowadays better than those in Hangzhou. I went to Hangzhou on the 9th together with friends and saw the houses there. Gosh. They are so old and broken. So different from the houses here. The experience that time makes me deeply feel that Tonglu is so good and develops rapidly. Houses here are high and new.

The social situation of this token is Bo in her fifties talking with the author in her acquaintance's inn one morning. Sitting on the stool, the speaker faces the hearer and talks about the development in Tonglu. On her left side are two women who occasionally listen to Mrs Bo and occasionally have their own chat. And on her right side are one grandfather, his daughter, and grandchild. Two adults stand facing the exit at first and then begin to play with the grandchild. Therefore, there are three activity types in this conversation, two of which overlap in a certain period of time.

There is one speaker, Bo, and several hearers in this illocutionary act token. With no presence of an explicit performative verb, this token is still judged as the act of *zanyang* according to the discourse content and prosodic features and gestures. More specifically, the discourse is about the sound development in Tonglu which has exceeded that in Hangzhou from the perspective of the speaker; she is intensely proud of the surging quantity and quality of property in Tonglu.

1 Emotions

The speaker compliments (*zanyang*) the sound development in Tonglu, particularly in the field of property. Judging from her discourse content, intonation, and expressions, her emotions are as follows:

Primary emotion: annotated as **delighted**. The speaker thinks the new apartments far exceed those in Hangzhou and is genuinely delighted about this.

Social emotion: Since it is the attitude and sentiment towards the city rather than herself, the emotion is self-directed and positive with the emotional factors of gratification and pride. Therefore, social emotion is labelled as **self-confident, positive, and proud**.

Background emotion: annotated as **energetic**. It is related to the physical health of the speaker. Judging from her movements, tone, and expressions, she is still in good health in her fifties.

2 Prosodic features

The speaker presents high intonation and ascending key during the illocutionary act token lasting for more than 50 seconds. To take a further look at where the high intonation appears, the author finds the following:

The first intonation group with high intonation is at the beginning of the topic, which is a method employed by the speaker when switching topics. Plus, the rather noisy background prompts the speaker to raise her voice to speak. This is not directly related to the attitudes and emotions when performing the act of *zanyang*. The subsequent high pitch appears in the intonation group of explaining that houses in Hangzhou are old, and, meanwhile, there are several ascending keys in the similar discourse content when the speaker emphasises how good houses in Tonglu are in comparison with those in Hangzhou. This shows that the speaker is in high spirits when performing the act and that she firmly believes that houses in her hometown far exceed those in Hangzhou. The speaker uses the descending key to express such firm belief, which matches with firm, positive belief and attitude that a felicitous performance of the act of *zanyang* requires.

In contrast, the speaker's prosodic features when performing the neutral illocutionary act are basically a low intonation (the pitch of the entire utterance is in the range of 150–260 Hz); descending key; normal tempo.

There are nine stresses in total when the speaker performs the act of *zanyang*. Study finds that the words bearing the speaker's subjectivity are more likely to be accented and at a higher intensity, including auxiliary verbs, adjectives, adverbs, quantifiers, or conjunctions. This implies the expression of the subjective information of the speaker (Zhang, 2014: 123). In this token, the accented words 'hang zhou' (Hangzhou), 'di di de' (low), 'xiao xiao de' (small), and '(jiu jiu de)' indicate that the speaker wants to set the houses in Hangzhou and in Tonglu in comparison and to emphasize how good the latter are; moreover, more stresses appear when describing Tonglu's houses 'hao' (good), 'gao fang zi' (tall house), and 'xian zai' (now) highlight the speaker's compliment and her underlying pride. Most of these accented words are the speaker's subjective judgements on the houses in Hangzhou and Tonglu, by which the speaker expresses her subjective attitude.

3 Gestures

Task-doing: The speaker sits on the stool and talks with the present author with one leg crossed over the other.

Non-verbal acts: The rich non-verbal acts presented by the speaker during the entire process can be analysed in Elan.

In sharp contrast, the speaker presents obvious gestures like pointing and shaking her head when speaking of the houses in Hangzhou being old and small in number.

When praising the houses in Tonglu, the speaker presents gestures like pointing, smiling, nodding, and laughing.

The rich non-verbal acts speak for the speaker's *zanyang* development in Tonglu and her pride. In contrast, such gestures as pointing and shaking her head show her complaint when talking about the worse condition of houses in Hangzhou. However, it should be noted that such temporal emotion embedded in the holistic illocutionary act token of *zanyang* (the social emotion is pride) is the concomitant emotion produced when the speaker believes that most of the houses in Hangzhou are older and the development of Tonglu in recent years outshines that of others. Gu (2013a) refers to such emotion as belief-mediated emotion, whose reason and purpose of occurrence lie in the fact that the speaker would like to express or highlight her praise of Tonglu and her corresponding positive emotions.

When the speaker performs the act of *pinglun* classified in the neutral class, there is no significant non-verbal act but only a smile and a sideways look that reflects the speaker's relaxation and calmness.

Therefore, in this token of *zanyang*, the speaker's emotions, intentions, etc. all match the illocutionary act type, accompanied by rich indicators like rhythm, movements, and expressions that prove the felicitous performance of the act of *zanyang*.

It can be seen from the analysis above that the positive emotion of the speaker when performing the act of *zanyang* is clearly mirrored in the prosodic features and gestures. Such phenomenon is not individual but general.

Token (2): Ms Shi – 110818 – erecting a tent – praise

huang lao shi hai guai hui mai de //mai de guai bian yi de //he he //hai guai piao liang

Teacher Huang is good at picking stuff to buy. The tent is pretty cheap (laughter). It is beautiful too.

This illocutionary act token happens in a campus office where the speaker Shi and two hearers (both conversation interlocutors) together erect the tent. Furthermore, there is another person who records the corpus and does not participate in the conversation. After managing to erect the tent, Ms Shi cannot help complimenting one hearer about this tent's purchase.

1 Emotions

Since the discourse content is interest-positive to the hearer, this token is classified as the beneficial class and the corresponding emotion is also positive.

Background emotion: **Energetic**
Social emotion: **Delighted**
Primary emotion: **Other-directed positive and respectful**

2 Prosodic features

The Praat annotation shows that the speaker's overall pitch is around 300 Hz, with the highest segment reaching 390 Hz. Compared with this, the speaker's pitch when performing the neutral illocutionary act ranges from 200 to 250 Hz, which is the typical low tonal mode.

The first intonational group 'huang lao shi hai guai hui mai de' (Teacher Huang is really good at making purchases) is an exclamatory sentence. In combination with the speaker's emotions, the exclamation tone pitch can be high or low, and the range can be wide or narrow. The emotion expression behind the tone of exclamation largely determines the pitch. In this sense, speakers in low spirits present lower intonation, while those in high spirits often present higher

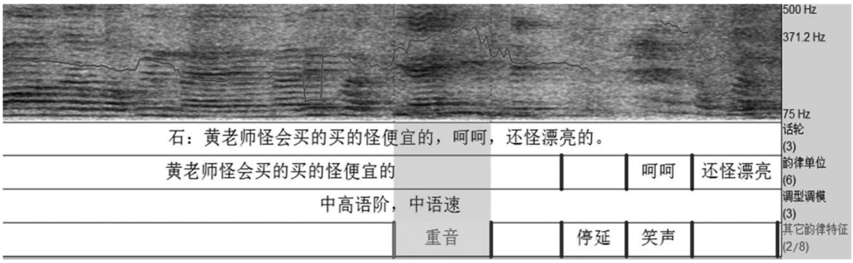


Figure 7.5 Analysis of prosodic feature.

ones (Liu, 2009: 22–23). This token has the middle–high intonation. Furthermore, the gestures (see the analysis below), the obvious smile shows his high excitement.

There are two obvious stresses in the following two intonational phrases, on '*pianyi*' (cheap) and '*piaoliang*' (pretty), respectively. The speaker tries to highlight the two advantages of the goods he purchased. The pattern is a convex tune as a whole with a peak. Such a pattern is often used in accented declarative sentences (Wu & Zhu, 2001: 334) that often underlies a certain type of emotion. Here the speaker presents joy and enthusiasm upon seeing a beautiful tent and *zanyang* the addressee for buying such a cheap and beautiful commodity. The Praat analysis diagram is shown in Figure 7.5.

3 Gestures

Prior to the performance of the illocutionary act of *zanyang*, three interlocutors set up a tent. After finishing the task-doing, the speaker looks back and forth around the tent; his non-verbal act of an obvious smile reflects his pleasure and appreciation to the tent purchaser.

Token (2): 20110912 mid-autumn festival 1 – beneficial class – praise
huang lao shi dui an men //na ge //man hao di //he he
Teacher Huang treats us very well (laughter).

This token's social situation is the doorman's room of a teaching building in a university where the speaker (Mr Guo in his forties) *zanyang* the hearer (the present author) for his care and kindness. This is only one activity type.

1 Emotions

The discourse content is the speaker *zanyang* the hearer's concern and care of him and is directly related to both parties' interests. So this token is classified as beneficial. And if the speaker *zanyang* with sincerity and genuineness, his attitude, belief, and emotions must be positive, therefore producing a felicitous illocutionary force of praise. In fact, it is possible to confirm the speaker's positive emotions from his prosodic features and gestures, etc.

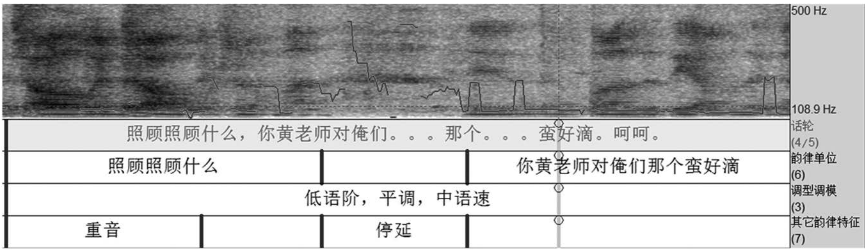


Figure 7.6 Analysis of prosodic feature.

Background emotion: **Energetic**

Primary emotion: **Delighted**

Social emotion: **Other-directed positive and grateful; other-directed negative and shy**

'Shy' is also included in social emotion because the speaker is found to have the non-verbal act of scratching the back of his head with his left hand, which is the expression of shyness. It could be that the speaker is rather shy of praising others in person.

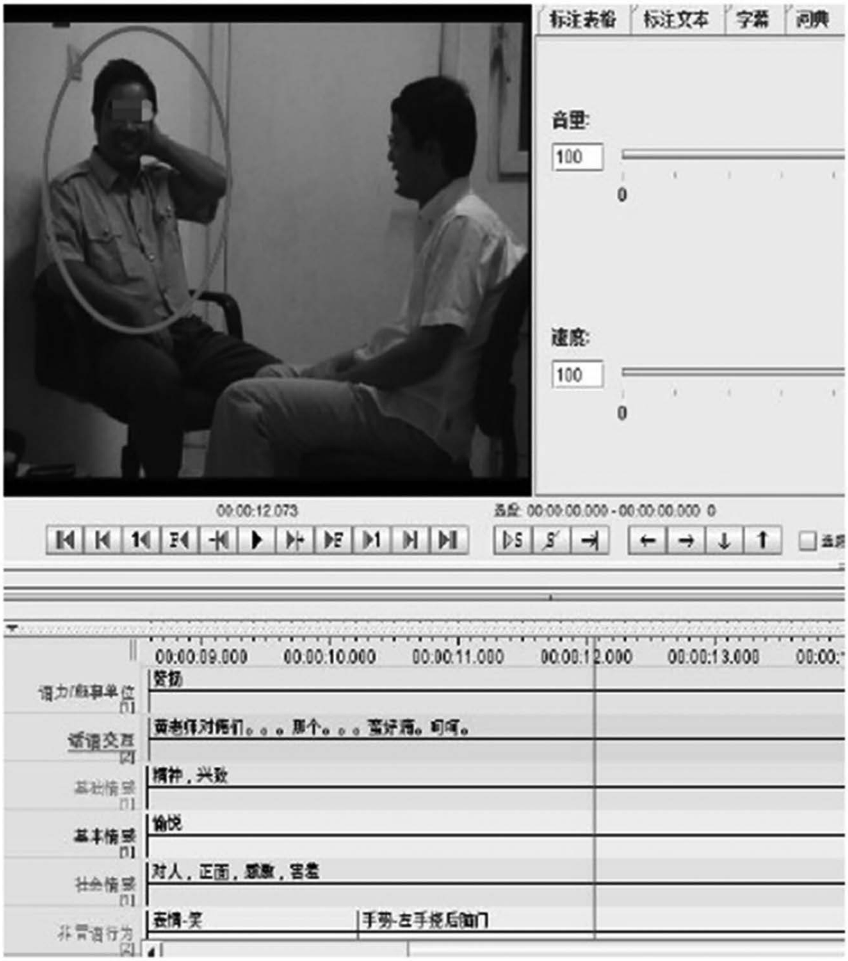
From the perspective of the hearer, the hearer laughs happily upon hearing what the speaker says, reflecting his positive emotion. This evidences the pleasant and relaxing atmospheric emotion of the entire activity type. The hearer happily acknowledges the speaker's act of *zanyang* so the perlocutionary act is performed and the purpose of *zanyang* is accomplished.

2 Prosodic features

In this token, the overall pitch is relatively low, hovering around 100 Hz. This is probably because the speaker is shy. The stress on 'man hao de' (pretty good) indicates that this speaker's subjective judgement of the hearer's daily behaviour is also the reason why the speaker *zanyang* him. The laughter in the end directly reflects the speaker's positive emotions as shown in Figure 7.6.

3 Gestures

The speaker sits on a chair and talks with the hearer throughout the entire process. His non-verbal acts include scratching the back of his head with his hand and laughing (its meaning has been analysed above). During the process of *zanyang*, the speaker does not look at the hearer (also the object of *zanyang*) all the time, but looks at other objects sometimes (watching at the TV diagonally opposite to the speaker), which should be related to the speaker's social emotion of shyness. See Figure 7.7.



音量: 100

速度: 100

选段: 00:00:00.000 - 00:00:00.000 0

语力/叙事单位 [1]	赞扬
话语交互 [2]	黄老师对我们。。。那个。。。蛮好滴。呵呵。
基础情感 [1]	精神，兴致
基本情感 [1]	愉悦
社会情感 [1]	对人，正面，感激，害羞
非言语行为 [2]	表情-笑 手势-左手挠后脑门

Figure 7.7 Analysis of illocutionary force.

7.3 Ontological properties and case study of the harmful illocutionary force

Harmful illocutionary force refers to the illocutionary force that the discourse content is negative to the interests of the speaker/hearer or will have a negative impact. Exemplars in the corpus include '*piping*' (批评, criticizing), '*baoyuan*' (抱怨, complaining), '*haipa*' (害怕, feeling scared), or '*danxin*' (担心, feeling concerned).

What echoes with the tokens in the beneficial class is that there are also two cases of the tokens in the harmful class. The first case is that the corresponding illocutionary act type is negative and negatively related to the interests of the interlocutors, such as '*piping*' (批评, criticizing), '*konggao*' (控告, accusing), '*haipa*' (害怕, feeling scared), or '*tiaoxin*' (挑衅, provoking). According to the Conceptual Model analysis, the essential content and intentions of these illocutionary act types should be negative; otherwise, the semantic meanings of these performative verbs will be violated and the unfit and unintended pragmatic meanings will be generated; the second case is that the corresponding illocutionary act type itself is neutral, such as '*jiashe*' (假设, hypothesizing), mentioned in the beneficial class. As long as the utterance content is interest-negative to the interlocutors, e.g. hypothesizing that someone will confront a negative event, the token of *jiashe* should be classified as harmful.

The five types of the harmful class in the multimodal corpus are '*baoyuan*' (抱怨, complaining), '*suku*' (诉苦, grumbling), '*piping*' (批评, criticizing), '*danxin*' (担心, concerning), and '*haipa*' (害怕, feeling scared). This section first analyses the ontological properties of different illocutionary act types based on the conceptual models in the Discovery Procedure of situated discourse and then studies them under different contexts. Only the analysis of *baoyuan* and *suku* are listed here.

7.3.1 The illocutionary act of baoyuan (complaining in Chinese Pinyin)

(1) Conceptual analysis

<The speaker's role>

The speaker performing the act of *baoyuan* can be an individual or a collective entity. In the latter case, multiple speakers complain about something together to produce multiple illocutionary act tokens.

<The role of the hearer>

For the illocutionary act of *baoyuan*, the hearer can be individual or collective, and there is a co-agency substantiated in the hearer's role that can be addressee, or audience, or both. The addressee may or may not overlap with the object complained about. The former refers to what the hearer says or causes the negative response from the speaker who then *baoyuan*; the latter refers to the speaker *baoyuan* other people (including him/herself) or other things to the hearer.

<Performativity>

The illocutionary act of *baoyuan* can be performed either performatively or non-performatively. The match between the act and force can be made explicit with the performative verb *baoyuan*. Alternatively, it can be made implicit when the speaker says a rather obscure thing and hardly touches upon the specific part that satisfies him/her. Sometimes the force of *baoyuan* can be expressed without the presence of the essential content, while the correct interpretation of the hearer is based on the common background of both sides. For instance, the wife says to her husband: 'ni ze me you zhe yang!' (Why you did this again!). In this case, the essential content (that is, what the wife is complaining about and why) is skipped, but the reference of '*zhe yang*' (this) can produce the illocutionary force of *baoyuan* since both parties in the same conversation know about the complaining content.

<Essential content>

The essential content is what the speaker complains about or feels dissatisfied with.

<Intentional states>

When performing the act of *baoyuan*, the speaker harbours dissatisfaction or resentment towards something or someone to different extents, yet also in the hope of the situation being somehow improved. Such negative attitude or belief is the speaker's subjective experience and it correlates directly with the speaker's emotions.

<Emotional states>

The background emotion hinges on the speaker's physical condition, while the primary emotion and social emotion are usually negative, such as resentment, fury, and dissatisfaction. This is because the discourse content, which is usually about some unsatisfactory thing or person, can easily trigger negative emotions.

<Occasions>

This illocutionary act token can occur in a variety of occasions. It is necessary to analyse a combination of such conditions as if the speaker's illocutionary act of *baoyuan* conforms to situational factors, the appropriateness of the speaker's identity and the interaction rules.

<Interdependency>

Illocution-and-reality interdependency: The person, event, or state makes the speaker feel unsatisfied, and resentfulness does occur. It can be a certain thing or state in the past, present, or future.

The doing-and-talking interdependency is basically consistent with the classification in Section 4.2.2.2. The speaker can solely express the illocutionary force of *baoyuan* when speaking or can express it while performing other task-doings.

The forward-and-backward interdependency should be analysed in its context. The before-and-after utterance may be the discussion on the subjective reality or certain event.

To sum up, the analysis results of the ontological properties of the illocutionary act of *baoyuan* can be presented in the formula of a set, shown in Figure 7.8.

(2) Case study

There are 14 tokens of the illocutionary act of *baoyuan* in the corpus.

Token 1–20110905 cleaning – harmful class – baoyuan, zhouma: shengqi2

shui xiao de ta si ji //mei yi tian dou shi de //san bai liu shi tian tian tian dou shi zhe yang de //fei ba zhe ge ce suo //chuang hu //wai bian chuang hu

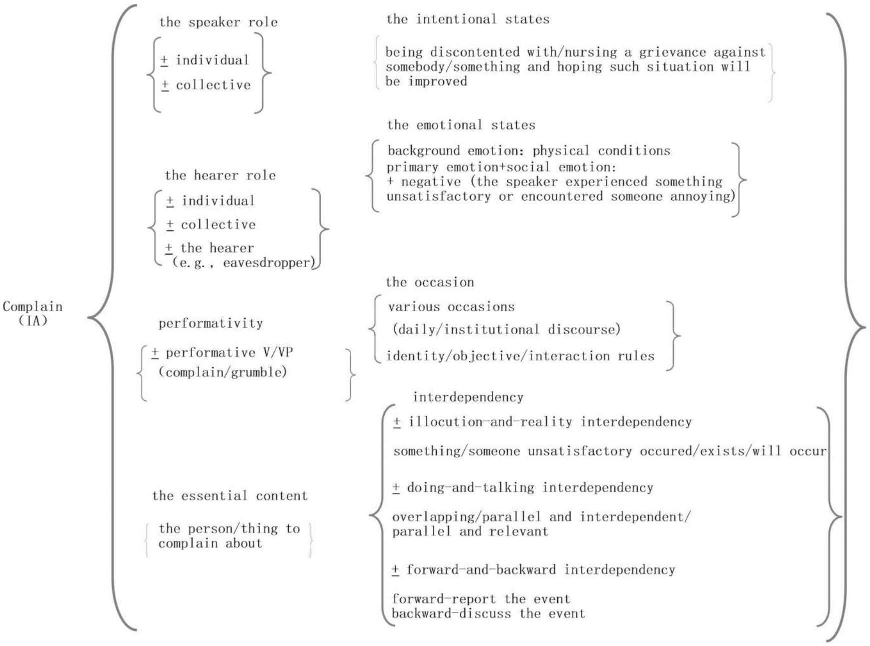

Complain (IA)
the speaker role
± individual
± collective
the hearer role
± individual
± collective
± the hearer (e.g., eavesdropper)
performativity
± performative V/VP (complain/grumble)
the essential content
the person/thing to complain about
the intentional states
being discontented with/nursing a grievance against somebody/something and hoping such situation will be improved
the emotional states
background emotion: physical conditions
primary emotion+social emotion:
+ negative (the speaker experienced something unsatisfactory or encountered someone annoying)
the occasion
various occasions
(daily/institutional discourse)
identity/objective/interaction rules
interdependency
± illocution-and-reality interdependency
something/someone unsatisfactory occured/exists/will occur
± doing-and-talking interdependency
overlapping/parallel and interdependent/parallel and relevant
± forward-and-backward interdependency
forward-report the event
backward-discuss the event

Figure 7.8 Analysis results of the ontological properties of the illocutionary act of baoyuan.

guan zhe //zhe men guan zhe //ni jiang jin lai yi gu chou wei yi gu sao wei // ni ba ta kai zhe //ta fei ba ta guan zhe //tuo ba ne //ni ba ta zhe yang liang qi lai //ta fei ba ta fang shang shui //ba ta yang zhe //tian tian tuo ba dou lan diao le

Who can imagine that he did this every day? Why on earth he would shut the toilet window and door? Wouldn't everyone feel the nasty smell at the entrance of the toilet? Even if I open them frequently, he keeps shutting them off. Not to mention his behaviour of soaking the mop in water. They would get rotten.

There are two activity types in the social situation of this token: One is the activity type where this illocutionary act token of *baoyuan* takes place, that is, at the entrance to the bathroom in a teaching building, the speaker Shi cleans the mop and mops the floor while complaining to the hearer (the present author) that someone always closes the door and windows of the bathroom upstairs causing a nasty smell and also that the mop is always soaked in the sink. In addition to this activity type, there is another cleaner besides her, working but not participating in the conversation.

The fact is that what the speaker Shi complains about is directly related to her work, and that the perpetrator's behaviours make it much more difficult for the speaker to do her cleaning work. So she embraces strong negative attitudes such as dissatisfaction and disgust. One feature of the illocutionary act token of *baoyuan* is that it often co-occurs with verbal acts such as threatening, accusing, and insulting. In this token, the act of '*ruma*' (辱骂, insulting) appears before and after the act of *baoyuan*.

1 Emotions

From the cues from utterance content, prosodic features, and gestures, it can be found that the speaker's negative emotions are quite intense. In this token, the primary emotion and social emotion need to be sub-classed into reported emotion and situated emotion (see the discussion of the difference in Section 4.2.3.5). This is because the interdependency of the essential content of the act of *baoyuan* points to the past and continues to the present, which infers that the perpetrator has always had such bad habits. When talking about such habits remaining from the past until now, the speaker presents reported emotion as well as situated emotion.

Primary emotion: **Reported – anger; situated – anger**

The speaker thinks that the perpetrator's behaviour has caused a lot of trouble and has remained unchanged. The bathroom where the perpetrator's behaviour directly causes the peculiar smell is exactly the place where the speaker cleans. Now it has become an unpleasant place so the emotion of anger has been triggered on the speaker's side.

Social emotion: **Reported – other-directed, negative, contempt; situated – other-directed, negative, contempt**

Since the speaker embraces negative emotion towards the perpetrator's behaviours and she deems such behaviour incomprehensible and despicable, she shows the social emotion of other-directed '*biyi*' (鄙夷, contempt), not to mention that she has the act of ruma prior to the act of *baoyuan*.

Background emotion: It is related to the speaker's physical condition and directly impacts the rhythm tone. Judging from her actions, demeanour, and speaking voice and strength, the speaker is physically healthy and in her forties. So the background emotion is annotated as **energy** and **well-being**.

2 Prosodic features

The speaker presents a very high pitch when expressing the illocutionary force of *baoyuan*, basically ranging from 300 to 400 Hz with the highest exceeding 500 Hz. According to the division of Wu and Zhu (2001: 396), this intonation group falls into the high pitch group. From the perspective of fundamental frequency, the prosody of anger generally is located in the high frequency range (Liu, 2011: 61). See Figure 7.9.

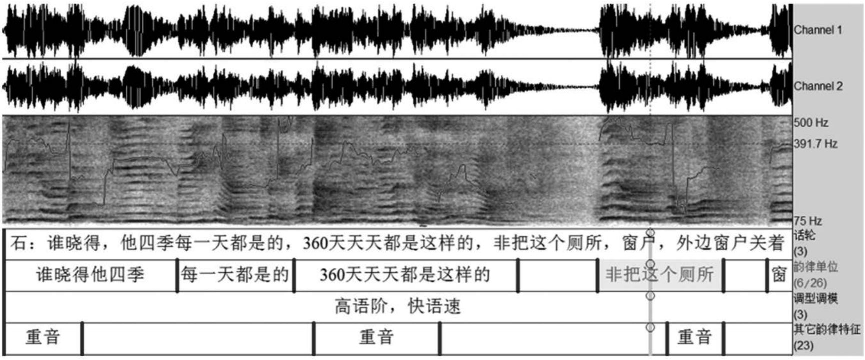


Figure 7.9 Analysis of prosodic feature.

For the speech rate, the whole token (duration 23 seconds, with 91 syllables in total) is calculated as 252ms per syllable. According to the standard stipulated by Wu and Zhu (2001: 398), the speech rate of this token is relatively fast. Compared with other emotions like feeling neutral, happy, sad, or surprised, the emotion of anger has the fastest tempo (Liu, 2011: 61) that reflects the speaker being rather emotional.

The high intonation presented in this token is in sharp contrast with that in the neutral class. When performing the neutral illocutionary act, the speaker maintains her intonation between 200 and 250 Hz, which is a typical low intonation mode. Here, the high intonation and fast tempo exactly show the speaker is filled with excitement or anxiety. Due to the fact that the speaker's negative emotions (anger, contempt, etc.) are relatively intense, her tonal mode instinctively mirrors such negativity and enables the hearer to feel the strength of them.

Furthermore, another obvious prosodic feature is the occurrences of pauses on 'san bai liu shi tian tian tian zhe yang' (every single day all the year around), 'jin lai yi gu chou wei' (the smell hit when entering the bathroom), and 'tuo ba dou yao lan diao le' (the mop is soaked to rot), etc. These stresses underscore the speaker thinking that the perpetrator's behaviour has caused dire consequences and he/she did so every day. Compared with the emotion of feeling neutral, happy, sad, and surprised, the emotion of anger includes more stresses (Liu, 2011: 61).

3 Gestures

Task-doing: The speaker is standing and mopping when performing the act of *baoyuan*.

Non-verbal acts: The speaker presents hand and head movements.

Among them, the expression of frowning exists when the act of *baoyuan* begins and lasts for six seconds. Such expression directly reflects the speaker's primary emotions (anger) and social emotion (contempt); the right-hand movement (twice) is employed when the speaker explains the specific behaviour of the perpetrator (closing the bathroom door and windows, soaking the mop in the sink) to emphasize the perpetrator's behaviour.

7.3.2 The illocutionary act of suku (grumbling in Chinese Pinyin)

(1) Concept analysis

<Speaker's role>

The speaker's role is generally individual. If collective or multiple speakers express the illocutionary force of *suku* (诉苦, grumble) together, multiple tokens will be produced.

<Hearer's role>

The hearer can be individual or collective. The latter can be at certain school conventions centring on the topic of 'recalling sufferings in the old society and contrasting them with happiness in the new society' when people who have lived in the old society or old revolutionaries are invited to share their life in the past with students.

<Performativity>

The chances are that the performative verb of *suku* would not appear. However, in most cases, the speaker sighs and emphasizes how hard his/her own life in the past was. So the discourse content can evidence this illocutionary act.

<Essential content>

In the past, the essential content is that the speaker thought it as *ku* (苦, miserable/bitter), such as not having enough food or clothes, or having no shelter.

<Intentional states>

The beliefs and attitudes of the speaker when expressing the illocutionary force of *suku* are negative. This is because the utterance content directly harms the speaker's interests (makes the speaker suffer). And the speaker is convinced that he himself/she herself is suffering from these tough and negative things.

<Emotional states>

The background emotion hinges on the physical condition of the speaker. Moreover, the primary emotion and social emotion are generally negative because the person or events adverse to the interests of the speaker (e.g. lack of food or clothes or being bullied by others) are basically the inducing condition of negative emotions.

<Occasions>

The illocutionary act of *suku* can occur on various occasions. It is necessary to analyse the combination of such conditions as the speaker's illocutionary act conforms to situational factors and association rules.

<Interdependency>

Illocution-and-reality interdependency: The event or state makes the speaker feel miserable. It can be a certain thing or state in the past or present (i.e. what the speaker is suffering from).

The doing-and-talking interdependency is basically consistent with the classification in Section 4.2.2.2. The speaker performs the act of *suku* solely when speaking or performs it while performing other task-doings.

The forward-and-backward interdependency should be analysed in its context. The previous utterance may be a discussion of the speaker's

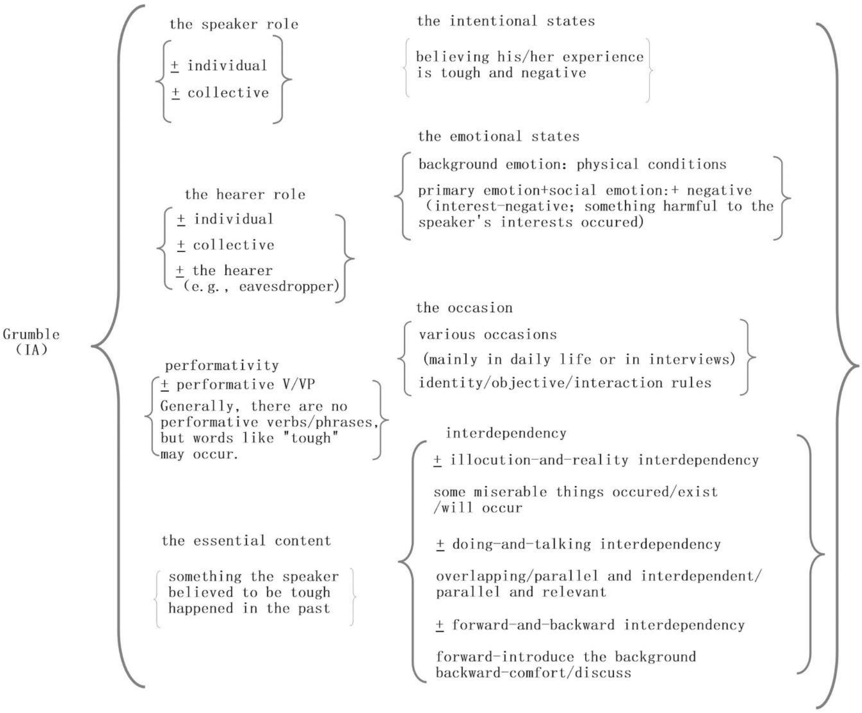

Grumble (IA)
the speaker role
± individual
± collective
the hearer role
± individual
± collective
± the hearer (e.g., eavesdropper)
performativity
± performative V/VP
Generally, there are no performative verbs/phrases, but words like "tough" may occur.
the essential content
something the speaker believed to be tough happened in the past
the intentional states
believing his/her experience is tough and negative
the emotional states
background emotion: physical conditions
primary emotion+social emotion:+ negative (interest-negative; something harmful to the speaker's interests occured)
the occasion
various occasions
(mainly in daily life or in interviews)
identity/objective/interaction rules
interdependency
± illocution-and-reality interdependency
some miserable things occured/exist /will occur
± doing-and-talking interdependency
overlapping/parallel and interdependent/ parallel and relevant
± forward-and-backward interdependency
forward-introduce the background
backward-comfort/discuss

Figure 7.10 Analysis results of the ontological properties of the illocutionary act of suku.

experience before, and the utterance after could be discussion of the speaker's concurrent state or the hearer's consolation of the speaker.

To sum up, the analysis results of the ontological properties of the illocutionary act of *suku* can be presented in the formula of a set in Figure 7.10.

(2) Case study

There are 11 tokens of the illocutionary act of *suku* in the corpus.

Token 1: The father-in-law of Li Shuangfu (before lunch) 2 – harmful group –suku

wo ku ri zi //jiang bu dao di le //ni kan //fu qin //shi jiu sui fu qin jiu si diao le //wu jiu nian //fu qin jiu si diao le //suo yi wo dai zhe mu qin //yi hou //dao //liu yi o liu er nian //zai //liu er nian zai yao de lao po //zai tan lian ai //yi hou nie //yao chu lai //yi jia jia man man ao ao ao ao //wo di di hai you yi ge di di yi ge mei mei //dou e si le //

e si diao le //dou e si diao le //na shi hou nong cun na shi hou e si de ren //ni zheng bu kai yan a //hao hao de ren a //zou zou jiu jiu mu you le

My miserable days are endless. I lost my father in 1959 when I was only 19. And then there were just two members in my family, my mother and me.

I married my wife in 1961, oh no, 1962. Later we just hang on there. I once had a brother and a sister, but they all died from hunger. People in the rural area then all starved extremely and died. Jesus Christ. Those are live people.

In this token, the speaker Tang in his seventies is from Anhui Province and makes a living by setting up a barber's stall in the community every day. The conversation venue is the community activity room. Due to the sudden rain, speaker Tang then moves his stall to the activity room with other community residents. At that time, Tang stands beside his stall and recalls his past sufferings with one hearer (the present author) and another (a resident of the same community). Therefore, in such a large social situation as the community activity room are included the activity type of speaker performing the act of *baoyuan* and other activity types like community residents playing mahjong or participating in other activities.

In this activity type, the speaker recalls tragic experiences including his father's death, his late marriage due to financial burdens, and his siblings starving to death due to famine. Judging from his words, gestures, and prosody, he feels very pained about these experiences.

It is worth noting that although the resident next to the speaker is merely listening to the speaker *suku*, he also presents frowning, staring, and opening his mouth slightly, which reflects that he also feels bad and sympathetic for the speaker.

According to the STFE-match principle discussed earlier in the present research, we can infer that the primary reason why that hearer shows sorrow and sympathy is that he sympathizes with the speaker and indeed thinks these experiences are very miserable. In talking with him afterwards, the author learned that this hearer was a believer in Buddhism and stressed that people should be compassionate and kind. Basically, it can be judged that the hearer indeed feels sympathetic and sorrowful, which forms a perfect match between what is thought and what is felt.

As mentioned above, the similar situation involves the atmosphere emotion (see Gu, 2013a). When the speaker is engaged in the activity type of performing the act of *suku*, the holistic atmosphere is sad or sentimental. Hatfield, Cacioppo, and Rapson (1994) believed that emotions can be 'contagious.' So what they are actually talking about is this phenomenon (Gu, 2013a). Judging from the speaker's and the hearer's performance, this activity type's atmosphere emotion is confirmed as sad.

Of course, other activity types in the same social situation, i.e. the residents playing mahjong and other activities, are relatively independent of the activity type of *suku*, and therefore, their dominant atmosphere emotions (should be relaxation and cheerfulness) are not affected.

1 Emotions

The cues from utterance content, prosodic features, and gestures manifest that this token is the speaker's recalling of his tragic and arduous past. The illocution-and-reality interdependency of the essential content points to the past (the speaker had been quite miserable before the act of *suku*). Therefore, the primary emotion and social emotion need to be sub-classed into report emotion and situated emotion in this token.

Primary emotion: **Report – sadness; situated – sadness**

Social emotion: **Report – self-directed, negative, contempt; situated – self-directed, negative, contempt.** It is the evaluation on the speaker's own experiences in the past, so the emotion is duiji. And it is negative in nature accompanied by the emotional factor of pity.

Background emotion: It is related to the speaker's physical condition and directly impacts the rhythm tone. Despite that the speaker is aged over 70 years, he is still healthy, judging from his other tokens recorded by the author. Furthermore, the prosodic features and gestures show that the speaker is quite **energetic** even though he is recalling certain sad experiences.

2 Prosodic features

It can be seen that what differs from the segmentation of the intonation units in this token and that of other tokens is more occurrences of pausing and relatively short intonation units with fewer syllables.

Except for four pauses caused by the speaker's thinking about words, the remaining pauses number 11, whose average duration exceed one second. Generally speaking, there are many occurrences of pauses in the emotion of sadness, mainly located in the middle and last long (Liu, 2011: 61).

What causes such a phenomenon? Generally speaking, more occurrences of pauses and smaller prosody units are justifiable when the speaker is elderly who gets breathy in his/her physical condition or becomes breathy after exercise or other reasons. However, the speaker is still in good health and talks with a loud voice and ample confidence when talking about other topics. So the above-mentioned possibility can be ruled out. We are encouraged to find the reason underlying this illocutionary act type and his emotion: The act of recalling past painful experiences often causes the speaker to talk more or less intermittently, and he/she may even choke up to some degree. Although the speaker in this token does not choke up, the past tragic experience discourse content still prompts him to show distinctive characteristics.

In addition, the overall intonation of this token is long, with the highest reaching roughly 200 Hz.

Tonal mode often has the function of conveying emotions and meanings. The low pitch and other prosodic features analysed in the part of prosodic feature demonstrate that the speaker bears the feelings of sadness, self-pity, and helplessness towards the discourse content (the past tragic experiences).

3 Gestures

Task-doings: The speaker is standing to have the conversation throughout the entire process.

Non-verbal acts: Include hand shaking and head shaking.

The most noticeable movement is the speaker shaking his head when talking about his siblings starving to death, which entirely tells how sad and helpless the speaker felt then.

Two more tokens are cited here to further illustrate the typical prosodic features and non-verbal acts of a speaker expressing the illocutionary force of *suku*.

The first one is the token already discussed in the multimodal segmentation and annotation scheme, that is, the old lady Chou in her seventies who lives in Tonglu, Zhejiang, recalling her hard childhood.

Token 2: Chou – harmful class – suku

ai //zhen ku ai //na shi hou shi ku a //ta men dou shuo de //ni ma ma ne // zai de hua duo hao a //ni ma ma //xiang fu mei you xiang fu //ku que chi gou le //wo men mei you de chi

na hui lai he wo ba ba jiang //ge jiang jiang wo men ba ba //yi jin bai mi liang ge （error）san ge ren （repair）chi liang tian //hu yi hu chi chi

duo shao can guo a //ku tou chi si lao zao

[sighs] Alas. really suffering at that time. They all said that it would be more consummate if your mother is still alive. Your mother didn't enjoy her life, but only suffered a lot. We suffered from hunger.

I told my father when I returned…told my father. Two [error] three persons [repair] took only about a pound of flour for two days. Stir with water

How miserable we were then.

1 Emotions

Primary emotion: **Report – sad, angry; situated – sad, angry*** (* means this emotion is weak or uncertain). The speaker is sad about her unfortunate past including suffering hunger due to poverty and experiencing her mother's death;

Social emotion: **Report – self-directed negative and pity + other – directed negative and resentful; situated – self-directed negative, pity + other-directed negative, resentful***. Since the speaker is recalling her own past miserable experiences, the emotion is self-directed. And the emotional nature is negative, misery, helplessness, and self-pity respectively in three segments;

Background emotion: It is related to the speaker's physical condition and directly impacts the tonal pattern. Although aged over 70 years old, the speaker is still emotional and talks with a powerful voice. So background emotion can be annotated as **energy** and **well-being**.

2 Prosodic features

The two intonation groups, i.e. 'zhen ku ai//na shi hou shi ku a' (so hard// life then was so hard) present a prominent falling tone. The tone appears to be weaker and lower as the speaker continues, which shows that the speaker feels sad for her tragic past. Before the speaker begins the act of *suku*, there is the prolonged (subordinate to the occurrence of pause) and accented prosodic unit of '*ai*.' This is obviously an intense

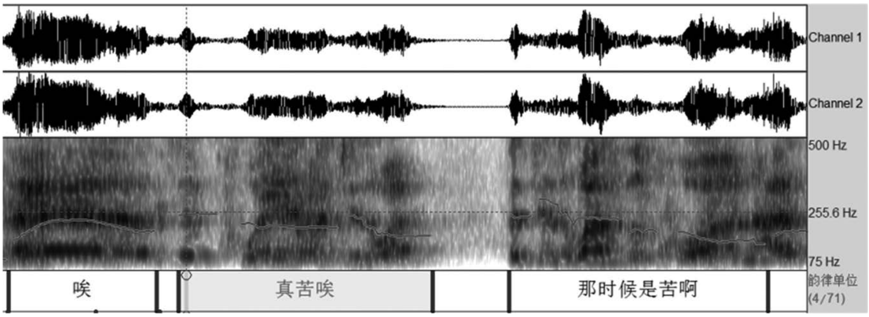


Figure 7.11 Analysis of prosodic feature.

sigh (modular particle) showing that the speaker is sad and helpless about the experiences she is going to converse about. Two exclamatory phrases follow:

zhen ku ai//na shi hou shi ku a

These two intonation groups are typically a falling tone. See Figure 7.11.

Liu (2009: 24) concluded from corpus analysis that in the corpus with rather intense emotion, exclamatory sentences still show a falling tone on the whole. In this token, the speaker's lamentation over the hardships, as well as facial features and other physical characteristics, suggest that her emotion of sadness is quite intense.

3 Gestures

In this token, this old lady presents rich non-verbal acts, including wry smiles, frowning, and head shaking, and waving hands. These are the message of her inner sadness. The following token is even more typical.

Token 3: Wu – harmful group – grumble

The social situation of this token takes place in a farmer's house in Shangdong. The speaker, Wu, a farmer in his forties, recalls his childhood of being bullied and escaping from being pushed into the well. And meanwhile, there are two hearers who perform the task-doing of cooking, whose activity type is independent from the speaker.

Due to the long duration of this token, only the discourse content before and after the non-verbal act of choking silence and covering the face to weep are picked out here:

bu suan wan //bu suan wan jiu jin jie zhe ba wo yi jiao yi jiao [qing]（yi zhi）[pai]（chuai）kai le //zhe jiu shi shuo an na shen zi //zhe jia li lian ge ge dai sao zi //lian lao de dou qu gei ta pei bu shi //zhe me pei dou bu hang //

[lowering head, touching face with the left hand, wiping tears]

zhe shi bu ying gai shuo //yi shuo jiu //

Not done yet. They then immediately kick me away. This means that my aunt, brother, and sister-in-law all went to apologize. That won't work.

[lowing his head, touching his face with left hand, wiping tears]
This should be kept a secret. But I just blurt it out.

1 Emotions

In this token, in addition to the sadness upon recalling the childhood memory of being bullied, the speaker also presents the belief-mediated emotion, that is, the emotion triggered by belief. Because the speaker believes that the people who bullied him (surprisingly his relatives) were aggressive and insistent on killing him regardless of his outcry, he deems them as 'ban dian ren xing dou mei you' (with no shred of humanity left), and the emotion of anger arises.

Primary emotion: **Report – sad, angry; situated – sad, angry**;

Social emotion: **Report – self-directed negative and pity + other-directed negative and resentful; situated – self-directed negative, pity + other-directed negative, resentful***;

Background emotion: **Fatigue**

Belief-mediated emotion in this token manifests that social emotion can be composite. When the speaker Wu recalls the time when as a child he was kicked and almost killed by his relatives, he presents self-directed compassion and the social emotion of contempt towards his relatives (other-directed). Therefore, this is a token of composite social emotions.

2 Prosodic features

The most significant prosodic feature is long pause occurrences (average duration 7.6 seconds) when the speaker sometimes chokes or sometimes buries his face in his hands to wipe tears. This finding is consistent with the feature of high frequency of the occurrence of longer pauses in the previous token; plus, there is no significant accent. The statistic of the emotion corpus in Liu (2011: 59) research shows that this speaker has presents the most pauses whose average duration is longer than that of other emotions. This statistical analysis results are concordant with the prosodic features of the speaker when expressing the illocutionary force of *baoyuan*. The pitch of this illocutionary act token is relatively low, hovering around 150 Hz. This is mainly because the speaker is filled with sadness.

3 Gestures

When the speaker performs the act of *suku*, he has movements such as choking in silence and many times covering his face to weep. Moreover, he keeps his head down during the whole conversation, accompanied by several occurrences of pausing (in silence).

7.4 Ontological properties and case study of the counterproductive illocutionary force

The counterproductive illocutionary force refers to the illocutionary force when the discourse content runs contrary to the primary expectation on the speaker/hearer's interests, or leads to a contradiction between the expectation and the reality. Exemplars include '*diulian*' (丢脸, feeling embarrassed) about things did not turn out as wished, '*weiqu*' (委屈, feeling aggrieved) for being wronged, '*wunai*' (无奈, feeling distressed) or '*yihan*' (遗憾, regretting) for something not working as expected.

The ontology of the counterproductive illocutionary act type is usually negative because the utterance content does not meet the interlocutors' initial expectations in the conversation. However, it is different from the harmful class. The reasons include the following: First, the failure to meet expectations does not necessarily produce a negative impact on the interests; second, the counterproductive class is usually based on the fact having already happened. And only after it happens can we judge whether the result is agreeable to the interlocutors' expectations, while the harmful class is not necessarily based on the facts that have occurred but may threaten to have a negative impact in the future, such as the act of '*weixie*' (威胁, threatening) and '*ansuan*' (暗算, plotting). Therefore, it is necessary to distinguish these two illocutionary act types.

The five types in the multimodal corpus are '*shiwang*' (失望, feeling disappointed), '*weiqu*' (委屈, feeling aggrieved), '*wunai*' (无奈, feeling distressed), '*yihan*' (遗憾, regretting), or '*jingya*' (惊讶, feeling astonished). This section first analyses the ontological properties of different illocutionary act types based on the conceptual models in the Discovery Procedure of situated discourse and then studies them in different contexts. Only the analysis of the illocutionary acts of *shiwang* and *weiqu* is presented here.

7.4.1 The illocutionary act of shiwang (being disappointed in Chinese Pinyin)

(1) Concept analysis

<Speaker's role>

The speaker can be an individual or a collective group expressing disappointment about something or someone.

<Hearer's role>

The hearer perceiving the illocutionary act can be an individual or a collective group. And there is a co-agency substantiated in the hearer's role that can be addressee, audience, or both (the first and second cases mean that the hearer is not the one to whom the speaker speaks, but who observes silently during the conversation). Sometimes, the addressee is the audience or the perpetrator causing the speaker *shiwang*, that is, the addressee's certain behaviour makes the hearer feel disappointed.

<Performativity>

Although the performative verb does not necessarily appear, the speaker sometimes uses the syntactic formula of 'wo shi wang' (I am disappointed) and sometimes explicitly expresses the essential content of *shiwang* accompanied with sighing, the expression of feeling a loss, or gestures like stretching out his/her hands.

<Essential content>

The essential content in the illocutionary act of feeling disappointed refers to what happened or what kind of behaviour causes the speaker to feel disappointed. According to Searle's classification of illocutionary acts, the illocutionary act of feeling disappointed can be classified as expressive, that is, the speaker expresses his/her inner state, attitude, or emotion.

<Intentional states>

The speaker believes that a certain objective fact or event that is about to happen has irreversibly violated his/her initial expectations, which has thereby produced a negative attitude and belief.

<Emotions>

The background emotion hinges on the physical condition of the speaker. And the primary emotion and social emotion are generally negative because the scenarios running contrary to expectation induce conditions of negative emotions.

<Occasions>

This illocutionary act can occur in a variety of occasions. It is necessary to analyse the combination of such conditions to ascertain whether the speaker's illocutionary act conforms to situational factors, his/her identity, or other factors.

<Interdependency>

Illocution-and-reality interdependency: The fact that already exists or will happen in the future, which makes the speaker disappointed and goes contrary to his/her expectations.

The doing-and-talking interdependency is basically consistent with the classification in Section 4.2.2.2, that is, one can express *shiwang* solely when speaking or expressing it when performing other task-doings.

The forward-and-backward interdependency should be analysed in its context. The discourse occurring before and after the illocutionary act token may be a discussion on an objective state or event, or another's consultation with or consolation of the speaker.

To sum up, the analysis results of the ontological properties of the illocutionary act of *shiwang* can be presented in the formula of a set (See Figure 7.12):

(2) Case study

There are five tokens of the illocutionary act of *shiwang* in the corpus.

Token: Li, father-in-law – counterproductive class – no labour insurance – disappointed

xian zai hao //xiang wo men nong cun //lao bao ye mei you //shi me ye mei you //guang //***[person's name]//jiang //duo shao nian la

jiang //gei nong cun //gei lao bao //dao xian zai gei na ge ne //wo men hai mei you ne //wo men dou qi shi duo sui le la hai mei you

dao xian zai mei you na dao lao bao

74nian //jiu jiang le //gei le //74nian o //jiu jiang le //ai ***[person's name] bu jiu jiang le me //nong cun dou yao you lao bao //wo men dao xian zai hai mei you

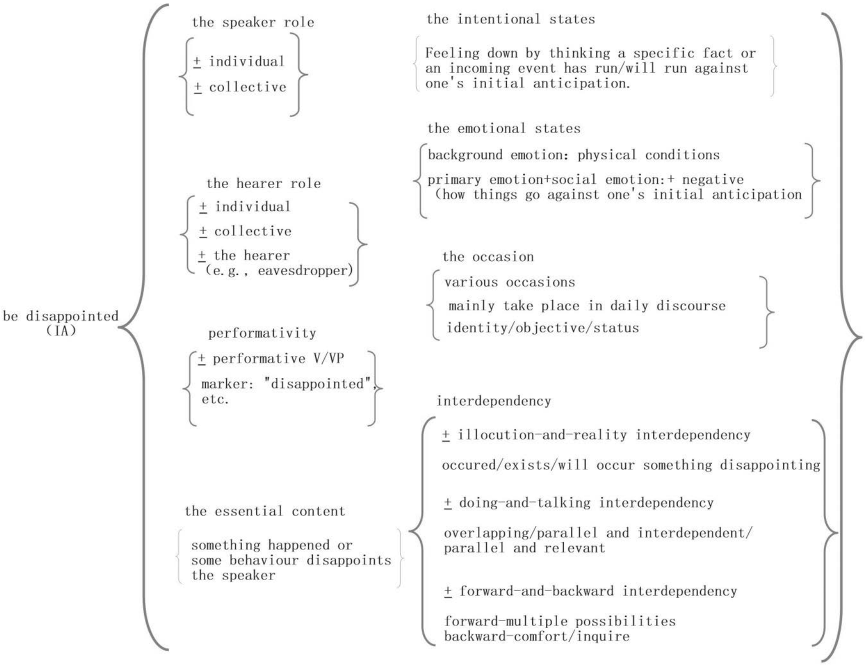

be disappointed (IA)
the speaker role
± individual
± collective
the hearer role
± individual
± collective
± the hearer (e.g., eavesdropper)
performativity
± performative V/VP
marker: "disappointed", etc.
the essential content
something happened or some behaviour disappoints the speaker
the intentional states
Feeling down by thinking a specific fact or an incoming event has run/will run against one's initial anticipation.
the emotional states
background emotion: physical conditions
primary emotion+social emotion:+ negative (how things go against one's initial anticipation
the occasion
various occasions
mainly take place in daily discourse
identity/objective/status
interdependency
± illocution-and-reality interdependency
occured/exists/will occur something disappointing
± doing-and-talking interdependency
overlapping/parallel and interdependent/ parallel and relevant
± forward-and-backward interdependency
forward-multiple possibilities
backward-comfort/inquire

Figure 7.12 Analysis results of the ontological properties of the illocutionary act of shiwang.

Alas now People in we rural areas has no labour insurance. We have nothing. ***[person's name] complains that for how many years people in the countryside do not have the labour insurance. We haven't got it yet even if we are all over 70 years old.

In 1974 word has it that we would get the labour insurance. Alas. *** [person's name] mentioned that there must be labour insurance in rural areas; however, we have not yet achieved it.

The social situation of this token takes place in the residents' activity room of some community. There is only one type of activity: The speaker Tang, an old man in his seventies from Anhui Province, expresses his disappointment that he has never received a pension. There are two addressees, one is the author (also the discourse recorder) and the other is the speaker's acquaintance in the same community who stands on the right. Another hearer who stands on the left of the speaker is his wife. In this token, the speaker speaks of the fact that he has been looking forward to applying for a pension that was said to be paid by the local authority. However, due to various reasons at the rural grassroots, he has not received the money and feels very disappointed about it.

1 Emotions

In this token, the speaker expresses his disappointment at not receiving the pension. The three tiers of emotions are analysed as follows:

Primary emotion: **Report – anger; situated – anger***. During the conversation, one of the hearers further asks why he has not received a pension and the speaker answers that it is likely to have been exploited by grassroots leaders. And the more he speaks, the more agitated he becomes and he begins to show a certain degree of anger.

Social emotion: **Report – other-directed – negative, dissatisfied; situated – other-directed – negative, dissatisfied**. The speaker expresses his disappointment with the disbursement of the pension at the grassroots level, thereby showing negative emotion of other-directed dissatisfaction.

Background emotion: **Energy**. The speaker speaks with confidence and is in good spirits.

2 Prosodic features

When the speaker says 'lao bao ye mei you' (I have no pension at all), his average pitch falls below 150 Hz, which is in the range of low pitch. And the low tempo indicates that he feels at a loss caused by the contradiction between the expectation and reality; when the speaker says 'shen me ye mei you' (there is nothing), there is a stress and a slight increase in the pitch. This means that the speaker is ruffled by such practices at the grassroots level in the rural area, depriving him of the benefits available to the elderly in urban areas.

Similarly, in the following discourse, the low pitch in several intonation groups like 'dao xian zai gei na ge ne' (which one do we give it now), 'wo men hai mei you ne' (we haven't had yet) ranging from 140 to 150 Hz as well as the low speech rate depict his feelings of helplessness and disappointment; in the following intonation group 'wo men dou qi shi duo sui le hai mei you' (We are all over 70 years old and have no labour insurance yet), the speaker suddenly raises the pitch and emphasizes it. The accent here highlights this part as the focus. The focus in the sentence is the concentrated expression of the speaker's intention, emotion, and attitude, which is subjective (Zhang, 2014: 91). From the perspective of sentence structure, it is an exclamatory sentence in which the speaker shows intense emotion about the status quo that he has no pension yet. Therefore, such prosodic features of the exclamatory sentence with intense emotion like stresses and high pitch all appear (Liu, 2009: 23–24). The speaker sighs that he has already moved into his seventies but has no pension yet and asks when the money will be available to him. His emotions become slightly agitated, and meanwhile his disappointment is revealed even more.

3 Gestures

Task-doing: The speaker is standing throughout his performance of the act.

Non-verbal acts: Apart from shaking his head, staring at others, and pointing, the speaker presents another two obvious non-verbal acts, i.e. stretching out his hands and grimacing.

There are two times of stretching out his hands, and the average duration is over six seconds. Such an act of stretching hands means indifference and nonchalance and, more importantly, helplessness and disappointment. Such an act lasting a long duration reflects the speaker's inner helplessness.

The wry smile appears in 'wo men dou qi shi duo sui le hai mei you' (We are all over 70 years old and have no labour insurance yet), accompanied by the stress and high pitch. These clues jointly reflect the speaker feeling dissatisfied and slightly agitated but also helpless and disappointed in the end.

7.4.2 The illocutionary act of weiqu (feeling aggrieved in Chinese Pinyin)

(1) Conceptual Analysis

<Speaker's role>

The speaker performing the act of *weiqu* can be individual or collective.

<Hearer's role>

The hearer perceiving the act can be individual or collective. And there is a co-agency substantiated in the hearer's role that can be addressee, audience, or both.

<Performativity>

Although a performative verb does not necessarily appear, the speaker sometimes uses syntactic structure such as '我受委屈了' (feel aggrieved) to perform the illocutionary act of *weiqu*. However, in most cases, the speaker will resort to the implicit strategy. The hearer can perceive the illocutionary force of *weiqu* upon seeing clues such as discourse content, prosodic features, and expressions.

<Essential content>

According to Searle's classification of illocutionary acts, the act of *weiqu* can be classified as expressive, that is, the speaker expresses his/her inner state, attitude, or emotion. The essential content of the act of *weiqu* includes the true situation that others have not understood, or the reason and process that lead to such negative cases as misunderstandings and distortions. In other words, the essential content is the difference between the speaker's expectation and the reality, or the difference between the speaker's real situation and others' beliefs and attitudes.

<Intentional states>

In the token of performing the act of *weiqu*, the speaker believes that the true situation is yet to be understood, or fails to accept the gap between the reality and the inner expectation, etc. which is the belief and attitude basis for the successful performance of the illocutionary act of *weiqu*. In the absence of such gap, such an illocutionary act is not genuine. This marks one of the important differences between the counterproductive and the harmful class.

<Emotions>

The background emotion hinges on the physical condition of the speaker. Furthermore, the primary emotion and social emotion are generally negative because the scenarios running contrary to expectation are always the inducing condition of negative emotions.

<Occasions>

The illocutionary act of *weiqu* can occur on a variety of occasions. It is necessary to analyse it in combination with such conditions to see whether the act conforms to situational factors, the speaker's identity, or other factors.

<Interdependency>

Illocution-and-reality interdependency: The contradiction between the speaker's expectation and the reality or the objective fact that is misunderstood can trigger the state of *weiqu*. Therefore, it can point to the past and the present.

The doing-and-talking interdependency is basically consistent with the classification in Section 4.2.2.2, that is, one can express *weiqu* solely or express it while performing other task-doings.

The forward-and-backward interdependency should be analysed in its context. The before-and-after discourse may be the discussion on a certain state or event, or someone consulting the speaker about something or comforting the speaker.

To sum up, the analysis results of the ontological properties of the illocutionary act of *weiqu* can be presented in the formula of a set (see Figure 7.13).

(2) Case study

There are five tokens of the illocutionary act of *weiqu* in the corpus.

Token: 20110914-110922 –_Ms Shi's father – drinking pesticide – weiqu

ni guang kao wo men wo men ye bu hang //wo men ye you fu dan //ye you [bu qing]//hai you shang ren //ni bu guang //

wo //ge ge yi ge dang bu le jia //ta ye bu guan bu wen di dou shi wo men // zi mei san ge //wo xiao mei zai jia li //dou bu chu lai //ta e //hai you //zai jia dai xiao hai //liang ge xiao hai shang xue //fan zheng jiu shi wo men san ge // en mei ban fa

wo men duo shi bu duo ya //zong gui ta jiang ni zai wai bian

You can't solely rely on us. // We also have our burdens and [unclear]//and parents and grandparents to take care of.//

Our brother can't do it alone. //He did not bother to take a look. It is always we three. //My little sister always stays at home to take care of children.// Two kids go to school. //Anyway, it's the three of us. //Hmm.//

We do not have much money.// He always says that I stay outside.

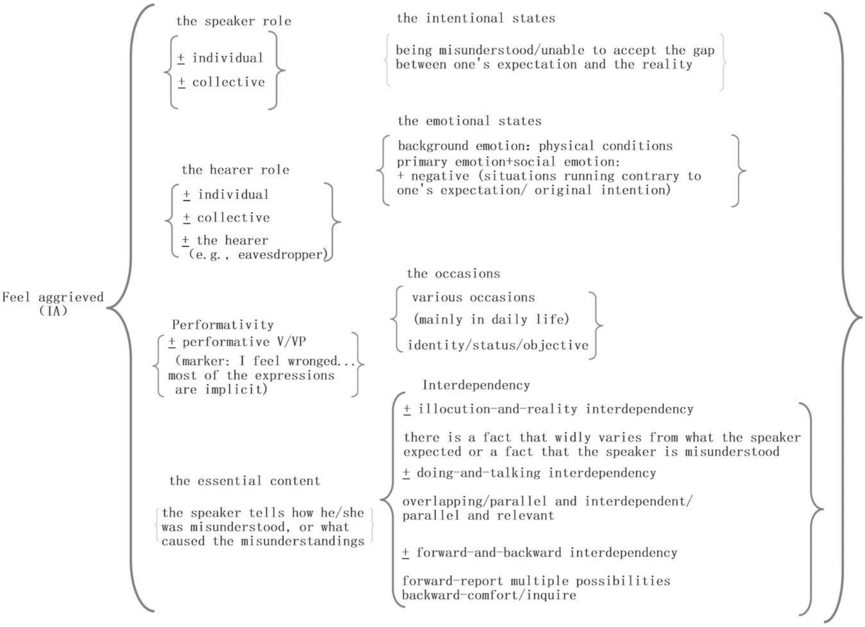

Feel aggrieved (IA)
the speaker role
± individual
± collective
the hearer role
± individual
± collective
± the hearer (e.g., eavesdropper)
Performativity
± performative V/VP
(marker: I feel wronged... most of the expressions are implicit)
the essential content
the speaker tells how he/she was misunderstood, or what caused the misunderstandings
the intentional states
being misunderstood/unable to accept the gap between one's expectation and the reality
the emotional states
background emotion: physical conditions
primary emotion+social emotion:
+ negative (situations running contrary to one's expectation/ original intention)
the occasions
various occasions
(mainly in daily life)
identity/status/objective
Interdependency
± illocution-and-reality interdependency
there is a fact that widly varies from what the speaker expected or a fact that the speaker is misunderstood
± doing-and-talking interdependency
overlapping/parallel and interdependent/ parallel and relevant
± forward-and-backward interdependency
forward-report multiple possibilities
backward-comfort/inquire

Figure 7.13 Analysis results of the ontological properties of the illocutionary act of weiqu.

This token's social situation is the guard room of a teaching building where the speaker converse with the hearers (only one activity type). The speaker is a school cleaner in her forties and the hearers include her sister sitting next to her, her husband, and the present author. The conversation is about the speaker's father trying to drink pesticide to commit suicide because he was suffering from ageing legs and the ensuing inconvenience in daily life, and that his indifferent children were always busy with their own affairs and did not care about him. The speaker's siblings living in their hometown took it for granted that their sister who worked in Shanghai would have better financial conditions and persuaded her to provide more money to their aged parents. But in fact, the speaker herself is under heavy financial burdens, and her economic condition is not as good as her siblings think; and meanwhile, the speaker hopes that they can also take on more responsibility for taking care of their parents, rather than solely depend on the speaker.

1 Emotions

The speaker presents rich facial expressions and head movements, including frowning and swaying her head. Figure 7.14 shows the facial expression and movements respectively at around the 1st, 25th, and

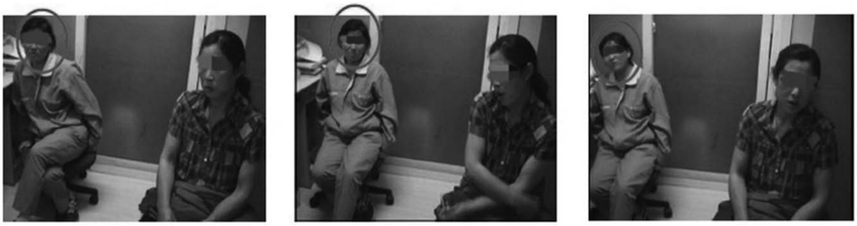


Figure 7.14 Analysis of facial expression.

32nd seconds of this token, all of which last for a long time. These fully deliver a message about her emotional state.

In combination with the discourse content and prosodic features, the different tiers of her emotion are concluded as below:

Primary emotion: **Report – upset; situated – upset**. Because of her father's incident, the speaker feels very upset.

Social emotion: **Report – self-directed, negative, depressed, and helpless; situated – self-directed, negative, depressed, and helpless**. The speaker feels very depressed and helpless at the lack of understanding. Therefore, the social emotion is self-directed negative emotion and the emotion type is depressed and helpless.

Background emotion: **Exhausted**. The reason why the speaker is exhausted is threefold. She has just finished a whole day of outdoor labour, not to mention that she has returned to her hometown to deal with something and rushed back to Shanghai. More importantly, she is still upset about her father's incident. Therefore, she seems very exhausted and discouraged during this conversation.

2 Prosodic features

The statistics of tonal patterns and modes show that there are both low and high pitches. The low ones generally appear when the discourse content is about how indifferent other children are towards their parents and affairs in the hometown, while the high ones appear when the speaker is arguing that working in Shanghai does not ensure her good economic conditions. Here the speaker presents her feeling of *weiqu*.

In addition, the speaker presents extended tone and stresses many times, which mirrors her depressed and helpless feelings underlying the act of *weiqu*.

3 Gestures

Task-doings: The speaker is seated to have a conversation without performing other task-doings throughout the entire process.

Non-verbal acts: Frowning and head shaking. See the analysis above for what they imply.

In short, the above-mentioned multimodal clues show that the speaker does bear the hope that her siblings will take care of their parents in the hometown but now they are always too busy to do so, and misunderstand their sister's economic situation and always ask for money from her. The inconsistency between her expectation and the reality has prompted the speaker's illocutionary act of *weiqu*. Her beliefs, attitudes, and related emotions are consistent with the act of *weiqu*. Therefore, this is a felicitous performance.

7.5 Summary

This chapter conducted the conceptual analysis of 20 types of illocutionary force classified into four classes (neutral, beneficial, harmful, and counterproductive) in the corpus; after collecting all the multimodal cues recorded after the investigation on the 134 tokens, the classified and complied statistics were presented with due consideration to the situation analysis. This was the preparation for further statistical analysis of IFIDs. To specify the method: First, concept modelling was employed to analyse the five illocutionary act types from each class; second, all the multimodal cues were presented in the formula of a set after the investigation on the 134 tokens, which prepared itself for further statistical analysis on IFIDs. Eventually, one or several illocutionary act tokens from the different 20 types of illocutionary forces were analysed in depth by using the data annotated by Elan and Praat to present their respective features.

Notes

[1] The statistical table does not list all the annotated information of the corpus (such as the frequency and duration) and only presents some representative information. Elan-based annotation information was utilized for further analysis of the tokens in a specific situation, and the statistical functions of Elan and other software were used (see further discussion in Chapter 8). In this study, following the idea of simulative modelling, the total saturated significance of simulating and mining the data from different dimensions varied: The concept modelling of a certain illocutionary act type was relatively abstract, with the lowest total saturated significance; the situational analysis was no longer abstract and had a certain degree of total saturated significance; the illocutionary act tokens annotated by Elan enjoyed the highest total saturated significance. However, as mentioned above, there was still considerable data remaining to be mined and presented in spite of our great effort in simulating the situated discourse in real life as much as possible.

8 Dynamic interaction of illocutionary forces in situated discourse

8.1 Statistical analysis of live speech acts and illocutionary force indicating devices

Chapter 7 discusses the fact that in situated discourse, the live illocutionary act is rooted in the whole multimodal interaction between the interlocutors. Such multimodality is fully displayed when the speaker utilizes such resources as speech, prosody, expressions, and gestures to perform illocutionary acts and therefore produce illocutionary force.

The multimodal corpus of this research included various cues and phenomenon, which can be reviewed from the following dimensions:

From the speaker's perspective, all the cues presented in the corpus, as suggested above, are expression devices of performing illocutionary acts. Aside from the frequently used devices, the speech act verbs/phrases or other syntactic structures, word order, intonation, facial expressions, and gestures are all useful devices at the speaker's command to express illocutionary force.

From the hearer's perspective, these multimodal cues help him/her to work out what illocutionary act the speaker is performing and what illocutionary force is intended to be expressed. Since real-life speech is a multimodal interaction, the hearer can make a judgement through his/her perception and interpretation of the speaker's discourse content, phrases and sentences, prosodic features, facial expressions, gestures, etc.

From the researcher's perspective, they are the speaker's multimodal cues of performing illocutionary acts, which has a certain logic of emergence and can be divided into several categories. Different categories imply different research perspectives on illocutionary force, and one perspective can point to one cue or even more. This is the basic thought of simulative modelling adopted in this research, i.e. investigating the tokens in the multimodal corpus from different perspectives. For example, if we want to analyse illocutionary force from the perspective of speech itself, the accessible cues include performative verbs, syntactic structure, word order, and intonation, and if we want to analyse from the perspective of gestures, then the cues could be head movements, hand movements, facial expressions, etc.

DOI: 10.4324/9781003251774-8

Certainly, different cues carry different functions in the performance of illocutionary acts in situated discourse, with some making a difference and some playing an auxiliary role. This is discussed in the following section. This section first conducts a statistical analysis of all the cues in the multimodal corpus constructed in this study.

8.1.1 Techniques of statistical analysis

Two techniques were adopted to carry out the statistical analysis of the data:

The first was to compile Elan and Praat annotation information into an Excel sheet named the 'analysis sheet of four classes of illocutionary act-types' (sheet omitted). In this sheet, 134 illocutionary act tokens divided into 20 types and classified as four classes were recorded one by one according to conceptual analysis and annotation tier in the Discovery Procedure of the situated discourse. Whenever a certain statistic was needed, the search function of Excel could help find its corresponding information. Also, information in this sheet was checked many times to ensure the accuracy of the data. The second is to use the search function in Elan to input multiple searching scope, including all the sample tokens, 4 classes, and 20 types in the corpus.

Elan's search function can support searching and counting according to different statistical requirements when a certain type of annotation is needed. Search results are directly linked to the corresponding segments ready for further analysis after clicking. For example, within the scope of four classes of illocutionary tokens, one sets the searching condition primary emotion as 'sadness' (present the frequency), and then the concordance generated provides data for subsequent statistical research.

8.1.2 Resource system of live speech acts and IFIDs in the *multimodal* corpus

Based on the detailed annotation of the 134 tokens in the corpus, annotation information related to prosodic features and gestures is presented in the formula of a table so that how various prosody and gesture strategies were used by each speaker when performing 134 tokens can be clearly understood.

This statistical method enumerates the speakers' devices in prosody and gestures in the corpus, which on the hearer's side were IFIDs that helped their interpretation and judgement of the illocutionary act-type. It should be noted that due to the limitation of the size of the corpus constructed in this research, certain cues in daily conversations, such as shaking legs and looking around, were not included in the corpus but they might also be some devices employed by the speaker to perform illocutionary acts due to his/her own pragmatic purposes, emotions, and intentions.

Therefore, based on both Austin's observation on the accompaniments of utterance when a speaker performed the illocutionary act and the cues

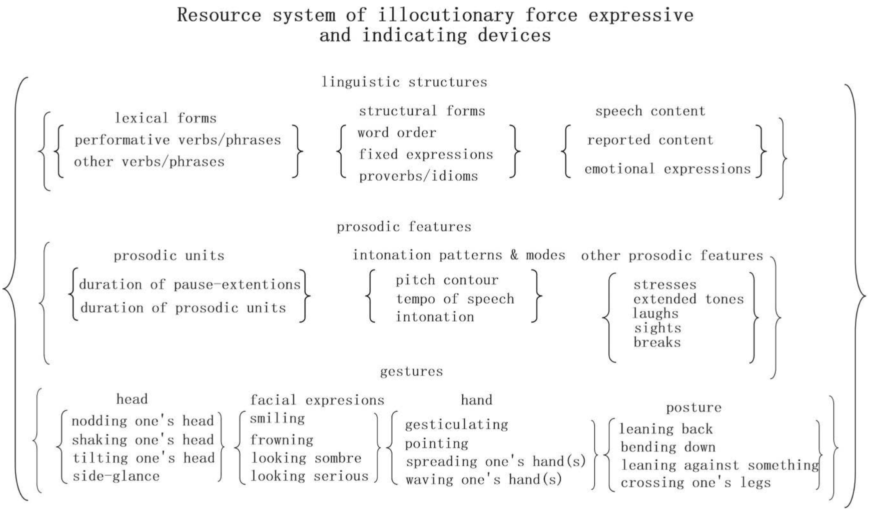


Figure 8.1 Resource system of illocutionary force expressive and indicating devices.

reported by the multimodal corpus, a diagram of the resource system[1] used to perform illocutionary act and express corresponding illocutionary forces in various contexts could be drawn (see Figure 8.1).

The resource system diagram represents the sum of various devices (IFIDs) that people of a certain ethnic group may use to perform illocutionary acts and express illocutionary forces, which is a hierarchical open set.

It can expand in layers and subsets while the subdivision depends on the research perspective and granularity. However, the methodology adopted this research, i.e. stimulative modelling of situated discourse, has a limited modelling perspective (observation perspective). Therefore, the diagram only reflects three aspects for the time being, i.e. language structure, prosodic features, and gestures, while in reality the resources that speakers can utilize are far more than these three types. Similarly, there are many subsets of language structure, prosodic features, and gestures in situated discourse. Most of the subsets enumerated here are what were presented in the corpus.

Moreover, it needs to be pointed out that within the scope of gestures, non-verbal acts turn out to be the device that has an auxiliary effect on the performance of illocutionary acts.

Language structure (including performative verbs and word and sentence forms) and prosodic features used by the speaker in situated discourse are language-specific to a great extent, while many forms in gestures are culture-specific; the functions undertaken by these forms vary across languages and cultures. These are questions about the extent to which these IFIDs have cross-language and cross-cultural universality and which forms are exclusive to a certain language or a certain culture. Their answers require cross-language and cross-cultural comparative research. In this research, the author drew on the concepts of 'linguistic behaviour potential' and 'actual linguistic behaviour' in system functional linguistics to distinguish the IFID system as potential and as choices of behaviour. The former should include all live speech acts and IFIDs at the command of the speaker in all cultural contexts, and the latter is the live speech act and IFIDs actually chosen and observed in a certain cultural context.

The resource system refers to the expression resources that a speaker could select to perform an illocutionary act. In situated discourse, speakers selected certain devices out of the resource system to perform an illocutionary act due to the social situation, pragmatic intention, and other influencing factors (see Section 8.4); Hearers then figured out the illocutionary force by analysing different IFIDs in the related social situation.

Indeed, the resource system was not confined to the individual speaker but was the sum of the expression devices that a certain ethnic group (such as the Chinese Han) could employ from the perspective of social semiotics. From the perspective of the individual speaker, he/she selected certain devices[2] out of the resource system to perform an illocutionary act in a specific context due to the social situation, pragmatic intention, and other influencing factors (such as personal style, expression capability, pragmatic intention, interdependency relation, and social situation), thereby making them the certain IFIDs to perform illocutionary acts under specific contexts.

Furthermore, the resource system devices varied in status and frequency, as some were at the core and used frequently while some were on the fringe and barely used. But its influencing factors were very complicated. This research focused on the actually used devices in situated discourse recorded into a multimodal corpus. They were closely related to the live illocutionary acts performed by a 'live, whole person' who interacted with the environment in a multimodal way. Multimodal cues from different levels interacted and collaborated with each other to jointly reflect the meaning of the utterance and helped the acknowledgement of a certain illocutionary force.

In addition, it should be noted that the various forms in prosodic features and gestures were auxiliary devices to the language structures and discourse content, which collaborated to help speakers perform illocutionary acts and express illocutionary force. And as mentioned above, this research did not examine speech activities carried out only through prosodic features and gestures without substantial discourse content.

8.1.3 Functions of language structure, prosodic features, and gesture in the corpus

It can be seen from Section 8.1.2 that the expression devices presented in live discourse for the speaker to perform illocutionary acts were various. In addition to such frequently used devices as speech act verbs/phrases, syntactic structures, or utterance content in the perspective of language structure, prosody, and gestures also made a difference in illocutionary force performance. In the same vein, the hearers also needed to determine the speaker's illocutionary force via perceiving his/her language structure, utterance content, prosody, and gestures. The author regards live speech acts in situated discourse as the multimodal interaction, which is also the process of cues working collaboratively to help the speaker's performance. The cues in prosodic features and gestures are embedded in the process of a speaker expressing illocutionary forces in a specific context. From the perspective of the pragmatic function of talking being the main constitutive part of the task, these forms presented themselves due to the speaker's choice, bearing the function of helping the speaker perform illocutionary acts and express illocutionary force, and meanwhile becoming the IFIDs that indicated the illocutionary act-types.

So how did various multimodal devices used by the speaker cooperate and interact with one another? What was their contribution in helping the speaker deliver illocutionary force? Different cues carried different functions and offer different evidence for the hearer to interpret and acknowledge these IFIDs. Next the functions of cues such as language structure, prosody features, and gestures are analysed.

8.1.3.1 Language structure

Section 4.2.1.2 refers to the concept of performativity when introducing the octet connotation of conceptual analysis, meaning that for every illocutionary act, there is a performance unit that can be performed performatively or non-performatively. Austin suggested that the standard for distinguishing between performative utterance and non-performative utterance is whether the verb of the illocutionary act is included in the discourse. And in the self-constructed corpus of situated discourse, there are fewer performative utterances and more non-performative utterances. In fact, the syntactic structures that can clearly indicate the illocutionary act-type in situated discourse are not limited to performative verbs/phrases. This section focuses on discussing various indicative language structures.

Regarding the assumption that the performative verb is one of the most essential IFIDs, many speech act theory researchers elaborated on this in their works. Austin also extended the classification of performative verbs to the classification of illocutionary act-types, and though criticized by scholars afterwards, he still managed to grasp the important role of performative verbs in classifying illocutionary act-types, upon which many scholars have agreed. Therefore, performative verbs and phrases are at the core of identifying the illocutionary act-type, thereby forming the central IFIDs. Previous studies paid attention to performative verbs, word order, and other IFIDs in language structure. After a concentrated investigation of a large number of speech act verbs, Wierzbicka (1987: 24) concluded that syntactic structure has its significance in judging the meaning of speech act verbs.

Here we focus on discussing the frequencies of performative verbs and other syntactic structures in the act-tokens from the corpus and their contribution in determining illocutionary act-types.

Although explicit performative verbs or phrases may be absent in situated discourse, some syntactic structures, though not at the predicate verb position and assuming other syntactic functions, often co-occur with a specific type of illocutionary force and can also be regarded as specific IFIDs to a certain extent. For example, a speaker does not use the performative verb *baoyuan* (complain) but uses the syntactic structure of *yuanyan* (complaints) in his/her discourse, which also clearly indicates the illocutionary act-type.

Therefore, explicit speech act verbs/phrases or other syntactic structures are significant devices in language structure. In this study, the author uses AntConc and other corpus tools as well as Elan to search for explicit speech act verbs and explicit syntactic structures in the 20 illocutionary act-types of 134 tokens. Their frequencies of occurrence are shown in Table 8.1.

The frequency of explicit speech act verbs/phrases in the 134 tokens is 0, and that of explicit syntactic structures is 18, with the portion of 0.1343. In view of this, we can see that in the tokens of the multimodal corpus, the speakers in most cases do not use explicit speech act verbs/phrases. As their intention to perform the illocutionary act is not indicated by speech act verbs, the performativity of the majority of tokens is non-performative. This finding is consistent with the statistics of Adolphs (2008: 55) on the use of the speech act verb *suggest*. In most cases, English speakers prefer not to directly use the word suggest when performing the illocutionary act of jianyi. Austin and Searle also discovered that as the sentence form does not directly reflect the type of illocutionary force, most speech acts are implicit (Adolphs, 2008: 27). Wierzbicka (1987: 16) also pointed out that the primary function of speech act verbs is to explain but not to perform the speech acts. In face-to-face interaction, people usually do not resort to speech act verbs to explain their nature.

There are multifold reasons for the speakers' adoption of a non-performative strategy.

First, this is because certain speech act verbs cannot be used in an explicit manner in discourse due to their own semantic nature and semantic logic, as well as usage habits. Gu (2015) referred to such verbs as potential illocutionary act verbs and referred to verbs that can be used explicitly as actual illocutionary act verbs. Descriptive verbs that cannot be explicitly stated because of people's habits include *baoyuan* (complain), *zhiwen* (question), *shandong* (incite), and *chuixu* (boast); implicit verbs that cannot be explicitly stated in semantics and logic are *sahuang* (lie), *chuiniu* (brag), and *anshi* (hint) (Xu, 2008). Searle (1979: 7) also pointed out that not all verbs can be performative verbs, such as *zikua* (brag) and *weixie* (threat). But the fact that these verbs cannot be performative verbs in performing illocutionary acts does not mean that the illocutionary acts (illocutionary force) cannot be named thus, or that there is no such speech act. When the speaker performs

Table 8.1 The frequencies of occurrence

Neutral class						
illocutionary act-type	*jieshi*/explain	*pinglun*/comment	*panduan*/judge	*yaoqiu*/require	*xunuo*/promise	**Total**
performative verbs	0	0	0	0	0	**0**
syntactic structures	0	0	2	0	0	**2**
Beneficial class						
illocutionary act-type	*manyi*/be satisfied	*zanyang*/praise	*cuiqing*/urge	*jiashe*/hypothesize	*yaoqing*/invite	**Total**
performative verbs	0	0	0	0	0	**0**
syntactic structures	0	0	0	2	0	**2**
Harmful class						
illocutionary act-type	*baoyuan*/complain	*suku*/grumble	*piping*/criticize	*danxin*/worry	*haipa*/be scared	**Total**
performative verbs	0	0	0	0	0	**0**
syntactic structures	1	7	0	1	3	**12**
Counterproductive class						
illocutionary act-type	*wunai*/feel distressed	*shiwang*/be disappointed	*weiqu*/feel aggrieved	*yihan*/regret	*jingya*/feel astonished	**Total**
performative verbs	0	0	0	0	0	**0**
syntactic structures	2	0	0	0	0	**2**

the act of *weixie* (threatening), he/she cannot use the performative act verb *weixie* in the utterance and utters 'I threaten you,' but this does not stop the type of illocutionary force being threatening.

Second, this is because some verbs are not performative act verbs but perlocutionary act verbs, including *shangdang* (being deceived), *shengsu* (winning a lawsuit), etc. If the speaker says, 'I was deceived,' he/she is actually performing the act of *chenshu* (stating) rather than *shangdang* (being deceived). The illocution-and-reality interdependency then is this: First, there is an objective incident of the speaker being deceived before the talking; second, the speaker discovers that he/she has been deceived (via certain behaviours or via listening to others' explanations) when he/she says so. Therefore, *shangdang* is a perlocutionary act.

Third, this is because there are self-referential verbs such as *tongzhi* (notice), *qingqiu* (request), *mingling* (command), *daoqian* (apologize), and *zhuhe* (congratulate) that can be expressed normally but are spontaneously omitted by the speaker out of his/her rhetorical or pragmatic intentions. Illocutionary act is related to two levels of rhetoric, namely, act rhetoric and expression rhetoric (Xu, 2008). From the perspective of talking being the main constitutive part of the task, the speaker can choose different illocutionary acts to achieve the same perlocutionary purpose, which is an act rhetoric. For example, to borrow money, the speaker can perform acts of begging, requesting, ordering, or even threatening. After determining the activity type of talking as the main constitutive part of the task, the speaker can make choices between different speech forms, such as whether to use performative verbs, what prosodic features to use, and how to make gestures, thereby forming an expression rhetoric. The speaker aims to choose the most suitable expression in the current situation, which includes what kind of performative verbs are used.

To explain the phenomenon that some verbs fail to be performative verbs, Verschueren (1999: 210–215) distinguishes linguistic action verbs (abbreviated as LAVs) and speech act verbs (i.e. performative verbs, abbreviated as SAVs). LAVs include and are bigger than SAVs in scope. The performance of a speech act requires a series of act conditions and features (referred to as A-condition and abbreviated as Ca) (the author deems that this is what Searle referred to as the act condition and constitutive rules); and to use language to describe speech acts also has its own description conditions or features (referred to as D-conditions and abbreviated as Cd, i.e. the rules for using LAVs). If we want to use a verb to describe a speech act, it involves both Cd and Ca. If Ud (the utterer of D) wants to use a LAV to describe the speech act A of Ua (the utterer of A, A is a felicitous speech act that meets Ca), then this description must meet Cd.

The distance between the Ca and Cd of LAV is called conceptual distance. When the distance is zero, that is, a LAV meets both Ca and Cd, then the LAV is a SAV; if the distance is not zero, that is, a LAV meets the Ca but has distance to the Cd, its performativity is relatively low and cannot be

Table 8.2 List of syntactic structures

Illocutionary act-type	**Syntactic structure**
Panduan	shi … ba (… is), kan xia lai (it seems that)
Jiashe	ru guo (if), … de hua (if …)
Baoyuan	yuan yan (complaint)
Suku	ku (bitter)
Danxin	pa (concerned)
Haipa	pa (concerned), xia (afraid)
Wunai	mei ban fa (there is no way)

counted as a SAV. Conceptual distance can also be divided into the interpretive distance, evaluative distance, and temporal distance. These distances are used to explain why various LAVs cannot meet the Cd, that is, why each LAV cannot become a SAV. For example, for *shuohuang* (lie) and *chuixu* (boast), there is an understanding distance between Ca and Cd because Ua does not want Ia (interpreter of A) to read his/her true intentions.

For instance, the performance of the act of *shuohuang* (lying) includes at least three of the following Ca:

1 Propositional content disagrees with the reality;
2 Ua contradicts the propositional content of the utterance;
3 Ua deliberately misleads Ia.

When Ud wants to use the verb *shuohuang* to describe the speech act A, the following two additional Cd need to be met:

1 Ud's negative judgement of A;
2 Ud's judgement of the truth of proposition in A.

It can be seen that if a LAV does not need to meet the additional Cd and fully expresses the meaning of the speech act, that is, Ca and Cd are consistent to the greatest extent, then the LAV can become a SAV. Therefore, to lie is not a SAV but a LAV. However, the LAV 'to lie' can still be used to describe a certain speech act.

It can be seen that some scholars have made explanations for the phenomenon of the speakers not explicitly using illocutionary act verbs and of certain verbs being unable to be used as illocutionary act verbs. Compared with the rarely occurring explicit performative verbs, some other syntactic structures occur. Table 8.2 shows a list of all the syntactic structures occurring in this corpus.

It should be noted that, unlike speech act verbs, these syntactic structures are not specifically directed towards a certain illocutionary act-type. The same syntactic structure may indicate multiple types of illocutionary force. For example, pa in the table above can indicate either *danxin* (feeling concerned) or *haipa* (afraid); shi … ba can be the IFID for *panduan* (judging) in the corpus, *chenshu* (stating), or *shuoming* (explaining). This is one of the distinctions between syntactic structures and performative verbs such as IFIDs for determining the illocutionary act-type.

Many corpus-based studies of English speech acts also show that certain syntactic structures rather than performative verbs are important evidence for certain illocutionary act-types, by searching the syntactic structure to conduct a statistical analysis of a certain type of speech act (Adolphs, 2008; McAllister, 2015). Sadock (1974) also put forward the concept of speech act idiom, which is used to refer to the expression form that shows the type of implicit speech act. For example, phrases like 'let's' and 'why don't you' are usually used to express the act of *jianyi* (suggesting) in English. Certainly, some scholars criticized this assumption that a certain expression form (or a so-called speech act idiom) is not one-way but points exclusively to a certain type of speech act and carries multiple functions (Levinson, 1983: 269). Adolphs (2008: 50) believed that the key to the appropriateness of the speech act idiom theory lies in the examination of how fixed the expression forms are, and the distribution frequency of their meanings and various influencing factors. The author believes that regardless of certain fixed speech act idioms or not, the speakers employ expression forms other than speech act verbs as a real happening phenomenon. This phenomenon happens frequently in situated discourse on the speakers' side to perform illocutionary acts. As for questions like whether certain language expressions are solidified as the manifestation of specific illocutionary act-type, how they evolve, and how they are used, they involve research in the fields of grammaticalization and discourse analysis.

In short, the reasons for whether and how explicit performative verbs or syntactic structures emerge are complex. And the analysis above just enumerates several common reasons when, in the expression of illocutionary force in situated discourse, the use of performative verbs/phrases and syntactic structures involves not only semantic logic, pragmatic habits, rhetorical intentions, cultural background, discourse context, etc. but also personal style, language type, emotional state, and other factors. Due to space limitations, they are not discussed here.

8.1.3.2 Prosodic features

Both Austin (1962: 73–74) and Searle (1969: 30) pointed out that prosodic features such as intonation, rhythm, and stress are the alternatives to illocutionary acts and are also IFIDs. From the perspective of the principle of the 'live, whole person' introduced previously, prosodic features such as intonation, rhythm, and stress, as components of the 'live, whole person' discourse, are also important means of expressing illocutionary force. Wennerstrom (2001) analysed the contribution of intonation to the expression of illocutionary force. For example, the grammatical form of statement but with an ascending intonation indicates that the speaker is implementing directive rather than representative illocutionary force; or it indicates commissive illocutionary force. Based on all the statistics, this study finds that in addition to intonation, many other prosodic features (including pitch, speech rate and pauses) also contribute to the speaker's expression of illocutionary force. Here is the analysis on some statistical data:

1 Tonal patterns and modes

In this layer, this study mainly focuses on the speaker's pitch, speech rate (tonal pattern), and tonal mode (flat tone, falling tone, or fluctuating tone) when expressing the illocutionary force.

i Pitch

After calculating the pitch of 20 illocutionary act-types in the four classes, the frequencies of low pitch are listed here:

neutral class	0.600
beneficial class	0.538
harmful class	0.457
counterproductive class	0.722

Among them, the frequencies of low pitch in the neutral and counterproductive class are relatively higher.

Low pitch frequently appears in the five types of neutral illocutionary force. This is because the speaker stays in a stable emotional state and presents no dramatic emotional feature when talking about things irrelevant to the interest of the speaker/hearer (such assumption does not exclude special cases in which a certain speaker shows significant emotions as sympathy/empathy arising from the bottom of the heart, i.e. the speaker shows ethnic emotion). Among the act-tokens in the neutral class, the frequency of annotating the speaker's primary emotion as default is 0.509. Since most speakers harbour rather neutral primary and social emotion without displaying drastic emotional features and, more importantly, there are no other situational factors conspiring to influence the speaker (e.g. a noisy background would prompt the speaker to raise the pitch or sound), they generally talk in a rather low pitch.

While in the counterproductive class, the low pitch mainly appears in the tokens of *wunai* (feeling frustrated), *shiwang* (feeling disappointed), and *weiqu* (feeling aggrieved), especially when it comes to the moment when the speaker is talking about the objective fact that runs contrary to his/her own expectation.

For example, in the token 'Father-in-law Li – lunchtime + afternoon – wunai,' the speaker Tang felt helpless about the incapacity at the grassroots level and therefore presented a low pitch when performing the illocutionary act.

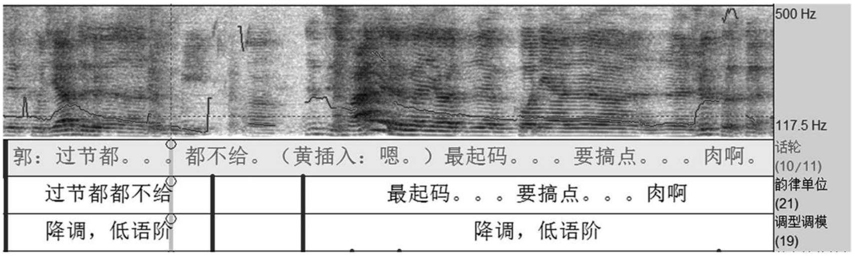


Figure 8.2 Analysis of prosodic features.

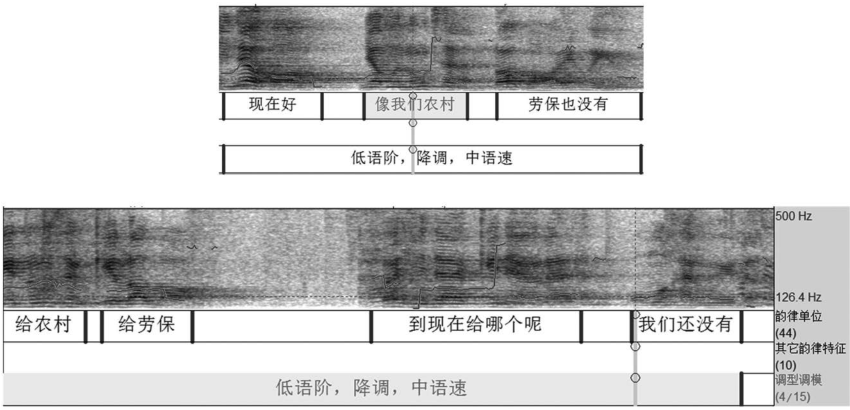


Figure 8.3 Analysis of prosodic features.

In the token '20110912 mid-autumn festival 1 – counterproductive class – shiwang,' the low pitch usually appears when the speaker Guo speaks of the injustice to his son when the canteen he worked in refuses to pay him a salary and provide a bad work lunch (see Figure 8.2). This goes against the speaker's expectation that the canteen should pay for the worker's extra work and provide a better work lunch, so he is rather disappointed about it:

Moreover, in the token 'Father-in-law Li – counterproductive class (no pension) – shiwang,' the speaker Tang is disappointed about the fact he still had no pension and displays a correspondingly low pitch (See Figure 8.3).

Among the tokens of *weiqu*, low pitches usually emerge when the speaker is expressing what makes him/her *weiqu*. The following token is '11y08m26 – talking about a theft– counterproductive class – weiqu' in which the speaker Guo feels *weiqu* about being dumped with work on his day off (See Figure 8.4).

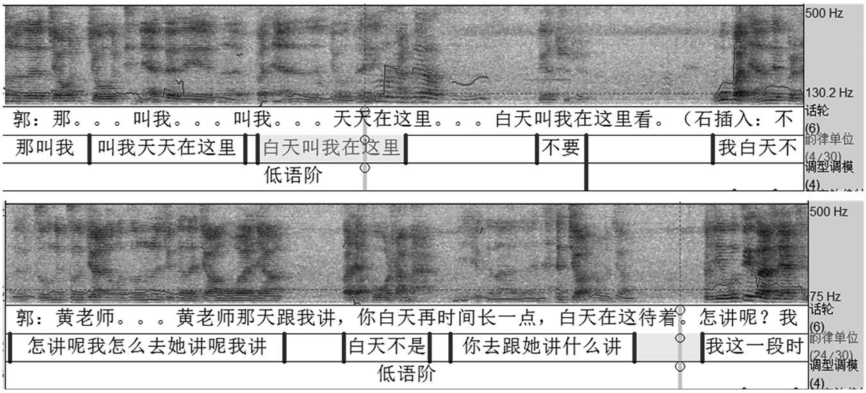


Figure 8.4 Analysis of prosodic features.

Generally speaking, in this class, speakers are usually filled with feelings of helplessness, frustration, and grievance. At this time, they usually speak faintly or in a low pitch.

Next is high pitch. Its frequencies in the four groups are respectively:

neutral: 0.272

beneficial: 0.538

harmful: 0.514

counterproductive: 0.667

Obviously, the frequency of high pitch in the beneficial class, harmful class, and counterproductive class far exceeds that of the neutral class. This is because the speaker is usually emotionally inclined due to the involvement of his/her own interest when performing the illocutionary acts of these three types.

First look at the high pitch of the harmful class. It mainly appears in the tokens of act of *baoyuan* (complaining), *piping* (criticizing), *danxin* (worrying), and *haipa* (feeling scared). This is induced by a speaker's strongly negative emotion, including the primary emotions of anger, concern, and scare, and the social emotions of discontent, contempt, and despising, in which the speaker is prone to present a high pitch to show such emotion.

An example in point is 'Ms Shi – 110912 wood carving, discussing his son's character – baoyuan 2.' In this token, Ms Shi complains about the high school teacher always taking revenge on the students in the class, and then he is indeed agitated and presents high pitch.

Additionally, in the token '20110905 cleaning – harmful class – baoyuan, zhouma:shengqi 2,' the speaker was agitated and strongly negative (evidenced by the accompanying gestures). Filled with anger and contempt about the person's unreasonable behaviour, she thereby presents high pitches during the whole process (see Figure 8.5).

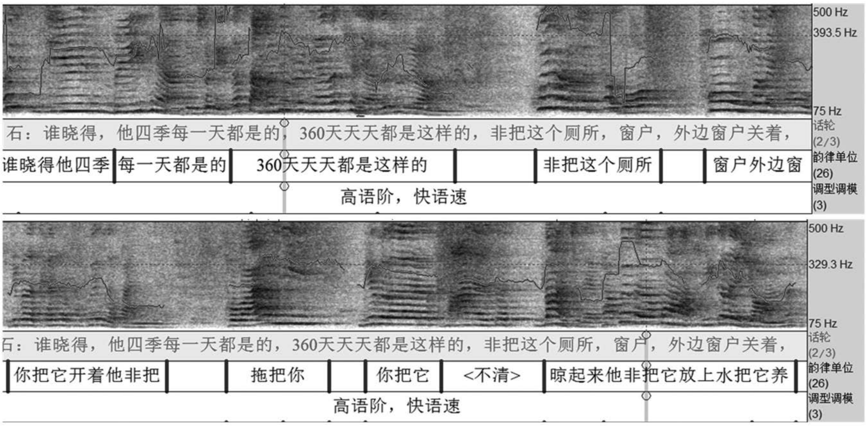


Figure 8.5 Analysis of prosodic features.

In the beneficial class, high pitch mainly emerges in the token of *manyi* (feeling satisfied) and *zanyang* (praising). As the two illocutionary act-types involve distinctively positive emotions, that is, if the speaker *manyi* or *zanyang*, his/her corresponding emotion belongs to the positive type such as happiness. And in this case, the speaker's pitch rises.

ii Speech rate

Among the four classes of illocutionary force, the medium speech rate accounts for the majority (respective values of 0.690, 0.846, 0.514, and 0.611 among the four illocutionary force classes). The frequency of a fast speech rate in the harmful class (0.343) is significantly higher than the other three classes (respective values of 0.038, 0.154, and 0.167). In the harmful class, a fast speech rate is mainly concentrated in the tokens of *baoyuan* (whose frequency is 0.5) when the speaker is usually complaining about the person's nasty behaviours and is emotionally engaged. This is reflected in the fast tempo in prosody.

For example, in the token '20110905 cleaning – harmful class – baoyuan, zhou ma emotion: anger 2,' the segment of performing baoyuan has 132 syllables in total and lasts for 32120 ms, which means the speech rate is 243.33ms/syllable. This can be counted as a fast speech rate according to the parameter of Wu (2001: 398) that 135–300 ms/syllable implies a fast speech rate.

Another token is 'father-in-law Li Shuangfu (before lunch) – harmful group –baoyuan.' There are in total 143 syllables in a duration of 39722 ms when the speaker Tang performs the act of *baoyuan*, which is 277.78 ms/syllable. According to Wu's division (2001: 398), the average time per syllable of fast speech rate ranges from 135 to 300 ms/syllable and that of medium speech rate ranges from 250 to 450 ms/syllable; the speech rate of this token is categorized as medium or fast.

In the token 'Zhang – talk about work – baoyuan,' there are in total 97 syllables in a duration of 20720 ms. The speech rate of 213.60 ms/syllable is typically fast according to Wu's division (2001: 398); the average time per syllable of fast speech rate ranges from 135 to 300 ms/syllable.

The frequencies of slow speech rate are low among the four classes of illocutionary force, 0.072, 0.076, 0.028, and 0.055 respectively. Upon searching all 134 tokens, we find that except for the token 'father-in-law Li Shuangfu (before lunch) – harmful class – baoyuan,' in which the background emotion of the speaker is 'energy and well-being' (the primary emotion is anger), the background emotion of all the remaining speakers is annotated as fatigue or relaxation. This phenomenon shows that the speaker's background emotion has a certain correlation with the occurrence of slow speech rate. Fatigue makes the speaker weak when speaking, while relaxation means that the speaker speaks in an unhurried way.

iii Tonal pattern

Among the four categories, a flat tone occupies the main position, whose emergence frequencies in the four classes are 0.672, 0.462, 0.686, and 0.5 respectively. But the frequency of a falling tone in the beneficial, harmful, and counterproductive classes is basically the same as (or even slightly higher than) the flat tone, that is, the ratio of falling tone/flat tone in these three classes is higher than that in the neutral class. Intonation changes as the emotional state changes. This study believes that the increasing frequency of a medium/falling tone in the beneficial, harmful, and counterproductive classes is mainly because the speaker usually presents more distinctively negative or positive emotions in performing illocutionary acts in these three classes, which cause fluctuations in the intonation.

For example, in the token 'Shandong Wu (work in an empty house) – beneficial class – manyi 2,' speaker Wu expressed his satisfaction with the current living conditions of farmers. In his discourse 'nong min de ri zi shi bi guo qu hao guo le' (farmers' life is better than before), he used a falling tone, which generally means the speaker is firm and endorsing. Here it means the speaker agreed with the improvement of the farmers' quality of life and manifested his satisfaction towards the current living conditions.

A convex tune emerges in the four classes of illocutionary force (but with a small number, nine times in total) and its frequencies of emergence in the four classes are 0.072, 0.076, 0.057, and 0.055 respectively. The pattern of the convex tune's peak being at the body and the trough at the beginning and end is common in statements with stresses or questions with interrogative pronouns (Wu & Zhu, 2001: 334). In the nine tokens in this corpus, stresses appear in the period of convex tune.

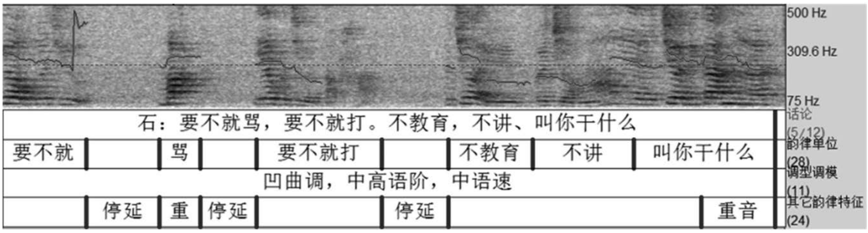


Figure 8.6 Analysis of prosodic features.

There is only one case of concave tune: The token 'Zhang – talking about children's education + about past experiences – s ku 2' of the harmful class. See Figure 8.6.

The pattern that the trough of concave tunes in the body and the peaks at the beginning and end is more common in sentences with sarcasm or sentences with stresses at the beginning and end (Wu & Zhu, 2001: 334). In this token, the utterance is about the speaker's father doing no teaching or educating at all and beating and scolding him when he was a child. Both the stresses at the beginning and end of the sentence and the discourse content of dissatisfaction and irony about his father's problems are the natural outpouring of the speaker's subjective emotion and attitude.

If we look at the three prosodic features above (pitch, speech rate, and tonal mode) as a whole, we can find that the relevant rules and conclusions in the aspects of emotion and prosody drawn by the predecessors still apply to these tokens. For example, the pattern of high pitch and fast tempo usually occurs when a speaker is emotionally excited. It is typical in the tokens of *baoyuan* in the harmful class of illocutionary force in the corpus.

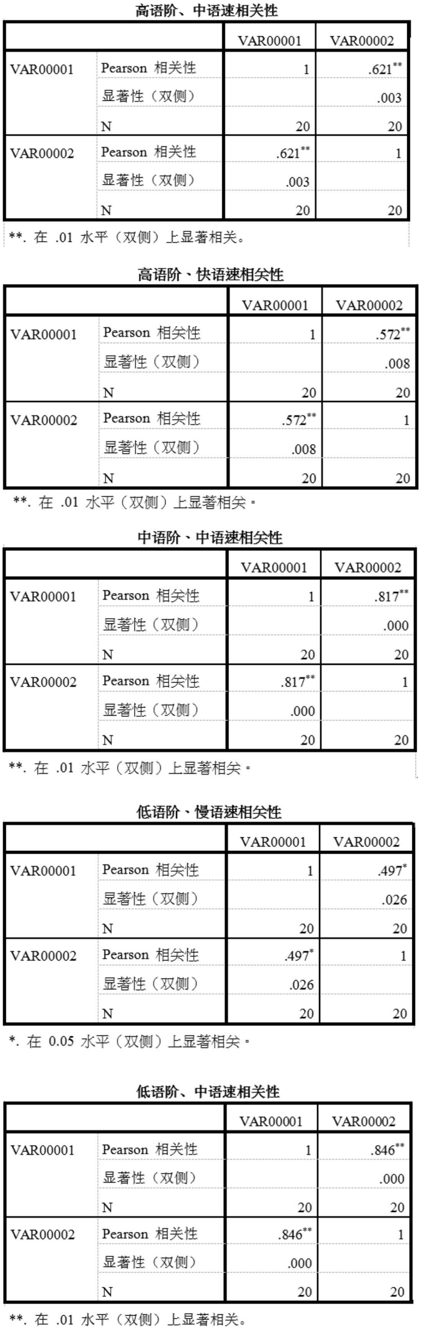
The author adopts the correlation analysis method in statistics to calculate the correlation between pitch and speech rate, and examines whether these correlations are reflected in the tokens of the four classes. More specifically: The method is to input the emergence frequencies of speech rate and pitch of 134 tokens into SPSS and calculate the respective correlation coefficients of the two in all tokens as well as within the four separate classes.

Results from SPSS are shown in Table 8.3.

Therefore, five types of pitch and speech rate are highly correlated among the 134 tokens (correspondingly 20 types) of the four classes, and these two-dimensional tonal patterns have the following communicative functions (Wu & Zhu, 2001: 404–409):

1 High pitch and medium speech rate: This two-dimensional tonal mode indicates that the speaker is impassioned and faces a large crowd as the audience. The tokens of the beneficial illocutionary force in the multimodal corpus are typical.

Table 8.3 SPSS calculation



高语阶、中语速相关性

		VAR00001	VAR00002
VAR00001	Pearson 相关性	1	.621**
	显著性（双侧）		.003
	N	20	20
VAR00002	Pearson 相关性	.621**	1
	显著性（双侧）	.003	
	N	20	20

**. 在 .01 水平（双侧）上显著相关。

高语阶、快语速相关性

		VAR00001	VAR00002
VAR00001	Pearson 相关性	1	.572**
	显著性（双侧）		.008
	N	20	20
VAR00002	Pearson 相关性	.572**	1
	显著性（双侧）	.008	
	N	20	20

**. 在 .01 水平（双侧）上显著相关。

中语阶、中语速相关性

		VAR00001	VAR00002
VAR00001	Pearson 相关性	1	.817**
	显著性（双侧）		.000
	N	20	20
VAR00002	Pearson 相关性	.817**	1
	显著性（双侧）	.000	
	N	20	20

**. 在 .01 水平（双侧）上显著相关。

低语阶、慢语速相关性

		VAR00001	VAR00002
VAR00001	Pearson 相关性	1	.497*
	显著性（双侧）		.026
	N	20	20
VAR00002	Pearson 相关性	.497*	1
	显著性（双侧）	.026	
	N	20	20

*. 在 0.05 水平（双侧）上显著相关。

低语阶、中语速相关性

		VAR00001	VAR00002
VAR00001	Pearson 相关性	1	.846**
	显著性（双侧）		.000
	N	20	20
VAR00002	Pearson 相关性	.846**	1
	显著性（双侧）	.000	
	N	20	20

**. 在 .01 水平（双侧）上显著相关。

2 High pitch and fast speech rate: The speaker is excited and anxious, particularly in the harmful class in the multimodal corpus.
3 Medium pitch and medium speech rate: The most frequently used tonal mode (in real-world scenarios and also in the self-constructed corpus in this study). It is often used in reports, lectures, and formal conversations in work and life.
4 Low pitch and slow speech rate: This tonal mode is associated with the speaker's deep and sincere emotions due to poor physical condition or fatigue. In this corpus, the background emotion when the speaker talks at a slow tempo (eight tokens in total) is basically relaxation or fatigue (occupying 0.75).
5 Low pitch and medium speech rate: This tonal mode is beyond the discussion of Wu and Zhu (2001). As mentioned above, a medium speech rate is the most frequent in this corpus and also the most common in daily conversation, while a low pitch mainly appears in the neutral and counterproductive classes.

2 Prosodic unit

As indicated above, the prosodic units in this study are mainly divided by pauses. Elan can calculate the average duration of pauses in multiple tokens. Among the four classes, the average duration of pauses in the neutral class is the longest (reaching 1.056s), among which, that of *jieshi* and *panduan* are the longest (reaching 1.138s and 1.129s respectively). This is in line with the specific conditions of the tokens of *jieshi* and *panduan* in the multimodal corpus, that is, there are accompanied task-doings when the speaker expressed the illocutionary force. More specific, the speaker was usually operating or fiddling with something while *jieshi* (explaining) or *panduan* (making judgements). From the perspective of interdependency, apart from overlapping, the doing-and-talking interdependency is mainly parallel and relevant (another case is talking and doing running in a conflicting parallel, e.g. the speaker is talking while eating).

For example, in the token 'Speaker Guo – 110912 – wood carving, discussing his son's character – panduan,' the speaker uses a knife to cut open the object in his hand to check the inner core and meanwhile judges whether the object is made from wood. There is a long pause during the entire process, for the speaker paused to attentively gaze at the object in his hand (see the picture below). So the doing-and-talking interdependency is parallel and related, that is, the speaker cuts the object in his hand to view the inner core and perform the act of *panduan*.

Among the tokens of the beneficial, harmful, and counterproductive classes, *cuiqing* (urging), *suku* (complaining), and *shiwang* (feeling disappointed) are the three types with relatively longer average duration of pause, with a duration of 1.656s, 1.099s, and 1.668s respectively.

In the beneficial class of illocutionary force, the pauses among five act-types vary, among which that of *cuiqing* is the longest (1.656s). Contrasting with the connection between Elan annotations, we can find that the

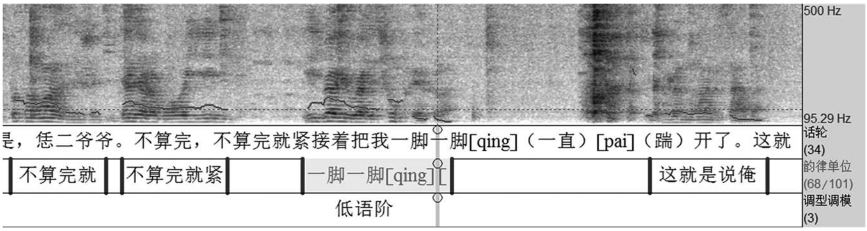


Figure 8.7 Analysis of prosodic features.

occurrence of a longer pause is because the speaker is performing the illocutionary act as well as other task-doings at the same time, such as picking a pomegranate and cutting a watermelon. For example, in the token 'eating at Guo's residence 1 + 2 eating pomegranate – cuiqing 2' and 'Zhang shopping at vegetable market, talking in the dormitory – cuiqing,' the talking is cut short abruptly because the speaker has other task-doings.

The average duration of pauses in the harmful class of illocutionary force is 0.879s. In this group, the average duration of pause of the four illocutionary act-types, i.e. *baoyuan*, *piping*, *haipa*, and *danxin*, ranging from 0.5s to 0.7s, is relatively shorter; and that of *suku* lasting 1.099s exceeds those of four illocutionary act-types. Looking closer at the tokens of *suku*, the speakers are usually found to harbour prominently negative emotions, including the primary emotion of sadness and the social emotion of self-compassion. Frequent gestures include wiping tears and shaking his head. Generally speaking, the speaker is always depressed and frequently pauses for a long time, and sometimes even chokes badly or fails to resume talking; while in the other four types, the speaker usually speaks fast and anxiously with the accompanying anger and fear, resulting in a shorter duration compared with that of the act of *suku*.

For instance, in the token 'Shangdong Wu – harmful class – suku,' the speaker Wu recalls all his sufferings in front of the hearer, and when speaking of the part where he feels so miserable, he presents a long occurrence of pause and choking as well as non-verbal acts like weeping. The pause here is due to his dramatic sadness that stops him from talking.

Wu and Lü (2011) put forward the concept 'sound stops while emotion continues' in the discussion of intonation; that is, though the speaker makes no sound during the speech, his/her emotion is continuous and affects the before-and-after prosody. For example, the deep intonation usually features a low-and-narrow tonal range with long duration and ultra-low pitch, which sometimes may even reach the extreme of the speaker being speechless. The token 'Shandong Wu – harmful class –suku' is a case in point, in which the speaker stops speaking for a while because of the forces of sadness and anger. And the whole interval is long, about 404 s; pitch low (about 150 Hz), and the tuning range narrow. At this point, although the speaker pauses speaking, his emotions linger and continue to project into gestures (such as wiping tears and bowing his head). See Figure 8.7.

In brief, there are two main situations of pausing for a long time in this corpus: The speaker is performing other related task-doings at the same time, or is interrupted when talking due to some significant emotional states (e.g. becoming choked when immersed in sadness).

3 Other prosodic features

The frequencies of stress in the harmful, beneficial, and counterproductive classes are basically the same: 0.962, 0.971, and 0.944 respectively. This is slightly lower in the neutral class, at a value of 0.872. Overall, the presence of stress is not significantly correlated to the distinction of illocutionary act-types. Certainly, when it comes to the analysis of each token, it is not difficult to find that the presence of stress is of great significance to the expression of illocutionary force. Stress is a coding method that helps speakers efficiently express their intentions, emotions, or attitudes and promote hearers' accurate and quick perception of these (Zhang, 2014: 122). Therefore, the more that a speaker expresses his/her subjective attitudes and emotions, the more likely stress will present in performing the illocutionary force.

The frequencies of laughter in the neutral, beneficial, and counterproductive classes are basically the same, at values of 0.218, 0.231, and 0.222 respectively, while that in the harmful class (a value of 0.143) is relatively lower, for, then the speaker usually harbours negative emotion when performing harmful illocutionary acts. In the corpus, there is laughter only in five tokens, one in the token of *baoyuan* and four in the token of *suku*. In the token '110914 discussing son's character – harmful class – baoyuan,' the speaker Guo complains about the restaurant where his son works and is unpaid for working overtime during national holidays and the food provided is poor, and he then says, 'The day you stop paying for overtime is the day I quit this job.' Laughter at this time shows his dissatisfaction and helpless ridicule. When performing the illocutionary act of *suku*, laughter usually appears after the speaker talks about a harsh experience, which is actually a bitter smile due to feeling helpless or a smile of relief.

Furthermore, among the tokens of the counterproductive illocutionary acts, prosodic features such as sighing, voice weakening, and pause are consistent with helplessness, disappointment, frustration, grievance, and depression when the speaker expresses the illocutionary force.

Other features such as coughing, wheezing, clearing one's throat, raising one's throat, and repetition are not universal but individual behaviours under specific situations, so they are omitted here.

8.1.3.3 Gestures

Some of the presented gestures are universal in the scope of gestures but not well classified and indicative, and some are frequently present in a certain class.

Head movements feature a host of forms and scattered occurrence. Swaying one's head, shaking one's head, head up, head down, looking at the other person, looking to the side, and looking down all occur in the four classes of illocutionary act-token; nodding is only absent in the counterproductive class; the gesture of thinking does not emerge in the neutral class; staring at something only emerges in the neutral class (0.218) and the beneficial class (0.230); leaning one's head back, tilting one's head, and looking upwards only occur in the beneficial class, and the frequency is only 1 or 2. The author did not find that the other head movements have a significant correlation with different classes and types of illocutionary forces.

Among various forms, smiling/laughing mainly occurs in the neutral and the beneficial classes, with the frequencies of 0.509 and 0.654 respectively. This is because in these two classes, the speaker's emotion is usually positive, which is easily expressed through smiling/laughing; while having a long face, frowning, wry smile, and rolling the eyes only emerge in the harmful and the counterproductive classes that the speaker's emotion is usually negative. Among them, frowning also appears in the tokens of *manyi* (feeling satisfied) and *jiashe* (hypothesizing) in the beneficial class. For example, in the token 'Zhou from Tonglu – beneficial class – manyi status quo,' the speaker compares her hard life in the past with the happy life now to highlight her contentment with her current life. Her frown appears in the utterance zao xie qiong o (so poor then).

In the token 'Zhou from Tonglu – beneficial class – jiashe 2,' the speaker Zhou assumes that her life would be very different if she had left her hometown to live in Beijing during her childhood. Therefore, her situated social emotion is self-directed negative and regret, that is, she regrets not following others to Beijing and presents frowning under such an emotional state.

Hand movements frequently emerge in the four classes of illocutionary force (neutral class: 1.327, beneficial class: 1.154, harmful class: 1.057, and counterproductive class: 1.500). The frequencies are all greater than one because the gestures within the range of hand movements are not exclusive but overlapped when they appear. For example, the speaker presents several hand movements such as shaking, pointing, and gesturing. This also evidences that hand movements are most frequently used by speakers when expressing illocutionary force. On the whole, there is no specific hand movement that has significant specificity in a certain type of illocutionary force, and the speaker presents them in multimodal interaction according to the situation, speech acts, or their personal style. However, the hand movements of 'fiddling, holding, placing, and selecting (items)' is significantly more frequent in the neutral class than in the other three classes (the frequency in the neutral class: 0.218, those in the other three classes: 0.115, 0.057, and 0.167). This is mainly because in the neutral class of illocutionary force, the speaker usually fiddles with the objects in hand and performs the corresponding illocutionary act, such as *jieshi* and *panduan* and the activity type directly affects the gestures.

A host of gestures in the corpus include leaning back, leaning forward and back, and sitting down and up, leaning against a wall, sitting down, bending over, and crossing one's legs/raising one's leg. As they are not exclusive to an illocutionary act-type, it is necessary to analyse them in combination with the speaker's individual factors and specific context.

8.2 Analysis of emotions, prosodic features, and gestures[3]

Prosodic features and gestures in helping speakers express illocutionary force were analysed in the previous section. However, it should be noted that each IFID of prosodic features and gestures may only appear once or multiple times in each token and as often is the case, IFIDs may co-occur. For example, when the speaker *baoyuan* or *piping*, he/she is likely to present high pitch, fast tempo, stress sounds, and non-verbal acts like frowning and swaying the head at the same time. These IFIDs collaborate to help the speaker's performance and the hearer's perception; however, in some cases (e.g. performing the act of *xunuo* or *yaoqiu*), besides certain discourse content and language structures, the speaker only employs a medium speech rate and medium pitch in prosody without any other non-verbal acts, still succeeding in articulating the corresponding illocutionary force to the hearer. In this corpus, in terms of non-verbal acts, frequently occurring ones in other illocutionary act-types are not found in the tokens of performing the act of *yaoqiu* (except for nodding twice and smiled/laughed twice) and the act of *xunuo* (except for one scratching and one folding hands).

In situated discourse, it is more often that multiple prosodic features and gestures appear together when the speaker expresses the illocutionary force. This involves how different IFIDs collaborate and interact.

Two items are necessary for analysing the relations between IFIDs:

1 Analysing the frequencies of emotion, prosodic features, and gestures in the four classes of illocutionary force;
2 Analysing the patterns of emergence and concurrence of emotions and multimodal cues in the four groups of illocutionary force.

When studying the relation between emotional state and IFIDs in prosody and gestures, this research adopts the methodology of inference studies (Scherer, 2003: 237–238). That is, in observable multimodal cues in prosodic features, gestures are measured and extracted in the first place, and then they are linked with the speaker's emotion and then by the annotator (verified by experts) to infer the correlation between emotions and these multimodal cues by statistical methods.

More specifically, uploading the frequencies of primary emotion, background emotion, social emotion, prosodic features, and gestures of four classes of illocutionary force into Minitab 17. Factor Analysis, a statistical method of inference, was adopted to explore the relationship between the speaker's emotions and multimodal cues as well as their influence on the performance; by interpreting the related parameter diagram like load graph and score graph generated by Minitab, the correlation between emotion, prosody, and gestures emerging in five illocutionary act-types in the same class of illocutionary force is investigated.

It needs to be pointed out that the predecessors' study of the relationship between emotion and prosody (particularly intonation) has been already discussed. This section does not use experimental methods to analyse the prosodic features when the speaker harbours a certain emotion but analyses whether there is a correlation between emotion and prosody when the speaker expresses the illocutionary force in the tokens recorded in the corpus, as well as whether such correlation can be explained through the existing findings. Similarly, the analysis of the correlation between emotion and gestures is also carried out via investigating the tokens in the self-constructed corpus.

8.2.1 Neutral class of illocutionary force

8.2.1.1 Distribution of the three layers of the neutral class

(1) Emotional state – primary emotion

The statistics show that in the neutral class, the most frequent case is that the speaker shows no distinctive primary emotion (28 times); followed by happiness (20 times); while in contrast, worry (5 times) and anger (2 times) are the least frequent.

The speaker is more inclined to show no distinctive primary emotion and second, inclined to show happiness in the neutral class. Since in this group the essential content is not directly related to the conversation interlocutors' interest, the speaker generally does not have a significant emotional tendency when performing the illocutionary act; certainly, as described in Section 3.4.2, a human being as a social animal will sympathize with another's circumstance and thus him/herself produce a corresponding emotional state. This is one of the reasons why the speaker also has the primary emotion of happiness, worry, or anger when expressing the neutral illocutionary force.

(2) Emotional state – background emotion

The statistics show that relaxation (31 times) is the background emotion type with the highest frequency in the neutral class, followed by energy or well-being (15 times), and finally fatigue (9 times), indicating that in most cases, speakers are relaxed or full of energy when performing the act of *jieshi*, *pinglun*, *panduan*, *yaoqiu*, and *xunuo*.

(3) Emotional state – social emotion

Three most frequent social emotions in the neutral class are positive, enthusiastic/positive/passionate (23 times), positive, confident (18 times), and neutral, indifferent/default (12 times). As the classification and annotation

Table 8.4 Five types of illocutionary acts in the neutral class

	Token number	*Other-directed*	*Self-directed*	*Neutral/default*
jieshi	19	19	17	0
pinglun	15	6	2	8
panduan	8	3	5	1
yaoqiu	6	4	2	1
xunuo	5	1	3	1

of social emotion may overlap (see the illustration and annotation scheme on social emotion in Sections 4.2.3 and 6.1.4) and does not meet the basic conditions of the chi-square test in statistics, e.g. the speaker has other-directed and self-directed social emotions, so here we simply use the method of sorting. Two social emotions, i.e. positive, enthusiastic/positive/passionate and positive, confident are mainly concentrated in the tokens of *jieshi*, which reflects that when the speaker *jieshi* to the hearer, he/she is self-directed confident and other-directed enthusiastic, positive, and passionate; and 'neutral, indifferent/default' mainly occurs in the tokens of *pinglun*, indicating that the speaker has a neutral and unbiased attitude when performing the act of *pinglun*. This is also because the interlocutors' interest is not directly involved; other positive and negative social emotions are scattered among the five illocutionary act-types with no special regular pattern, which also underline the importance of analysing them in their context.

From the perspective of the direction of social emotion, the statistic of five illocutionary act-types in the neutral class are shown in Table 8.4.

The statistics above indicate that the social emotion of the speaker performing the illocutionary act of *jieshi* is both other-directed and self-directed. Such two-way emotional direction is because a felicitous act of *jieshi* requires the speaker to be enthusiastic or passionate towards others and meanwhile being confident about him/herself to articulate things to others; and in the neutral class, as the essential content of *pinglun* is interest-neutral to the interlocutors, the speaker tends to be indifferent. So the proportion of being neutral/default is higher; meanwhile, the emotion is always other-directed as *pinglun* is expressing opinions on other people or things; the act of *panduan* is the speaker assessing other people or things, so mostly it is self-directed emotion; the emotion of *yaoqiu* is mostly self-directed; and the act of *xunuo* is the speaker expressing some constraint on him/herself, so the emotion is mostly self-directed.

8.2.1.2 Correlation of emotional state and prosody and gestures in the neutral class

We input the frequencies of each emotion when the speakers express the neutral illocutionary force into Minitab and the frequencies of cues like

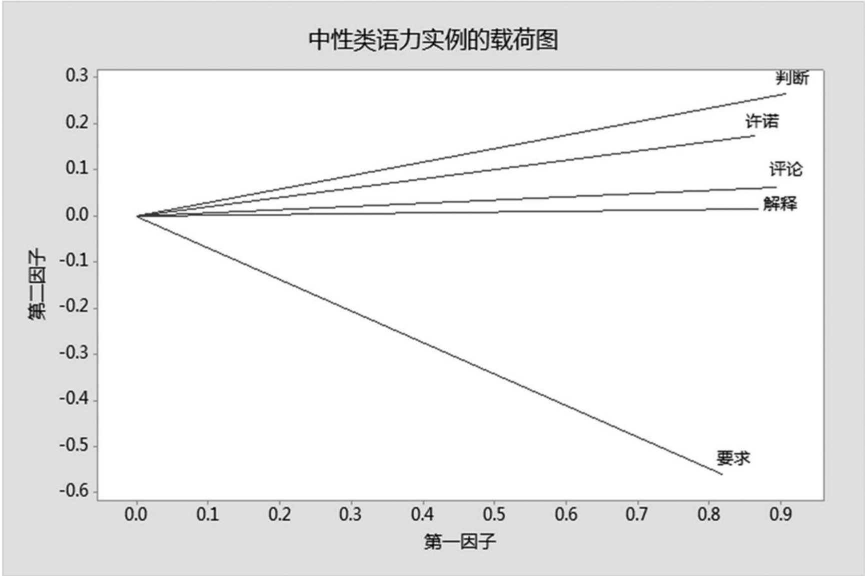


Figure 8.8 Load diagram from Minitab.

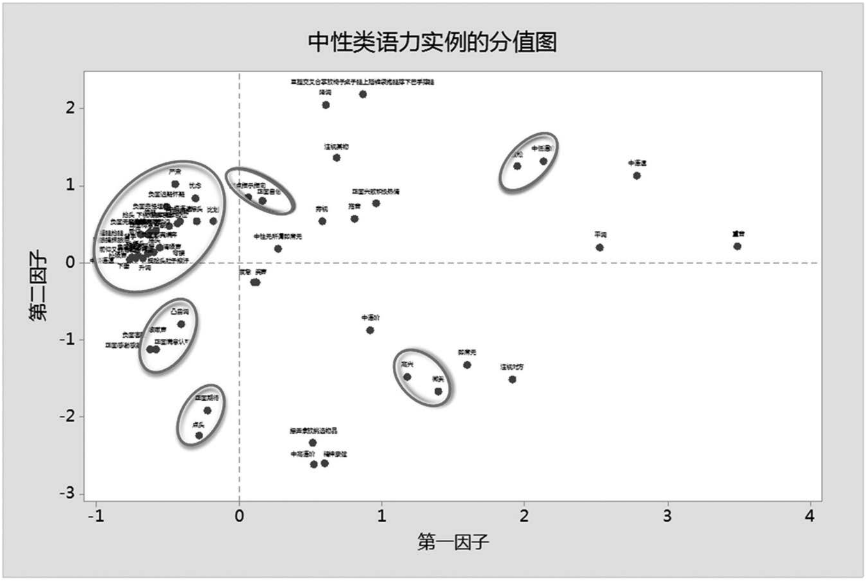
prosodic features and gestures. We use the factor analysis function to output the load diagram and distribution diagram:

As we can see from Figure 8.8, among the tokens of neutral illocutionary force, *panduan*, *xunuo*, *pinglun*, and *jieshi* belong to the same category (located in the first quadrant), that is, the speaker's emotion, prosody, and gestures are relatively changing in the same direction; while *yaoqiu* differs from those four illocutionary act-types.

Relevant multimodal cues can be analysed from this figure (Figure 8.9).

Let us look at the first quadrant. Combined with the previous load diagram, the four illocutionary act-types, i.e. *panduan*, *xunuo*, *pinglun*, and *jieshi*, all appear in the first quadrant, indicating that when the speaker expresses these four illocutionary forces, the value points in the same direction as the lines are closely related. More specifically, prosodic features such as flat tone and stress sound will emerge frequently when the speaker performs the illocutionary act of *jieshi* and *pinglun*; when the speaker *panduan* or *xunuo*, medium to low pitch and medium speech rate frequently appear. And meanwhile, relaxation is the correlated background emotion. In terms of social emotion, positive and confident social emotion are closely correlated to pointing with hands, instructing, and gesturing. Such correlation is particularly obvious in the illocutionary act of *jieshi*.

Now the third quadrant. The social emotion of positive, grateful/appreciative, positive, satisfactory/endorsing, and negative, shy are correlated to



Result	Factor 1	Factor 2
Variance	3.7896	0.4158
Variance contribution rate	0.758	0.083
Cumulative contribution rate	0.758	0.841

Figure 8.9 Distribution diagram from Minitab.

* The cumulative contribution rate of the first and the second factors is 84.1%, which can represent the overall sample situation.

convex tunes, which means that the alteration of positive or negative social emotion will cause changes in the intonation. Among the four illocutionary act-types of *panduan*, *xunuo*, *pinglun*, and *jieshi*, intonation is an important clue of social emotion when the speaker acts. In addition, the social emotion of positive and hopeful is also correlated to nodding.

Combining the load and the distribution diagrams, we can see that the second quadrant belongs to the scope of the illocutionary act-type of *yaoqiu*, but many multimodal clues do not appear in the tokens of *yaoqiu* in this corpus. This may be a biased error caused by the small-sized statistical sample. However, the correlation between some emotions and multimodal cues has certain universality, such as contemptuous/despised, helpless/disappointed/frustrated/depressed in negative social emotion and (medium) slow speech rate, fast speech rate, light sound/weakening sound, and intermittent sound are highly correlated, which is in line with the general law of intonation. Furthermore, they may be the expression devices in prosody when the speaker has the corresponding emotion.

The fourth quadrant shows that the primary emotion of happy is related to smiling in gesture; and the background emotion energy/well-being is related to the medium and high pitch, that is, the speaker generally presents a high pitch when full of energy or in a good physical state.

In addition, positive, negative, and neutral social emotions may all emerge in the neutral class. Pointing/instructing/gesturing, swinging (swaying), leaning back the hand, dropping it, and other hand movements are gestures that frequently appear together with positive social emotions; having a long face (being serious) is highly correlated to negative social emotion; other highly correlated gestures cannot be explained in their universal meaning.

8.2.1.3 Summary

In the neutral class, the most frequent type of emotion is that the speaker does not show significant primary emotion, followed by the primary emotion of happiness; while in contrast, worry and anger are less frequent; the top three social emotions are: positive, enthusiastic/positive/passionate, positive, confident, and neutral, indifferent/default.

On the whole, prosodic features do not reflect significant specific tendencies, low pitch, middle pitch, high pitch, flat tone, and stresses all appear; other prosodic features such as lowering sound, weakening sound, and intermittent sound are all correlated with the negative social emotions; laughter is not universal because it is specific to the occasion.

The speaker's hand movements such as poking/instructing/pointing, swaying, putting hands back, and lowering them are relatively more frequent than other types.

***8.2.2 Beneficial* class of illocutionary force**

8.2.2.1 Distribution of the three layers of emotional state in the beneficial class

(1) Emotional state – primary emotion

Statistics indicates that in the beneficial class, there are two cases of the primary emotion, i.e. happy whose frequency is the highest (21 times) and default (the speaker does not show significant primary emotion) whose frequency is fewer (five times). There is no other primary emotion. This is related to the definition of the beneficial illocutionary force. Since its essential content is positive or it has a positive impact on the interest of the speaker/hearer, the speaker generally shows a positive emotional inclination. This explains why speakers generally do not have negative primary emotions, such

Table 8.5 Five types of illocutionary acts in the beneficial class

	Token number	*Other-directed*	*Self-directed*	*Neutral/ default*
manyi	10	4	6	1
zanyang	7	6	2	0
jiashe	3	0	3	0
cuiqing	5	5	0	0
yaoqing	1	1	0	0

as worry, anger, and fear. And it is justifiable that sometimes the speaker maintains relatively calm although his/her interest is positively related.

(2) Emotional state – background emotion

Statistics tell us that in the beneficial class, energy/well-being has the highest frequency (12 times), followed by relaxation (9 times), fatigue (3 times), and discouragement (2 times). This is related to the speaker's physical state when expressing the illocutionary force (such as after a long time of work or labour).

(3) Emotional state – social emotion

Among the tokens of beneficial illocutionary force in this corpus, social emotions are mainly positive, including satisfactory/endorsing, enthusiastic/positive/passionate, proud/conceited, expectation, grateful/appreciative, respectful/admiring, and praising. They total 28 times, accounting for 84% of the social emotions in the beneficial class; the social emotion of 'neutral, indifferent' emerges once; negative (shy, contemptuous/despising, regretful) three times. This says that when performing the illocutionary act that is positively related to his/her own interest, the speaker usually has positive social emotion. The emergence of some negative social emotions is also the accompaniment of the speaker's positive emotion.

From the perspective of the directivity of social emotion, the statistics of five illocutionary act-types in the beneficial class is shown in Table 8.5.

The statistics above say that when the speaker expresses the illocutionary force of *manyi* (feeling satisfied), it can be either *manyi* other people or other things, or *manyi* him/herself, so the illocutionary force can be other-directed and self-directed; as the illocutionary force of *zanyang* (praising) is mainly aimed at others, it is other-directed; the illocutionary force of *jiashe* (hypothesizing) means the speaker has interest-positive assumptions about the past or the future, and in this corpus, the speaker generally harbours self-directed regret or expectation; *cuiqing* (urging) and *yaoqing* (inviting) is towards others so they are other-directed.

8.2.2.2 Relationship between emotional state and prosodic features and gestures in the beneficial class

We input the frequencies of various emotional states as well as those of prosodic features and gestures presented by the speakers expressing the

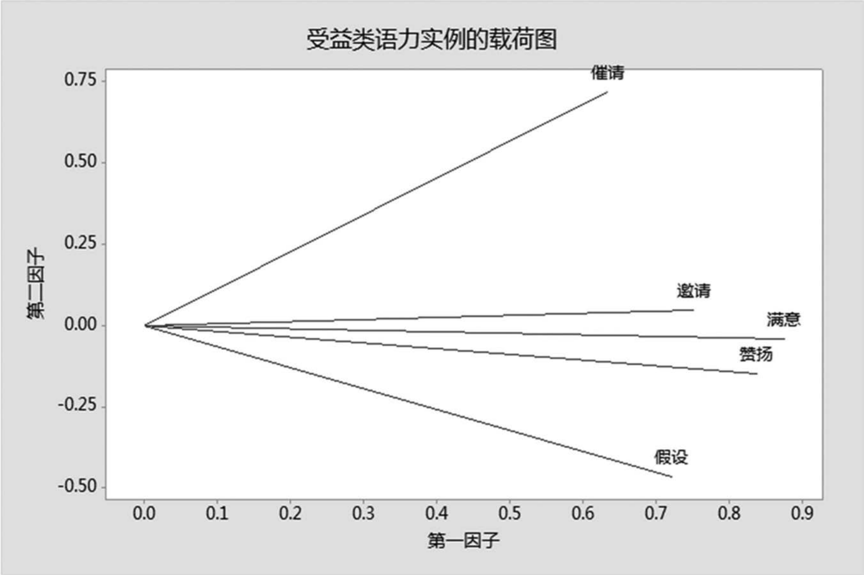


Figure 8.10 Load diagram from Minitab.

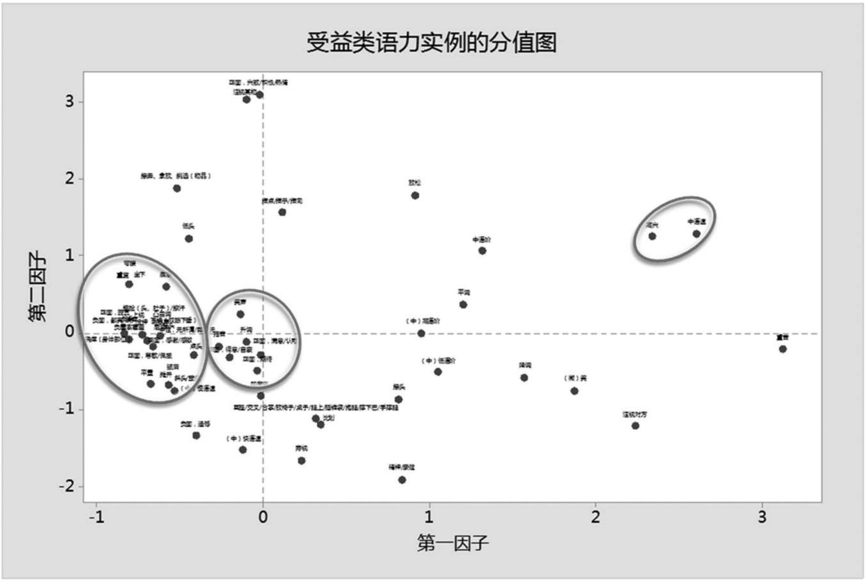
beneficial illocutionary force in the corpus to Minitab, and use the factor analysis function to output the load diagram and distribution diagram.

As we can see from Figure 8.10, among the tokens of beneficial illocutionary force, *cuiqing* and *yaoqing* belong to the same category (located in the first quadrant), that is, the speaker's emotion, prosody, and gestures are relatively changing in the same direction; while *manyi*, *zanyang*, and *jiashe* differ from these two illocutionary act-types (Figure 8.11).

As mentioned above, in the beneficial class, the social emotion significantly correlated to prosody is mainly positive and neutral or indifferent, including grateful/appreciative, conceited/proud, respectful/admiring, praising, and satisfied/endorsing. Two negative emotions, shy and contemptuous/despised appeared in '20110912 mid-autumn festival 1 – beneficial class – zanyang' and 'Bo from Tonglu – beneficial class – zanyang,' but they accompany the speaker's other-directed, positive, and thankful social emotions and self-directed, positive, conceited, and proud social emotions.

Happiness and relaxation are the type of primary emotion and background emotion that closely correlate with *cuiqing* and *yaoqing*.

Energy/well-being is the main type of social emotion. It is significantly correlated to (medium) high pitch, which implies that the speaker's well-being has a tangible influence on his/her prosodic features; positive, enthusiastic/positive/passionate is correlated with the speaker's gesture of staring at a certain object.



Result	Factor 1	Factor 2
Variance	2.9546	0.7564
Variance contribution rate	0.591	0.151
Cumulative contribution rate	0.591	0.742

Figure 8.11 Distribution diagram from Minitab.

* The cumulative contribution rate of the first factor and the second factor is 74.2%, which can represent the overall sample situation.

Medium speech rate, flat tone, medium pitch, and (middle) high pitch are the prosodic features of *cuiqing* and *yaoqing*; stress, laughter, etc. are also common prosodic features; other tonal modes (rising tone, falling tone, flat tone) and breath sound, extended tone, etc., although correlated, do not have typical meaning here.

On studying the adjacent dots in the red circle on the left corner of Figure 8.20: In the beneficial class, positive social emotion (including conceited, proud, respectful, admiring, hopeful, satisfactory, endorsing, grateful/appreciative, and praising) is the main type of emotion significantly correlated to gestures; combined with the analysis of specific tokens, negative emotions (shy, despised, regretful) or neutral, indifferent/default are the accompaniment of the positive social emotions. For example, in the token 'Bo from Tonglu – beneficial class – zanyang,' the speaker praises the rapid development in her hometown Tonglu in recent years, with self-directed, positive, conceited, and proud social emotion. At the same time, she also makes a comparison with other cities that have developed slowly in recent years, so she has other-directed, negative, and despising social emotions. This can be inferred from her expressions and shaking her head; except for (slight) laughter as the significantly correlated gestures with the positive social emotion, other head and hand movements are not so typical.

8.2.2.3 Summary

In the beneficial class, the speaker's emotion is usually positive with primary emotion being mainly happy and social emotion being positive, including satisfactory/endorsing, enthusiastic/positive/passionate, conceited/proud, hopeful, and grateful/appreciative, and praising.

In the beneficial class, (middle) high pitch, stress sound, and smiling/laughing are prosodic features significantly correlated to social emotion.

And in terms of gestures, with the exception of smiling/laughing frequently appearing together with positive social emotion, other head and hand movements are not typical.

8.2.3 Harmful class of illocutionary force

8.2.3.1 Distribution of three layers of emotions in the harmful class

(1) Emotion – primary emotion

The statistics show that in the harmful class, there are two cases: One is the negative primary emotion, including: anger (11 times), sadness (6 times), worry (2 times), fear (once), and disgust (once); the other is that the speaker does not show distinctive primary emotion ('default' for 14 times) and negative emotion accounts for 0.6. This is related to the definition of the harmful class of illocutionary force. Since its essential content is negative or will have a negative effect on the speaker/hearer's interest, the speaker tends to have negative emotion. Certainly, although the speaker's interest is at risk, he/she may also try to stay calm. This explains why the primary emotion of default has been annotated in several tokens of the harmful class.

(2) Emotion – background emotion

In the harmful class, five types of background emotion all appear: energy/well-being (18 times) in the dominant position, relaxation (ten times), fatigue (four times), composure (two times), and discouragement (once).

(3) Emotion – social emotion

Social emotion is basically negative in the harmful class. Except for one occurrence of positive, conceited/proud and one neutral, indifferent, the others are negative, including: negative-unsatisfactory (14 times), negative-compassionate (11 times), negative-blame/blaming (five times), negative-contemptuous/despised (four times), negative-helpless/disappointed/frustrated/depressed (four times), negative-resentful (three times),

Table 8.6 Five types of illocutionary acts in the harmful class

	Token number	*Other-directed*	*Self-directed*	*Neutral/ default*
baoyuan	14	14	1	0
suku	11	4	10	0
piping	4	4	0	0
danxin	3	2	0	1
haipa	3	1	0	2

negative-hesitated/sceptical (once), and negative-awe (once). It shows that when the speaker performs an interest-negative illocutionary act, his/her social emotions are mainly negative. The only occurrence of positive, conceited/proud is in the token 'Interviewing Pan from Fengxian (after lunch) – harmful class – baoyuan.' When Pan talks about his sufferings in the past, he became conceited and relieved that he can lead an optimistic life now regardless of all the sufferings in the past. These two emotions are accompanied by the social emotion of compassionate.

From the perspective of the direction of social emotion, the statistics of five illocutionary act-types in the harmful class are shown in Table 8.6.

The statistics above imply that the act of *baoyuan* is the speaker commenting negatively on others' words and deeds, so the emotion is usually other-directed; *suku* is the speaker talking about his/her own experience, so the emotion is self-directed. However, the speaker can also breed other-directed social emotion. For example, in the token 'Shandong Wu –harmful class–suku' already discussed in Section 7.3.2, the speaker has compound accompanying social emotions as he acts. When he recalled that he was almost kicked to death by a relative when he was a child, he not only has self-directed compassion but also other-directed contempt (towards his relative); the act of *pinglun* is the speaker commenting negatively of others, so the emotion is other-directed; the act of *danxin* and *haipa* are expressive illocutionary acts according to Searle's classification. The speaker presents the negative primary emotion as well as other-directed awe and complaints.

8.2.3.2 Correlation of emotional state and prosodic features and gestures in the harmful class

We input the frequencies of each emotion when the speakers express the harmful illocutionary force into Minitab and the frequencies of cues like prosodic features and gestures. We use the factor analysis function to output the load diagram and distribution diagram.

As we can see from Figure 8.12, among the tokens of harmful illocutionary force, *danxin*, *suku*, and *haipa* belong to the same category (located in the first quadrant); that is, the speaker's emotion, prosody, and gestures are

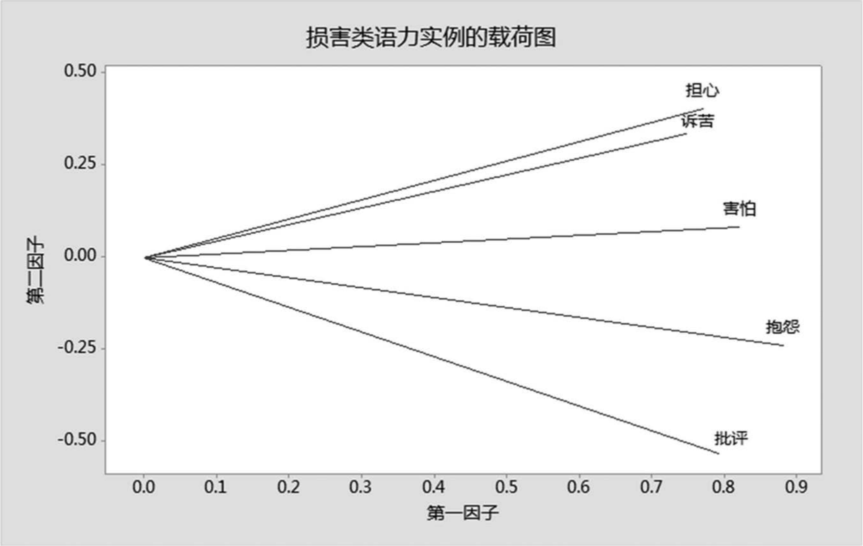


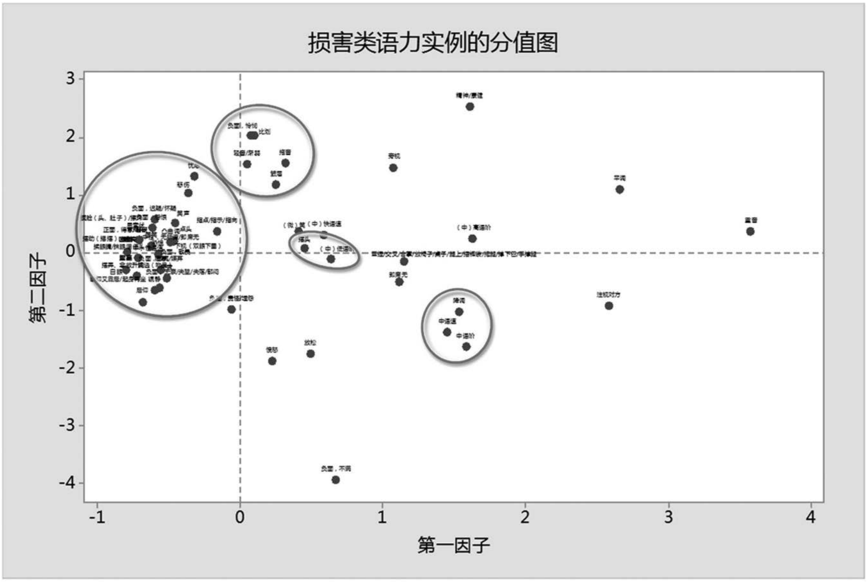
Figure 8.12 Load diagram from Minitab.

changing in the same direction; while *baoyuan* and *piping* differ from those three illocutionary act-types (Figure 8.13).

In the first quadrant, among the tokens of *danxin*, *suku*, and *haipa*, (medium) high pitch, flat tone, and stress sound are all correlated prosodic features; swaying one's head and (medium) low pitch are the correlated gesture and prosody; light sound/weakening sound and extended tone are closely correlated to the gestures of frowning and the social emotion of negative, pity in line with the general rules of the harmful class of illocutionary force.

In the harmful class, fear, sadness, and disgust are primary emotions that are significantly correlated to gestures; negative social emotion is still the main emotion type that is significantly correlated to prosody; social emotions such as helpless/disappointed/frustrated/depressed, hesitated/sceptical, and despised are correlated to each other. They frequently appear with gestures like looking down (eye dropping), rubbing one's eyes/wiping tears, touching the face, rolling one's eyes, indicating that these gestures are clues for negative social emotion. The remaining gestures (such as looking at the other, raising one's head, fiddling, holding and placing, and picking (items)) also appear in other classes of illocutionary force, so they have no specific meaning.

Positive social emotion of conceited/proud only emerges in the token 'Interviewing Pan from Fengxian (after lunch) – harmful class – suku,' in which the speaker is proud of the hardship he has been through. His social emotion of self-directed, negative, and pity accompanies it. Pitch, speech rate,



result	factor 1	factor 2
variance	3.2339	0.6210
variance contribution rate	0.647	0.124
cumulative contribution rate	0.647	0.771

Figure 8.13 Distribution diagram from Minitab.

* The cumulative contribution rate of the first factor and the second factor is 77.1%, which can represent the overall sample situation.

and intonation are still the most prominently correlated prosodic features; other prosodic features (such as laughter and coughing) are not distinctive.

In the harmful class, pitch, speech rate, intonation, and stress sound are significantly related to background emotion; other gestures like laughter and coughing have little universal meaning.

8.2.3.3 Summary

In the harmful class, the speaker's emotion is usually negative, particularly when the speaker is performing the acts against his/her own interest, and the social emotion is negative.

Intonation, speech rate, and stress sound are prosodic features significantly correlated to background emotion.

And in terms of gestures, except for such gestures as looking down (dropping one's eyes), rubbing one's eyes/wiping tears, touching one's face, and rolling one's eyes that are strongly correlated with negative emotions, other gestures are not very typical.

Swaying one's head and (medium) low pitch are the correlated gesture and prosody; in addition, light sound/weakening sound and extended tone are closely related to frowning and the negative, compassionate social emotion, which is in line with the general law of the harmful type of illocutionary force.

8.2.4 Counterproductive class of illocutionary force

8.2.4.1 Distribution of three layers of emotions in the counterproductive class

(1) Emotional state – primary emotion

Statistics show that in the counterproductive class of illocutionary force, the top primary emotion is 'default' (i.e. the speaker does not show any significant primary emotion). This is related to the properties of the counterproductive illocutionary force, that is, the speaker's inner expectation contradicts with reality. Therefore, it is mostly reflected in the tier of social emotion, and he/she may maintain a relatively calm primary emotion; and the primary emotions followed are sad (4 times), angry (2 times), surprised (once), and worried (once).

(2) Emotional state – background emotion

Statistics show there are three background emotions in the harmful class of illocutionary forces: relaxation (8 times), energy/well-being (5 times), and fatigue (5 times).

(3) Emotional state – social emotion

In the harmful class of illocutionary force of this corpus, social emotion is all negative. The ranking is negative, helpless/disappointed/frustrated/depressed (8 times) and negative, dissatisfied (8 times), negative, compassionate (4 times), negative, contemptuous/despising (4 times), and negative, blaming/complaining (once). The top two types of social emotion accord with the fact that the speaker will show helplessness, disappointment, disappointment, or depression, and dissatisfaction when his/her own expectations are different from reality.

From the perspective of the direction of social emotion, the statistic of five illocutionary act-types in the counterproductive class are shown in Table 8.7.

The statistics above imply that that the act of *wunai* usually involves other-directed social emotion of compassion, *shiwang* involves other-directed dissatisfaction, *weiqu* involves other-directed despise and contempt, and *yihan* involves self-directed compassion, disappointment, and frustration, as well as other-directed blaming, and *jingya* involves other-directed despise.

Table 8.7 Five types of illocutionary acts in the counterproductive class

	Token number	*Other-directed*	*Self-directed*	*Neutral/ default*
wunai	7	6	1	0
shiwang	4	3	1	0
weiqu	4	3	1	0
yihan	2	1	2	0
jingya	1	1	0	0

8.2.4.2 Correlation of emotional state and prosodic features, gestures in the counterproductive class

We input the frequencies of each emotion and prosodic features and gestures when the speakers are expressing the counterproductive illocutionary force into Minitab. We use the factor analysis function to output the load diagram and distribution diagram (see Figures 8.14 and 8.15).

As we can see from Figure 8.14, among the tokens of counterproductive illocutionary force, *jingya*, *weiqu*, and *wunai* belong to the same category (located in the first quadrant); that is, the speaker's emotion, prosody, and gestures are relatively changing in the same direction; while *shiwang* and *yihan* differ from those three illocutionary act-types.

In the first quadrant, the multimodal cues are closely related to the tokens of weiqu and *wunai* are medium speech rate, (medium) high pitch, and gazing at the other; the social emotions of negative, unsatisfactory are related and head down, swaying one's head, and extended tone; in addition, relaxation and wry smile are also related, but they do not have typical meanings.

In the third quadrant, the primary emotion sad, social emotion negative, pity, and negative, blaming/complaining are related to gestures such as frowning, touching one's face, shaking one's head, and leaning back one's head, which indicates that these clues are an important means of expressing negative emotions for the speaker to express the counterproductive type of illocutionary force.

In the second quadrant, sigh, intermittent sound, (medium) slow speech rate, light sound/weakening sound, and convex tune are distinctive multimodal clues. They are related to anger, fatigue, and worry, which are the main types of primary emotion relating to gestures. Frequent gestures are mainly head movements, especially swaying one's head, looking down (dropping one's eyes), shaking one's head, swaying one's head, and thinking. They can reflect the speaker's anxiety; there are also hand movements but with little significant specificity that also appear in other classes.

In addition, there is a case where the falling tone is related to the (middle) low pitch in the fourth quadrant.

Fatigue is the only type of background emotion that is significantly related to prosodic features, represented in four tokens of *wunai* and one

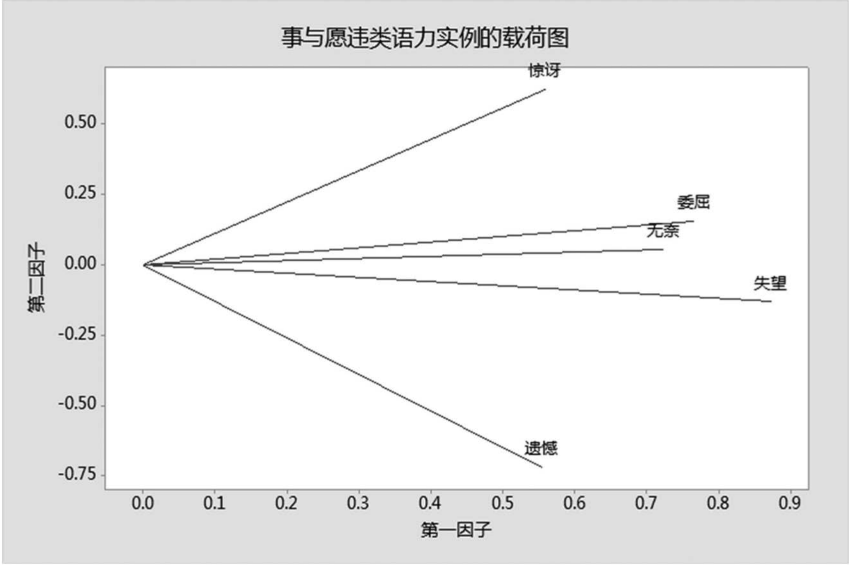
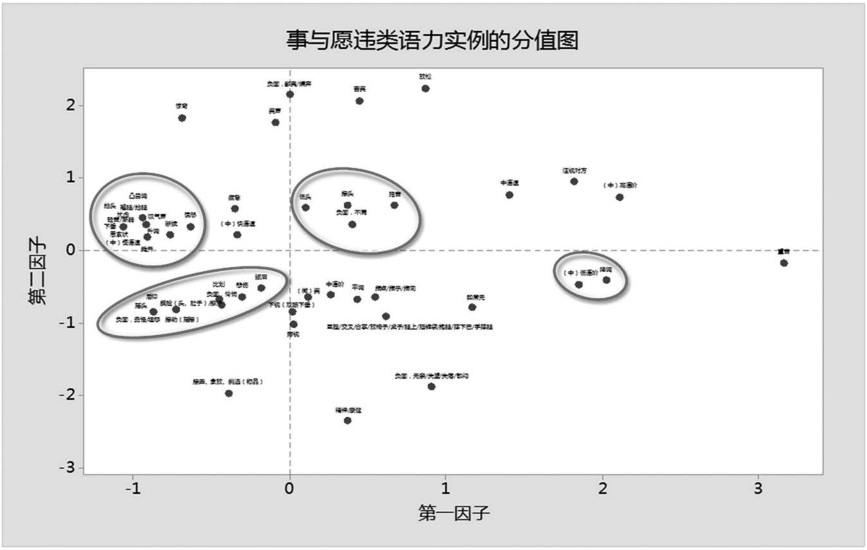


Figure 8.14 Load diagram from Minitab.



result	factor 1	factor 2
variance	2.4904	0.9476
variance contribution rate	0.498	0.190
cumulative contribution rate	0.498	0.688

Figure 8.15 Distribution diagram from Minitab.

* The cumulative contribution rate of the first and second factors is 68.8%, which can represent the overall sample situation.

token of *weiqu*. Prosodic features are also mainly manifested in speech rate, intonation. Features such as light sound/weakening sound and sighing can also reflect the speaker's state of depression and fatigue.

It can be seen that in the counterproductive class of illocutionary force, negative emotions are mainly related to speech rate, intonation, pitch, and stress sound, which can be regarded as core IFIDs. This is similar to the situations of the other three classes.

8.2.4.3 Summary

In the counterproductive class, the speaker's emotion is usually negative because the speaker will breed helplessness, disappointment, disappointment, depression, and dissatisfaction when reality goes against his/her own expectations.

The primary emotion that is significantly correlated to prosodic features is worry and anger; these types of primary emotions are correlated to speech rate, intonation, pitch, stress sound, light sound/weakening sound, and sighs.

And in terms of social emotion, it is mainly compassionate, blaming/complaining, unsatisfactory. Gestures like looking down (dropping one's eyes), shaking one's head, swaying one's head, and thinking well reflect the speaker's negative emotional state; there are also hand movements that also emerge in other classes and with little significant specificity.

8.3 Emergence patterns of speech acts and IFIDs

Statistical methods have been adopted to conduct statistical analysis on the relevant clues in the multimodal corpus. Various forms in prosody and gestures as IFIDs in expressing illocutionary force have been presented, and the correlation between emotion and prosody and gestures has been discussed previously. This section mainly summarizes the emergence pattern of multimodal clues and IFIDs on the basis of all statistical analyses.

8.3.1 Multidimensionality of IFIDs in situated discourse

Section 8.1.3.1 discussed how in situated discourse, the emergence frequency of performative verbs/phrases or other syntactic structures is very low and, in other words, it is rare that the type of illocutionary force is implied through performative verbs/phrases or other syntactic structures. Generally speaking, although various devices like performative verbs, phrases, and syntactic structures in the scope of language can clearly display the illocutionary act-type, statistics of this corpus show that they are not frequently occurring IFIDs in situated discourse.

In contrast, through the investigation on the prosody, gestures, and other clues in this corpus, we can find that such prosodic features as intonation, pitch, speech rate, and stress sound and gestures such as frowning and swaying are all correlated to the performance of the illocutionary act. According to the token analysis in Chapter 7, these prosodic features and gestures make a difference in speaker's successful and felicitous performance. This is because in live discourse how a 'live, whole person' performs an illocutionary act and expresses illocutionary force is a multimodal interaction with multiple devices involved including prosody, gestures, and language structures. And of course, this also validates the viewpoint of Austin (1962: 73–76) that mood, intonation, rhythm, stress (tone of voice, cadence, emphasis), and accompanying gestures (such as blinking, pointing, shrugging, and frowning) are all alternatives for performing illocutionary acts and they are all IFIDs.

Therefore, in situated discourse, the devices used by the speaker to perform illocutionary acts are multidimensional. It is inaccurate to determine the type of illocutionary force only through hooks like performative verbs, intonation, expressions, or acts. This marks the necessity of examining it from the multimodal perspective and at the level of the 'live whole person.' In fact, this is also the mechanism of the hearer judging the illocutionary force in live speech. Judging from the self-constructed multimodal corpus, the speaker has employed at least multimodal cues in the three dimensions of language structures, prosodic features, and gestures to act. However, the patterns of how multimodal cues emerge and concur vary under different situations, which must be analysed in the related social situation.

8.3.2 Different effects of different IFIDs on judging illocutionary force and their collaboration

In situated discourse, speech act verbs/phrases or other syntactic structures in the perspective of language structure, pitch, speech rate, pitch, and stress in prosodic features and facial expression and movements in gestures can all be counted as IFIDs. From the interaction between the speaker and hearer in situated discourse, different IFIDs have different references for the hearer to judge the illocutionary act-type. On the observer's side, judgement is made in view of all the IFIDs, but each of the contributions varies.

As for which IFIDs are at the core and which are auxiliary ones, the statistics extracted from the corpus tell us that from the perspective of statistics, some forms can be deemed core IFIDs, and others only play an auxiliary role. For example, by analysing the correlation between each IFID in the four classes and the speaker's emotions, pitch, speech rate, intonation, and stress sound are found to be the frequently appearing prosodic features; expressions and head movements are also frequently appearing gestures. In this sense, these forms outweigh the others in helping the speaker perform the illocutionary act and then express illocutionary force (on the side of the speaker) and hinting at the illocutionary act-type (on the side of the hearer).

The contribution of these IFIDs in determining the type of illocutionary force in situated discourse can be concluded as follows:

(1) One or several IFIDs occupy the dominant position and other IFIDs play a supporting role in the performance.

In this case, the commonest is the illocutionary force by performative utterance, that is, there are explicit performative verbs/phrases or other syntactic structures that can indicate the illocutionary act-type. Other prosodic features including tone, pitch, intonation, and speech rate and gestures including expressions and head or hand movements all become IFIDs that assist the hearer in judging the illocutionary act-type. Starting from the original intention of speech act theory, this study believes that IFIDs in language, including performative verbs/phrases or other discourse syntactic structures that reveal the illocutionary act-type are always at top priority. In other words, if they occur, these IFIDs usually play a central role. In this corpus, seven of the 11 tokens of *suku* witness the emergence of the syntactic structure ku(bitter); two of the three tokens of *haipa* witness the emergence of the syntactic structure *pa* (scared); and two of the three tokens of *jiashe* witness the emergence of the syntactic structures *ruguo* (if) and … *dehua* (if). Certainly, their occurrence does ensure that the illocutionary act is definitely *suku* or *haipa*. They may also emerge in other illocutionary act-types. However, to no small extent, these syntactic structures play a central role in determining the type of illocutionary force, and render hearers more accurate judgements of the illocutionary act-type. In other words, the performative utterance can be clearly and explicitly perceived and judged by the hearer. Together with their occurrence are other types of IFIDs. For example, in tokens of *suku*, shaking one's head, frowning, and having a long face frequently appear, and the significant increase in pause (reaching the average duration of 1.099s) and low pitch also concur. These prosodic features and gestures also play an important auxiliary role in the speaker's expressing illocutionary force.

Certainly, performative verbs/phrases or syntactic structures are the prioritized IFIDs, which does not necessarily mean that they are the key for determining the illocutionary act-type in all cases. Top priority and at the core are two definitions. The former is targeted at the original intention of speech act theory that language is always the crux to 'doing by talking;' the latter is targeted at judging the specific act-token. In many cases, one or more forms of prosodic features or gestures can also become core IFIDs, such as intonation or expression. For example, if someone says 'ni zhen cong ming!' (You are so smart!), in which there are no performative verbs or syntactic structures to clearly indicate the illocutionary act-type, it is non-performative utterance. However, the speaker's specific intonation can give clearer hints on the type of illocutionary force, e.g. whether it is *zanyang* (praising) or *fengci* (mocking). At this time, as the hearer distinguishes the type through intonation, then intonation is classified as the core IFID in this token; sometimes, if the speaker distinguishes via other forms, such as frowning and eyes full of contempt, then they can be listed as the core IFIDs; any other forms that help the speaker to express illocutionary force and assist the hearer in judging are auxiliary IFIDs.

(2) No IFIDs in the dominant position and multiple IFIDs complement and collaborate

In this case, the speaker uses neither performative verbs/phrases nor other syntactic structures (non-performative utterance), nor presents one or more prosodic features or gestures that could clearly indicate the illocutionary act-type. The hearer needs multiple IFIDs in combination with the specific situation to judge the illocutionary act-type. For example, in the token '20110914–110922_Ms Shi's father drank the pesticide – expressive – weiqu,' there are no performative verbs/phrases or other syntactic structures and prosody or gestures that clearly indicate the illocutionary act-type. But via the discourse content, prosody (low pitch, high pitch, convex tune, falling tone, pause, extended tone), gestures (swaying one's head, frowning) and other clues, the speaker is judged to express the illocutionary force of *weiqu*. She believes that caring for the elderly is not only her responsibility but that it also falls on her siblings' shoulders; however, the harsh reality is that she is the only one to do so.

There is also a special case in this category. Sometimes, even if the speaker does use certain performative verbs/phrases or syntactic structures, the hearer takes other IFIDs (forms of prosodic features and gestures) into consideration and determines the type of the illocutionary force as a different one from what the performative verb/phrases and syntactic structures represent. In this, the performative verb/phrases or syntactic structures are not the core IFIDs mentioned in the first case. For example, the speaker smiles and says: 'xian zai wo jiu chui xu zi ji yi xia zi ji de ai guo xing wei' (now let me brag about my patriotic behaviour…) (Xu, 2008: 16). The verb chui xu (boast) here is a modest strategy of the speaker and does not indicate an illocutionary act-type. The speaker could be stating, introducing, or explaining. Therefore, the performative verb here is not an IFID, let alone occupying a core position. The hearer needs to comprehensively determine the illocutionary act-type based on the situation, the utterance content, prosody and gestures, etc.

From the perspective of multimodal discourse analysis based on social semiotics, the function of IFIDs to help the hearer determine the type of illocutionary force boils down to the synergy between modalities. Multimodal discourse analysis regards prosodic features (pitches, speech rate, intonation, etc.), gestures (expressions, hand movement, gestures, etc.) and language itself as the expression devices and also as the symbol resources. Head movements, gestures like hands and legs shaking, body movements like bending and stretching, as well as tone and sound all constitute a multimodal discourse resource system, which have expressive meaning after

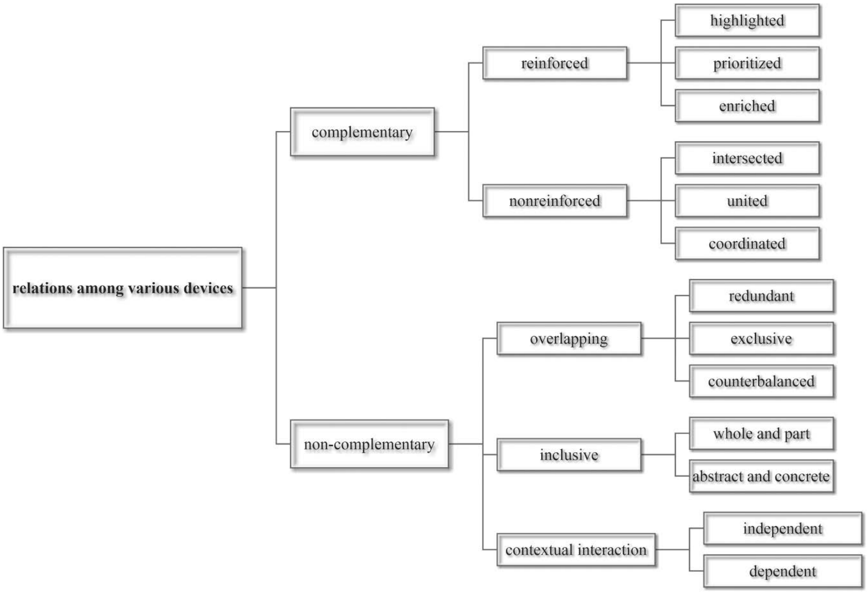


Figure 8.16 Relations among various expressive devices.

being sorted and patterned to form multimodal discourse (Zhang, 2009: 25–26). Discourses with different modalities are intended to reflect the overall meaning of the speaker. A variety of expression devices are used just to help the speaker's effective act.

As mentioned above, the definition of modality in this study is different from that of multimodal discourse analysis, but the latter's research framework of the relations between various devices can still be inherited. Zhang (2009: 27) divides the relationship between various forms into complementary and non-complementary relations. The former means that a modality (such as the language form itself) cannot fully express its meaning and needs to rely on another supplement that can be subdivided into reinforced and nonreinforced relation; the latter means that other mode does not contribute much to the expression of the first mode but still emerges. It can be subdivided into overlapping, inclusive, and contextual interaction. See Figure 8.16 (modified from Zhang, 2009: 27).

Certainly, the analysis framework above is only a theoretical construction of all the relations of multimodal discursive formations. And in situated discourse, not all of the above relations between various IFIDs are be presented. For example, although the speaker's devices may somewhat contradict one another (i.e. what is said, what is thought, what is felt, and what is embodied do not match, e.g. the expression and prosody fail to match when the speaker utters his/her sincere congratulations that makes the hearer feel the act of *zhuhe* is disingenuous), they would not offset one another.

Here, a number of tokens in the multimodal corpus have been selected to present the synergy of various devices with the aid of the analysis framework.

Let us first look at the complementary relation between multiple devices.

The first is reinforced relation, which means that one IFID occupies the dominant position and another IFID or other IFIDs strengthen its dominance. This relation is the first case discussed in the previous section, that is, one or several IFIDs occupy the dominant position, and other IFIDs play the supporting role in the performance. For example, to perform *jieshi*, the speaker usually uses speech act verbs/phrases or other syntactic structures and uses hand movements to explain, which is precisely the supplement that strengthens the dominant IFID.

For example, in the token 'Interview Pan from Fengxian (after lunch) – explaining the processing of cornflour' that lasts 143 seconds, the speaker explains to the hearer how to make dumplings with corn, during which several gestures are involved to strengthen the act of *jieshi* and to let the speaker understand the steps.

For example, when the speaker explains how to process dumplings made of cornflour, he demonstrates with one or both hands, pointing, gesturing, etc. Among them, demonstrating how to process corn with both hands takes up to 22.98 seconds, and gesturing lasts more than ten seconds. They emerge when the speaker introduces the processing steps.

When the speaker explains the steps of harvesting corn, processing corn, and grinding cornflour respectively, hand movements or gestures all emerge. These devices as the powerful supplement to the speaker's performance allow the hearer to more effectively perceive his explanation, which evidences the complementary relation between language structure and gestures.

The nonreinforced relation in the complementary relation means that two or more devices are indispensable to and complement each other. This belongs to the second case discussed above: No IFIDs in the dominant position and multiple IFIDs complement and collaborate. There are three relations between these IFIDs: synergy, correlate, and overlap.

Synergy means the coordination of different devices in delivering the speaker's overall meaning, all of which are indispensable. Just as the token discussed before '20110914-110922_Shi's father drank the pesticide – expressive class – weiqu' in which there are neither performative verbs/phrases, other syntactic structures, nor prosody or gestures clearly shows the type of illocutionary force. But the utterance content, prosodic features (low pitch, high pitch, convex tune, falling tone, pause, extended tone), gestures (swaying one's head, frowning) and other clues evidence that the speaker is expressing *weiqu*. These devices collaborate with and complement each other so that the hearer can perceive the illocutionary force of *weiqu*.

Overlap means the speaker is talking while doing something, such as explaining the theory and reason of something. From the perspective of interdependency, its doing-and-talking interdependency is parallel and relevant. In the token 'Speaker Guo – 110912 wood carving, talking about son's personality – panduan' in the corpus, the speaker Guo talks about the material of the item while lowering his head and carefully cutting open the objects. During the token of performing panduan that lasts 24.29 seconds, the utterance overlaps with the hand movements.

Analysing the relationship between various IFIDs is based on the investigation of the tokens under Discovery Procedure. In the same token, how various IFIDs collaborate is sometimes not limited to one relation but there could be multiple collaborative relationships. For example, when the speaker talks and performs task-doings, the utterance content overlaps with hand movements. But due to the fact that what is said, what is thought, what is felt, and what is embodied fail to match (there may be multiple patterns proposed by Gu (2013: 8)), the prosodic features and facial expressions somehow contradict each other. For example, the smile of the speaker seems to reflect positive emotions such as happiness, but the prosodic features (including pitch and intonation) speak for him/her that he/she feels ironic and dissatisfied. At this time, there is a contradictory and repulsive relationship between expression and prosody, which suggests that the speaker has another extraneous meaning.

The token 'Speaker Zhang – about job 1 – yaoqing:'

Turn-taking:

da ni dian hua zheng hao mai le ge cai //hui jia //shao cai de shi hou //e // dai ni jie shao jie shao //gan cui wo shao cai ni dao wan shang //shao dian jian dan de zai wo jia chi yi dun fan

I have just got some grocery and went home when you called. //How about you come to my place to have dinner and I will introduce some people to you.//

The speaker Zhang, the hearer (also the one who is invited: the author), and another audience-member (Shi who is the security man) were seated to have a conversation, which is the only activity type. Speaker Zhang invited the hearer to have a meal in her campus dorm. The forward-and-backward interdependency is that the hearer proposed seeing the speaker's campus dorm, and naturally the speaker accepted and invited him to do so.

First, we focus on the IFIDs in the range of prosodic features:

The speaker used medium and relatively stable pitch and medium tempo during the entire token; and pragmatic stresses and pause occurred. These prosodic features indicate that the speaker's emotional state is relatively stable, with no significant happiness or expectation.

Next, we focus on the IFIDs in the range of gestures:

The speaker's smile and staring at the hearer when expressing *yaoqing* (inviting) reflect her emotions to a certain extent. However, in the later stage, movements occur like shaking her head, rolling her eyes, putting her hands on her legs, and making gestures with her left hand.

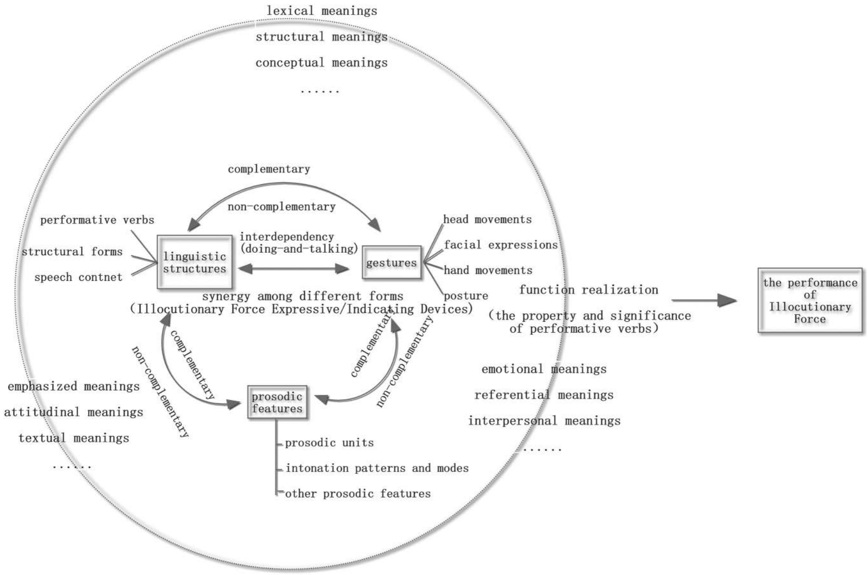


Figure 8.17 Synergy mechanism of various speech acts in situated discourse.

Here the smile seems to have the implication that the speaker is happy and willing to invite the hearer, while it is somewhat contradicted by shaking her head and rolling her eyes afterwards. Moreover, her prosodic features are relatively stable (medium pitch, medium tempo, and stable tone) with no presence of significant happiness, which fail to synchronize with the previous smile.

Therefore, in this token, the sincerity of the speaker's invitation should be questioned. From the perspective of its doing-and-talking interdependency, the hearer had already proposed to go to the speaker's residence before this illocutionary act-token. In this sense, it is hard to judge whether Zhang's invitation was spontaneous and genuine, or whether she did it to save face. And judging from the above-mentioned contradictions multiple IFIDs, the speaker may perform the act of *yaoqing* for fear of not losing face rather than speaking spontaneously.

Figure 8.17 describes the synergy mechanism of various speech acts in situated discourse.

With regard to form, IFIDs including language structure, prosodic features, and gestures are at the command of the speaker who in a specific situation selects certain speech acts in situated discourse because of specific emotions, intentions, and other influencing factors. Multimodal cues from different levels correlate and collaborate to express the meaning of the utterance, helping the speaker perform the illocutionary act and thereby express the illocutionary force.

8.3.3 Unification of the exclusive tendencies and cross-class universality of IFIDs

As indicated in Section 8.1.3.1, the speakers in the corpus are found to adopt syntactic structures to accomplish certain illocutionary forces. These syntactic structures make a difference in determining the type of illocutionary force. Some of them are quite exclusive to a certain type of illocutionary force, e.g. *ruguo* (if) and … *dehua* (if …) determine the type as *jiashe* (hypothesizing), while some of them do not exclusively point to one type. Furthermore, exactly the same syntactic structure could indicate several illocutionary act-types, e.g. the performative verb *pa* (be afraid of/be concerned about) could direct one to the illocutionary act of *haipa*(scare) or *danxin*(concern); the syntactic structure shi … ba(… is/could be …) could mean the illocutionary act of *panduan* (judging), *chenshu* (reporting), or *shuoming* (illustrating).

Through the previous analysis, we can see that there is a certain degree of correlation between prosody, gestures, and illocutionary act-types. For example, in the tokens of the beneficial class, the speaker's emotion is usually positive, and the speaker's non-verbal acts are basically smiling, nodding, etc.; in the harmful class, the speaker's emotion is usually negative with non-verbal acts of frowning and shaking one's head. But more often they present a complex relationship of interlacing and overlapping, lacking specificity and exclusivity to a certain type of illocutionary force. Prosodic features (such as intonation and pause) and gestures (smiling, crossing hands, etc.) are not exclusive to a certain type but may correspond to multiple types of illocutionary forces.

For example, the prosodic feature of high pitch emerges both in the beneficial class and the harmful class, for the speaker usually has a relatively significant emotional state that can lead to an increase in intonation when performing these two classes of illocutionary act. Furthermore, the gesture of frowning is the external manifestation of disgust, dissatisfaction, or contempt, but its appearance does not necessarily point to the harmful and counterproductive type of illocutionary force; for example, the speaker frowns when performing a neutral or beneficial illocutionary force. In the token 'Tonglu Zhou–beneficial class – manyi –status quo,' the speaker performs the act of *manyi*. Her essential content is that she is satisfied about her current life. The primary emotion is happiness, and social emotion is self-directed, positive, and satisfied. Approaching the end, the speaker says, 'xian zai dou hao le, chi chi, zao xie qiong o' (It's so nice now, eat. We were so poor then) and then frowned. This expression here is because the speaker mentioned her poor life in the past to highlight the current situation. Therefore, although the token of *manyi* is classified in the beneficial class, this does not deny the possibility that gestures or expressions related to negative emotions may occur.

As the analysis above indicates, pitch, speech rate, intonation, and stress sound are the main forms in prosody significantly correlated with the emotional states of the speaker and that frequently emerge in various illocutionary act-types; furthermore, the range and level of intonation (the fundamental frequency curve) are closely correlated to the emotion intensity. The more intensive the emotion, the more significantly the tonal mode (fundamental frequency curve) changes, but the tonal mode of intonation does not have a significant correlation with any type of emotion (Bänziger & Scherer, 2005). The frequencies of head movements, hand movements, and expressions in gestures are relatively higher, while forms in gestures are less common in comparison with various forms in prosodic features, as not all illocutionary act-tokens witness the emergence of IFIDs; some IFIDs have a certain degree of tendencies, e.g. wiping tears tends to appear in the harmful class of illocutionary force, while smiling/laughing mostly appear in the beneficial and neutral classes.

In general, IFIDs such as syntactic structures, prosody, and gestures tend towards a specific class of illocutionary force-specific and cross-class universality.

8.3.4 Dynamic emergence of IFIDs

The process of the speaker using a variety of resources and presenting multiple IFIDs in situated discourse is dynamic yet complex. The emergence order, the beginning and ending of different speech acts and IFIDs, is not identical in each token, which needs to be analysed in consideration with different speakers in their respective social situations. From the set analysis framework perspective, this research mainly focuses on analysing doing-and-talking interdependency, prosodic features, and gestures.

The doing-and-talking interdependency mainly refers to the interdependency between the talking and doing when the speaker performs an illocutionary act. Its analysis needs to combine the speaker's speech and gesture cues in a specific situation. Here, the dynamic analysis between doing-and-talking interdependency, prosody, and gestures can refer to Gu's configuration analysis method (2006).

In this study, the layer of interdependency is specially set when annotating each token in Elan to analyse its interdependency. Here, the analysis of two tokens is elaborated to illustrate the dynamic analysis of the speaker's illocutionary force in situated discourse.

Token 1: 20120229 – 10:30 Guo hanging calligraphy in the new building of the foreign language school – neutral class – jieshi

Turn-taking:

Guo: ta zhe wai bian hai you yi ceng de na ge, hao xiang zhe ni kan duo hou, ni kan duo hou na ge shi gao, zhe ge shi gao fen

Huang: ni shi gao fen tai hou le, zhe ge ding zi cha bu duo li mian

Guo: en, gen ben jiu ding bu dao na me duo, zhe ge ding zi tai duan

Guo: There is another layer outside. You can see how thick the plaster is.

Huang: Your plaster powder is too thick that this nail can't be inserted.

Guo: Well, I can't nail too deep because the nail is too short.

Interdependency analysis:

Illocution-and-reality interdependency: The wall and nail are objective reality.

Forward-and-backward interdependency: Before – talking about this wall is not suitable for hanging heavy stuff; after – promise to take other type of nails.

Doing-and-talking interdependency: Parallel and relevant (explaining why the wall cannot be nailed, while touching to check the wall).

Duration: 14.435 seconds

Figure 8.18 is the original figure annotated by Elan.

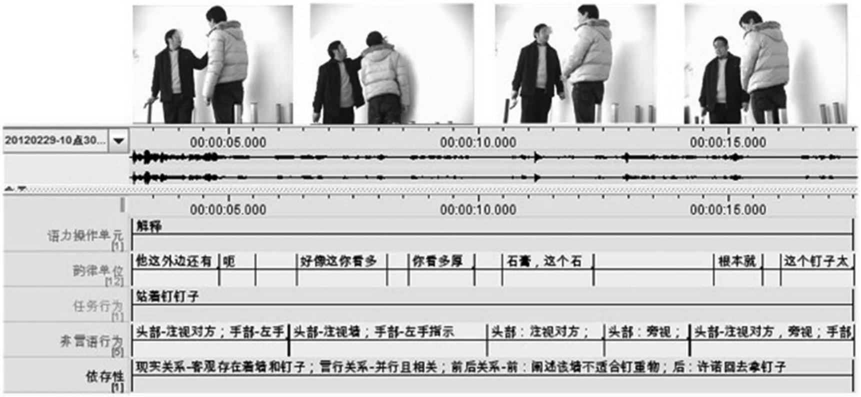
Based on the Elan annotation, we can analyse the dynamic doing-and-talking interdependency. Figure 8.19 shows the dynamic timeline of doing-and-talking interdependency.

Speaker Guo maintains various gestures during the whole process (including looking at the wall, looking aside, pointing with his hand, and dropping his hands). And there were five long pauses in total, among which, the fourth pause is because of the hearer's interruption.

After this token, the speaker Guo performs the illocutionary act of *chennuo* (promising), that is, he promises to go back and try another nail and at the same time, he lowers the table and leaves.

Token 2: 20110905 – cleaning – harmful class – baoyuan, zhou ma emotion: anger 2

Turn-taking:



语力操作单元 [1]: 解释

韵律单位 [12]: 他这外边还有 | 呃 | | 好像这你看多 | | 你看多厚 | | 石膏，这个石 | | 根本就 | | 这个钉子太

任务行为 [1]: 站着钉钉子

非言语行为 [6]: 头部-注视对方；手部-左手 | 头部-注视墙；手部-左手指示 | 头部：注视对方； | 头部：旁视； | 头部-注视对方，旁视；手部

依存性 [1]: 现实关系-客观存在着墙和钉子；言行关系-并行且相关；前后关系-前：阐述该墙不适合钉重物；后：许诺回去拿钉子

Figure 8.18 Analysis of illocutionary force.

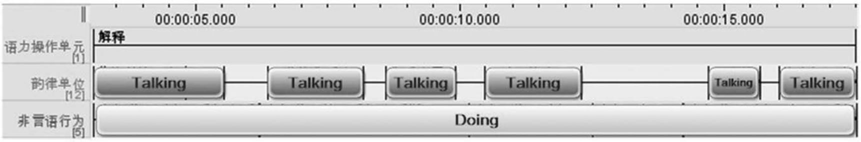


Figure 8.19 The dynamic timeline of doing-and-talking interdependency.

Shi: shui xiao de, ta si ji mei yi tian dou shi de, san bai liu shi tian tian tian du shi zhe yang de, fei ba zhe ge ce suo, chuang hu, wai bian chuang hu guan zhe, zhe men guan zhe. ni jiang jin lai yi gu chou wei, yi gu sao wei. ni ba ta kai zhe, ta fei ba ta guan zhe. tuo ba ne? ni ba ta zhe yang liang qi lai, ta fei ba ta fang shang shui, ba ta yang zhe. tian tian tuo ba dou lan diao le. huang cha ru: en. ta… ta bu zhi dao zong gao di nao dai you ge… da gai you ge wen ti.

Who can imagine that he did this every day? Why on earth would he shut the toilet window and door? Wouldn't everyone feel the nasty smell at the entrance of the toilet? Even if I open them frequently, he keeps shutting them off. Not to mention his behaviour of soaking the mop in water. They would get rotten. (Huang interpreted: I agree) His behaviour … What was he thinking … He is just insane.

Interdependency analysis:

Illocution-and-reality interdependency: The perpetrator indeed did something incomprehensible in the bathroom (shutting the window and soaking the mop in the sink).

Forward-and-backward interdependency: Before – the hearer asked why the perpetrator soaked the mop in the sink; after – the hearer agreed with the speaker.

Doing-and-talking interdependency: parallel and independent (complaining while preparing to mop the floor).

Duration:27.348 seconds

The Elan annotation diagram is shown in Figure 8.20.

On the basis of Elan annotation, we can analyse the dynamic doing-and-talking interdependency. Figure 8.21 is the dynamic timeline of doing-and-talking interdependency.

In the whole process, due to the relatively strong emotional state (pronounced anger (primary emotion) and other-directed contempt (social emotion)), the speaker talks at a fast tempo and keeps talking (the duration is 27.348 seconds with only nine obvious pauses); she also presents rich gestures, including facial expressions (frowning), head movements (swaying, looking at the hearer, looking aside), and other physical movements. After performing the act of *baoyuan*, the speaker enters the toilet and starts mopping the floor.

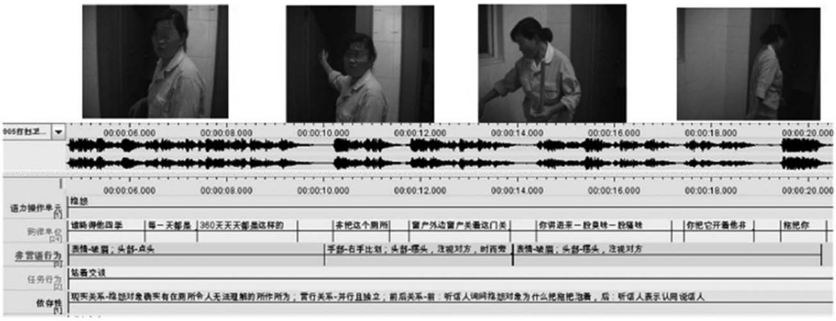


Figure 8.20 Analysis of illocutionary force.

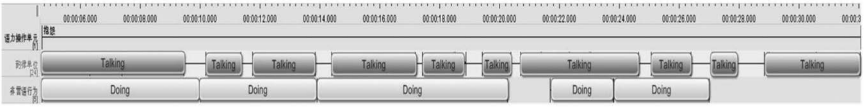


Figure 8.21 The dynamic timeline of doing-and-talking interdependency.

The interdependency analysis of these two tokens indicates that the sequence, beginning, and ending of the employed devices are all dynamic, and talking and doing usually overlap and co-occur. The analysis of this dynamic process needs to be carried out in conjunction with each token. This process is completed during the annotation and multimodal analysis in this research.

8.4 Influencing factors of live speech acts and IFIDs

As mentioned above, from the perspective of talking is doing, the speaker can achieve the same perlocutionary intention by choosing different illocutionary acts. The choice of the illocutionary act-type belongs to the level of acting rhetoric (Xu, 2008: 11). Once the speaker has determined the type of illocutionary force, he/she can use various expressions to act, including different forms of discourse, prosodic features, and gestures. This belongs to the level of expressive rhetoric. As the speaker performs, various IFIDs emerge. Then comes the question: What are the influencing factors working on the speaker's ability to employ speech acts and the looming of IFIDs? In Sections 4.2.3 and 6.1.4, the author discusses the emotional state being an important influencing factor for the speaker's expression of prosody, gestures, and other resources. However, in situated discourse, emotion is not the only factor that affects the speaker's expression of multimodal illocutionary force as well as the emergence of various IFIDs. This section focuses on analysing the possible influencing effects besides emotional state. It should be clarified that 'influencing factors' here means correlation rather than causality. In other words, this study only discusses the multimodal speech acts and IFIDs correlating to multiple factors, but whether these factors directly cause their emergence remains to be further studied.

Three perspectives for situated discourse modelling, i.e. the perspective of the actor, the activity, and the system, were introduced in Section 3.2.1. Furthermore, in this section, such a modelling perspective is adopted as a basic thought when investigating the influencing factors of live speech acts and IFIDs. In combination with the self-constructed multimodal corpus, this study finds that there are many other influencing factors of live speech act and IFIDs.

8.4.1 Influencing factors from the actor's perspective

This study adopts the actor's perspective to do modelling and researching on situated discourse. Emotion is an important dimension of the actor's perspective. Previous sections elaborated on emotion's influence on the speaker's performance of the illocutionary act; in other words, emotion is an important influencing factor for live speech acts and IFIDs. In general, certain emotions have a significant correlation with certain prosodic features and gestures. This chapter calculates the correlation between various emotional states and prosodic features, gestures, and other IFIDs, and the results show that certain IFIDs are significantly correlated with certain types of emotions. In the neutral class of illocutionary force, as the conversation interlocutors are interest-unbiased with the essential content, the speaker generally does not have any significant emotional tendency. In the neutral class, the most common situation is that the speaker does not show significant basic emotion (28 times), followed by happy (20 times), worried (5 times), and angry (2 times).

This chapter mainly calculates the correlation between various emotional states and prosodic features, gestures, and other IFIDs. The results show that certain IFIDs have a significant correlation with certain types of emotions. Among the tokens of neutral illocutionary force, as its essential content is not directly related to the speaker and hearer's interest, the speaker generally does not have a significant emotional tendency. Among the tokens of the neutral class, the most common case is that the speaker does not show significant primary emotion (28 times), followed by happy (20 times), worried (5 times), and angry (2 times). Speakers who express the neutral class language skills seldom have physical features such as frowning, as well as prosodic features with high pitch, rapid speech speed, or intonation fluctuations, and cross-category, more common IFIDs (such as medium pitch, medium speech rate, flat tone, and pointing/instruction/pointing) appear frequently; and in the harmful class of language, negative and negative primary emotions include angry (11 times) and sad (6 times). The speaker often makes gestures such as frowning and wiping tears, and these are sometimes accompanied by prominent prosodic features such as whispering, slow speaking, and more pauses, and these features are less common in the neutral class.

Speakers who express the neutral class of illocutionary force seldom show the gesture of frowning and prosodic features of high pitch, fast tempo, or fluctuations in intonation, but these IFIDs frequently emerge among all the classes (such as medium pitch, medium tempo, flat tone, and pointing/instructing/gesturing), and in the harmful class, negative primary emotions including angry (11 times), sad (6 times) as well as gestures such as frowning and wiping tears frequently emerge, sometimes accompanied by low pitch, slow speech rate, and more pauses that are less common in the neutral class.

However, the author finds that prosodic features and gestures do not present an exclusive correspondence to 20 illocutionary act-types in this corpus. Even in the same token, the tonal pattern and mode do not remain unchanged but fluctuate according to the utterance content and corresponding emotion. Certainly, if we classify all the tokens according to the four classes of this research, it can be found that some IFIDs in prosody and in gestures appear frequently in a certain class. In other words, the speaker's prosodic features and gestures when expressing the illocutionary force have certain tendency towards a specific class of illocutionary force. But the underlying reason is that the classification of illocutionary force in this research is based on the relevance between the core content and the interlocutors' interest, which is the important influencing factor for emotional state.

Such classification standards of illocutionary act-tokens render each class a higher correlation with certain emotions (in other words, certain emotions tend to appear in one class of act-tokens more frequently). Therefore, this study actually introduced the dimension of emotional state when examining the correlation between the four classes with IFIDs in prosodic features and gestures. And emotions also affect a speaker's utterance content, prosody, and gestures, etc., and render them as emerging more frequently. Therefore, the correlation between the class of illocutionary force and certain prosodic features and gestures is, in essence, the correlation between the emotion of that class and the prosodic features and gestures.

It should be noted that, on the one hand, the speaker may have multiple emotional states when expressing illocutionary force (one of the reasons why this study sets the three layers of emotion), so the emotions do not have a one-to-one correspondence to the IFIDs in prosodic features and gestures. The correlation between emotional state and prosody and gestures means that a single-layered emotion is being investigated with prosody and gestures. On the other hand, even if the speaker has only one emotion or has multiple roughly consistent emotions, the correspondence between them and prosody and gestures is not unique. They are correlated rather than cause and effect (Planalp & Knie, 2002: 59).

In addition, the emergence of cues (such as speech, prosody, and gestures) is closely correlated with the speaker's intention, type of speech act, and pragmatic purpose. Situated emotion is embedded in the context of the speech act, and the presentation of its cues should also be adapted to the social situation of the speech act (Planalp & Knie, 2002: 67). In short, emotional state is an important but not the sole influencing factor for IFIDs (the two features have a quite complex relationship).

Apart from emotion, personal style and expression capability from the actor's perspective must also be analysed.

The selection of devices to express illocutionary acts and the various IFIDs presented are related to the speaker's personal style, which refers to the individual having an introverted or extroverted, simple or sophisticated personality. It can make a difference in expressions, e.g. the introverted person is inclined to conceal his/her own feelings and present few IFIDs in conversation, while the extroverted person is willing to or tends to employ more devices and therefore present more speech acts and IFIDs. Different personalities and talking styles directly cause different means and forms in verbal communication. For example, in the third chapter of *A dream of red mansions* by Cao Xueqin in China's Qing Dynasty, there is a poem: 'tai sheng liang ye zhi chou, jiao xi yi shen zhi bing. lei guang dian dian, jiao chuan wei wei. xian jing shi ru jiao hua zhao shui, hang dong chu si ruo liu fu feng' (Her dusky arched eyebrows were knitted and yet not frowning; her speaking eyes held both merriment and sorrow; her very fragility had charm. Her eye sparkled with tears, her breath was so soft and faint. In response, she was like a lovely flower mirrored in the water; in motion, a pliant willow swaying in the wind). The poem is used to describe how sentimental and weak Lin Daiyu is, which is also her own unique personal style when talking. Another example is the description of Wu Zixu whose 'eyes glare like the lightning, sound resonant like the bell' in *Annals of the kingdoms in the East Zhou Dynasty* by Feng Menglong in the Ming Dynasty of China. In this corpus, different speakers have different styles in expressing illocutionary force: The speaker Wu, a male farmer in his forties from Ju County, Shandong, rarely speaks in a high pitch. His tempo is slow and the expression barely changes. In contrast, speaker Guo, a male also in his forties from Lu'an, Anhui, has more changes in tonal patterns and modes and presents rich expressions. In this sense, personal style determines to a large extent the number, concurrence, and order of IFIDs in the illocutionary act-tokens.

Similarly, expression capability is also an influencing factor to be examined. Expression capability is multi-level, which refers to the individual competence of utilizing diverse resources in language structures, rhetoric skills, and other non-verbal expressions. Illiterate people were defined in Section 5.1.1 as a group of people who cannot read or write. However, the inability to read and/or write does not mean that their verbal expression skills are also weak. Whether there is a significant correlation still needs to be studied.

But generally speaking, illiterate people who have limited knowledge and literacy capability will also be weak in using vocabulary phrases, language structures, and rhetoric in language. This will to a certain extent affect their choice of performative verbs/phrases or syntactic structures and also affect the utterance content in situated discourse. For example, some complex vocabulary or utterance hooks are relatively less likely to appear. However, illiterate people can use rich non-verbal acts, including prosodic features and gestures to help his/her performance of illocutionary acts. Judging from the multimodal corpus in this study, these devices are also extremely rich without any influence of limited literacy and knowledge. Speakers with different expression capabilities and knowledge levels will also show individual differences in the emergence of performative verbs, syntactic structures, prosody, and gestures.

In addition, personal intentions and pragmatic purposes will also affect the emergence of IFIDs. The felicitous performance of one illocutionary act requires the corresponding intentional state and then the corresponding emotional state. For example, to express the illocutionary force of *zanyang*, the speaker is happy and indeed has the intention of thinking another's behaviour is commendable, without which the illocutionary act-token is infelicitous.

In situated discourse, whether the speaker has the corresponding intentional state and emotional state when performing the act and whether he/she intends to let the hearer know is complex and will be split up into different cases; different cases are closely related to the speaker's intention and pragmatic purpose. In view of the STFE-match principle, only the first case in which what is said, what is thought, what is felt, and what is embodied by the speaker are highly matched, and the other cases are against this principle and produce additional meaning.

1 The speaker has the intention and emotion of the corresponding illocutionary act and does hope that the hearer can perceive this. So the speaker usually resorts to the corresponding discourse content, prosody, and gestures. For example, when the speaker performs the act of *zhuhe* (congratulating), his/her emotion is happy with the intention that he/she genuinely believes that the hearer deserves some kind of achievement or reward, and the pragmatic intention is hoping that the hearer will know and accept his/her *zhuhe*. Under normal circumstances, the speaker will present a smile and other corresponding prosodic features. At this time, what is said, what is thought, what is felt, and what is embodied are highly unified.
2 The speaker has the corresponding intentions and emotions but does not want the hearer to know because of a certain communicative purpose. For example, in spite of the intentional state of deeming it natural that the object should get some achievement or reward when performing

the act of *zhuhe*, the speaker does not want the hearer to know about it. Just like a strict mother who is really happy in the bottom of her heart about the child's progress or achievement but would rather her children do not know such intentions and emotions, she usually does not smile and may have a long face or even frown when expressing the illocutionary act of congratulating. She intends to present prosodic features that make people perceive her indifference or understatement.

3 The speaker does not have the corresponding illocutionary force's intention and emotion but hopes that the hearer thinks that he does so. At this time, the speaker is not performing a felicitous performative act while, driven by his/her own pragmatic purpose, he/she still shows certain prosody or gestures in the hope that the hearer can buy his/her story and perceive its sincerity. In this case, the speaker is performing a deceptive act, such as falsely inviting, falsely promising, and false apologizing. At this time, certain strong and exaggerated prosody and gestures during the speaker's overacting may make the hearer think him/her less sincere.

4 The speaker does not have the intention and emotion, and hopes the hearer knows this. Then the speaker performs the illocutionary act as a last resort, involuntarily, or on the surface due to the restriction of the situation, saving face, culture, etc. Both parties know well that the act is ingenious. So the prosody and gestures are usually not as natural as the felicitous performance of the illocutionary act.

5 The speaker may or may not have the intention and emotion, and is indifferent about whether the hearer knows. At this time, the speaker's prosodic features, gestures, etc. will not be identical with the illocutionary act's felicitous performance, which can be noticed from some other clues.

In each of the situations above, the speaker will employ different devices because of specific intentions and purposes, thus presenting different IFIDs.

8.4.2 Influencing factors from the activity's perspective

As indicated in Section 3.2.1, the activity's perspective can be adopted to conduct modelling and investigation on situated discourse. The perspective of activity is subdivided into pattern and configuration modelling, and role and relation modelling. The former includes turn pattern and configuration, task pattern and configuration, and activity pattern and configuration. The latter include familial relation, role relation, power relation, and interdependency relation. For example, for the same act of *ganxie* (appreciating), different role relations and power relations between the interlocutors prompt the speaker to employ different speech acts and present varying IFIDs. When thanking a boss or client, the speaker may smile more, speak in melodious rhythms, and even bow. And when thanking someone in a similar position, close friends, or family members, the speaker's relevant gestures may be simplified and the intonation is slower and flatter.

The influence of interdependency on IFIDs is emphasized here. In this study, interdependency is divided into three tiers, i.e. forward-and-backward interdependency, illocution-and-reality interdependency, and doing-and-talking interdependency.

1 Forward-and-backward interdependency determines the illocutionary act-type and influences the selection of performative verbs and phrases, syntactic structures, and other devices. For example, in the token 'Zhou from Tonglu – beneficial class – jiashe 1,' the utterance before was the speaker talking about someone planning to take her away from her hometown, and the utterance after is the hearer asking where to go. In the interval, the speaker performs the act of *jiashe*, assuming how she would be if she left her hometown. The before-and-after utterance is logically justifiable for the act of *jiashe*: The statement that someone planned to take the speaker away from home before the illocutionary act of *jiashe* led the speaker to hypothesize her situation after leaving home; after the speaker's act of *jiashe*, the hearer continues to inquire about the same topic. Therefore, the forward-and-backward interdependency logically connected with the utterance corresponding to illocutionary force, otherwise the utterances would end up being inconsistent and incoherent. In this sense, the forward-and-backward interdependency then determines the choice of syntactic structures. In this token, the speaker used the explicit syntactic structure of ru guo (if) to express the illocutionary force of *jiashe*.
2 Illocution-and-reality interdependency is also an important influencing factor. For example, in the token '20110926 Guo – neutral class – yaoqiu,' the illocution-and-reality interdependency is that the speaker fiddles with the items and performs the act of *yaoqiu* with the item to be repaired. During this process, obviously presented IFIDs are his hand movements, expressions (lowering his head and staring at the item) and hand movements (fiddling with the item).
3 Doing-and-talking interdependency correlates with the speaker's non-verbal acts. If the speaker is performing a certain illocutionary act and meanwhile performing other task-doing, his/her hand/head movements and gestures will be constricted to a certain extent, thereby affecting the non-verbal acts. In the token '20120229 – 10:30 Guo hanging calligraphy in the new building of the foreign language school – neutral class – jieshi,' the speaker's doing and talking are parallel and relevant. He hammers a nail into the wall while explaining his illocutionary act, during which his head movements, expressions (staring at others, looking to the side, looking at the wall, smiling), and hand movements

(dropping, waving, instructing) are closely related with the illocutionary act of *jieshi*.

8.4.3 Influencing factors from the system's perspective

The system's perspective, an important modelling perspective in situated discourse, includes the space–time dimension and environmental layout, etc., and these influencing factors are referred to as 'situational factors' in this study. In Section 4.2.2, the author discusses how important and necessary situational factors are in the investigation of illocutionary force in situated discourse. In fact, situations, occasions, and locations all somewhat regulate and restrict how the speaker performs illocutionary acts and expresses illocutionary force. For example, during a meeting, the speaker usually lowers his/her voice, whispers, or uses gestures to perform illocutionary acts; while in a crowed and intense football match scene or dance hall, the speaker tends to harbour excitement or other specific intentional states when performing illocutionary acts, during which his/her live speech acts and IFIDs are greatly affected. Sometimes, the different perceptions of the hearer and the speaker towards the same social situation result in a deviation in the hearer's understanding of the speaker's behaviour and intentions. For example, the speaker (a boss) just wants to have some casual chat with the hearer (a subordinate) in the office, while the hearer perceives it as a formal conversation or a promotion interview. In this sense, there is a possibility for the successful implementation of the speaker's illocutionary act.

In addition, cultural convention is also considered an influencing factor in the expression of illocutionary force and IFIDs in situated discourse from system modelling. Obviously, the entire system where the situated discourse is embedded is under the influence of different languages and cultures, so the speaker's cultural tradition and conventions also influence the implementation of performative acts and the expression of illocutionary force. Coming from different cultural backgrounds, the speakers can either speak frankly or implicitly in a specific pragmatic context, that is, the expectation of them varies from culture to culture. For example, people tend to show off their feelings and intentions in some cultures; in contrast, other people may be more reserved (Wierzbicka, 1999). When expressing the same type of illocutionary force, speakers' prosodic features and gestures vary in different cultural backgrounds. In terms of prosodic features, intonation presented to reflect happiness or sadness vary across different languages and cultures. That explains why language and culture are set as one variable in the study of emotion and intonation. For example, in the category of gestures, Austin (1962: 76) proposed that shrugging that accompanies an illocutionary force's performance is also an alternative to accomplishing certain performative acts and hence belongs to IFIDs. But shrugging is so rare in Chinese culture that such an IFID is not found in this multimodal corpus. Therefore,

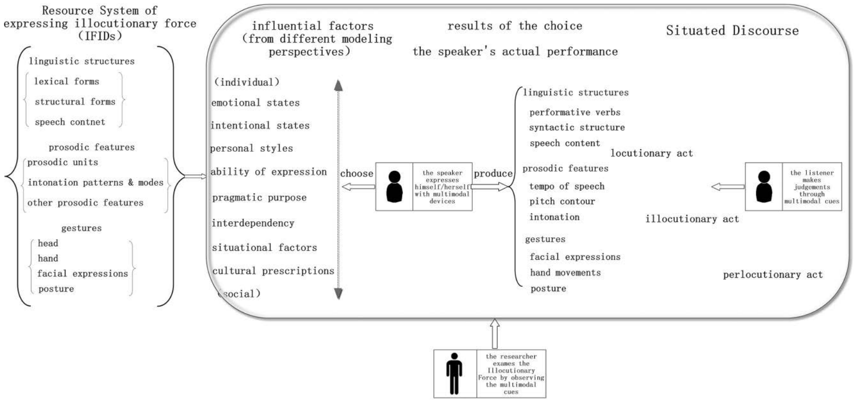


Figure 8.22 Relationship between live speech acts and influencing factors.

cultural conventions affect and even regulate the speaker employing various devices (especially gestures) when performing illocutionary acts. This is also an important topic that needs to be investigated in cross-cultural pragmatics.

It has been discussed from the perspective of the actor, activity, and system that the influencing factors on the speaker's choice of live speech acts and the result of IFIDs include personal style, expression capability, pragmatic intention, interdependency relation, social situation, and cultural conventions. Figure 8.22 shows the relationship between the speaker's use of live speech acts, influencing factors, and indications.

In brief, how the speaker manipulates live speech acts and produces possible IFIDs when performing illocutionary acts and expressing illocutionary force is affected by various influencing factors, such as personal style, expression capability, pragmatic intention, interdependency relation, social situation, and cultural conventions. These influencing factors intertwine and work collaboratively in the situated discourse. The assumption of the live speech acts and IFIDs is merely theoretical deduction that requires further investigation on the corpus, due to the non-negligible fact that live speech acts are rather complicated. This could also be the research orientation in the future.

8.5 Comparison of the cluster pattern of the four classes

The set formula is adopted here to conclude and present the cluster pattern of illocutionary forces in the neutral, beneficial, harmful, and counterproductive classes. It intends to articulate their differences and similarities, by which this study can show the differences and similarities in emotional state, intentional state, situation distribution, interdependency, prosodic features, and gestures. See Figures 8.23–8.26.

Also, it should be pointed out that the set formula adopted here to present the cluster pattern is not directly related with concept modelling discussed before. Having no intention to establish any new model, the author does so to present the differences and similarities of the illocutionary forces of the four classes as clearly as possible.

8.5.1 Cluster pattern of the neutral illocutionary force

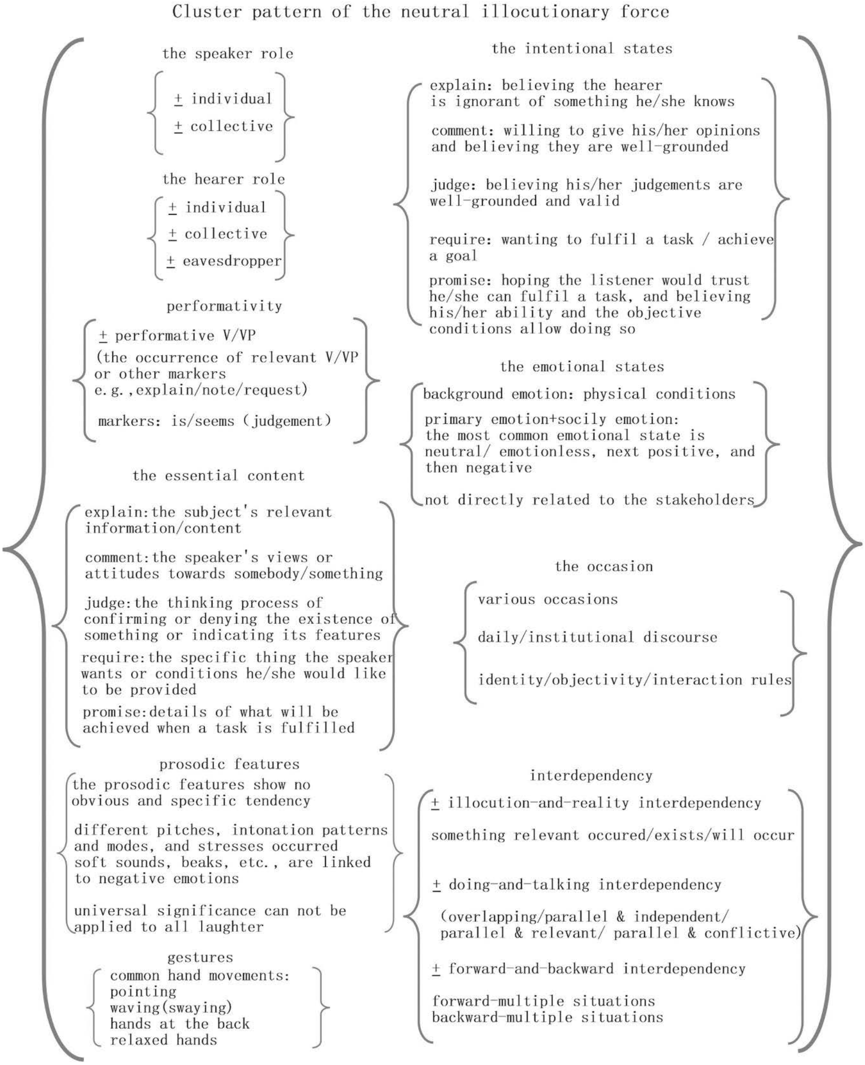

Cluster pattern of the neutral illocutionary force
the speaker role
± individual
± collective
the hearer role
± individual
± collective
± eavesdropper
performativity
± performative V/VP
(the occurrence of relevant V/VP or other markers e. g. ,explain/note/request)
markers: is/seems (judgement)
the essential content
explain:the subject's relevant information/content
comment:the speaker's views or attitudes towards somebody/something
judge:the thinking process of confirming or denying the existence of something or indicating its features
require:the specific thing the speaker wants or conditions he/she would like to be provided
promise:details of what will be achieved when a task is fulfilled
prosodic features
the prosodic features show no obvious and specific tendency
different pitches, intonation patterns and modes, and stresses occurred
soft sounds, beaks, etc., are linked to negative emotions
universal significance can not be applied to all laughter
gestures
common hand movements:
pointing
waving(swaying)
hands at the back
relaxed hands
the intentional states
explain: believing the hearer is ignorant of something he/she knows
comment: willing to give his/her opinions and believing they are well-grounded
judge: believing his/her judgements are well-grounded and valid
require: wanting to fulfil a task / achieve a goal
promise: hoping the listener would trust he/she can fulfil a task, and believing his/her ability and the objective conditions allow doing so
the emotional states
background emotion: physical conditions
primary emotion+socily emotion: the most common emotional state is neutral/ emotionless, next positive, and then negative
not directly related to the stakeholders
the occasion
various occasions
daily/institutional discourse
identity/objectivity/interaction rules
interdependency
± illocution-and-reality interdependency
something relevant occured/exists/will occur
± doing-and-talking interdependency
(overlapping/parallel & independent/ parallel & relevant/ parallel & conflictive)
± forward-and-backward interdependency
forward-multiple situations
backward-multiple situations

Figure 8.23 Cluster pattern of the neutral illocutionary force.

8.5.2 Cluster pattern of the beneficial illocutionary force

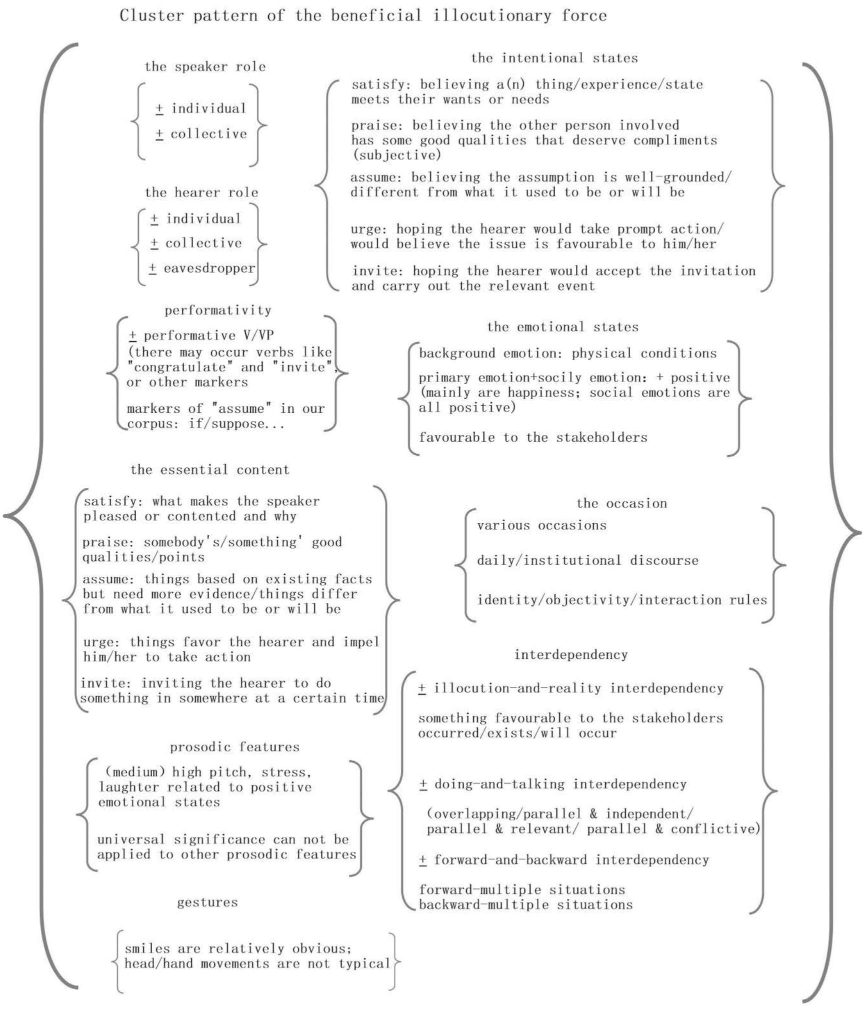
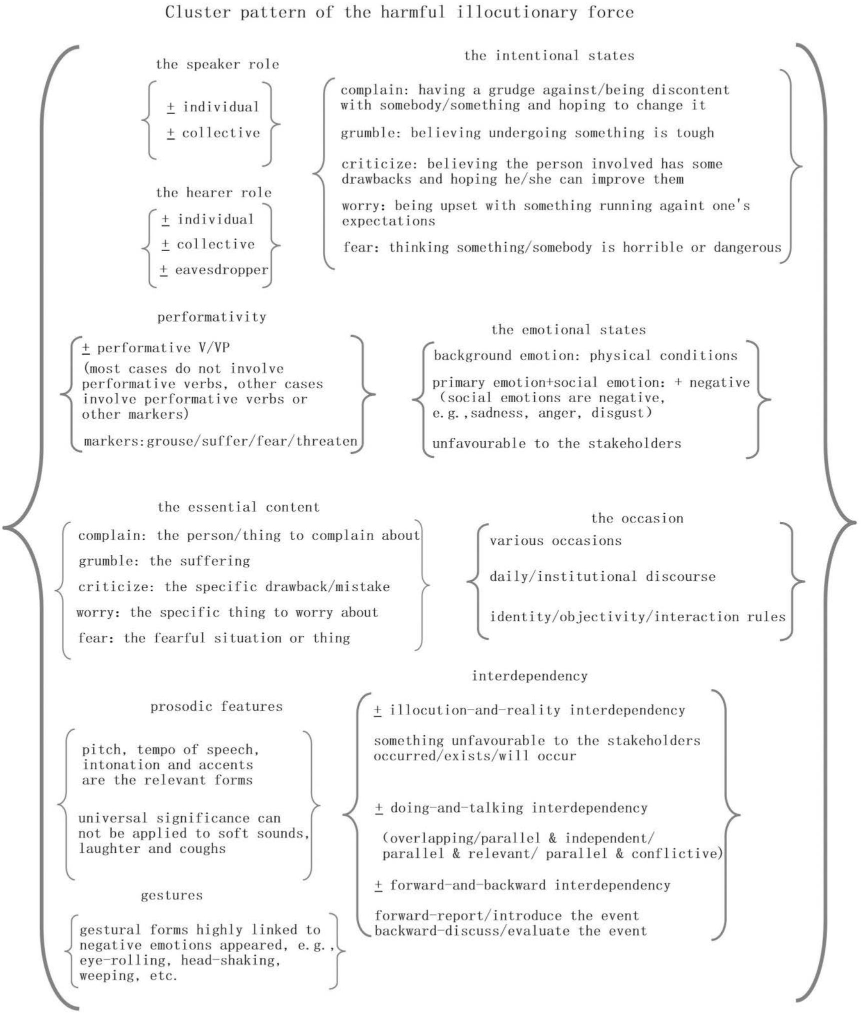

Cluster pattern of the beneficial illocutionary force
the speaker role
± individual
± collective
the hearer role
± individual
± collective
± eavesdropper
performativity
± performative V/VP
(there may occur verbs like "congratulate" and "invite",
or other markers
markers of "assume" in our corpus: if/suppose...
the essential content
satisfy: what makes the speaker pleased or contented and why
praise: somebody's/something' good qualities/points
assume: things based on existing facts but need more evidence/things differ from what it used to be or will be
urge: things favor the hearer and impel him/her to take action
invite: inviting the hearer to do something in somewhere at a certain time
prosodic features
(medium) high pitch, stress, laughter related to positive emotional states
universal significance can not be applied to other prosodic features
gestures
smiles are relatively obvious;
head/hand movements are not typical
the intentional states
satisfy: believing a(n) thing/experience/state meets their wants or needs
praise: believing the other person involved has some good qualities that deserve compliments (subjective)
assume: believing the assumption is well-grounded/ different from what it used to be or will be
urge: hoping the hearer would take prompt action/ would believe the issue is favourable to him/her
invite: hoping the hearer would accept the invitation and carry out the relevant event
the emotional states
background emotion: physical conditions
primary emotion+socily emotion: + positive (mainly are happiness; social emotions are all positive)
favourable to the stakeholders
the occasion
various occasions
daily/institutional discourse
identity/objectivity/interaction rules
interdependency
± illocution-and-reality interdependency
something favourable to the stakeholders occurred/exists/will occur
± doing-and-talking interdependency
(overlapping/parallel & independent/ parallel & relevant/ parallel & conflictive)
± forward-and-backward interdependency
forward-multiple situations
backward-multiple situations

Figure 8.24 Cluster pattern of the beneficial illocutionary force.

8.5.3 Cluster pattern of the harmful illocutionary force



Cluster pattern of the harmful illocutionary force

the speaker role

± individual
± collective

the hearer role

± individual
± collective
± eavesdropper

the intentional states

complain: having a grudge against/being discontent with somebody/something and hoping to change it

grumble: believing undergoing something is tough

criticize: believing the person involved has some drawbacks and hoping he/she can improve them

worry: being upset with something running againt one's expectations

fear: thinking something/somebody is horrible or dangerous

performativity

± performative V/VP
(most cases do not involve performative verbs, other cases involve performative verbs or other markers)

markers: grouse/suffer/fear/threaten

the emotional states

background emotion: physical conditions

primary emotion+social emotion: + negative (social emotions are negative, e. g., sadness, anger, disgust)

unfavourable to the stakeholders

the essential content

complain: the person/thing to complain about

grumble: the suffering

criticize: the specific drawback/mistake

worry: the specific thing to worry about

fear: the fearful situation or thing

the occasion

various occasions

daily/institutional discourse

identity/objectivity/interaction rules

interdependency

± illocution-and-reality interdependency

something unfavourable to the stakeholders occurred/exists/will occur

± doing-and-talking interdependency

(overlapping/parallel & independent/ parallel & relevant/ parallel & conflictive)

± forward-and-backward interdependency

forward-report/introduce the event
backward-discuss/evaluate the event

prosodic features

pitch, tempo of speech, intonation and accents are the relevant forms

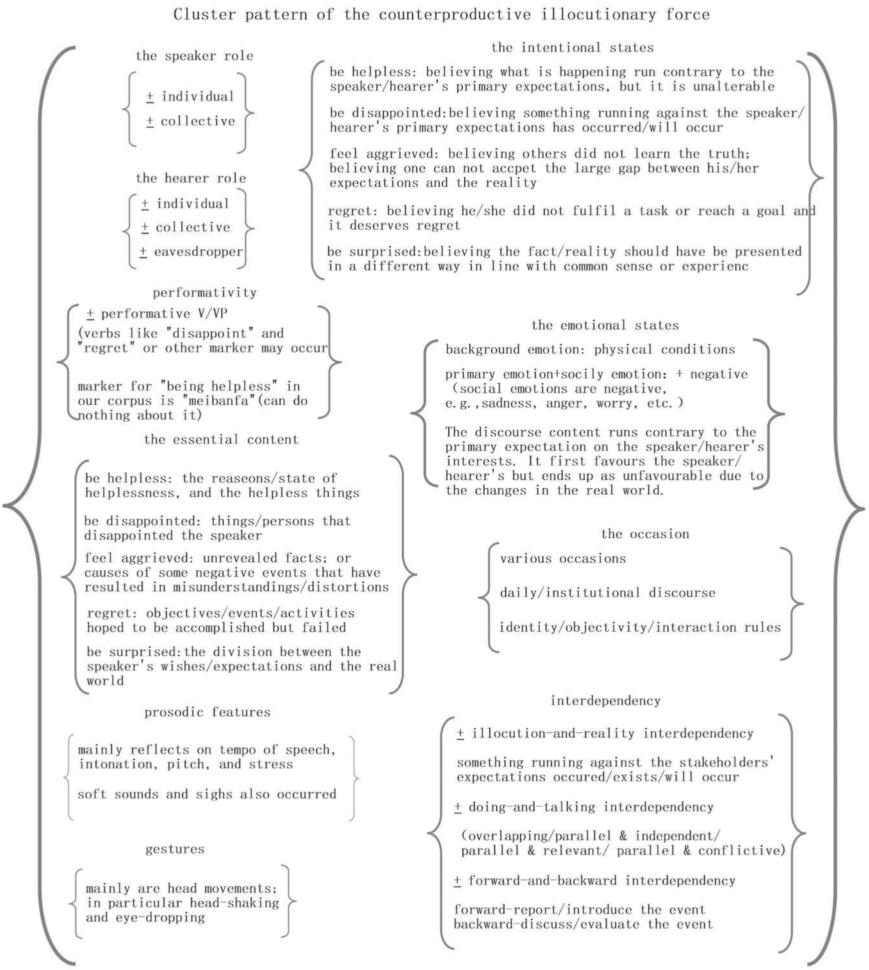
universal significance can not be applied to soft sounds, laughter and coughs

gestures

gestural forms highly linked to negative emotions appeared, e. g., eye-rolling, head-shaking, weeping, etc.

Figure 8.25 Cluster pattern of the harmful illocutionary force.

8.5.4 Cluster pattern of the counterproductive illocutionary force



Cluster pattern of the counterproductive illocutionary force

the speaker role

± individual
± collective

the hearer role

± individual
± collective
± eavesdropper

performativity

± performative V/VP
(verbs like "disappoint" and "regret" or other marker may occur

marker for "being helpless" in our corpus is "meibanfa"(can do nothing about it)

the essential content

be helpless: the reaseons/state of helplessness, and the helpless things

be disappointed: things/persons that disappointed the speaker

feel aggrieved: unrevealed facts; or causes of some negative events that have resulted in misunderstandings/distortions

regret: objectives/events/activities hoped to be accomplished but failed

be surprised:the division between the speaker's wishes/expectations and the real world

prosodic features

mainly reflects on tempo of speech, intonation, pitch, and stress

soft sounds and sighs also occurred

gestures

mainly are head movements; in particular head-shaking and eye-dropping

the intentional states

be helpless: believing what is happening run contrary to the speaker/hearer's primary expectations, but it is unalterable

be disappointed:believing something running against the speaker/hearer's primary expectations has occurred/will occur

feel aggrieved: believing others did not learn the truth; believing one can not accpet the large gap between his/her expectations and the reality

regret: believing he/she did not fulfil a task or reach a goal and it deserves regret

be surprised:believing the fact/reality should have be presented in a different way in line with common sense or experienc

the emotional states

background emotion: physical conditions

primary emotion+socily emotion: + negative (social emotions are negative, e.g.,sadness, anger, worry, etc.)

The discourse content runs contrary to the primary expectation on the speaker/hearer's interests. It first favours the speaker/hearer's but ends up as unfavourable due to the changes in the real world.

the occasion

various occasions

daily/institutional discourse

identity/objectivity/interaction rules

interdependency

± illocution-and-reality interdependency

something running against the stakeholders' expectations occured/exists/will occur

± doing-and-talking interdependency

(overlapping/parallel & independent/ parallel & relevant/ parallel & conflictive)

± forward-and-backward interdependency

forward-report/introduce the event
backward-discuss/evaluate the event

Figure 8.26 Cluster pattern of the counterproductive illocutionary force.

8.6 Summary

This section draws on the multimodal nature of illocutionary force and the multidimensionality of IFIDs. In situated discourse, the live illocutionary act is rooted in the whole multimodal interaction between the speaker and the outsider (including the physical surroundings). In return, the illocutionary force is correspondingly multimodal in nature, and thus IFIDs are multidimensional. According to the cues from the multimodal corpus, the speakers employ at least three types of devices, i.e. language structure, prosodic features, and gestures. These three, all indicating devices for illocutionary forces, bear with different function and collaborate to aid in the performance of illocutionary acts and then the production of illocutionary forces.

This section discusses the triadic interaction between emotions, prosodic features, and gestures. The four classes of neutral, beneficial, harmful, and counterproductive illocutionary acts are established according to the relation between what is discussed by speakers and the interest of discourse participants, since emotion is usually related to the self-interest of discourse participants and further influences the performance of prosodic features and gestures. Certain emotions are highly correlated to some types of illocutionary forces, and these emotions determine the configuration of speech content, prosodic, and gestural features. Moreover, some common features exist among different illocutionary forces, while this also shows the distinctive properties in the interaction of emotional states, prosodic features, and gestures in the four classes.

The section also concludes the influencing factors on the performance of illocutionary acts. Speakers' emotions are necessary in performing a felicitous illocutionary act, and they determine the production of IFIDs. This study considers the influencing factors from the three perspectives also included in the real-life social activity modelling, i.e. the perspective of the actor, activity, and system. The possible influencing factors of live speech acts and IFIDs include the following: personal style, expression capability, pragmatic intention, interdependency relation, social situation, and cultural conventions. Therefore, the performance of illocutionary act in situated discourse is collaboratively influenced by many different factors.

To clearly present the differences and similarities of the illocutionary forces in the neutral, beneficial, harmful, and counterproductive classes, four cluster patterns are concluded and drawn in the set formula.

Notes

[1] The resource diagram is presented in the formula of a set.

[2] Such selection is not entirely rational. The collaboration principle proposed by Grice is based on the assumption of a rational person, which differs from the STFE-match principle based on the 'live, whole person' proposed by Gu (2013b). In situated discourse, the speaker spontaneously presents cues (including prosody, gestures, and movements) due to the emotion or intention rather than rational choice. Of course, the author does not deny its possibility. For example, because of certain pragmatic or rhetorical intentions, the speaker may or may not use certain devices deliberately after rational thinking. However, whether natural emergence or rational choice, they are the actual situation of the 'live, whole person' and adapt to the STFE-match principle proposed by Gu (2013).

[3] Since it is situated emotion that directly influences the speaker's prosody and gestures, reported emotion is omitted in the calculation.

9 A multimodal study of illocutionary force

What has been found?

9.1 Main work of this study

9.1.1 Improving the Discovery Procedure of situated discourse

By adopting the basic thought of Simulative Modelling as well as the four analytical perspectives of what is said, what is thought, what is felt, and what is embodied, this study refers to and then revises the Discovery Procedure of situated discourse proposed by Gu (2013a), as well as introducing its design principles, basis, and characteristics to establish specific analysis procedure of live speech act in situated discourse. The Discovery Procedure is based on the 'live, whole person' principle proposed by Gu (2013), that is, the 'live whole person' performing live illocutionary acts in situated discourse and interacting with the environment in a multimodal way is the object of Simulative Modelling; its related justification is Simulative Modelling and STFE-match assumption; it features in the openness of the analysis procedure itself, which is determined by the methodology of Simulative Modelling. Subsequently, this study introduces the major structure and connotation of the Discovery Procedure, including conceptual model, situational factors, emotional states, prosodic features, and gestures.

The octet formula of conceptual analysis is used to model the properties of different types of illocutionary forces (i.e. corresponding to concept modelling, the first stage of Simulative Modelling), and it includes eight subsets: (1) speaker's role, (2) hearer's role, (3) performativity, (4) essential states, (5) intentional states, (6) emotional states, (7) occasion, and (8) interdependency.

Situational factors provide evidence for analysing the influence factors for the speaker's performance, including **contextual analysis** and **interdependency analysis**. Elements like speakers, hearers, time, place, and other factors related to speech acts in the context of act-tokens are taken into account.

Emotion, as the constitutive condition of illocutionary acts and correlating to its appropriateness, also functions as one of the influencing factors of IFIDs. In this study, three tiers of emotion are: (1) **background emotion**, (2) **primary emotion**, and (3) **social emotion**. Background emotion, the type

DOI: 10.4324/9781003251774-9

of emotion correlated to the speaker's physical condition, directly affects the speaker's prosodic features, facial expressions, etc.; primary emotion is a universal emotion across ethnicities, and social emotion is the outcome of the influence of different cultural conventions and social situations.

Prosodic features analysis refers to intonation group, prosodic pattern, and other prosodic features, with the focus set on the speech rate, tone, pitch and prosody, pause, stress, sound quality, etc. as well as some other secondary language information related to prosody.

Gestural analysis refers to **task-performances** and **non-verbal acts**. The former examines the interdependency of speech acts and talking within the framework of behaviour theory, such as the speaker performing the act of explaining while fiddling with certain objects; while the latter examines other synergistic acts under the influence of emotion, such as expressions (furious and then frowning) and postures (laughing and then leaning back).

In short, the analysis of the illocutionary force in situated discourse in this study was carried out according to this Discovery Procedure, which attempts to extend the methodological path and vision of the traditional speech acts study.

9.1.2 Construction of a multimodal corpus of Chinese situated discourse

After collecting, processing, and compiling multimodal data, this study constructed the multimodal corpus of Chinese illiterate people's situated discourse, with a whole scope of 24.43 hours as well as the transcription of 364,498 Chinese characters. Based on this mother corpus, another mini-corpus, the multimodal corpus of Chinese illiterate people's illocutionary force with 134 tokens in total was developed. The segmentation and annotation scheme for Chinese illiterate people's live illocutionary force was designed with the aforementioned Discovery Procedure. In addition, the working definition, segmentation standard, and annotation method of all the 13 tiers were organized: (1) performance unit of illocutionary force, (2) activity type, (3) turn-taking, (4) background emotion, (5) primary emotion, (6) social emotion, (7) intonation group, (8) prosodic pattern, (9) other prosodic features, (10) tasking-performance, (11) non-verbal act, (12) intentional state, and (13) interdependency were introduced in detail. The pilot annotation with Elan and Praat software and its data validation by laypeople was implemented. The pilot annotation provides the models for the later large-scope annotation of the whole corpus. The linguistic experts evaluated all the data and annotation in this corpus for their reliability and validity; the results show that the data of 134 tokens is reliable and valid for further analysis. Meanwhile, this study introduces the construction principle, constitution information, and presented the form of the corpus.

9.1.3 Multimodal corpus-based study of live illocutionary force

In this study, an illocutionary act is defined as a property exemplified by the speaker in producing an illocutionary act-token in a given social situation. Therefore, from the perspective of discourse, 134 tokens of illocutionary force (correspondingly 20 types) were classified into four groups (neutral, beneficial, harmful, and counterproductive) in the corpus according to the influence of discourse content in the conversation interlocutors' interest. All the tokens underwent conceptual analysis, and the respective conceptual models of the corresponding types were established regarding ontological properties, performance conditions, or regulations. Then, all the multimodal cues were recorded after the investigation of the 134 tokens, which prepared itself for further statistical analysis. Using the data annotated by Elan and Praat, the sample tokens from the different 20 types of illocutionary force were analysed in depth in the Discovery Procedure framework.

This study analysed the multimodal interaction in situated discourse after the classification and analysis of illocutionary force in situated discourse. The work includes the following:

1 Summarizing the type and functions of live speech acts and IFIDs in situated discourse;
2 Describing the interaction pattern of live illocutionary force and emotions in situated discourse;
3 Analysing the influence factors for the utilization and co-occurrence of live speech acts and IFIDs.

9.2 Major conclusion

Adopting the multimodal corpus linguistics approach, this research focused on figuring out how speakers use diverse devices to implement illocutionary acts and then express live illocutionary force in situated discourse. In this sense, it shed light on the interaction between illocutionary force and emotions, prosody, and gesture. Additionally, the rule of these factors integrating to produce a variety of live illocutionary forces as well as the emergence pattern of different IFIDs are elaborated. The major findings can be summarized in the following aspects.

9.2.1 Multimodal nature of illocutionary force and the multidimensionality of IFIDs in live speech

In situated discourse, the illocutionary act is rooted in the whole multimodal interaction between speakers and outsiders (including physical surroundings). Thus, the illocutionary force is multimodal in nature. Such multimodality is fully displayed when speakers utilize other resources than speech to perform illocutionary acts and therefore produce illocutionary force. More specifically, prosodic features (including pitch, speech rate, stress sound, and pause) and gestures such as pointing, frowning, and swaying work collaboratively to help speakers perform illocutionary acts effectively. Therefore, by taking the viewpoint of systematic theory, we can conclude that all the expressive devices and IFIDs during this process in various contexts comprise a resource system of conveying illocutionary force. Based on the multimodal corpus of illocutionary force constructed in this study, the resources employed by the speakers can be basically divided into three categories: (1) linguistic structure (speech act verbs, verb phrases, discourse markers, etc.), (2) prosodic pattern (including pitch, speech rate, pitch, and pause), and (3) gestures (pointing, facial expressions, postures, etc.). This conclusion to a great extent validates the claim of Austin (1962: 73–74) that alternatives to implement illocutionary forces like mood, tone of voice, cadence, emphasis, and accompanying gestures (such as blinking, pointing, shrugging, and frowning) all prove to be IFIDs. Meanwhile, it also validates the concept of the ‘whole man’ in Firth’s sense with emotion, which can be embodied in speech, prosody, and gesture. Clearly, this statement is also concordant with the principle of the ‘live, whole person’ proposed by Gu (2013b), that is, the subject engaged in the situated discourse activities is a ‘live, whole person’ with dynamic sounds, emotions, and gestures, rather than an idealized ‘talking head’ who can only exchange information via language.

In the category of prosodic features, pitch, speed rate, pitch, and emphasis are the main tactics for speakers performing various illocutionary forces; thus, they frequently emerge in various groups of illocutionary forces. While in the category of gestures, the frequency of head and hand movements as well as facial expressions is relatively higher when performing certain illocutionary forces. However, when compared with the prosodic pattern, gestures turn out to be less general due to the fact that apparent movements or expressions will not appear in all classes of illocutionary force.

As demonstrated in various illocutionary force cases, the contribution of IFIDs on the hearer’s perception and interpretation of illocutionary forces varies in different aspects. Generally speaking, it can be concluded into two patterns: (1) One or several IFIDs occupy the dominant position in the performance to determine the type of certain illocutionary force, and (2) several IFIDs collaborate in the performance without a dominant device.

In short, the speakers can resort to multiple tactics to perform illocutionary acts and thus produce illocutionary forces. The speech act verbs/phrases, syntactic structures, or utterance content are frequently used devices in the perspective of language structure. Meanwhile, prosodic pattern and physical gestures also make a difference in the performance of illocutionary forces. In the same vein, the hearers also need to perceive the speaker’s illocutionary force from language structure, utterance content, prosody, and gesture. However, the frequency, concurrence, and order of these IFIDs are affected by different situations, speakers, and intentions and are therefore complicated. In this sense, the function and effect of certain illocutionary force productions should be analysed in specific cases.

9.2.2 Interaction pattern of emotion, prosody, and gestures

The speaker's emotional state is an indispensable component in performing felicitous speech acts, and in return, it can be demonstrated by IFIDs like speech content, prosody, and gestures. The way that emotions, prosody, and gestures interact and correlate somehow varies between different classes of illocutionary forces. To further illustrate this, the classification of illocutionary force in this research is set according to the relation between the core content of illocutionary force and the interest of the speaker as well as the hearer, which might influence the emotional state of the speaker. After such categorization, we found that the act-tokens in a specific classification are highly correlated with a certain emotion. In other words, certain emotions tend to appear in a specific class of illocutionary force. Furthermore, correspondingly, these emotions are significantly related to the frequent presentation of certain discourse content, prosodic features, and speaker gestures.

This research describes the similarities and differences in the interaction between emotions, prosody, and gestures of the four groups of illocutionary force and lists them through the ensuing feature distribution maps.

1 Commonality of four groups of illocutionary forces

The speaker and hearer can be either individual or collective. Plus, the hearer may not be the one to whom the speaker talks, which means there are observers or hearers.

In four groups of illocutionary force, speech act verbs/phrases or other syntactic structures can all be used. However, there are no explicit speech act verbs/phrases in most cases, and speakers occasionally utilize some set syntactic structures explicitly to convey certain types of illocutionary force.

In situated discourse, the social situations where most of the tokens of illocutionary forces are embedded are diverse. The scenarios can be in our daily life or institutions. Moreover, the expression of illocutionary force should be consistent with the identity, goals, and acquired social relationship roles of the speaker/hearer.

In terms of the interdependency relation of illocutionary force, most of the instances of illocution-and-reality interdependency are that 'the relevant events have occurred/exist/will occur'; and the doing-and-talking interdependency of speech acts include overlapping, parallel and independent, parallel and related, and parallel and conflicting, while their forward-and-backward interdependency is diverse and complex.

2 Distinctions of the four groups of illocutionary force

In the neutral group, speakers in most cases show no certain dominantly presented basic emotions, followed by the emotion of happiness; logically, the frequency of anxiety and anger is relatively low; in terms of the social emotions, positive-interested/proactive/enthusiastic, positive-confident, neutral-indifferent/default rank in top three; and prosodic features do not show any distinctive evidence compared with the other three groups, but hand movements like pointing/directing, swaying, putting one's hands to the back, letting one's hands drop appear more frequently than other groups.

In the beneficial group, the speakers usually show positive emotions. Their dominantly presented basic emotions are happiness and social emotions are almost positive, including satisfied/approving, interested/proactive/enthusiastic, pleased/proud, expectant, appreciative/grateful, respectful/admiring, etc. And (medium) high pitch, stress, and laughing are prosodic features significantly related to these social emotions in the beneficial group; in terms of gestures, smiling/laughing are highly relevant to positive social emotions, while head and hand movements are not very typical.

In the harmful group, the speakers usually present negative emotions, especially when s/he is performing illocutionary force against his/her own interest. Their social emotions are generally negative. Pitch, speech rate, intonation, and stress are still the main prosodic features significantly correlated to the speakers' background emotions; gestures such as rolling the eyes, shaking the head, and wiping tears appear with negative emotions.

In the counterproductive group, the negative social emotions reflect the change of emotions caused by the discrepancy between the speakers' expectations and reality. Speakers display helplessness, frustration/depression, and dissatisfaction when performing this group of speech acts. The basic emotions that are significantly related to prosodic features include anxiety, anger, or default-situated-emotion. These emotions are mainly related to speech rate, intonation, pitch, and stress. Besides, other cues such as soft sounds, sighs, and head movements also appear with the negative emotion. The gestures related to the negative emotion are mainly displayed with the head, particularly shaking one's head, looking down (dropping the eyes), and pondering. Hand movements also exist, but these are not significantly distinctive.

Undeniably, the conclusions drawn are suggestive in nature. In many cases, they are complicatedly intertwined and overlapped while the exclusive, specific one-to-one relation is absent. This is because, on the one hand, a certain prosodic feature (e.g. intonation, occurrence of pause) and gestures (e.g. smiling, crossing one's hands) are not exclusive to a certain type of illocutionary force. On the other hand, one type of illocutionary force might present more prosodic features and gestures under different circumstances. Besides, three types of emotional states (background emotion, social emotion, basic emotion) have different degrees of correlation to prosodic features and gestures in the four groups of speech acts. This suggests that it is inadequate and inaccurate to judge the type of illocutionary force from single cues like intonation, expressions, and postures, since IFIDs are multidimensional in situated discourse. Therefore, it is imperative to examine IFIDs from the multimodal perspective and judge the type of illocutionary force, interaction, and correlation between these devices from all the multimodal cues. In the same vein, the hearers in authentic interaction also apply this mechanism to interpret the conveyed illocutionary force. To sum up, from the perspective of system theory, speakers employ a variety of modal resources (elements) that interact and correlate with each other to perform infelicitous illocutionary force under the influence of a variety of complex factors.

9.2.3 Influence factors of multimodal illocutionary force and IFIDs

Emotional state is a crucial criterion for judging the felicity of the speaker's illocutionary force, and it is also an important factor that influences such IFIDs as speech content, prosodic features, and gestures. However, in situated discourse, emotional state is not the only factor that affects the speaker's multimodal illocutionary force. Moreover, the emergence of different IFIDs is under the co-influence of emotional state and other factors. To better illustrate such a complicated mechanism, this research applies three perspectives (i.e. the perspectives of the actor, activity, and system) proposed by Gu (2009) to situated discourse modelling. It is inferred that possible influencing factors of multimodal illocutionary force and IFIDs include personal style, expressive capability, pragmatic intention, interdependency relation, situational factors, and cultural conventions.

1 Personal style refers to the individual introverted or extroverted, simple or sophisticated personalities, which can influence different expressions in the same type of speech act.
2 Expression capability refers to the individual competence of utilizing diverse resources in words and phrases, language structures, rhetoric skills, and other non-verbal expressions, which can be understood as multimodal capacity.
3 Pragmatic intention means whether and how the speaker would like the hearer to understand his/her real intention or emotion of the interaction.
4 Interdependency relation includes three aspects: (1) forward-and-backward interdependency, i.e. before-and-after relation of the utterance, (2) illocution-and-reality interdependency, i.e. the relation between illocutionary force and what is happening in the here-and-now behaviour setting, and what happens beyond the here, and (3) doing-and-talking interdependency, i.e. the relation between doing and talking.

5 The social situation refers to different situations, activity types, or surroundings, which confine and shape the way that the speaker performs the illocutionary force to a certain extent.
6 Cultural conventions refer to the manner of expression that is direct and explicit or indirect and implicit and more appreciated in a particular culture.

In short, how the speaker utilizes the multimodal resource system and performs illocutionary acts in situated discourse is under the influence of multiple factors. These factors work in a collaborative way to help the speaker perform felicitous illocutionary force.

9.3 Innovation and value of the research

9.3.1 Innovation

9.3.1.1 Theoretical innovation of speech act theory

Taking an ever-pioneering perspective and methodology, this study attempts to deepen the understanding of speech act theory, which is the classic topic in pragmatics. Despite the fact that a few researchers have alluded to the importance of prosody and gestures to the performance of speech acts, systematic research on this topic has always been absent since then. This study believes that in situated discourse, the live illocutionary act is one of the behaviours in the whole multimodal interaction between speakers and outsiders (including physical surroundings). In return, the illocutionary force is correspondingly multimodal in nature, and thus IFIDs are multidimensional. Moreover, speakers can resort to multiple tactics to perform illocutionary acts and thus produce illocutionary forces. Additionally, emotional state is an indispensable component in performing speech acts, and in return, it can be demonstrated by IFIDs like speech content, prosody, and gestures. This study also concludes that IFIDs in live speech range in their scope from language structures and prosodic features to non-verbal gestures, etc. Some common features exist among different types of illocutionary forces while they also show distinctive properties among the interaction of emotional states, prosodic features, and gestures in neutral, beneficial, harmful, and counterproductive groups. Meanwhile, the study deduces the possible influential factors for the performance of illocutionary acts in situated discourse through the observation of live data. All the arguments above are based on the assumption that the speech act is one type of ordinary human behaviour and is drawn with various methods in the multimodal corpus. They are expected to deepen the understanding of speech act theory and expand the scope of pragmatics.

9.3.1.2 Innovation in research content and data resource

This study focuses on exploring how the triadic interaction of emotion, prosody, and non-verbal acts produces a variety of live illocutionary forces, that is, describing the mechanism of the devices carrying various illocutionary forces adopted by the speakers and their characteristics, displaying regularity, and possible influential factors of the IFIDs. Although there has been a considerable amount of study on speech acts that are invariably the core of research in pragmatics, the existing research mostly starts the discussion from the internal structure of language (such as vocabulary and syntax) and fails to reflect the truth that authentic communication is a multimodal interaction with the 'live, whole person' engaging. Several scholars have mentioned the contribution of other cues in live speech (such as prosody, expressions, and actions) to the conveyance of discourse meaning; however, this is rather rare when researching speech acts. Therefore, it is not too bold to claim this study as an innovation in the field of pragmatics to investigate the relationships between emotions, prosody, gesture, and the performance of the illocutionary force. Furthermore, this study takes illocutionary forces produced by Chinese illiterate people in live speech as the research object and then transcribes the multimodal interactions of the illiterate people using modern audiovisual technology in the first place. Such a relatively rare corpus is of high value to further analysis and discussion, which will open up a new dimension for situated discourse and multimodal linguistic study, and accumulate experience for following research.

9.3.1.3 Innovation in research approach and methodology

This study is an attempt to study pragmatic issues with a multimodal corpus approach. Currently, the corpus approach is a growing adoption in the field of pragmatics. Meanwhile, multimodal corpus pragmatics has developed prosperously with the development of corpus linguistic study and the multimodal approach. Therefore, it is of great significance to introduce multimodal corpus linguistic methods into pragmatics with such a background. By adopting the technology of multimodal corpus linguistics and the basic thought of Simulative Modelling, this study endeavours to break the ceiling of traditional speech act study to develop an analysis framework for live illocutionary forces investigating their relationship with emotions, prosody, and gestures in a dynamic and multidimensional way. The application of a multimodal corpus linguistics method to examine illocutionary force is a pioneering action to solve classic pragmatic issues from the corpus. It is of great value to bring out a more comprehensive and accurate interpretation of the significance of language usage, and provides a reference to future theoretical discussion and case study in pragmatics. Of course, the adoption of multimodal corpus methods to study such pragmatic issues as speech acts is still in its infancy, and more empirical research is needed to further elaborate it.

In addition, this study treats language communication as an intricate system, and launches qualitative and quantitative analysis afterwards inspired by mixed-method research, which is an innovation to employ in pragmatic study.

9.3.1.4 Establishing multimodal pragmatics

Although a pilot study on speech acts is illustrated in this book, we should bear in mind that multimodal pragmatics is far more than a speech act study. As a matter of fact, inspired by the integration of multimodality and corpus pragmatics, this book is concerned with a more general multimodal framework that facilitates the exploration of pragmatic questions by using corpus methods; this further provides an approach in pragmatics to verify that human interaction is multimodal in nature, and that meaning in discourse is created through an interplay of an array of modalities. The relationship between verbal and non-verbal cues in relaying pragmatic meaning is still very intractable, while the multimodal approach can provide a novel perspective to handle this intriguing issue. Based on this, empirical research claims that the scope and methods of pragmatic studies will be enriched and that classic pragmatic theories could be further developed toward *multimodal pragmatics*, by demonstrating how the study on speech acts in situated discourse benefits from the multimodal approach in the field of pragmatics. Certainly, we have to admit that multimodal pragmatics is still in its infancy and that more theoretical and practical investigation is needed. More traditional research topics in pragmatics could be reinvestigated in the multimodal corpus approach framework, including conversation implicature, presupposition, (im)politeness, and identity construction. Additionally, pragmatics is now to be seen not as a mere component of language, but rather as 'a general cognitive, social, and cultural perspective' (Verschueren, 1999: 7). In this way, multimodal pragmatics can empower researchers with the tools to explore more cognitive, social, and cultural issues.

In summary, by including non-verbal resources in the observatory data, the scope and methods of pragmatic studies will be enriched and the classic pragmatic theories will be enhanced and developed towards authentic face-to-face interaction. This is what multimodal pragmatics is about.

9.3.2 Research value

9.3.2.1 Theoretical value

Starting from the viewpoint of multimodal illocutionary force, this study has developed a multimodal analysis framework of live illocutionary force and investigated it by employing a multimodal corpus approach. The major findings can be concluded in the following three aspects: (1) illocutionary force is multimodal in nature and the IFIDs are multidimensional; (2) there exist not only some common features but also distinctive properties among the interaction of emotions, prosody, and gestures in different groups of illocutionary force; and (3) speakers' emotions are necessary in performing a felicitous illocutionary act, which also determines the production of IFIDs. And other possible influential factors of live speech acts and IFIDs include personal style, expression capability, pragmatic intention, interdependency relation, social situation, and cultural conventions. This study is designed to investigate illocutionary force/act and IFIDs from both original and extended viewpoints, and to examine speech acts from the perspective of the 'live, whole person' in live speech and at the height of ordinary act theory, which further justifies establishing the field of multimodal pragmatics. More generally, this study explores a more general multimodal framework that facilitates the exploration of pragmatic questions by using corpus methods, verifying human interaction as multimodal in nature and meaning in discourse as created through an interplay of an array of modalities.

9.3.2.2 Method value

Multimodal corpus-based pragmatic study has become the research frontier in international linguistics academia. However, it is such a regret to see far too little relevant study in China. More research and practice in related fields are urgently needed presently to provide reference and experience. As this study uses the idea of a multimodal corpus approach to study pragmatic problems, a multimodal analytic framework is constructed, which is expected to give more reference to similar research. Meanwhile, the multimodal segmentation and annotation scheme is developed. It should be noted that the multimodal segmentation and annotation scheme on illocutionary act/force is rare in this domain. The scheme developed in this study is a further advancement of speech behaviour annotation in corpus pragmatics research, and it is a typical study of speech behaviour using multimodal corpus methods. Therefore, this study's approach and methodology expect to inspire future inquiry on pragmatic and multimodal studies in China.

9.3.2.3 Application value

This research integrates pragmatics, phonetics, and non-verbal communication studies and is designed to investigate how speakers employ various modalities to perform illocutionary acts and then convey illocutionary force. Multimodal HCI is managed based on tracking and recognizing the speaker's sound, facial expression, posture, and emotion (Tao et al., 2011: 31). Therefore, the related research outcomes in this study can provide more pragmatic evidence for developing the framework for multimodal HCI (including speech, expression, and posture) and enhancing speech recognition, the naturalness of virtual speakers, and user experience.

Since speech acts are regarded by Searle as the 'minimum units in human communication' (1969: 21) and other researchers in speech act theory, this has offered new ideas for the development of HCI systems. Relevant research on speech act has laid a theoretical basis for conversational agents in the domain of natural language processing. Some computational linguists have put forward the belief, desire, and intention model to conceptualize the performance of speech acts to make them computable and improve conversation quality (see Feng, 2014; Feng & Yu, 2015). From the perspective of some international R&D projects on HCI, it is relatively common to use speech acts (also referred to as 'dialogue acts') as the basic unit of communication for segmentation and processing. In the research and development of dialogue acts, a series of elements such as participants, task, activity types, and scenario (see Gibbon, Mertins, & Moore, 2000: 5–11) should all be taken into consideration and be further processed. In view of this, the author includes all the above-mentioned elements in the multimodal speech annotation and analytic framework in this research, which is conceptually consistent with that in the R&D of HCI. Additionally, the rules of how emotion and illocutionary force interact revealed in this study are helpful for the R&D of human–computer multimodal interactions, primarily for providing some linguistic evidence for configuration recognition. Moreover, this research also endeavours to provide some linguistic reflection on the promising transition from the simple dialogue between virtual human (which is featured by multimodal content and monomodal carrier) to multimodal interaction (such as robots) in this aspect.

9.3.2.4 Social value

This research takes illocutionary forces produced by Chinese illiterate people as the research object and explores their speech acts in live speech. This has certain significance for enriching the research and practice on helping this special group of people become more literate and competent in communication, which exactly represents the author's concern about the social signification of linguistic study. It should be pointed out that people need to learn a variety of grammatical rules and vocabulary structures in different contexts to achieve different communicative purposes. However, when it comes to the colloquial context, to what extent the language knowledge is conventional and formalized does not secure the dominant position in human communication, the reason for which is that the hearer can infer the speaker's intention via analysing various resources and thereby comprehend the underlying intention of the speaker even if the latter has used the wrong vocabulary or syntactic structure (Olson, 2013: 22). Therefore, in situated discourse, illiterate people can effectively convey their intentions and achieve verbal communication by employing a variety of resources and tactics other than speech itself. Although the illiterate and peoples without letters are somewhat hindered, they still manage to pass down their civilization and culture by the method of talking and remembering. In this sense, the traditional methods of oral communication are worthy of further research, and the persistent negativity and even discrimination towards these groups should be corrected.

Additionally, in terms of method, the multimodal corpus linguistic method used in this study also shows social concerns and a broader application value to some extent. Since the information stored in a multimodal corpus is enormous, the problem of transforming information into data according to research purposes and effectively storing, retrieving, and applying this data calls upon the collaboration of scholars from numerous disciplines, such as linguistics, information technology, and sociology. From the perspective of future development, the framework and methods of multimodal corpus have the potential to be applied not only to research fields of linguistics but also to other disciplines in humanities and social sciences. It enjoys theoretical value and application value simultaneously and is expected to make a difference in fields like theatre performance, social psychology, political behaviour, commercial behaviour, mass media, and military criminal investigation. At present, international academic research already utilizes big data and corresponding analytic methods in the fields of humanities and social sciences, which are referred to as digital humanities and e-social science. This research method and corresponding computing technology help researchers investigate people's interaction patterns and social behaviours through large-scope data analysis. In this sense, if we build a massive database with the multimodal data acquired by investigations on people's interaction behaviour, and analyse this data with the help of effective and reasonable analysis technology, there is every likelihood that people's behaviour patterns can be further studied.

9.4 Limitations of the study and reference for future research

9.4.1 Limitations

9.4.1.1 Limited recording quality and corpus scope

This research is to construct an analysis framework of live illocutionary forces in situated discourse to describe and analyse the interaction pattern for emotional states, prosodic features, and gestures in live speech. As the corpora collected in this research are all live speech and both the recording equipment and experience are limited, it is unavoidable that noises other than the conversation itself mixed in the audio make the audio less than ideally pure and clear. Although it has little influence on the analysis of specific token, it might cause some interference to such large-scope quantitative statistical analysis. In addition, due to the complexity of multimodal corpus construction, the corpus sample of this study is relatively small. Therefore, the findings drawn in this study are suggestive or indicative, which require further verification from more finely recorded corpus at a larger scale. The author believes that quantitative analysis based on large-scope multimodal corpus seems to be worthy of future research.

9.4.1.2 No involvement in perlocutionary acts

As this study has paid more attention to talking and doing in actual corpus analysis and more importantly, there is no relevant corpus of the audiences' situated discourse, there is almost no analysis of perlocutionary acts. This aspect could be marked in the agenda for future research in an effort to improve the completeness of the research object. At present, many international scholars have used multimodal corpora to study hearers' responses (Buschmeier et al., 2014; Knight & Adolphs, 2008), which provides reference for future research on perlocutionary acts. Gu (2013) conducted a preliminary exploration on perlocutionary acts via analysing students' expressions and other responses during lectures in the multimodal corpus to investigate teachers' perlocutionary acts including lecturing.

9.4.1.3 No consideration of dialect factors

The corpora collected in this research are mostly dialects or Mandarin with dialect accents. The regional dialects include Hongchao of Jianghuai Mandarin (Lu'an, Luzhou, Anhui), Taihu of Wuhu Dialect (Tonglu, Zhejiang, Fengxian, Shanghai), and Qinglai of Jiao Liao Mandarin (Shaoxian, Rizhao, Shandong). Although distributed geographically, they are not typically representative. Despite no suggestion from previous studies that dialects might have some variation in intonation and emotion, whether dialect factors affect the expression of illocutionary force requires further verification in more corpora.

9.4.1.4 Other limitations

In Section 8.2, the author uses factor analysis to investigate the correlation between the emergence frequency of various emotional states, prosodic features, and gestures. Some inexplicable related items occur in the distribution maps of four classes of illocutionary force, such as the disheartened background emotion co-occurring with coughing in prosody, the sad background emotion co-occurring with the movement of stretching forward one's hands, the positive, conceited/proud social emotion co-occurring with rubbing one's eyes/wiping tears in gestures, the negative, contemptuous/despising social emotions co-occurring with fiddling, holding, and picking (items) through gestures. The reasons for this could be two-fold: First, the overall statistical sample is relatively small. Although there are 134 act-tokens in this corpus that meet the sample minimum of being generally statistically significant, when it comes to specifically analysing the correlation between emotion, prosody, and gestures in the four classes, the statistical sample is further reduced; second, in each classes of illocutionary force, the emergence frequency of some IFIDs (prosodic features, gestures) is low (all single digits); the same happens to certain emotional states (all single digits), so there could be correlations beyond common sense when performing a factor analysis on them. Therefore, although the statistical data of this study is strictly operated in accordance with the actual situation, the results can only be indicative. In the future, a large-scope corpus will be used for stricter and more scientific statistics analysis. Furthermore, as the exploration of the influencing factors for live speech acts and IFIDs in situated discourse, this study focuses on a detailed examination of emotional states while lacking sufficient examples to support it. Questions like exactly how many influencing factors there are and how they each function remain unsolved and underline the necessity of conducting more scientific and objective research with a larger-scoped corpus.

9.4.2 Future agenda for this study

9.4.2.1 Research on complete speech acts based on the multimodal corpus

This research mainly examines illocutionary acts and corresponding illocutionary forces, and involves perlocutionary acts (represented by the conceptual analysis of interdependency of illocutionary force) and locutionary acts. The illocutionary act has always been the focus of traditional pragmatics research, since Austin first proposed speech act theory that said that the complete speech act includes locutionary act, illocutionary act, and perlocutionary act. In the future, modern audiovisual technology could be used to record the hearers' multimodal corpora, including their facial expressions, physical responses, and prosodic features. And then by analysing these clues to see how the perlocutionary act is performed, a complete multimodal study of the speech act in situated discourse could be conducted. In addition, for the synergy and weight among various IFIDs when the speaker is expressing illocutionary force in the situated discourse, and their role when the hearer is perceiving and determining the illocutionary force, this study as preliminary research is in need of further investigation.

9.4.2.2 Comparative study of illiterate and literate people's speech

The multimodal corpus of Chinese illiterate people's situated discourse constructed in this study is relatively rare and enjoys high linguistic value, which can be used to study illiterates' language in the future. Through quantitative analysis, this study finds that illocutionary act verbs/phrases and other syntactic structures are not the core IFIDs that indicate the type of illocutionary force. The speaker uses a variety of devices such as prosodic features and gestures to perform the illocutionary act and express illocutionary force in a coordinated and comprehensive manner. Is such a pattern of less usage of performative verbs/phrases and other syntactic structures exclusive to illiterate people or is it the same as that of literate people? In addition, illiterate people have no Chinese character representation for their psychological vocabulary system but can sense meaningful chunks in their speech stream. Assuming that their psychological vocabulary corresponds to the phonetic vocabulary, will literate people have one more subsystem (Chinese character system) after learning the Chinese characters? Or there is only one system of psychological vocabulary, and phonetic words and Chinese characters are two parallel representations corresponding to one system (that is, both phonetic words and Chinese characters have a direct correspondence to their psychological vocabulary)? Or is the text only the secondary representation of speech flow, that is, the character and psychological vocabulary have indirect correspondence that needs the speech flow to break? What are the similarities and differences between the acquisition and access of psychological vocabulary by illiterate and literate people? These issues need to be addressed by subsequent comparative studies on illiterate people's and literates' discourse (Gu, 2011: 36–38; Huang, 2014b).

9.4.2.3 Multimodal corpus approach to human interaction research

As mentioned in Section 4.2.4, prosodic features are an important focus of interactional sociolinguistic study. Gumperz and other scholars consider prosody as an important clue in social interaction. In fact, methodologically, an important concern of interactional sociolinguistics is finding a verifiable method for qualitative analysis to explain the linguistic and cultural differences in real social communication (Gumperz, 2003). The basic ideas and techniques of multimodal corpus linguistics are an exact and verifiable research method that can provide verifiable quantitative methods for a linguistic research field such as sociolinguistics and can be applied to other humanities and social sciences research. Therefore, using the multimodal corpus approach to open up a venue for humanities and social discipline study, including social interaction and human behaviour is an important path for multimodal corpus linguistics. In the future, on the basis of existing results and methodologies in this research, we can use the theoretical basis and analytical methods of interactional sociolinguistics to conduct linguistic communication research, including Goffman's (1974) empirical description of interactive practice, Garfinkel's (1967) ethnomethodology, and the methodology and research findings of Sacks (1984) and Schegloff (1984, 2007). From the perspective of interactional linguistics, the discourse in people's interactive communication is rooted in certain when-and-where circumstances. Therefore, research on linguistic phenomena should not be isolated but should connect the before-and-after colloquial context and speakers' performance, which means at the level of data collection, it is live speech in daily activities that should be recorded. Research combining syntax, semantics, pragmatics, and prosody are not in a minority while it adopts the multimodal corpus approach to systematically considering gestures has much room for development. In this type of research, scholars pay attention to how interlocutors' communicative intentions and locutionary acts are realized through modal resources. At present, foreign multilingual corpus-based interactive linguistic studies are gradually emerging (e.g. Li, 2014). In short, such theoretical foundations and research paths not only focus on the understated clues in linguistic study including prosody, gestures, and movements, but more importantly, insist that daily discourse is an important component of social activities and are in line with the basic thoughts of this research, that is, addressing language issues under the basic framework of ordinary speech act theory. Studying linguistic communication issues from the perspective of ordinary speech acts and adopting the novel multimodal corpus approach to carry out research will definitely promote the development of related theories.

9.4.2.4 Further development on the corpus construction

To fully explore and tap into a multimodal corpus is also an important aspect of multimodal linguistic research. So far, the relative boost in corpus pragmatics shows no sign of pausing, which is worthy of more attention and exploration. In this study, the processing of the multimodal corpus of Chinese illiterate people's situated discourse whose whole scope is 24.43 hours meets the needs of this research but still falls short of more detailed annotation and in-depth utilization, especially in the aspect of improving the annotation schemes on prosody and gestures. Therefore, we will continue to refer to more international experience on corpus processing, deepen the self-construct multimodal corpus, and expand and share the corpus if possible, so as to meet future development. And we will also make full use of modern audiovisual technology to serve the humanities and social sciences and provide more multimodal data for digital humanities and e-social science.

9.5 Summary

This chapter summarizes this research's work and conclusions, analyses its innovation, research value, and limitations, and proposes possible future agendas.

This research's innovative work is as follows: revising the Discovery Procedures of illocutionary force in situated discourse; constructing the multimodal corpus of situated discourse, and conducting analysis on live speech acts and IFIDs based on the self-constructed multimodal corpus.

This study's conclusions include the following: The illocutionary force is multimodal in nature and IFIDs have multidimensionality; speakers' emotions are necessary in performing a felicitous illocutionary act and determine the production of IFIDs (discourse content, prosodic features, gestures, etc.). In different classes of illocutionary force, the interaction pattern between emotional states, prosodic features, and gestures varies; the possible influencing factors of live speech acts and IFIDs include personal style, expression capability, pragmatic intention, interdependency relation, social situation, and cultural conventions. And from the perspective of systems theory, speakers use various modal resources (elements) that all work in a collaborative way to help speakers perform illocutionary acts.

As an attempt to use the multimodal corpus approach to study pragmatic issues from the perspective of situated discourse and with the illiterate people's data, this study locates itself at the frontier of pragmatics. And it has its unique value in theory, practice, and social concern, and provides important reference for further study of pragmatic issues based on multimodal corpus.

This study's limitations are the limited recording quality, overall corpus scope, no involvement in the perlocutionary acts study, and no consideration of dialect factors. Further study could be conducted in the following aspects: multimodal corpus-based study of the complete speech acts (including perlocutionary acts), comparative study of the discourse of literate and illiterate people, multimodal corpus approach to human interaction, and further development in corpus construction.

10 Developing multimodal pragmatics

10.1 Starting from multimodal corpus pragmatics

Pragmatics believes in attaching more attention to daily authentic conversation data. Levinson's 1983 foundational text of pragmatics has a whole chapter on conversation analysis, which blurs the borders between traditions. Representative journals in this field such as *Pragmatics* and *The Journal of Pragmatics* have published several papers by conversation analysts. In conversation analysis, researchers believe in the importance of gestures in interpreting interlocutors' meaning and hearers' feedback. Therefore, the real analysis of 'non-verbal cues in authentic conversation' has due attention paid to it by analysts.

Corpus pragmatics, combining both key methodologies of corpus linguistics and pragmatics, is a relative newcomer. The inquiries adopting this approach relocate themselves along the data of natural conversation instead of self-constructed examples. However, analyses of pragmatic functions in the spoken corpus have not included the quantitative calculation of the interplay between gesture and speech in human communication, largely due to the lack of an adequate annotated corpus. Fortunately, the multimodal corpus approach prompts a better channel to represent more information and contextual cues in natural interaction, and provides the possibility of quantitative studies.

Therefore, this author attempts to elaborate on how pragmatics studies could benefit from adopting a multimodal corpus approach and what specific methods are exploited in such studies. We prefer to adopt the name 'multimodal corpus pragmatics' to refer to this inquiry. By analysing speech acts in situated discourse as shown in this book, it aims to present the practice in which pragmatic functions can be further investigated with the consideration of gestures and prosody. All these elements, fundamentally speaking, should be included in studies towards authentic communication.

10.1.1 What can a multimodal corpus provide?

Using the technologies facilitated by the multimodal corpus, researchers can explore many speech issues with relatively rich resources beyond what

DOI: 10.4324/9781003251774-10

the traditional monomodal corpora can provide. This idea can be summarized as a multimodal corpus linguistics method of pragmatics problems. In this sense, multimodal corpus pragmatics investigates how language users integrate different symbols in a certain context to make meaning using multimodal corpora. Information provided by a corpus is inevitably 'partial' as it is impossible to include everything in one single dataset. The methodological and practical processes of recording and documenting natural language are selective, therefore the data is always 'incomplete' (Knight, 2011: 17). This idea is especially true for multimodal corpora. We should note that even though a multimodal corpus provides linguists with rich information in a specific time and place for analysis, including speakers' utterance content, prosodic features, gestures, space layout, and background sound, the corpus can only reinstate partial elements of the reality of natural interaction. This fact is claimed by Gu (2006b) as a partial reinstatement of 'total saturated signification,' which refers to 'the total of meanings constructed out of the face-to-face interaction with naked senses and embodied messages in Goffman's sense by the acting co-present individuals' (Gu, 2009: 436).

Realistically, the current technique for the study of the total saturated signification in real-life face-to-face interaction is audio and video recording and the relevant annotated multimodal corpus. But we should always bear in mind that the audio and video recording, if realistic, is a compromise for the study of Gu's term 'total saturated signification.' The information for other human modalities, e.g. olfactory and haptic, cannot be captured through today's recording devices (Gu, 2009: 436). But with the advancement of techniques, it is reasonable to believe that more types of data could be collected in the future.

Understandably, the benefit of analysing audio and video data for pragmatic function in conversation is quite obvious and substantial: Not only in terms of discovering new patterns between the different channels of conveying meaning, but also in terms of adding to the description and identification of patterns that have been derived on the basis of textual analysis of the transcripts (Adolphs, 2008: 117).

10.1.2 Context as an indispensable term in multimodal corpus

Pragmatics in nature

> does not assume a one-to-one relationship between language form and message function. Rather, it attempts to account for the reason and processes behind the phenomenon that certain linguistic forms might be interpreted as carrying a particular function in a particular context.
>
> (Knight & Adolphs, 2008: 176)

Therefore, context is an indispensable issue in pragmatics studies and corpus approach. The development of a multimodal corpus provides us with better accessibility to context information in human communication. In the studies based on text corpus, even if linguists take discourse context into consideration, much information (such as space layout, activity types, background sound, and relationship between speakers and hearers) is unavailable for pragmatics studies unless the metadata of the corpus is richly provided. Meanwhile, some pragmatics studies have relied mainly on invented contexts to illustrate the interdependency between speech act function and its place in the wider context of use (Adolphs, 2008: 31). This is problematic, or at least inadequate, because a better understanding of discourse context is vital for the interpretation of participants' utterance content, prosodic features, gestures, and the whole integrated meaning. If we presume that the relative inaccessibility of a multimodal corpus is the main reason for this kind of insufficiency, then recently, the prevalent utility of multimodal datasets can provide us with an ideal approach to further exploration on discourse context.

There are various approaches and theories that have been developed to account for the patterns of context in discourse. Hymes (1972), for example, proposed a distinction between speech situation, speech event, and speech act. The speech act encodes social norms in linguistic form. The interpretation of speech acts is thus greatly dependent on an analysis of the sequential organization in discourse and the speakers' social roles in the context (Adolphs, 2008: 32).

Similarly, Gu (2006b) argued that the multimodal analysis of pragmatics should adopt a top-down strategy, rather than hold a sentence or utterance as the entrance point of an illocutionary act. In Gu's model, if we regard a social situation per se as the unit of analysis, then the social situation is actually a configuration of activity types, which in turn are the configuration of tasks or episodes, and the latter are the configuration of participants' individual behaviour. Furthermore, individual behaviour can be segmented into the configuration of talking and doing, of which the lower unit 25 is the speech act since it is regarded as the 'minimum units in human communication' (Searle, 1969: 21). In the author's study, therefore, an illocutionary act is defined as a property exemplified by the speaker in producing an illocutionary act-token in a given social situation.

10.1.3 Multimodal corpus construction for pragmatic studies

After development over the years, many issues in multimodal corpus design and building have been discussed. Kipp et al. (2009), Knight (2011), and Huang (2015c, 2019) introduced some general issues of the multimodal corpus of natural discourse, including data recording, processing, segmentation, and annotation from both theoretical and practical perspectives. In addition, some existing problems and potential research agenda are also introduced. Except for some common properties shared by text corpus and multimodal corpus, the markup schemes and process in multimodal corpus are relatively customized. It is largely because annotating a corpus is dependent on the purpose of the corpus construction, which is 'hypothesis-driven' (Rayson, 2003: 1). This is especially true for the study of pragmatics. For instance, the annotation scheme in a speech act study is quite different from that in the study of backchannelling phenomena.

Apparently, the segmentation, annotation, and representation of multimodal data are quite different from the traditional text corpus. This is partially because today's more comprehensive understanding of 'meaning-making' requires us to work beyond linguistic forms. We largely agree that 'automatic parsing of textual syntagmas on the basis of formal criteria alone is not a panacea for solving all of the problems' (Baldry & Thibault, 2006: 181). More consideration of discourse infrastructure, correspondent novel segmentation, and annotation schemes should be developed in this sense.

For the orthographic corpus, conventional concordance software such as 'Wordsmith' and 'Textsmith' can clearly represent the frequency counts for given linguistic forms in a selected corpus. But the aforementioned software is not eligible for the representation of multimodal data, not to mention pragmatic meaning. A few corpora are semantically or pragmatically marked up, and many pragmatic meanings are without any formal 'hook.' This kind of 'form-function mismatch' in most pragmatic phenomena leads to the consequence that the automatic assignment of tags will often lack precision, and therefore manual implementation is unavoidable (Rühlemann & Aijmer, 2014: 11). In speech act study, this problem is predominant because there are no constant and explicit lexico-grammatical forms associated with speech act-types. Therefore, researchers are not able to annotate, represent, or search a certain type of speech act just by one or several single forms. In most cases, the speech acts have to be tagged manually through line-by-line analysis and the corpus needs to be tailored.

On the other hand, searching in a multimodal corpus is also different from one in a text corpus. A researcher might simply type a certain word or expression into the search bar in a text corpus, while having to manually search for all the possible forms of multimodal information in a multimodal corpus. Such concordances are also limited to present transcripts and text files, rather than multimodal datasets (Knight, 2011: 48).

Unlike the lexico-grammatical studies based on a corpus with a relatively universal data representation format, pragmatic studies usually customize their representation. Orthographic corpora are often equipped with similar representation formats, while multimodal corpora do not even have a widely accepted format or criterion to represent different multimodal datasets. In this sense, a series of represented multimodal data is not necessarily suitable for other multimodal corpus-based pragmatic studies since segmentation and annotation should be tailored for different research purposes. Therefore, just as Knight (2011: 49) claimed, the requirements for representing multimodal corpora for accurate and appropriate analysis and re-use are still being answered by researchers in this field. Some attempts, such as the Digital Replay System (Knight et al., 2010), have been made to provide a possible solution (Knight, 2011: 187–193). In this sense, although many multimodal corpus tools exist, e.g. Anvil, Elan, and MMAV, the integrating search, concordance, and representation functions still need enhancing.

The general pattern in corpus linguistics is identified as 'highly quantitative' of given data represented in a corpus. By using corpora, researchers are concerned with 'the patterns of language, determining what is typical and unusual in given circumstances' (Conrad, 2002: 77). With the use of statistics in corpora, researchers can clearly expose the regular, frequent, and more generalizable patterns of meaning in use. This is also true for the multimodal corpus approach. The only difference in the statistics and quantitative concern between the traditional text or audio corpus and the multimodal corpus is that more information is involved in calculation. Sometimes the algorithm is really complicated and the presentation pattern is beyond the scope of the traditional text corpus.

Sinclair (1996) claimed that corpus-driven and corpus-based are the two basic approaches in corpus linguistics. This is especially so for the corpus-based approach, as it has now been commonly used in a wide range of both linguistic fields as well as humanities and social sciences. For multimodal corpus studies, most researches are corpus-based. This is not only due to the relative shortage of large-scale multimodal corpora for deep quantitative calculation, but also because by using 'corpus-driven,' linguists can be informed by the corpus itself which allows it to lead us in all sorts of directions (Rayson, 2003: 1).

While aiming to promote the general idea of 'multimodal corpus pragmatics,' this book has to be modest while introducing feasible multimodal datasets for the study of live speech acts in natural discourse.

As a functional unit, the speech act connects text and utterance as well as vocabulary and syntax. It is an ideal conjunction for the study of discourse function and syntactic structure in human interaction. Combining quantitative statistics of corpus linguistics with qualitative analysis of individual cases, and re-examining the classic topics of pragmatics through innovative research methods, it can expand the horizon of speech act research, and develop concepts and the scope of illocutionary force and IFIDS. In particular, it provides a new approach in the investigation of indirect speech acts, which can deeply reveal the nature and characteristics of speech acts.

We should realize that the multimodal research approach expands the perspective of pragmatic research, including speech acts. Chinese scholars have acknowledged the importance of the multimodal corpus in pragmatic analysis and proposed a multimodal pragmatic analysis framework with the multimodal research path of semiotic attributes (Chen & Qian, 2011). But as it is based on part of the multimodal corpus or the constructed multimodal corpus, the space for researchers to investigate the expression rules of pragmatic meaning from such perspectives as prosody, expression, gesture, and posture has been expanded. A multimodal corpus can reflect plenty of previously unrecorded information and present the relationship between speech and context or other factors, which is useful for researchers to more comprehensively and accurately interpret speech meaning at the level of common behaviour. In addition, its adoption in studying illocutionary force in live speech can provide more linguistic evidence for the development of human–machine dialogue systems, especially for the research and development of the dialogue framework of multimodal fusion (including voice, expression, and posture).

At present, pragmatics study based on a certain number of corpora that develop into a multimodal corpus has risen as a new field for scholars (Romero-Trillo, 2008), and some scholars have even proposed multimodal corpus pragmatics (Knight & Adolphs, 2008). Given its irreplaceable advantages, some scholars also predict that multimodal corpus-based research will become the mainstream of corpus-based pragmatic research (Rühlemann, 2010).

The adoption of the approach of multimodal corpus linguistics in studying pragmatic issues could not only make pragmatic research more objective in quantitative analysis and erase the shortcomings of traditional pragmatic research to a greater extent, but could also rely on the momentum of multimodal corpus linguistics to upgrade pragmatic study in methodologies, thereby expanding and revising related theories.

Multimodal research is an integrative paradigm. Its application to investigating pragmatic issues has carved out the research field of multimodal pragmatics. As the multimodal corpus approach is one of the approaches, others can also be used to study pragmatic topics, including multimodal discourse analysis and multimodal interaction research. Using the various approaches of multimodal paradigms to revisit pragmatic topics so as to see discoveries and the generation of new theories is exactly the logic of constructing multimodal pragmatics.

Apart from the speech acts discussed in this research, multimodal methods could also be used to study other pragmatic issues like pragmatic hooks, pragmatic identity construction, power operation, politeness, discourse meaning, presupposition, and pragmatic competence. A new way is provided for investigating pragmatics from the perspectives of cognition, society, and culture. Another example is illocutionary meaning, which is a deeply concerning topic in pragmatics, rhetoric, and even literary creation. We can start from the STFE-match assumption to compare its similarities and differences with Grice's conversational principle: The latter is based on the rational human, that is, the two parties engage in the conversation with a common purpose and then cooperate with each other rationally if they both want to exchange as much information as possible. Otherwise, the conversation will slip away from the original purpose, in which the conversation interlocutors are inferred as being irrational. However, the STFE-match assumption is based on a live whole person. Once the speaker violates such a principle in situated discourse, the hearer will make an unintended interpretation, thereby generating illocutionary meanings and new pragmatic meanings (Gu, 2013b: 11–16). Such a way of thinking has expanded the research horizon of pragmatic research of illocutionary meaning.

As far as we know, there are two main views of pragmatics study with two origins. One is the British-American school that regards pragmatics as a branch of linguistics and instruction, conversation meaning, and speech acts are clearly its research units; speech act and illocutionary force discussed in this study are classic topics of this school. The other is the continental European school (e.g. Jef Verschueren's general perspective on pragmatics) that believes that pragmatics is the general perspective all levels of language functions. With a general understanding of the scope of pragmatic research that may include discourse analysis, conversation analysis, anthropology, sociolinguistics, and psycholinguistics, this school believes that pragmatics should be related to such fields as society, culture, psychology, and cognition (He, 2000). At present, multiple approaches under the multimodal research paradigm match the two pragmatics studies, but the question of which pragmatics research is more accessible to the multimodal research paradigm needs further demonstration.

10.1.4 Future agenda of multimodal corpus pragmatics

The integration of the multimodal corpus approach into pragmatics has been increasingly recognized today. This trend is also reflected in a growing number of multimodal corpus-based studies presented in pragmatics academic conferences. Recent years' panels and keynote speakers, in the 'International Pragmatics Conference' for example, have presented a growing number of multimodal corpus-based studies of pragmatics issues.

Despite an array of relevant studies, there still are many challenges and opportunities in today's multimodal corpus pragmatics:

1 Most corpora are relatively small in size and not accessible to other researchers. There are still no publicly available large-scale multimodal corpora. This is partially because building multimodal corpora is really painstaking and time- and cost-consuming. It would take a very long time to mark up all the gestural instances and validate the annotation across multiple annotators before a final version of the annotated corpus is completed. Besides, there are no widely accepted or agreed coding schemes. The segmentation and annotation of video data with prosodic and gestural information for pragmatics study purposes vary from one project to another, which makes a multimodal corpus quite unique and exclusive in the studies. In pragmatics, the annotation would be quite research-specific rather than generally shared, and the corpus design, analytic frameworks, and many other aspects are different. Certainly, an ideal coding scheme should be developed so that they can be shared across different communities, all likely to have different analytic needs. Therefore, a multimodal corpus for pragmatics study is sufficiently

balanced to achieve the aims of a particular linguistic enquiry, which might be quite adequate for other users with different research goals. Nevertheless, we can 'use corpora in full awareness of their possible shortcomings' (Sinclair, 2008: 30) 'because there is no better alternative resource for analyzing real-life language-in-use than a corpus offers' (Knight, 2011: 18).

2 The need for an integrated approach and the tools necessary for the representation of data are still a work in progress. Though many multimodal processing tools have been developed, including multimodal annotators, concordancers, and automatic parsing, a more integrated tool still remains elusive. Based on the assumption that multimodal concordancers can reveal recurrent patterns that constitute important building blocks in the construction of meaning (Sinclair, 1991), Baldry and Michele (2005) developed the multimodal corpus authoring (MCA) system to advance multimodal concordancing. Similarly, Kay O'Halloran's group developed the multimodal video analysis (MMVA) for multimodal discourse analysis in the framework of systemic functional linguistics. These attempts enhance the explanatory power and analytic accessibility to multimodal corpora, but more data processing and statistics tools are still needed.

3 More traditional research topics in pragmatics could be reinvestigated in the framework of the multimodal corpus approach. Only through rich and large-scale studies can this new inquiry flourish. Novel data and observation methods can promote the extension, if not revision, of classic theories in pragmatics, including conversation implicature, presupposition, (im)politeness, and identity construction. Taking Brown and Levinson's verbal politeness strategies as a starting point, for example, Pennock-Speck and Saz-Rubio (2013) found that 'politeness' strategies in charity advertisements on British television are constructed through both paralinguistic and extralinguistic modes of communication. Forceville (2014) outlined, for example, that discussions of visual and multimodal discourse can be embedded in Sperber and Wilson's relevance theory (RT) in pragmatics, which extends the applicability of RT since it claims itself to be ready for all forms of communication even though it has traditionally been applied to the spoken verbal varieties. Meanwhile, discourse markers (DMs) have drawn much attention from pragmaticians. DMs are believed to be multifunctional and to play communicative roles in different dimensions simultaneously. However, most studies on DMs predominantly investigate their use and function in text-based frameworks, and therefore the researchers neglect the coexistence of DMs and speakers' gestures (Hata, 2016). Similarly, Knight (2011) provided a systematic analysis of the pragmatic functions of backchannelling phenomena in English conversation based on the Nottingham multimodal corpus (NMMC), which extends traditional

studies on the relationship between verbalization and gesture in people's active listening.

Multimodal corpus pragmatics can be also applied to many different fields, e.g. gerontolinguistics (Gu & Huang, 2020). Bolly and Boutet (2018) initiated a project based on the multimodal CorpAGEst corpus to study the pragmatic competence of old people by exploring their use of verbal and non-verbal pragmatic markers in real-life conversations, which is both methodologically innovative and socially significant in pragmatic studies. Similarly, Huang (2017b) initiated a multimodal corpus-based study on speech acts of elderly patients with Alzheimer's disease (AD). The preliminary observation concludes that the number of certain types of speech acts has a significant change, e.g. emotion-expressive speech acts will drop in the moderate to severe AD patient. With the decline of linguistic competence, AD seniors are inclined to 'compensate' their pragmatic communication from other resources, including gestures, gaze, and body pose. This observation further justifies the adoption of multimodal data and Perkins's emergentist model for clinical pragmatics (Perkins, 2007). To further explore this issue, the author and his cooperators are establishing the working group for Archives and Resources for Ageing Studies (AR4AS) in China, in which the building of Multimodal Corpus for Gerontic Discourse (MCGD) is the core content.

Additionally, with multimodal data, we can even conduct a multimodal investigation of emotional expression in the field of intercultural pragmatics, and provide more convincing evidence to investigate the issues in both pragmatics and rhetoric (Huang, 2018b). Traditionally, both pragmatics and rhetoric attach importance to verbal cues only, neglecting the important role of non-verbal cues in communication. Under such influence, traditional Chinese rhetoric, including phonetic rhetoric, is mainly concerned with the study of written-word-borne rhetorical figures or the stylistics of written text in the aesthetic perspective, rather than a full consideration of live rhetoric in situated discourse. As a matter of fact, rhetoric deals with the use of language, which can be regarded as a human behaviour. Suppose rhetorical behaviour is regarded as an action reflecting speakers' effectiveness of speech acts; in this case, rhetorical research must start from the nature and characteristics of speech acts in situated discourse to establish methods and analysis. Therefore, the author argues that speech acts should be also treated as the minimum unit of rhetoric behaviour.

In this sense, the speech act is a natural link between rhetoric and pragmatics. In a real-life situation, speakers utilize varied devices in the hierarchies of lexicon, syntax, prosody, and gesture to perform speech acts, which are influenced by the speaker's intentions, emotions, and other factors. From the perspective of the multimodal nature of human interaction, the whole person produces live rhetorical acts in communication. Therefore, multimodal rhetoric behaviour can be defined as a person's all-media-borne effective communication activity in different temporal and spatial dimensions, using varied modalities and taking speech act as the basic unit. By including vocal-auditory and gestural-visual information in the observatory data, rhetoric's conceptualization and scope will be enriched, and the classic rhetoric theories could be developed more towards authentic face-to-face interaction. The conception of multimodal rhetoric aims to study the mechanism and effectiveness of such multimodal rhetoric behaviour, and is also a further development of non-verbal and multimodal metaphor study of texts (see Forceville, 2009; Forceville & Urios-Aparisi, 2009; Huang, 2018b). Undeniably, multimodal rhetoric is a research perspective and method, which does not stand opposed to the traditional rhetoric research of rhetoric and sentence formation. Traditional rhetoric research can be regarded as a level of multimodal rhetoric research.

10.2 Developing multimodal pragmatics

Multimodal corpus pragmatics is an integrated field of inquiry that includes both corpus linguistics and multimodal studies from the disciplinary perspective. Multimodal studies on communication can be traced back to the video recording method adopted in discourse analysis as a pioneering attempt. Though not using the word 'multimodality,' some pioneering researchers in the very early stages already made good use of audio and video technologies to record discourse from the face-to-face interaction of speakers and hearers (Kendon, 1967, 1972; Scheflen et al., 1970), drawing particular attention to gaze, facial expression, and the movement of body parts.

In the recent two decades, some scholars realize that pragmatic study should not constrain itself on the verbal domain of communication (Hoye & Kaiser, 2006, 2007; Kwiatkowska, 2005). To further develop this idea, Hoye (2009) puts forward the concept of 'visual pragmatics.' Different from the connotation advocated in this book, it is still one of the pioneers to expand pragmatic studies into the multimodal horizon. What the author attempts in this book is to upgrade pragmatics studies into 'multimodal pragmatics' after the inspiration of the previous discussion and multimodal corpus-based research in pragmatics.

10.2.1 What's new?

Using a purely textural type of corpus and traditional analysis can only lead to a very incomplete description of the conveyance of meaning in natural conversation. Therefore, a growing number of scholars in different fields agree that conducting linguistic research should use multimodal as the benchmark and study language as its multimodal manifestation in contexts of face-to-face interaction. Many linguistic theories or interaction rules could be enriched or even revised if the fresh methodology and new kinds of data are provided.

This is how today's pragmatics is reflected with the introduction of both the corpus method and multimodal data. The traditional dichotomy between verbal and non-verbal information in pragmatics and other linguistic branches seems to be inadequate if the linguistic theory goal is to target genuine human communication. The relationship between verbal and non-verbal cues in relaying pragmatic meaning is still very intractable, but the multimodal corpus approach can provide a novel perspective to handle this intriguing issue, although we have to admit that multimodal corpus pragmatics, or the multimodal approach, is still in its infancy. More theoretical and practical investigation is needed. The multimodal corpus approach bridges the linguistic gap between the verbal and visual, which has been neglected far too long (see Huang, 2016, 2017a, 2018c, 2021a).

Attention to new materials or data often brings out a shift in research perspectives, an expansion of research fields, innovation in research methodologies, and ultimately new discoveries. If situated discourse can be researched with the multimodal corpus approach, more inspiration on topics or focus includes the following:

1 Realizing that situated discourse is interconnected with people's behaviour, and that related linguistic research should be conducted at the level of ordinary behaviour theory (Gu, 2009, 2012b) drawing on the research methods of time geography (see Hagerstrand, 1975; Thrift, 1977), as well as the theory of time and space in human geography, especially social time and space to solve the problem of situated discourse embedded in physical and social when-and-where circumstances. It discusses the relationship between people's dynamic daily behaviours and their discourse, constructs the workplace discourse geography, and designs corresponding analysis methodologies. In this way, time geography provides an approach for investigating situated discourse. In addition, the basic theory of knowledge ontology and related technologies could also be connected to build a situated discourse database and do research at the level of ordinary human behaviour, which has strong engineering significance for AI and other fields (Gu, 2015c).

2 Re-examining classic topics in pragmatics by using novel multimodal situated discourse corpus. Classical pragmatics, inheriting the traditions of Western philosophy and logic, investigates the law of language-in-use developed from the monomodal (visual sensation of the text or auditory sensation of the audio) analysis of information exchange. However, pragmatics theories and principles constructed by ignoring the meaning contained and conveyed by other modalities are inadequate and inaccurate. Therefore, only by analysing the communication process of total saturated experience can we establish actual pragmatic laws and principles. Furthermore, the different spaces, times, and ways in which speech occurs will affect the applicability of classical pragmatic principles. Through the approach of multimodal linguistic research and

conversation analysis, research on language-in-use can go back to the authentic situation in which talking is intertwined with social activities. They could revisit classic topics in pragmatics such as politeness, cooperation, instruction, and speech act. In addition, multimodal corpus linguistics has also expanded the research horizons of interactional linguistics and conversation analysis.

3 Using the multimodal situated discourse corpus to carry out linguistic research on special populations, including the illiterate, elderly with language decline (e.g. caused by neuro-degenerative disease) (Huang, 2018a, 2021c) and other mental disability, and infants or children with autism spectrum disorder. For example, with no written representations in their brains, the illiterate only has speech acts in situated conversations. A study of the discourse on this group of people helps to understand the nature of initial pragmatic communication of humans and further understand various questions in language information processing of the illiterate (Huang, 2014b). As language loss of the elderly goes side by side with various linguistic and pragmatic features in different situated discourses, multimodal corpus outperforms monomodal corpus (such as audio) in providing researchers with more information (Gu & Huang, 2020; Huang, 2015b). In addition, the modality synergy and evolution of infants' and children's occurrent experience, cognition, and multimodal meaning modules is also an important issue in human language development research.

4 Protecting endangered languages or dialects through the construction of a multimodal situated discourse corpus. Many linguists studying endangered languages have adopted practices consistent with language documentation. The goal of such documentation goes beyond the phonetic analysis of oral production and grammatical analysis of written language (Pernnis, 2018). Instead, the goal is to create a 'lasting, multipurpose record of a language' (Himmelmann, 2006: 1) if linguists want to inherit and protect 'live' language. Therefore, apart from recording the phonetic features of endangered languages or dialects, the multimodal corpus could also effectively capture abundant information of speakers in situated communication, such as the various cultural customs and communication habits in the carrier of language or dialect. A relatively complete record of language activities in social and cultural practices is of great significance for studying and protecting endangered languages and dialects. And many institutions at home and abroad have established relevant multimodal digital archives.

5 Studying the phenomenon of multimodal rhetoric based on the multimodal corpus. From the perspective of the multimodal nature of human communication, live rhetorical behaviour is formed in communication by the 'live, whole person.' Researchers can investigate how live, whole people perform rhetorical behaviours in different when-and-where circumstances, through different carriers and modals to

> make communication more effective. Certainly, multimodal rhetoric is not limited to situated discourse but also in off-site discourse. For example, multimodal metaphors emerge in off-site discourse (Huang, 2015a). With the aid of a multimodal corpus, the rhetoric between different languages could be compared and analysed and the rules can be summarized.

The multimodal corpus could also contribute to innovative humanities and social science research to a certain extent. At present, some researchers in this field have used large-scale text (monomodal) corpora to research historical and cultural, sociopolitical, and other related topics. Research field like digital humanities and e-social research have also emerged at home and abroad. After establishing the modelling and retrieval technology of a multimodal corpus, they could provide scholars with more information than monomodal ones, which is conducive to relevant human and social science research based on objective facts and the effective combination of qualitative and quantitative research.

Of course, although the multimodal corpus is at the current stage, it should be pointed out that it is a relatively advanced type of corpus in terms of concepts and data, its development will never stop. People's deepening knowledge of the brain and language will breed novel research models, which encourage researchers to continue with breakthroughs in technical limitations, exploration of data types, fine modelling granularity, and the construction of composite corpus of big data. In this sense, human efforts in exploring the nature of speech and the cognitive mechanism will witness more sufficient and profound evidence.

In short, the attention paid to situated discourse and multimodal research methods in linguistic research is closely related to technological progress. On the one hand, technological progress provides more ways for people to effectively observe and study language phenomena previously underestimated or failing to be captured, thereby enriching and expanding the research interface and scope; on the other hand, relevant research results will absolutely promote the development of related technologies, such as linguistic preparation for the development of human–computer dialogue and AI systems. Moreover, the thoughts, methods, or paths adopted could also enlighten other humanities and social science research. Multimodal research with profound theoretical value and application value inspires linguistic research and should become one of the emerging fields of linguistic research in the future.

10.2.2 Some preliminary attempts

This novel field has attracted the attention of international academia. Many scholars have been sharing their pragmatic study from a multimodal perspective at the International Pragmatics Conference for many years. In July

2005, Charles Forceville and Eduardo Urios-Aparisi set up the Pragmatics of Multimodal Representations seminar at the 9th International Pragmatics Conference in Italy and called on scholars to focus on pragmatics rhetorics in the multimodal text. At the 12th International Pragmatics conference in the UK, Jean-Marc Colletta and Heather Brookes set up the session 'Multimodality, discourse and speech acts: new insights in pragmatics'. They point out that scholars in pragmatics study have not fully studied how such modal resources as gestures serve the speaker to express intention, construct discourse, and help the hearer make sense via these clues. All contributions to this international panel include multimodal data to tackle these questions, 'looking for more appropriate ways to describe action and discourse in everyday social interactions, either from an empirical and ethnographic background in pragmatics or from more theoretical perspectives'; Alexa Bódog expresses his thinking on the issue of speakers comprehensively using resources such as speech, prosody, and gestures to preserve politeness and cooperation in communication in the post titled 'Strategic speech in formal discourses – A multimodal corpus-based study at the intersection of pragmatics and human ethology.' At the 14th Pragmatics Conference held in Belgium in 2015, Inés Olza shared her thesis 'Co-speech gesture: A pragmatic approach based on big multimodal data' about her exploration seeks to formulate an explanatory model for the pragmatic anchoring and contextual variation of co-speech gesture based on large multimodal corpora of TV news. Rosalice Pinto discussed the phenomenon of multimodal argumentation in the enterprise promotion text and examined how to produce indirect speech acts to persuade consumers and conduct public welfare propaganda in multimodal argumentation and indirect speech acts. At the 15th Pragmatics Conference in Northern Ireland in 2017, a panel studying (im)politeness and turn-taking from the multimodal perspective was established.

Currently, Chinese linguists are also showing growing interest in multimodal pragmatics. At the 15th National Pragmatics Symposium held at Beijing Normal University in August 2017, the present author and his colleagues established a dependent panel on multimodal pragmatics research for the first time in China, which attracted the attention of many pragmatics scholars. Here is what was stated in the panel introduction:

> The research object of pragmatics is language-in-use, and human communication is multimodal in nature. However, traditional pragmatics study as limited in research horizons and techniques only focuses on the use of text while ignoring the fact that modalities construct meaning.
>
> The definition of pragmatics in this topic is mainly based on the micro-perspective, which is a comprehensive study of speech from the perspective of cognition, society, and culture. By observing how people select and use various modalities in their speech that reflect much unrecorded information and the relationship between language and context

> or other factors, it has reference value for more comprehensive and accurate interpretation of speech meaning facilitates comprehensive research on meaning construction. The shift in paradigm and perspective could expand the concepts and scope of traditional pragmatics, develop related theories, enrich the objective description of language-in-use in pragmatics analysis, and enhance the explanatory power of existing pragmatics theories.

The author believes that by including non-verbal resources in the observatory data, the scope and methods of pragmatic studies will be enriched, and classic pragmatic theories will be enhanced and developed more towards authentic face-to-face interaction. This exploratory study hopes to inspire further theoretical discussion and case study on pragmatic questions with the multimodal corpus approach, which will carry forward a novel domain of inquiry: multimodal corpus pragmatics.

10.3 Rethinking multimodal research paradigm

Multimodal linguistic study entails two basic questions: What is modal(ity)? And how to use multimodal theory to study linguistics? Several Chinese scholars such as Yang and Xin (2010), Zhang and Chen (2011), Li (2013), Gu (2006b, 2009, 2013b, 2015a), Feng, Zhang, and O'Halloran (2014), and Huang (2021a) have summarized its research approaches. There are similarities and differences between their classifications in the reviews, but most of them are concentrated in multimodal discourse analysis and do not touch upon basic concepts, research approaches, and overall paradigms. Starting with the definition, this section intends to conduct a more comprehensive review of the multimodal research approaches in linguistics and related fields, in an effort to establish a classification based on the ontology of modality and to discern the modal- and multimodal-related issues in linguistic study.

Multimodal research is a paradigm that integrates multiple approaches. As Kuhn pointed out in *The Structure of Scientific Revolutions*, the changes in paradigm mean changes in how scientists observe the world. That is to say, scientific revolutions are triggered by the displacement of the conceptual network through which scientists view the world (Kuhn, 1996: 102). Therefore, changes in horizon and paradigm foster growth points in a research field. The development of science features the very existence of groups of determined researchers working in their fields, and sufficient research space and unsolved problems for new researchers (Kuhn, 1996: 10). So far, there are many scholars who have been researching in the field of multimodal study under a common belief and witnessing emerging research objectives remaining to be solved.

First, different approaches of the multimodal research paradigm will be introduced. Linguistic-related multimodal research generally includes multimodal research in HCI, multimodal discourse analysis, multimodal corpus research, multimodal learning research in a multimedia environment, and multimodal technique application research. Most of them are interdisciplinary research. Therefore, when referring to multimodal research, we should be clear about what theoretical basis, research ideas, and approaches have been applied, and we should avoid blurring all the research as multimodal research. However, many Chinese studies unfortunately fail to clarify the relationship between different approaches (Huang & He, 2013) or have the simple idea that multimodal research equals studying graphic-text relations or non-verbal acts.

Different approaches have different definitions on multimodality. There are three main perspectives in general: (1) The definition from brain science or physiology in which modality is regarded as sense organs and the interconnected neural networks, and therefore multimodality is a way of interaction between humans and the external environment; (2) the definition from semiotics that regards modality as a symbolic resource that creates meaning in social culture, so multimodal discourse is composed of multiple ideographic symbols; and (3) the definition that modality is the way that information presents in HCI, and hence multimodal HCI means that people use multiple sensory modalities to interact with computers and other machines through multiple physical media.

Research based on different definitions of modality adopts different approaches. As an integrated and increasingly enriched paradigm, the multimodal study consists of several different approaches and fields, including: (1) multimodal discourse analysis rooted in semiotics, (2) multimodal corpus-based study, and (3) multimodal study in neuroscience, HCI, and learning science (Huang & Zhang, 2019). Today's researchers may vary in their interpretation of the word 'multimodality,' but all of them can reach the following consensus: Verbalization in face-to-face interaction is not the only generation of meaning. In other words, human interaction is multimodal in nature. Speakers usually make the most of all available resources to relay their intentions. Hearers likewise comprehend meaning from a wide range of accessible cues or resources.

10.3.1 Multimodal study in semiotics

Both mode and modality can be used to refer to the concept of modality[1] in a multimodal study rooted in semiotics. Its definition in semiotics is 'semiotic resources with meaning created in social culture' (Kress, 2010: 79). In this vein, mode refers to different forms, ways, or channels of information transmission such as sound, image, colour, and action, while language is only one of them. It should be noted that modality is different from the medium, in that the former is the semiotic system that can be compared and opposed, while the latter is physical technique distributed with semiotics. The tools we utilize when producing discourse, such as paper, ink, blackboards, tape recorders, televisions, and projectors belong to media; modality is the way of expressing information through a certain medium; the same modality could be expressed via different media (Hu, 2007: 2).

Linguistic study embedded in semiotics enjoys a long history since Halliday (1978) regarded language as a social symbol. In functional linguistic study represented by Halliday, mode is the selection people make between oral or written ways of communicating in a certain context; while expanding the social symbols theory, Hodge and Kress (1988) inherited his idea and further expanded it to non-verbal modes; Kress (2009, 2010) later used mode to refer to any symbolic resource that produces meaning, such as text, images, colours, typography, gestures, music, cartoons, and even hairstyles and makeup. How these meanings are presented is called multimodal. Systematic analysis on the discourse with these symbolic resources is referred to as multimodal discourse analysis (Gu, 2015a). As discourse in daily life is multimodal, previous studies, however, analysed each mode separately and failed to grasp their correlation in the process of expressing meaning. Their corresponding discourse analysis is confined to language while omitting the meaning of images, sounds, colours, animations, etc. which render the analysis itself rather shortsighted. In the multimodal discourse analysis of semiotics, researchers pay more attention to non-verbal yet important devices like sound, images, layout, gestures, etc. and choose symbolic modes and their relations in both static and dynamic multimodal discourses like advertisement, web-pages, film, literature, and teaching as the research object.

Multimodal study rooted in semiotics includes the school of social semiotic analysis represented by Kress and van Leeuwen (e.g. Kress & van Leeuwen, 2001; van Leeuwen, 2005), the school of discourse analysis under the framework of system-functional grammar represented by O'Toole, Baldry, Thibault and O'Halloran (O'Halloran, 2005), and the school of multimodal interaction analysis represented by Scollon and Norris (Norris, 2004; Scollon & Scollon, 2004) (see Jewitt, 2009). In spite of different focuses, what the three schools have in common is that the 'modal' in multimodal discourse analysis refers to semiotic mode with semiotic attributes. Such semiotic attributes refer to the internal structure of the modes and the way people usually use them to construct meaning (Yang & Xin, 2010: 24). These perspectives on discourse are also broad, including daily communication and other forms of text like printed ads, web-pages, and videos. Traditionally speaking, discourse analysis is rooted in semiotics, as various ideographic channels or resources in the discourse are deemed a symbolic system that carries meaning. It examines the relations between one symbol and another, symbol and user, and symbol and referenced concept. As system-functional linguistics treats language as a social symbol, multimodal study extends the theory of social symbol to other modalities, including images, sounds, gestures, and layouts. All symbolic modalities develop into an alternative and interconnected network that produces meanings (Hu, 2007: 5). Furthermore, multimodal study examines how these discourses convey meaning through different combinations of signs, layouts, and changes, which is different from the traditional linguistic opinion of meaning.

The school of social semiotics analysis highlights the tradition of social semiotics, focusing on how people use various symbolic modes as meaning potential to achieve specific social meaning; the school of discourse analysis under the framework of system-functional grammar emphasizes the meta-functional theory of Halliday (Li, 2013: 22). Both perspectives believe that language and other modalities are multifunctional (Kress & van Leeuwen, 2006; Martinec, 2000; O'Toole, 1994; van Leeuwen, 1999). Moreover, in the interpretation of multimodal discourse, researchers focus on situational factors and examine the respective characteristics and combination rules of symbolic modalities in multimodal discourse and how they construct meaning in a certain context. Researchers apply analytic methods developed from linguistic studies to multimodal discourse analysis, such as metafunction, stratification, and coherence. Representative researchers include Kress and van Leeuwen (2001), van Leeuwen (2005), O'Halloran (2005), and Jewitt (2009). In addition, with the development of modern technology and the shift in reading habits, multiliteracy has risen to being one of the research focus areas.

The school of multimodal interaction analysis (Norris, 2004; Norris & Jones, 2005; Scollon & Scollon, 2004) expands to social action when analysing discourse, believing that symbolic modality, user, and context are closely correlated, emphasizing the interaction between contextual concepts and contextualization (Li, 2013: 22). Having absorbed the research results of interactive sociolinguistics, intermediary discourse analysis, and multimodal research, this school adopts the modal density foreground–background continuum framework.

Another crucial achievement related to this approach is multimodal metaphor study based on cognitive perspective, which also includes images, gestures, and sounds as symbolic resources in the study of metaphor mechanisms and cognitive features (see Forceville & Urios-Aparisi, 2009). This is the metaphor that the source domain and target domain are represented separately or mainly by different modals (Forceville, 2009: 24). As the human senses are information receivers of various symbol systems and the brain acts as the processor, various symbol systems can be connected with the brain's senses and functional areas in the interaction between people and the external environment. Certainly, modals are intertwined in real-world interaction. Multimodal metaphor is embodied in such levels as rhetoric, pictures, or sounds and is a cognitive psychological mechanism in essence (see Huang, 2015c). In recent years, scholars such as Gibbons have also proposed the concept of multimodal cognitive poetics, citing research results in neurocognitive science and visual perception research to explore the relation between multimodal representation and reader cognitive processing. Such efforts facilitated multimodal study on the background of the ongoing digital revolution and promoted more investigation on the processing mechanism (see Gibbons, 2012; Zhao, 2013).

10.3.2 Multimodality study in neurolinguistics, HCI, and learning science

Multimodal research also occupies an important position in neurolinguistics, which needs to examine the brain mechanism of language understanding and output, and the role of other functional areas of the brain in verbal communication. The multimodal nature of communication also requires neurolinguistic research focusing on the role of multimodal coordination in speech. At present, empirical studies featuring eye movement experiments or brain imaging techniques to examine readers' cognition of multimodal discourse have gradually emerged (Feng, Zhang, & O'Halloran, 2014: 89). In addition, the multimodal discourse of people suffering from aphasia and other speech disorders, ageing, and the mentally disabled elderly is also a research topic with practical value.

Modern imaging technology and information technology facilitate researchers in acquiring, processing, analysing, and retrieving multimodal corpora and carrying out related research; in return, multimodal linguistic research promotes research in the field of speech engineering, HCI, AI, etc. As the development of information technology and AI continuously expands the channels and methods for humans to interact with the world, the multimodal nature of human communication becomes the crux of the former. In HCI research, a modal is defined as a method of information presentation (Bernsen, 2008: 7). The so-called multimodal HCI is people using multiple sensory modals to conduct multi-channel information interaction through multiple physical media with computers and other machines. Relevant research entails how AI recognizes the human multimodal information input, including face detection and recognition, expression analysis, speech emotion analysis, gesture recognition, motion analysis, etc.; how to comprehensively analyse and judge multi-channel information input so as to realize multimodal information fusion and improve the comprehensive cognitive capability of the machine; and how to produce multimodal information such as virtual humans or robots generating voice, actions, and even expressions to interact with humans.

Early HCI research focused more on single-channel information processing (single modality), e.g. limiting the interaction between human and computer as keyboard text input. In authentic communication, information transmission takes the form of multiple channels at the same time, including at least voice, expressions, gestures, and postures. These channels are involved in the transmission of information and meaning, and single-channel transmission often results in ambiguity. Therefore, with the aim of actually improving the authenticity and effectiveness of human–machine dialogue, it is necessary to take the multi-channel and multi-form meaning of transmission in live speech into consideration. From the perspective of multimodal HCI structure, multimodal is embodied by multiple information channels of the dialogue system. The system can generally be divided into three modules, i.e. multi-channel information acquisition, multi-channel information analysis and fusion, and multi-channel information expression (see Tao et al., 2011): (1) Multi-channel information acquisition mainly depends on modern imaging technology, such as microphones, cameras, or other input devices to receive the speaker's voice (including prosody), expression, posture, and other modal information; (2) multi-channel information analysis and fusion module to understand the speaker's semantics and to manage dialogue; and (3) multi-channel information expression to produce virtual human expressions, including voice (prosody), action, expression, and other modal information.

Multimodal linguistic study is the theoretical preparation for the research and development of multimodal HCI and AI, and one of the goals of developing AI into multimodal information output and human real speech simulation (Bernsen, 2002; Jaimes & Sebe, 2007). For example, researchers need to train machines in 'learning' various multimodal utterances in authentic interactions to make speech recognition and speech synthesis more accurate. The information input channel of machine corresponds to the sensory modality of human, e.g. camera to vision, touching sensor to touch, microphone to audition, and even electromagnetic skin sensors and other biological sensors (Jaimes & Sebe, 2007: 118). Currently, speech recognition is more common. Some research and development personnel take into account such prosodic features as intonation, but technology that can fully recognize the difference in speech meanings due to the changes in prosody is still rare. Indeed, human communication is multimodal in nature. And questions like how to make AI recognize multimodal cues (expression, gesture, posture, prosody, etc.) and understand their meaning, how to recognize emotions by analysing paralingual phenomena, how to comprehensively analyse multichannel information input instead of doing separately so as to realize multimodal information fusion, and improve the comprehensive cognitive ability of the machine – these are insurmountable challenges as climbing Mount Everest in the research and development of AI. At present, many researchers have conducted extensive research on the core content of HCI, such as multimodal information processing, dialogue management, and emotion recognition (Bengio & Bourlard, 2005; Gibbon, Mertins, & Moore, 2000; Maragos, Potaminos & Gros, 2008; Tao et al., 2011; Taylor, Néel, & Bouwhuis, 2000; van Kuppevelt, Dybkjær, & Bernsen, 2005). The development of multimodal HCI and AI has been regarded as a strategic frontier around the globe, e.g. the representative SmartKom project in Germany (Wahlster, 2006), the multifunctional perception technology of China's 863 Program that has successively realized a series of multimodal interpersonal interaction prototypes since the 1990s. Its related research includes research on the relationship between posture and emotion by the Institute of Automation of the Chinese Academy of Sciences, research on visual speech synthesis by the Department of Computer Science and Technology of Tsinghua University, and the multilingual spoken language system developed by the Systems Engineering and Engineering Management of the Chinese University of Hong Kong. The European Sixth Framework Project Cross Modal on Verbal and Non-Verbal Communication and the Oxygen Project of Massachusetts Institute of Technology are also important research in this field.

Another related research is multimodal learning. The development of multimedia and HCI technology has fuelled multimodal learning research. It is known that multimodal information both promote and interfere with one another in the cognitive process (Yang & Fu, 2001). The following research topics are important: how learners use modalities, brain cognitive processes, learning rules, multimodal learning courseware development, and the teacher's role and development in the learning process, how to make full use of multimedia materials such as network for multimodal learning, and verify the internalization of the learned content, memory persistence, the improvement of learning efficiency, etc. (see Baldry, 2000; Gu, 2007a, 2007b; Stefanucci & Proffitt, 2005), and how to use multimodal theory to serve the teaching and learning of foreign languages (see Gu, 2012a; Huang, 2014c, 2018d, 2021b; Li, 2015; Zhang, 2012; Zhang & Wang, 2010), multimodal interactive technology (such as wearables and network information platforms), etc. And its impact on learning methods and effects are all important topics worthy of in-depth study.

10.3.3 Prospects of multimodal research and application fields

It should be noted that scholars' classifications of multimodal research approaches vary due to different standards or perspectives. For example, Jewitt, Bezemer, and O'Halloran (2016: 6–13) believe that there are three main approaches to multimodal research: systemic functional linguistics, social semiotics, and discourse analysis. In addition, there are five other approaches whose partial concepts or methods overlap with these three, carving out a relatively independent research field, including the following: geo-semiotics, multimodal (inter)action analysis, multimodal ethnography, multimodal corpus analysis, and multimodal reception analysis. The author believes that some of these approaches can be merged, while others apply multimodal approaches in other fields. Therefore, the author's classification is mainly based on the ontology of modality.

Despite different definitions and different approaches, the overall goal of various research boils down to exploring the phenomenon of multimodal interaction between humans and the world, employing such multimodal interaction law to conduct research, developing technology, and serving human development. Therefore, they can be integrated under the paradigm of multimodal research. Figure 10.1 shows the interconnection and logic of various approaches.

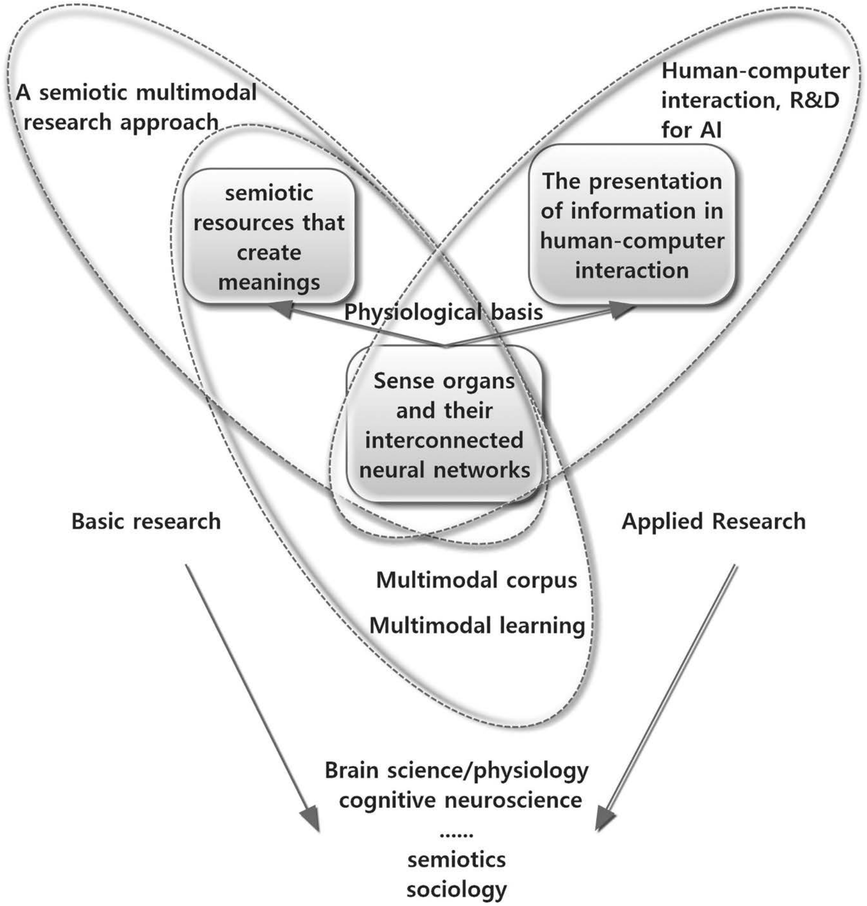

A semiotic multimodal research approach
Human-computer interaction, R&D for AI
semiotic resources that create meanings
The presentation of information in human-computer interaction
Physiological basis
Sense organs and their interconnected neural networks
Basic research
Applied Research
Multimodal corpus
Multimodal learning
Brain science/physiology
cognitive neuroscience
......
semiotics
sociology

Figure 10.1 The relation between three definitions of modals and approaches.

Currently, multimodal research has continually explored the basic research of brain mechanism and behaviour law of human multimodal interaction, and research and development research based on modern technology and future development (Huang & He, 2013: 95). It can be said that multimodal research is both basic research and applied research, which involves theoretical and technical issues and can even rise to a philosophical level.

Multimodal study embedded in semiotics investigates language communication from the perspective of social communication and interpersonal relations, seeks clues from real-world examples of multimodal interaction between human and the world, and explores the laws behind them. We should realize that multimodal communication is essentially supported by the multimodal cooperative mechanism of human brain, and the latter is the physiological basis of the former. The 'multimodal' in multimodal corpus-based linguistic study refers not only to the explored ideographic symbol resources in the corpus, but also to a variety of sensory modals that researchers need to use for corpus processing and retrieval. Researchers resort to modern imaging technology (and developing the means of data acquisition) to examine multimodal interactions between humans and the world; multimodal research in neurolinguistics, HCI, and learning science will explore the multimodal collaborative mechanism of the human brain and ponder how to use such mechanisms to serve human interaction and its own development, in which its definition of mode includes both information presentation and knowledge in brain science, psychology, and other fields.

From the perspective of basic research, multimodal research is connected with such subjects as brain science, psychology, and neurocognitive science, trying to explore the multimodal physiological and psychological basis of human interaction. Relevant studies show that when the brain processes single-modal system representation, the functional area employed is monomodal, and when processing mixed multimodal information, multimodal functional areas are employed (Gu, 2013b: 3). Researchers use fMRI, ERP, PET, EEG, and other technologies to prove that different information processing in speech is dealt with by different functional areas of the brain. In natural conversation, the speaker gets functional areas of his/her brain and corresponding information processing systems working simultaneously, and these processing areas compensate, support, and connect with each other to perform multimodal processing during multimodal interaction. Physiologically, amygdala cells feature multimodal working, in that they can respond to more than one sensory modality, including visual, auditory, tactile, taste, or smell (Kolb & Whishaw, 2005: 411); on the other hand, multimodal research is also connected with semiotics and social communication that endeavours to investigate the phenomenon of multimodal interaction and believes that multimodal speech is a manifestation of the multimodal social activities of humans. These studies focus on investigating the temporal and spatial dimensions of verbal communication, exploring the role of body experience in meaning expression, examining meaning and its laws of expression, transmission, and perception when interacting with other modes in the framework of human communication activities, and thinking what will happen to people's social interaction and meaning expression with the development of science and technology.

From the perspective of applied research and development, people use the achievements of multimodal basic research to reflect on utilizing the coordination of their own brains and multimodal interaction rules between people and the world to develop technologies that serve human development (e.g. multimodal learning) and improve the dimension and level of such interaction (e.g. multimodal HCI). With the development of science and technology, multimodal technology has surpassed the barriers of the subject, bonding humanities and social sciences, natural sciences, engineering technology, and other fields. In the process of research and development, linguistic research has secured its own position in such advanced fields as brain and cognition, HCI, and AI, some of which even manage to fight their way to the frontier. At present, multimodal interaction technologies between people and machines are thriving, including wearables, intelligent robots, and virtual reality. In addition, the development of digital technology is of great significance to multimodal research itself, e.g. the development of multimodal analysis tools or platforms facilitates the transcription, processing, and analysis of complex multimodal corpus (see Allwood et al., 2003; LeVine & Scollon, 2004; O'Halloran, 2009).

Many international academic conferences on multimodal research have been held recently, such as *the ACM International Conference on Multimodal Interaction (ICMI)*, *the International Conference on Multimodality (ICOM)*, *the European and Nordic Symposium on Multimodal Communication*, *the Bremen Conference on Multimodality*, *the International Workshop Series on Multimodal Corpora*, *Tools and Resources*, *the International Conference on Language Resources and Evaluation (LREC)*, and *the Multimodal Studies Panel of International Pragmatics Conference*. Additionally, some Chinese conferences have also begun to establish multimodal research seminars. The author and his colleagues established *China Multimodality Forum* and Tongji Institute of Linguistics and Multimodality (TILM), one of China's first independent institutions. Since multimodal research is a comprehensive paradigm, integrating multiple approaches to various spheres, the institute is inter-disciplinarily sustained, problem-orientated, and its development is distinctively featured. This institute is currently carrying out fundamental research by further strengthening its distinctive characteristics and actively exploring the fields of multimodal semiotics, social semiotics, multimodal teaching, and multimodal studies of language ageing.

Moreover, international multimodal research publications, i.e. *Multimodal Communication* (ISSN 2230–6579) (edited by Norris), *Journal of Multimodal Communication Studies* (ISSN 2391–4033) (edited by Bonacchi and Karpiński), and the Routledge Series in Multimodality Studies published by Routledge have all made a difference in integrating academic resources and showcasing relevant academic progress.

In short, the ascendant multimodal research opens up an avenue for various interdisciplinary research and comprehensive applications, with the ultimate goal of revealing the essential law of human interaction and developing advanced technology for human development. Future multimodal linguistics research and research in related fields should focus on these issues:

> What is the brain mechanism of multimodal interaction between humans and the world?
>
> What are the characteristics of multimodal interaction among people with speech impairment, other people with impaired brain function, or special populations (such as the speech interaction of the elderly with

mental disability, see Gu & Xu, 2013; the speech interaction of the illiterate, see Huang, 2014b)?

How does multimodal interaction technology help disabled people better interact with the world (see Edwards, 2002)?

What is the role of multimodal discourse in social communication?

How has the nature of multimodal interaction changed or developed classic theories or the research span of pragmatics and even linguistics?

What role does multimodal interaction play in the inheritance of language and culture and the development of human civilization?

How can we use multimodal big data to serve human learning?

How can we improve the accuracy of AI's multimodal interaction?

The questions above are important topics projecting into the future. From the perspective of language researchers, the vigorous development of multimodal research allows us to see that linguistics undertakes the academic mission of revealing the nature of human behaviour and promoting the development of related applied technologies, and that linguistics is among the advanced disciplines and enjoys a brighter future.

10.4 Summary

Humans are born multimodal. We combine multimodalities in input and output when experiencing the world. The neuroscientific investigation of language processing claims that the distinction between 'language' and 'communication' and between 'linguistic' and 'non-linguistic' elements in language use has been undermined (Perniss, 2018). Additionally, research shows that our brain does not privilege linguistic information in processing. Instead, multimodal cues are processed simultaneously and immediately in all kinds of contexts (Hagoort & van Berkum, 2007).

Although a pilot study on speech acts forms the illustration in this book, we should bear in mind that multimodal corpus pragmatics is far more than a speech act study. Pragmatics is to be seen not as a mere component of language, but rather as 'a general cognitive, social, and cultural perspective on linguistic phenomena in relation their usage in forms of behaviour' (Verschueren, 1999: 7). In this sense, it is unwarranted to restrict pragmatics to specific phenomena (e.g. speech acts). The benefit of a multimodal corpus approach to pragmatics is far more than a better or more comprehensive understanding of meaning in use in speech, but a promotion of re-evaluation of 'claims and concepts that originate in more philosophical traditions where the conceptualization of pragmatic functions has arguably received most attention' (Knight & Adolphs, 2008: 175). We also believe that such studies are of great significance to AI development and HCI systems, since speech acts (or dialogue acts) are believed to be the minimum units in human communication; that is, they are multimodal in nature. Meanwhile, we also claim that the multimodal corpus, as a novel and promising approach to linguistics, can be applied in many sub-disciplines in language studies, such as conversation analysis and interactional linguistics.

The nature of human discourse activities is multimodal. It is no longer justifiable to ignore 'the non-linguistic aspects of communication – including gesture, facial expression, body movement – have primarily been studied separately from language proper' (Perniss, 2018: 1). Researchers need to examine the meaning expression process at the level of behavioural research, which will greatly expand the horizons of related research and deepen existing theories. This is also the fundamental logic that exists in multimodal pragmatics. This study further motivates the need for a new paradigm – multimodality – in the study of pragmatics. Advocating for a multimodal conception of pragmatics, or 'multimodal pragmatics' in his term, the author attempts to redefine the parameters, upgrade the theory by forging a multimodal research paradigm into a pragmatic study, and align pragmatics study with real-world use of language.

Note

[1] Some scholars distinguish mode from modality (e.g. Kress & van Leeuwen, 2006: 155–174; van Leeuwen, 2005: 160–177), while others do not.

References

Adolphs, Svenja. (2008). *Corpus and context: Investigating pragmatic functions in spoken discourse*[M]. John Benjamins.

Adolphs, Svenja & Carter, Ronald. (2013). *Spoken corpus linguistics: From monomodal to multimodal*[M]. Routledge.

Aijmer, Karin. (1996). *Conversational routines in English*[M]. Longman.

Allwood, Jens. (2002). Bodily communication dimensions of expression and content[A]. In B. Granström et al. (Eds.) *Multimodality in language and speech systems*[C] (pp. 7–26). Kluwer Academic.

Allwood, Jens. (2008). Multimodal corpora[A]. In Anke Lüdeling & Merja Kytö (Eds.) *Corpus linguistics: An international handbook*[C] (pp. 207–225). Mouton de Gruyter.

Allwood, Jens, Cerrato, Loredana, Dybkær, Laila, Jokinen, Kristiina, Navarretta, Costanza, & Paggio, Patrizia (2004). *The MUMIN multimodal coding scheme*[R]. Technical report Retrieved from ww.cst.dk/mumin/stockholmws.html.

Allwood, Jens, Gronqvist, Leif, Ahlsen, Elisabeth, & Gunnarsson, Magnus (2003). Annotations and tools for an activity based spoken language corpus[A]. In Jan C. J. van Kuppevelt & R. W. Smith (Eds.) *Current and new directions in discourse and dialogue*[C] (pp. 1–18). Kluwer Academic.

Archer, Dawn, Aijmer, Karin, & Wichmann, Anne (2012). *Pragmatics: An advanced resource book for students*[M]. Routledge.

Argyle, Michael, Furnham, Adrian, & Graham, Jean A. (1981). *Social situations*[M]. Cambridge University Press.

Aston, Guy. (1997). Large and small corpora in language learning[A]. In Barbara Lewandowska-Tomaszczyk & Patrick J. Melia (Eds.) *PALC97: Practical applications in language corpora*[C] (pp. 51–62). Lodz University Press.

Austin, John L. (1962). *How to do things with words*[M]. Oxford University Press.

Austin, John L. (1979). Performative utterance[A]. In James O. Urmson & Geoffrey J. Warnock (Eds.) *Philosophical papers* (3rd ed.)[C] (pp. 233–252). Oxford University Press.

Baader, Franz, McGuinness, Deborah L., Nardi, Daniele, & Patel-Schneider, Peter F. (Eds). (2003). *The description logic handbook: Theory, implementation, and applications*[C]. Cambridge University Press.

Bach, Kent & Harnish, Robert M. (1979). *Linguistic communication and speech acts*[M]. The MIT Press.

Baken, Ronald J. & Orlikoff, Robert, F. (2000). *Clinical measurement of speech and voice*[M]. Singular.

Baldry, Anthony. (2000). *Multimodality and multimediality in the distance learning age*[C]. Palladino Editore.

Baldry, Anthony & Michele, Beltrami, et al. (2005). The MCA Project: Concepts and tools in multimodal corpus linguistics[A]. In Maj A. Carlsso, Anne Løvland, & Gun Malmgren (Eds.) *Multimodality: Text, culture and use. Proceedings of the Second International Conference on Multimodality. Kristiansand*[C] (pp. 79–108). Agder University College/Norwegian Academic Press.

Baldry, Anthony & Thibault, Paul. (2006). Multimodal corpus linguistics[A]. In Geoff Thompson & Susan Hunston (Eds.) *System and corpus: Exploring connections*[C] (pp. 164–183). Equinox.

Bänziger, Tanja, Pirker, Hannes, & Scherer, Klaus R. (2006). GEMEP-Geneva multimodal emotion portrayals: A corpus for the study of multimodal emotional expressions[A]. In Laurence Deviller et al. (Ed.) *Proceedings of LREC'06 Workshop on Corpora for Research on Emotion and Affect*[C]. Genoa , Italy. vol. 6, pp. 15–19.

Bänziger, Tanja & Scherer, Klaus R. (2005). The role of intonation in emotional expressions[J]. *Speech Communication*, *46*(s 3–4), 252–267.

Barth-Weingarte, Dagmar, Reber, Elisabeth, & Selting, Margret. (2010). *Prosody in interaction*[C]. John Benjamins.

Bar-Yam, Yaneer. (1997). *Dynamics of complex systems*[M]. Addison-Wesley.

Bateman, John, Wildfeuer, Janina, & Hiippala, Tuomo. (2017). *Multimodality*[M]. Walter de Gruyter GmbH.

Batliner, Anton, Fischer, Kerstin, Huber, Richard, Spilker, Jörg, & Nöth, Elmar. (2003). How to find trouble in communication[J]. *Speech Communication*, *40*, 117–143.

Batliner, Anton & Huber, Richard. (2007). Speaker characteristics and emotion classification[A]. In Christian Müller (Ed.) *Speaker classification I: Fundamentals, features, and methods*[C] (pp. 138–1510). Springer.

Bengio, Samy & Bourlard, Hervé. (2005). *Machine learning for multimodal interaction*[C]. Springer.

Bernsen, Niels O. (2002). Multimodality in language and speech systems – From theory to design support tool[A]. In Björn Granström, David House, & Inger Karlsson (Eds.) *Multimodality in language and speech systems*[C] (pp. 93–148). Kluwer.

Bernsen, Niels O. (2008). Multimodality theory[A]. In Dimitrios Tzovaras (Ed.) *Multimodal user interfaces: From signals to interaction*[C] (pp. 5–30). Springer.

Bernsen, Niels O. & Dybkjær, Laila. (2007). Annotation schemes for verbal and non-verbal communication: Some general issues[A]. In Anna Esposito et al. (Eds.) *Verbal and nonverbal communication behaviours*, *LNAI 4775*[C] (pp. 11–12). Springer.

Biber, Douglas. (1993). Representativeness in corpus design[J]. *Literary and Linguistic Computing*, *8*(4), 243–257.

Bing, J. (1985). *Aspects of English Prosody*[M]. New York: Garland.

Binmore, Ken. (2005). *Natural Justice*[M]. Oxford: Oxford University Press.

Blache, Philippe, Bertrand, Roxane, & Ferre, Gaëlle. (2009). Creating and exploring multimodal annotated corpora: The ToMA Project[A]. In Michael Kipp et al. (Eds.) *Multimodal Corpora*[C] (pp. 38–53). Springer.

Blaha, Michael & Rumbaugh, James. (2006). *Object-oriented modeling and design with UML*[M]. Posts & Telecom Press.

Blum-Kulka, Shoshana, House, Juliane, & Kasper, Gabriele. (1989). *Cross-cultural pragmatics: Requests and apologies*[M]. Ablex.

Bolinger, Dwight. (1972a). Accent is predictable (if you're a mind-reader)[J]. *Language*, *48*(3), 633–644.

Bolinger, Dwight. (1972b). Around the edge of language: Intonation[A]. In Dwight Bolinger (Ed.) *Intonation*[C]. Selected readings (pp. 19–29). Penguin.

Bolinger, Dwight. (1986). *Intonation and its parts: Melody in spoken English*[M]. Stanford University Press.

Bolinger, Dwight. (1989). *Intonation and its uses: Melody in grammar and discourse*[M]. Edward Arnold.

Bolly, Catherine & Boutet, Dominique. (2018). The multimodal CorpAGEst corpus: Keeping an eye on pragmatic competence in later life[J]. *Corpora*, *13*(3), 279–317.

Brazil, David. (1975). *Discourse intonation I. Discourse analysis monograph 1*[M]. University of Birmingham English Language Research.

Brazil, David. (1978). *Discourse intonation II. Discourse analysis monograph 2*[M]. University of Birmingham English Language Research.

Brazil, David. (1985). *The communicative value of intonation. Discourse analysis monograph 8*[M]. University of Birmingham English Language Research.

Brazil, David. (1997). *The communicative value of intonation in English*[M]. Cambridge University Press.

Bunnin, Nicholas & Yu, Jiyuan Y. (2001). *Dictionary of western philosophy: English–Chinese*[K]. People's Publishing House.

Burgoon, Judee K., Buller, David B., Hale, Jerold L., & de Turck, Mark A. (1984). Relational messages associated with nonverbal behaviors[J]. *Human Communication Research*, *10*(3), 351–378.

Burkhardt, Armin. (1990). *Speech Acts, meaning, and intentions: Critical approaches to the philosophy of John R. Searle*[C]. Walter de Gruyter.

Buschmeier, Hendrik, Malisz, Zofia, Skubisz, Joanna, Włodarczak, Marcin, Wachsmuth, Ipke, Kopp, Stefan, & Wagner, Petra. (2014). ALICO: A multimodal corpus for the study of active listening[A]. *Proceedings of the 9th Language Resources and Evaluation Conference*[C]. Reykjavík, Iceland: 3638–3643.

Cameron, Deborah. (2001). *Working with spoken discourse*[M]. Sage.

Cao, Jianfen. (2001). Phonetic and linguistic clues of Chinese prosody segmentation[A]. *Modern phonetics in the new century: Proceedings of the 5th National Conference on Modern Phonetics*[C]. Tsinghua University Press, Beijing: 176–179.

Carney, Dana R., Hall, Judith A., & LeBeau, Lavonia S. (2005). Beliefs about the nonverbal expression of social power[J]. *Journal of Nonverbal Behavior*, *29*(2), 105–123.

Carretero, Marta, Maíz-Arévalo, Carmen, & Martínez, M. Ángeles. (2013). "Hope this helps!" An analysis of expressive speech acts in online task-oriented interaction by university students[A]. In Jesús Romero-Trillo (Ed.) *Yearbook of corpus linguistics and pragmatics 2013*[C] (pp. 261–290). Springer.

Carter, Ronald & Adolphs, Svenja. (2008). Linking the verbal and visual: New directions for corpus linguistics[J]. *Language and Computers*, *64*(1), 275–291.

Cavicchio, Federica, & Poesio, Massimo. (2009). Multimodal corpora annotation: Validation methods to assess coding scheme reliability[A]. In Randy Goebel, Jörg Siekmann, & Wolfgang Wahlster (Eds.) *Multimodal Corpora*[C] (pp. 109–121). Springer-Verlag Berlin Heidelberg.

Cerrato, Loredana. (2007). *Investigating communicative feedback phenomena across languages and modalities*[D]. Ph.D. dissertation, Göteborg University.

Chafe, Wallace L. (1979). The flow of thought and the flow of language[A]. In Talmy Givón (Ed.) *Discourse and syntax*[C] (pp. 159–181). Academic Press.

Chafe, Wallace L. (1980). The deployment of consciousness in the production of a narrative[A]. In Wallace L. Chafe (Ed.) *The pear stories*[C] (pp. 9–50). Ablex.

Chafe, Wallace L. (1988). Linking intonation units in spoken English[A]. In John Haiman & Sandra A. Thompson (Eds.) *Clause combining in grammar and discours*e[C] (pp. 1–27). John Benjamins.

Chafe, Wallace L. (1994). *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*[M]. University of Chicago Press.

Chen, Xinren, & Qian, Yuehong. (2011). The application of multimodal discourse analysis in pragmatic research[J]. *Foreign Languages in China*, *8*(5), 89–93.

Chomsky, Noam. (2000). *New Horizons in the Study of Language and Mind*[M]. Cambridge: Cambridge University Press.

Code, Chris. (2005). First in, last out? The evolution of aphasic lexical speech automatisms to agrammatism and the evolution of human communication[J]. *Interaction Studies*, *6*(2), 311–334.

Conrad, Susan. (2002). Corpus linguistic approaches for discourse analysis[J]. *Annual Review of Applied Linguistics*, *22*, 75–95.

Cook, Guy. (1990). Transcribing infinity: Problems of context presentation[J]. *Journal of Pragmatics*, *14*(1), 1–24.

Couper-Kuhlen, Elizabeth. (2009). Prosody[A]. In Sigurd D'hondt, Jan-Ola Östman, & Jef Verschueren (Eds.) *The pragmatics of interactio*n[C] (pp. 174–189). John Benjamins.

Couper-Kuhlen, Elizabeth & Ford, Cecilia E. (2004). *Sound Patterns in interaction: Cross-linguistic studies from conversatio*n[C]. John Benjamins.

Couper-Kuhlen, Elizabeth & Selting, Margret (Eds.) (1996). *Prosody in conversation: Interactional Studies*[C]. Cambridge University Press.

Cowie, Roddy & Cornelius, Randolph R. (2003). Describing the emotional states that are expressed in speech[J]. *Speech Communication*, *40*(1), 5–32.

Creswell, John W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.)[M]. Sage.

Cruttenden, Alan. (2002). *Intonation* (2nd ed.)[M]. Peking University Press & Cambridge University Press.

Culpeper, Jonathan. (2011). *Impoliteness: Using language to cause offence*[M]. Cambridge University Press.

Cutting, Joan. (2001). The speech acts of the in-group[J]. *Journal of Pragmatics*, *33*(8), 1207–1233.

Dahlmann, Irina & Adolphs, Svenja. (2007). Pauses as an indicator of psycholinguistically valid multi-word expressions (MWEs)?[A]. In Nicole Gregoire, Stefan Evert & Su N. Kim (Eds.) *Proceedings of the ACL Workshop on A Broader Perspective on Multiword Expressions* (pp. 49–56)[C]. Association of Computational Linguistics.

Damasio, Antonio, (1999). *The feeling of what happens: Body and emotion in the making of consciousness*[M]. Heinemann.

Darwin, Charles R. (1872). *The expression of the emotions in man and animals*[M]. John Murray.

De Felice, Rachele, Darby, Jeannique, Fisher, Anthony, & Peplow, David. (2013). A classification scheme for annotating speech acts in a business email corpus[J]. *ICAME Journal*, *37*, 71–105.

de Moraes, João A. & Rilliard, Albert. (2014). Illocution, attitudes and prosody[A]. In Tommaso Raso and Heliana Mello (Eds.) *Spoken corpora and linguistic studies*[C] (pp. 233–270). John Benjamins.

Deng, Fanglin, et al. (2009). *Complex engineered systems modeling and simulation*[M]. National Defense Industry Press.

Devillers, Laurence & Vidrascu, Laurence. (2007). Real-life emotion recognition in speech[A]. In Christian Müller (Ed.) *Speaker classification II: Selected projects*[C] (pp. 34–42). Springer.

Dore, John. (1973). *The development of speech act*s[D]. Ph.D. dissertaion, City University of New York.

Drew, Paul. (2013). Conversation analysis and social action[J]. *Journal of Foreign Languages*, *36*(3), 2–19.

Duan, K. C. (1988). *Searle's speech act theory*[M]. Foreign Language Teaching and Research. (4), 29–33.

Edwards, A. D. N. (2002). Multimodal interaction and people with disabilities[A]. In Björn Granström, David House, & Inger Karlsson (Eds.) *Multimodality in language and speech system*s[C] (pp. 73–92). K1uwer Academic Publishers.

Edwards, Jane A. (1992). Design principles in the transcription of spoken discourse. In Jan Svartivik (Ed.) *Directions in corpus linguistics: Proceedings of Nobel Symposium 82* (Stockholm, August 4–8, 1991)[C] (pp. 129–144). Mouton de Gruyter.

Ekman, Paul. (1972). Universals and cultural differences in facial expressions of emotion[A]. In James Cole (Ed.) *Nebraska symposium on motivation*[C], Vol. 19 (pp. 207–284). Lincoln University of Nebraska Press.

Ekman, Paul. (1973). *Darwin and facial expression*[M]. Academic Press.

Ekman, Paul. (2003). *Emotions revealed: Recognizing faces and feelings to improve communication and emotional life*[M]. Times Books.

Ekman, Paul & Davidson, Richard J. (Eds.) (1994). *The nature of emotion: Fundamental questions*[C]. Oxford University Press.

Ekman, Paul & Friesen, Wallace V. (1975). *Unmasking the face*[M]. Prentice Hall.

Ekman, Paul & Rosenberg, Erika L. (2005). *What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system*[C]. Oxford University Press.

Elster, Jon. (1999). *Alchemies of the mind: Rationality and the emotions*[M]. Cambridge University Press.

Erman, Britt. (2007). Cognitive processes as evidence of the idiom principle[J]. *International Journal of Corpus Linguistics*, *12*(1), 25–53.

Escalera, Sergio, Pujol, Oriol, Radeva, Petia, Vitria, Jordi, & Anguera, M. Teresa. (2010). Automatic detection of dominance and expected interest[J]. *EURASIP Journal on Advances in Signal Processing*, Volume 2010, Article ID 491819, 1–12, doi:10.1155/2010/491819.

Fanelli, Gabriele, Gall, Juergen, Romsdorfer, Harald, Weise, Thibaut, & Van Gool, Luc. (2010). 3D vision technology for capturing multimodal corpora: Chances and challenges[A]. In *Proceedings of the LREC Workshop on Multimodal Corpora*[C], Mediterranean Conference Centre, Malta, May 18, 2010, pp. 70–73.

Fann, K. T. (1969). S*ymposium on J. L. Austin*[C]. Routledge and Kegan Paul.

Felder, Ekkehard, Müller, Marcus, & Vogel, Friedemann (Eds.). (2011). *Korpuspragmatik: Thematische korpora als basis diskurslinguistischer analyse*n[C]. DeGruyter.

Felix-Brasdefer, J. Cesar. (2010). Data collection methods in speech act performance: DCTs, role plays, and verbal reports[A]. In Alicia Martíner-Flor & Esther Usó-Juan (Eds.) S*peech act performance: Theoretical, empirical and methodological issues*[C] (pp. 41–56). John Benjamins.

Feng, Shengli. (1995). *Research on Chinese prosody grammar*[M]. Peking University Press.

Feng, Shengli. (1997). *Prosody, morphology and syntax of Chinese*[M]. Peking University Press.

Feng, Shengli. (2000). *Chinese prosody syntax*[M]. Shanghai Education Press.

Feng, William D. & O'Halloran, Kay L. (2013). The multimodal representation of emotion in film: Integrating cognitive and semiotic approaches[J]. *Semiotica*, *197*, 101–122.

Feng, William D. & Qi, Yujie. (2014). The multimodal construction of attitudinal meaning: An analytical model based on cognitive appraisal theory[J]. *Modern Foreign Language*, *37*(5), 585–596.

Feng, William D., Zhang, D. L., & O'Halloran, Kay. (2014). The development and frontiers of multimodal discourse analysis[J]. C*ontemporary Linguistics*, *16*(1), 88–99.

Feng, Z. W. (2014). Speech act theory and conversation agent[J]. *Foreign Languages*, *37*(1), 21–35.

Feng, Z. W., & Yu, W. H. (2015). BDI model in conversation agent system[J]. *Foreign Languages*, *38*(2), 2–14.

Firth, John R. (1957). *Papers in linguistics, 1934–1951*[M]. Oxford University Press.

Fónagy, Ivan, Bérard, Eva, & Fónagy, Judith. (1983). Clichés mélodiques[J]. *Folia Linguistica*, *17*(1–4), 153–186.

Forceville, Charles. (2009). Nonverbal and multimodal metaphor in a cognitivist framework[A]. In Charles Forceville & Eduardo Urios-Aparisi (Eds.) *Multimodal metaphor*[C] (pp. 19–42). Walter de Gruyter.

Forceville, Charles J. (2014). Relevance theory as model for analyzing visual and multimodal communication[A]. In David Machin (Ed.) *Visual communication*[C] (pp. 51–70). Mouton de Gruyter.

Forceville, Charles & Urios-Aparisi, Eduardo. (Eds.) (2009). *Multimodal metaphor*[M]. Mouton de Gruyter.

Foster, Mary E. & Oberlander, Jon (2007). Corpus-based generation of head and eyebrow motion for an embodied conversational agent[J]. *Language Resources and Evaluation*, *41*(3/4), 305–23.

Frijda, Nico. (1986). *The emotions*[M]. Cambridge University Press.

Fu, A. P., & Song, P. Y. (2005). *Metadata and the construction of Chinese corpus*[R]. JSCL-2005, August 27, 2005, Nanjing. pp. 573–575.

Fu, X. T. (2005). Should "Illocutionary Logic" be translated into "语用逻辑"?[J]. *Journal of Hunan First Normal University*, *5*(2), 4–8.

Fussell, Susan R. (Ed.) (2002). *The verbal communication of emotion: Interdisciplinary perspectives*[C]. Lawrence Erlbaum.

Gao, H., & Yan, C. S. (2004). A review of the twenty years' development of pragmatics in China[J]. *Journal of PLA University of Foreign Languages*, *4*, 12–19.

Garcia, Paula. (2007). Pragmatics in academic contexts: A spoken corpus study[A]. In M. C. Campoy & M. J. Luzón (Eds.) *Spoken corpora in applied linguistics*[C] (pp. 97128). Peter Lang.

Garfinkel, Harold. (1967). *Studies in ethnomethodology*[M]. Prentice-Hall.

Garside, Roger, Leech, Geoffrey, & McEnery, Tony. (1997). *Corpus annotation*[C]. Longman.

Georgila, Kallirroi, Lemon, Oliver, Henderson, James, & Moore, Johanna D. (2009). Automatic annotation of context and speech acts for dialogue corpora[J]. *Natural Language Engineering*, *15*(3), 315–353.

Gibbon, Dafydd, Mertins, Inge, & Moore, Roger K. (2000). *Handbook of multimodal and spoken dialogue systems: Resources, terminology and product evaluation*[M]. Kluwer.

Gibbons, Alison. (2012). *Multimodality, cognition, and experimental literature*[M]. Routledge.

Gibson, James J. (1986). *The ecological approach to visual perception*[M]. Lawrence Erlbaum.

Giddens, Anthony. (1987). *Social theory and modern sociology*[M]. Stanford University Press.

Goffman, Erving. (1974). *Frame analysis: An essay on the organization of experience*[M]. University of Pennsylvania Press.

Goldman, Alvin I. (1970). *A theory of human action*[M]. Prentice-Hall Inc.

González-Lloret, Marta. (2010). Conversation analysis and speech act performance[A]. In Alicia Martíner-Flor & Esther Usó-Juan (Eds.) S*peech act performance: Theoretical, empirical and methodological issues*[C] (pp. 57–74). John Benjamins.

Goodwin, Charles. (1981). *Conversational organization: Interaction between speakers and hearers*[M]. Academic Press.

Grice, Herbert P. (1957). Meaning[J]. *Philosophical Review*, *66*, 377–388.

Grice, Herbert P. (1975). Logic and conversation[A]. In Peter Cole & Jerry Morgan (Eds.) *Syntax and semantics*[C]: Vol. 3, Speech acts (pp. 41–58). Academic Press.

Grice, Herbert P. (1978). Further notes on logic and conversation[A]. In Peter Cole (Ed.) *Syntax and semantics 9: Pragmatics*[C]. Academic Press.

Gu, Yueguo. (1989). Austin's speech act theory[J]. *Foreign Language Teaching and Research*, *77*(1), 30–39.

Gu, Yueguo. (1993). The impasse of perlocution[J]. *Journal of Pragmatics*, *20*, 405–432.

Gu, Yueguo. (1994a). Pragmatics and rhetoric: A collaborative approach to conversation[A]. In Herman Parrate (Ed.) *Pretending to communicate*[C] (pp. 173–195). Walter de Gruyter.

Gu, Yueguo. (1994b). John Searle's speech act theory and philosophy of mind[J]. *Contemporary Linguistics*, (2), 1–8.

Gu, Yueguo. (1994c). John Searle's speech act theory: Criticism and reference[J]. *Contemporary Linguistics*, (3), 10–16.

Gu, Yueguo. (1996). Doctor-patient interaction as goal-directed discourse in Chinese sociocultural context[J]. *Journal of Asian Pacific Communication*, *7*(3), 156–176.

Gu, Yueguo. (1997). Five ways of handling a bedpan[J], *Text*, *17*(4), 457–475.

Gu, Yueguo. (1999a). Linguistic 'facts' and paradigms[J]. *Contemporary Linguistics*, *1*(3), 3–14.

Gu, Yueguo. (1999b). Towards a model of situated discourse[A]. In Ken Turner (Ed.) *The semantics and pragmatics interface from Different Points of View*[C] (pp. 149–178). Elsevier Science.

Gu, Yueguo. (2002a). Introduction on *How to do things with words*[A]. In John L. Austin (Ed.) *How to do things with words*[M]. Foreign Language Teaching and Research Press: F23–F36.

Gu, Yueguo. (2002b). Sampling Situated Discourse for Spoken Chinese Corpus[A]. In Chinese Academy of Social Sciences (Ed.) *Globalization and the 21st century*[C] (pp. 484–500). Social Sciences Archives Press.

Gu, Yueguo. (2002c). Towards an understanding of workplace discourse[A]. In Christopher Candlin (Ed.) *Research and practice in professional discourse*[C] (pp. 137–185). The City University of Hong Kong Press.

Gu, Yueguo. (2006a). Agent-based modeling language (AML) 1: Dynamic behavior modeling[A]. The Office of Informationization Leading Group of Chinese Academy of Social Sciences/Computer Network Information Center (Ed.) *The Proceedings of the 20th International CODATA Conference: Scientific Data and Knowledge within the Information Society*[C] (pp. 21–47).

Gu, Yueguo. (2006b). Multimodal text analysis – A corpus linguistic approach to situated discourse[J]. *Text and Talk*, *26*(2), 127–167.

Gu, Yueguo. (2007a). On Multimedia Learning and Multimodal Learning[J]. *Technology Enhanced Foreign Language Education*, 114(4), 3–12.

Gu, Yueguo. (2007b). Learning by multimedia and multimodality: Some critical reflections on web-based courseware design in the Chinese context[A]. In Helen Spencer-Oatey (Ed.), *e-Learning Initiatives in China: Pedagogy, Policy and Culture*[C] (pp. 37–56). Hong Kong University Press.

Gu, Yueguo. (2009). From real-life situated discourse to video-stream data-mining[J]. *International Journal of Corpus Linguistics*, *14*(4), 433–466.

Gu, Yueguo. (2010). *Gu Yueguo's selected overseas linguistics works*[C]. Foreign Language Teaching and Research Press.

Gu, Yueguo. (2011). Exploring the development of contemporary linguistics, Part 3: Language, medium, and technology[J]. *Contemporary Linguistics*, *13*(1), 22–48.

Gu, Yueguo. (2012a). A chess master model for classroom teaching and teacher/researcher development[J]. *Chinese Journal of Applied Linguistics*, *35*(1), 5–23.

Gu, Yueguo. (2012b). Discourse geography[A]. In James Paul Gee & Michael Handford (Eds.) *The Routledge Handbook of Discourse Analysis*[C] (pp. 541–557). Routledge.

Gu, Yueguo. (2013a). A conceptual model of Chinese illocution, emotion and prosody[A]. In Zheng, Yuqiu (Ed.) *Human language resources and linguistic typology*[C] (pp. 309–362). Academia Sinica.

Gu, Yueguo. (2013b). The STFE principle in situated discourse: A multimodal corpus linguistics approach[J]. *Contemporary Rhetoric*, *180*(6), 1–19.

Gu, Yueguo. (2015a). Chinese speech acts and speech act verbs[A]. In Rint Sybesma, Wolfgang Behr, Yueguo Gu, Zev Handel, C.-T. James Huang, & James Myers (Eds.) *Encyclopaedia of Chinese language and linguistics*[C]. Brill. http://dx.doi.org/10.1163/2210-7363_ecll_COM_00000394

Gu, Yueguo. (2015b). Multimodal sensory system and linguistic research[J]. *Contemporary Linguistics*, *17*(4), 1–22.

Gu, Yueguo. (2015c). *Multimodality and big data research on emotion and learning*[R]. Academic report of the Institute of Language Studies, Shanghai International Studies University, Shanghai, October 23, 2015.

Gu, Yueguo. (2015d). *Ontology-supported special corpus of illocution, emotion and prosody*[R]. Keynote speech at the 18th Oriental COCOSDA/CASLRE 2015, Shanghai Jiaotong University, October 28–30, 2015.

Gu, Yueguo. (2016). Multimodal experiencing, situated cognition and big data with a demonstrative analysis of a newborn baby[J]. *Contemporary Linguistics*, *18*(4), 475–513.

Gu, Yueguo & Huang, Lihe. (2020). *Gerontolinguistics and multimodal studies*[C]. Tongji University Press.

Gu, Yueguo & Xu, X. F. (2013). *Alzheimer's disease patient discourse: A multimodal corpus linguistics approach*[R]. Plenary speech delivered at the 5th Symposium on Functional Linguistics and Multimodality, July 20–22, 2013, HKPU.

Gu, Yueguo & Zhang, Yunwei. (2013). OWL-based ontology in still image segmentation and annotation[J]. *Contemporary Linguistics*, *15*(2), 214–229.

Gumperz, John J. (1982). *Dsicourse Strategies*[M].Cambridge University Press, Cambridge.

Gumperz, John J. (2003). On the development of interactional sociolinguistics[J]. *Language Teaching and Linguistic Studies*, (1), 1–10.

Gumperz, John J. & Berenz, Norine. (1992). Transcribing conversational exchanges[A]. In Jane A. Edwards & Martin D. Lampert (Eds.) *Talking data: Transcription and coding in discourse research*[C] (pp. 91–121). Lawrence Erlbam.

Gunes, Hatice, Piccardi, Massimo, & Pantic, Maja. (2008). From the lab to the real world: Affect recognition using multiple cues and modalities[A]. In Jimmy Or (Ed.) *Affective computing, focus on emotion expression, synthesis and recognition*[C] (pp. 185–218). I-Tech Education and Publishing.

Gussenhoven, Carlos. (1984). *On the grammar and semantics of sentence accents*[M]. Foris.

Hagerstrand, Torsten. (1975). Space, time and human conditions[A]. In Anders Karlqvist, L. Lundqvist, & Folke Snickars. (Eds.) *Dynamic allocation of urban space*[C] (pp. 3–12). Saxon House.

Hagoort, Peter & Van Berkum, Jos J. A. (2007). Beyond the sentence given[J]. *Philosophical Transactions of the Royal Society B*, *362*(1481), 801–811.

Halliday, Michael A. K. (1967). *Intonation and grammar in British English*[M]. Mouton.

Halliday, Michael A. K. (1970). *A course in spoken english intonation*[M]. Oxford University Press.

Halliday, Michael A. K. (1978). *Language as social semiotic: The social interpretation of language and meaning*[M]. Arnold.

Hancher, Michael. (1979). The classification of cooperative illocutionary acts[J]. *Language in Society*, *8*(1), 1–14.

Hata, Kazuki. (2016). On the importance of the multimodal approach to discourse markers: A pragmatic view[J]. *International Review of Pragmatics*, *8*(1), 36–54.

Hatfield, Elaine, Cacioppo, John, & Rapson, Richard L. (1994). *Emotional contagion*[M]. Cambridge University Press.

He, Z. R. (2000). Introduction to *understanding pragmatics*[A]. In Jef Verschueren (Ed.) *Understanding pragmatics*[M] (pp. F13–F18). Foreign Language Teaching and Research Press, Edward Arnold.

He, Z. R. (2011). Context revisited[A]. *Foreign language studies (revised)*[C] (pp. 1–13). Yilin Press.

He, Z. X. (1984). The indirectness of English language[J]. *Journal of Foreign Languages*, *31*(3), 9–13.

He, Z. X. (1988). Indirect requests in English and their classification[J]. *Journal of Foreign Languages*, (4), 22–26.

He, Z. X. (1989). *A survey of pragmatics*[M]. Shanghai Foreign Language Education Press.

Heath, Christian. (1986). *Body movement and speech in medical interaction*[M]. Cambridge University Press.

Himmelmann, Nikolaus. (2006). Language documentation: What is it and what is it good for?[A] In Jost Gippert, Nikolaus Himmelmann, & Ulrike Mosel (Eds.) *Essentials of language documentation*[C] (pp. 1–30). Mouton.

Hirose, Keikichi, et al. (2000). *Analytical and perceptual study on the role of acoustic features in realizing emotional speech*[R]. ICSLP 2000.

Hirschberg, Julia. (2002). The pragmatics of intonational meaning[A]. In Bernard Bel & Isabelle Marlien (Eds.) *Proceedings of the Speech Prosody 2002 conference*[C], Aix-en-Provence: Laboratoire Parole et Langage, April 11–13, pp. 65–68.

Hodge, Robert & Kress, Gunther. (1988). *Social semiotics*[M]. Polity Press.

Hou, Lizhen, Han, Demin, Xu, Wen, & Zhang, Li (2002). Study on voice characteristics of people with different sexes and ages[J]. *Journal of Clinical Otorhinolaryngology Head and Neck Surgery*, *16*(12), 667–669.

Hoye, Leo F. (2009). Pragmatics: Chasing the sky or another way of seeing[A]. In Bruce Fraser & Ken Turner (Eds.) *Language in life, and a life in language: Jacob Mey-a Festschrift*[C] (pp. 187–192). Emerald Group Publishing.

Hoye, Leo F. & Kaiser, Ruth. (2006). The migration and reconstruction of a symbol: An exploration in the dynamics of context and meaning across cultures[A], *Linguistic LAUD Agency*[C], Essen, Universitat Duisberg-Essen, Series A. Paper No. 662: 1–29.

Hoye, Leo F. & Kaiser, Ruth. (2007). Branding a symbol: Context and meaning across cultures[J]. *Intercultural Pragmatics*, *4*(1), 51–69.

Hu, F. Z. (2009). Analysis of speech subject: An important category of pragmatics in daily language[J]. *Journal of East China Normal University (Humanities and Social Sciences)*, *41*(9), 66–72.

Hu, Z. L. (1980). Pragmatics[J]. *Contemporary Linguistics*, (3), 1–10.

Hu, Z. L. (2002). A pluralistic approach to the study of context[J]. *Foreign Language Teaching and Research*, *34*(3), 161–166, 239.

Hu, Z. L. (2007). Multimodalization in social semiotics[J]. *Language Teaching and Linguistic Studies*, (1), 1–10.

Huang, Gale Y. (2012). *Measurement and evaluation of education*[M]. East China Normal University Press.

Huang, Lihe. (2014a). The study on speech acts from the perspective of situated discourse[J]. *Journal of Zhejiang International Studies University*, *3*, 45–53.

Huang, Lihe. (2014b). Linguistic research on illiterates: Retrospective and prospective[J]. *Journal of Nan Jeon*, *17*, A3,1–13.

Huang, Lihe. (2014c). A multimodal approach to the design of extended learning systems for college English[J]. *Language Education*, *3*, 11–16, 21.

Huang, Lihe. (2015a). Synaesthesia, metaphor and multimodal metaphor[J]. *Languages and International Studies*, *13*, 91–104.

Huang, Lihe. (2015b). The study of language aging in last decade: Retrospective and prospective[J]. *Journal of Beijing International Studies University*, *10*, 17–24.

Huang, Lihe. (2015c). Corpus 4.0: Multimodal corpus building and related research agenda[J]. *Journal of PLA University of Foreign Languages*, *38*(3), 1–7, 48.

Huang, Lihe. (2016). Situated discourse and multimodal studies: Significance, theory and methodology[J]. *Journal of University of Science and Technology Beijing (Social Sciences edition)*, *32*(5), 29–36.

Huang, Lihe. (2017a). Speech act theory and multimodal analysis: The logic in multimodal (corpus) pragmatics[J]. *Journal of Beijing International Studies University*, *3*, 12–30, 133.

Huang, Lihe. (2017b). *The design and agenda of multimodal corpus-based study of language ageing and related issues: From the perspective of speech act*[R]. Speech at The 4th Sino-UK Symposium of Corpus Linguistics, Qingdao, August 5, 2017.

Huang, Lihe. (2018a). *Developing study of language aging for a better aging society*[N]. Chinese Social Sciences Today, March 20.

Huang, Lihe. (2018b). Establishing multimodal rhetoric: Also looking for a link between rhetoric and pragmatics[J]. *Contemporary Linguistics*, *20*(1), 117–132.

Huang, Lihe. (2018c). Issues on multimodal corpus of Chinese speech acts: A case in multimodal pragmatics[J]. *Digital Scholarship in the Humanities*, *33*(2), 316–326.

Huang, Lihe. (2018d). Intercultural education on the theme of the belt & road initiative: A multimodality-oriented pedagogical design[A]. In Yu Cheng, Lilei Song, & Lihe Huang (Eds.) *The belt and road initiative in the global arena: Chinese and European perspective*[C] (pp. 5–54). Palgrave.

Huang, Lihe. (2021a). Towards multimodal corpus pragmatics: Rationale, case and agenda[J]. *Digital Scholarship in the Humanities*, *36*(1), 101–114.

Huang, Lihe. (2021b). Foreign language teaching and research in post-pandemic era: A mindset of multimodality[J]. *Contemporary Foreign Languages Studies*, *451*(1), 75–85.

Huang, Lihe. (2021c). *Language issues in ageing society and gerontolinguistics in China*[N]. Guangming Daily, March 21.

Huang, Lihe & He, J. H. (2013). A review on the Routledge handbook of multimodal analysis[J]. *Language Education*, *1*, 92–95.

Huang, Lihe & Zhang, Delu. (2019). A multi-core parallel system: Paradigm, approaches and fields in multimodal study[J]. *Foreign Language Education*, *40*(1), 21–26.

Huang, Sihong & Zhan, Hongwei. (2011). The latest development of formulaic sequences processing[J]. *Journal of Foreign Languages*, *34*(2), 64–71.

Huang, Yushun. (2007). On 'compassion' of confucianism and 'sympathy' of Scheler's 'A comparison between confucianism and phenomenology of feelings'[J]. *Journal of Graduate School of Chinese Academy of Social Sciences*, *159*(3), 33–40.

Hymes, Dell. (1972). On communicative competence[A]. In J. B. Pride & Janet Holmes (Eds.) *Sociolinguistics*[C] (pp. 269–293). Penguin.

Ireland, Kathleen & Tenenbaum, David. (2008). *Visualizing human biology*[M]. John Wiley.

Jaimes, Alejandro & Sebe, Nicu. (2007). Multimodal human-computer interaction: A survey[J]. *Computer Vision and Image Understanding*, *108*, 116–134.

Jewitt, Carey. (2009). *The Routledge handbook of multimodal analysis*[C]. Routledge.

Jewitt, Carey, Bezemer, Jeff, & O'Halloran, Kay. (2016). *Introducing multimodality*[M]. Routledge.

Jiang, Wangqi. (2014). A personal view on pragmatic inference[J]. *Modern Foreign Languages*, *37*(3), 293–302.

Jin, S. (1992). Tone and intonation of Beijing dialect[J]. *Studies of the Chinese Language*, (2), 113–123.

Kallen, Jeffrey L. & Kirk, John. (2012). *SPICE-Ireland: A user's guide*[R]. Cló Ollscoil na Banríona.

Keltner, Dacher & Ekman, Paul. (2000). Facial expression of emotion[A]. In Michael Lewis & Jeannette M. Haviland-Jones (Eds.) *Handbook of emotions* (2nd ed) [C] (pp. 236–251). Guilford Publications.

Kemper, Theodore D. (1987). How many emotions are there? Wedding the social and autonomic components[J]. *American Journal of Sociology*, *93*(2), 263–289.

Kendon, Adam. (1967). Some functions of gaze direction in social interaction[J]. *Acta Psychologica*, *26*, 22–63.

Kendon, Adam. (1972). Some relationships between body motion and speech[A]. In Aron Seigman & Benjamin Pope (Eds.) *Studies in dyadic communication*[C] (pp. 177–216). Pergamon.

Kendon, Adam. (1990). *Conducting interaction: Patterns of behavior in focused encounters*[M]. Cambridge University Press.

Kendon, Adam. (1992). The negotiation of context in face-to-face interaction[A]. In Alessandro Duranti & Charles Goodwin (Eds.) *Rethinking context: Language as an interactive phenomenon*[C] (pp. 323–332). Cambridge University Press.

Kendon, Adam. (1995). Gestures as illocutionary and discourse structure markers in Southern Italian conversation[J]. *Journal of Pragmatics*, *23*(3), 247–279.

Kendon, Adam. (2004). *Gesture: Visible action as utterance*[M]. Cambridge University Press.

Kipp, Michael & Martin, Jean-Claude. (2009). Gesture and emotion: Can basic gestural form features discriminate emotions?[A]. In Cohn, J., Nijholt, A., & Pantic, M. (Eds.) Proceedings *of* 2009 3rd International Conference on Affective Computing and Intelligent Interaction, ACII: September 10–12, 2009, Amsterdam, The Netherlands. IEEE Computer Society. IEEE Press: 505–512.

Kipp, Michael, Martin, Jean-Claude, Paggio, Patrizia, & Heylen, Dirk (Eds.) (2009). *Multimodal Corpora: From models of natural interaction to systems and applications*[C]. Springer.

Knight, Dawn. (2009). *A multi-modal corpus approach to the analysis of backchanneling behaviour*[D]. Ph.D. dissertation, University of Nottingham, pp. 16–28.

Knight, Dawn. (2011a). *Multimodality and active hearership: A corpus approach*[M]. Bloomsbury.

Knight, Dawn. (2011b). The future of multimodal corpora[J]. *Brazilian Journal of Applied Linguistics*, *11*(2), 391–415.

Knight, Dawn & Adolphs, Svenja. (2008). Multi-model corpus pragmatics: The case of active hearership[A]. In Jesús Romero-Trillo (Ed.) *Pragmatics and corpus linguistics*[C] (pp. 175–190). Walter de Gruyter.

Koester, Almut. (2010). Building small specialised corpora[A]. In Anne O'Keeffe & Michael McCarthy (Eds.) *The Routledge handbook of corpus linguistics*[C] (pp. 66–79). Routledge.

Kohnen, Thomas. (2000). Corpora and speech acts: The study of performatives[A]. In Christian Mair and Marianne Hundt (Eds.) *Corpus linguistics and linguistic theory. Papers from the twentieth international conference on english language research on computerized Corpora (ICAME 20)*[C]. Freiburg im Breisgau 1999 (pp. 177–186). Rodopi.

Kolb, Bryan & Whishaw, Ian Q. (2005). *Fundamentals of human neuropsychology*[M]. Worth.

Kress, Gunther. (2009). What is mode?[A] In Carey Jewitt (Ed.) *The Routledge handbook of multimodal analysi*s[C] (pp. 54–67). Routledge.

Kress, Gunther. (2010). *Multimodality*[M]. Routledge.

Kress, Gunther & van Leeuwen, Theo. (2001). *Multimodal discourse: The modes and media of contemporary communication*[M]. Arnold.

Kress, Gunther & van Leeuwen, Theo. (2006). *Reading images: The grammar of visual design* (2nd ed.)[M]. Routledge.

Kuang, C. (2013). *A corpus-based study on thanking speech acts in Chinese*[D]. Master dissertation. University of Electronic Science and Technology of China, Sichuan.

Kuhn, Thomas S. (1996). *The structure of scientific revolutions* (3rd ed.)[M]. University of Chicago Press.

Kwiatkowska, Alina. (2005). Pictorial acts of communication[A]. In Piotr Cap (Ed.) *Pragmatics today*[C]. Peter Lang.

Ladd, D. Robert. (1980). *The structure of intonational meaning: Evidence from English*[M]. Indiana University Press.

Ladd, D. Robert. (1996). *Intonational phonology*[M]. Cambridge University Press.

Lazarus, Richard S. (1991). *Emotion and adaptation*[M]. Oxford University Press.

Ledoux, Joseph E. (1992). Emotion and the amygdala[A]. In John P. Aggleton (Ed.) *The amygdala*[C]. Wiley-Liss.

Leech, Geoffrey. (1983). *Principles of pragmatics*[M]. Longman.

Leech, Geoffrey. (1993). Corpus annotation schemes[J]. *Literary and Linguistic Computing*, *8*(4), 275–281.

Leech, Geoffrey & Weisser, Martin. (2003). Generic speech act annotation for task-oriented dialogues[A]. In Dawn Archer, Paul Rayson, Andrew Wilson & Tony McEnery (Eds.) *Proceedings of the Corpus Linguistics 2003 Conference*[C]. Lancaster University: UCREL Technical Papers, vol. 16.

Leech, Geoffrey & Weisser, Martin. (2014). *The SPAADIA Annotation Scheme*[K] [online] http://www.martinweisser.org/#spaadia

Lepore, Ernest & Van Gulick, Robert. (Eds.) (1991). *John Searle and his critics*[M]. Blackwell.

Levelt, Willem J. M. (1989). *Speaking: From intention to articulation*[M]. The MIT Press.

LeVine, Philip & Scollon, Ron. (2004). *Discourse and technology: Multimodal discourse analysis*[C]. Georgetown University Press.

Levinson, Stephen C. (1979). Activity types and language[A]. In Paul Drew & John Heritage (Eds.) *Talk at work: Interaction in institutional settings*[C] (pp. 66–100). Cambridge University Press.

Levinson, Stephen C. (1983). *Pragmatics*[M]. Cambridge University Press.

Li, Aijun. (2002). Prosodic analysis on conversations in Standard Chinese[J]. *Studies of the Chinese Language*, (6), 525–535.

Li, Aijun, et al. (2008). Relationships between gestures and speech in spontaneous Chinese speech[J]. *Journal of Tsinghua University (Sci & Tech)*, *48*(S1), 627–634.

Li, B. T. (1985). *A brief history of Chinese textbooks of primary school*[M]. Shandong Education Press.

Li, H. B. (2013). Multimodal research methods and research fields[J]. *Journal of Xi'an International Studies University*, *21*(3), 21–25.

Li, J. Q. (2002). *Introduction of non-verbal communication*[M]. Peking University Press.

Li, Xiaoting. (2014). *Multimodality, interaction and turn-taking in mandarin conversation*[M]. John Benjamins.

Li, Y. X. (2015). *Construction of multimodal corpus of classroom teaching (MCCT) and its application with a special reference to teacher self-development*[D]. Ph.D. dissertation, Tongji University, Shanghai.

Liao, S. Y., Wang S. C., & Zhang J. S. (2015). *Study on complex adaptive system and agent-based modeling and stimulation*[M]. National Defense Industry Press.

Lin, M. C. (2000). Breaks and prosodic phrases in the utterances of standard Chinese[J]. *Contemporary Linguistics*, *2*(4), 41–50.

Lin, M. C. (2001). *Prosodic structure and hierarchy of accent in Chinese*[C]. Festschrift in honor of the 80th birthday of Professor Li Rong. The Commercial Press.

Lin, M. C. (2002). Prosodic structure and lines of F0 top and bottom of utterances in Chinese[J]. *Contemporary Linguistics*, *4*(4), 254–265.

Lin, Phoebe M. S. & Adolphs, Svenja. (2009). Sound evidence: Phraseological units in spoken corpora[A]. In A. Barfield & H. Gyllstad (Eds.) *Researching collocations in another language: Multiple interpretations*[C] (pp. 34–48). Palgrave Macmillan.

Lindström, Jan. (2009). Interactional linguistics[A]. In Sigurd D'Hondt, Jan-Ola Östman, & Jef Verschueren (Eds.) *The pragmatics of interaction*[C] (pp. 96–103). John Benjamins.

Liu, C. Y. (2013). Study on the acquisition of modal sequences as speech act formulas: Using corpus-based methods[J]. *Journal of PLA University of Foreign Languages*, *36*(1), 62–66.

Liu, H. X. (2009). *The experimental study of Chinese Mandarin Chinese emotional intonation*[D], Master dissertation. Shanghai Normal University, Shanghai.

Liu, L. L. (2007). A review of Chinese prosody research in the past eight decades[J]. *Linguistic Research*, *103*(2), 5–11.

Liu, Q. & Pan, M. W. (2010). Construction of a multimodal corpus of oral English for Chinese science and engineering majors[J]. *Modern Educational Technology*, *20*(4), 69–72, 119.

Liu, Y. (2011). *Prosody analysis of Mandarin emotional speech*[D], Master dissertation. Nanjing Normal University, Nanjing.

Maragos, Petros, Potamianos, Alexandros, & Gros, Patrick. (2008). *Multimodal processing and interaction: Audio, video, text*[C]. Springer.

Martin, Jean-Claude M. & Devillers, Laurence. (2009). A multimodal corpus approach for the study of spontaneous emotions[A]. In Jianhua Tao & Tieniu Tan (Eds.) *affective information processing*[C] (pp. 267–292). Springer.

Maynard, Carson & Leicher, Sheryl. (2007). Pragmatic annotation of an academic spoken corpus for pedagogical purposes[A]. In Eileen Fitzpatrick (Ed.) *Corpus linguistic beyond the word. Corpus research from phrase to discourse*[C] (pp. 107–115). Rodopi.

McAllister, Paula G. (2015). Speech acts: A synchronic perspective[A]. In Karin Aijmer & Christoph Rühlemann (Eds.) *Corpus pragmatics: A handbook*[C] (pp. 29–51). Cambridge University Press.

McBurney, Susan L. (2002). Pronominal reference in signed and spoken language: Are grammatical categories modality-dependent?[A]. In Richard P. Meier, Kearsy Cormier, & David Quinto-Pozos (Eds.) *Modality and structure in signed and spoken languages*[C] (pp. 329–369). Cambridge University Press.

McCarthy, Michael. (1991). *Discourse analysis for language teachers*[M]. Cambridge University Press.

McEnery, Tony, Xiao, Richard, & Tono, Yukio. (2006). *Corpus-based language studies: An advanced resource book*[M]. Routledge.

Mcneill, David. (1992). *Hand and mind: What gestures reveal about thought*[M]. University of Chicago Press.

Meng, P. Y. (1998). *Spiritual transcendence and State of Mind*[M]. People's Publishing House.

Meng, P. Y. (2002). *Emotion and reason*[M]. China Social Sciences Press.

Meng, Z. L. (1989). *Human emotions*[M]. People's Publishing House.

Mey, Jacob L. (2001). *Pragmatics: An introduction* (2nd ed.)[M]. Blackwell.

Meyer, Charles F. (2002). *English corpus linguistics: An introduction*[M]. Cambridge University Press.

Meyer, Roland & Mleinek, Ina. (2006). How prosody signals force and focus – A study of pitch accents in Russian yes–no questions[J]. *Journal of Pragmatics*, *8*(10), 1615–1635.

Mubenga, Kajingulu S. (2008). *Film discourse and pragmatics in screen translation: A contrastive analysis of speech acts in French and English*[D], Ph.D. dissertation. University of the Witwatersrand, Johannesburg.

Mubenga, Kajingulu S. (2009). Towards a multimodal pragmatic analysis of film discourse in audiovisual translation[J]. *Meta: Journal des traducteurs*, *54*(3), 466–484.

Nespor, Marina & Vogel, Irene. (1986). *Prosodic phonology*[M]. Foris.

Norris, Sigrid. (2004). *Analyzing multimodal interaction: A methodological framework*[M]. Routledge.

Norris, Sigrid. (2019). *Systematically working with multimodal data: Research methods in multimodal discourse analysis*[M]. Wiley Blackwell.

Norris, Sigrid & Jones, Rodney H. (2005). *Discourse in action: Introducing mediated discourse analysis*[C]. Routledge.

Ochs, Elinor. (1979). Transcription as theory[A]. In Elinor Ochs & Bambi Schieffelin (Eds.) *Developmental pragmatics*[C] (pp. 43–72). Academic Press.

O'Connell, Daniel & Kowal, Sabine. (2009). Transcription systems for spoken discourse[A]. In Sigurd D'hondt, Jan-Ola Östman, & Jef Verschueren (Eds.) *The pragmatics of interaction*[C] (pp. 240–254). John Benjamins.

O'Connor, Joseph D. & Arnold, Gordon F. (1961). *Intonation of colloquial English*[M]. Longman.

O'Halloran, Kay L. (2005). *Mathematical discourse: Language, symbolism and visual image*[M]. Continuum.

O'Halloran, Kay L. (2009). Multimodal analysis and digital technology[A]. In Anthony Baldry and Elena Montagna (Eds.) *Interdisciplinary perspectives on multimodality: Theory and practice*[C], *Proceedings of the Third International Conference on Multimodality*, Palladino, Campobasso: 1–26.

O'Keeffe, Anne, Clancy, Brian, & Adolphs, Svenja. (2011). *Introducing pragmatics in use*[M]. Taylor & Francis.

Olson, David R. (2013). From utterance to text: The bias of language in speech and writing[A]. In Mastin Prinsloo & Mike Baynham (Eds.) *Literacy studies I*[C] (pp. 1–28). Sage.

Ortony, Andrew, Clore, Gerald L., & Collins, Allan. (1988). *The cognitive structure of emotions*[M]. Cambridge University Press.

O'Toole, Michael. (1994). *The language of displayed art*[M]. Leicester University Press.

Pennock-Speck, Barry & de Saz-Rubio, Ma Milagros. (2013). A multimodal analysis of facework strategies in a corpus of charity ads on British television[J]. *Journal of Pragmatics*, *49*, 38–56.

Perkins, Michael R. (2007). *Pragmatic impairment*[M]. Cambridge University Press.

Perniss, Pamela. (2018). Why we should study multimodal language[J]. *Frontiers in Psychology*, *9*, 1109–1109.

Pijper, Jan Roelof & Sanderman, Angelien A. (1994). On the perceptual strength of prosodic boundaries and its relation to suprasegmental[J]. *Journal of the Acoustical Society of America*, *96*(4), 2037–2047.

Planalp, Sally. (1998). Communicating emotion in everyday life: Cues, channels, and processes[A]. In Peter A. Andersen & Laura K. Guerrero (Eds.) *Handbook of communication and emotion: Research, theory, applications, and contexts*[C] (pp. 29–48). Academic Press.

Planalp, Sally & Knie, Karen. (2002). Integrating verbal and nonverbal emotion(al) message[A]. In Susan R. Fussell (Ed.) *The verbal communication of emotions: Interdisciplinary perspectives*[C] (pp. 55–79). Psychology Press.

Plutchik, Robert. (2001). The nature of emotions[J]. *American Scientist*, *89*(4), 344–350.

Plutchik, Robert & Kellerman, Henry. (1980). *Emotion: Theory, research, and experience: Vol. 1. Theories of emotion*[M]. Academic.

Qian, Y. H. & Chen, X. R. (2014). A survey of the use of corpus-based approach in pragmatics research[J]. *Foreign Language Teaching*, *146*(2), 15–20, 26.

Rayson, Paul. (2003). *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison*[D], Ph.D. dissertation. Lancaster University.

Robin, Donald A., Tranel, Daniel, & Damasio, Hanna. (1990). Auditory perception of temporal and spectral events in patients with focal left and right cerebral lesions[J]. *Brain and Language*, *39*(4), 539–555.

Romero-Trillo, Jesús. (2008). *Pragmatics and corpus linguistics*[C]. Walter de Gruyter.

Ross, James R. (1970). On declarative sentences[A]. In Roderick A. Jacobs & Peter S. Rosenbaum (Eds.) *Readings in transformational grammar*[C] (pp. 222–272). Ginn-Blaisdell.

Rühlemann, Christoph. (2010). What can a corpus tell us about pragmatics?[A] In Anne O'Keeffe & Michael McCarthy (Eds.) *The Routledge handbook of corpus linguistics*[C] (pp. 288–301). Routledge.

Rühlemann, Christoph & Aijmer, Karin. (2014). Corpus pragmatics: Laying the foundations[A]. In Karin Aijmer & Christoph Rühlemann (Eds.) *Corpus pragmatics: A handbook*[C]. (pp. 1–28). Cambridge University Press.

Rühlemann, Christoph, Bagoutdinov, Andrej, & O'Donnell, Matthew B. (2013). Windows on the mind: Pauses in conversational narrative[A]. In Gaëtanelle Gilquin & Sylvie De Cock (Eds.) *Errors and disfluencies in spoken corpora: Cross-linguistic perspective*s[C] (pp. 59–91). John Benjamins.

Russell, James A. & Fernandez-Dols, José-Miguel (Eds.) (1997). *The psychology of facial expression*[C]. Cambridge University Press.

Sacks, Harvey. (1984). Notes on methodology[A]. In J. Maxwell Atkinson & John Heritage (Eds.) *Structures of social action: Studies in conversation analysis*[C] (pp. 21–27). Cambridge University Press.

Sacks, Harvey, Schegloff, Emanuel A., & Jefferson, Gail. (1974). A simplest systematics for the organization of turn-taking for conversation[J]. *Language*, *50*(4), 696–735.

Sadock, Jerrold M. (1974). *Toward a linguistic theory of speech acts*[M]. Academic Press.

Sagisaka, Yoshinori, Campbell, Nick, & Higuchi, Norio. (1997). *Computing prosody: Computational models for processing spontaneous speech*[M]. Springer.

Sargent, Robert G. (2009). Verification and validation of simulation models[A]. In M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, & R. G. Ingalls (Eds.) *Proceedings of the 2009 Winter Simulation Conference*, IEEE, Piscataway, NJ: 162–176.

Scheflen, A. E., Kendon, Adam, & Schaeffer, J. A. (1970). A comparison of videotape and moving picture film in research in human communication[A]. In Milton M. Berger (Ed.) *Videotape techniques in psychiatric training and treatment*[C]. Brunner.

Schegloff, Emanuel A. (1984). On some questions and ambiguities in conversation[A]. In J. Maxwell Atkinson & John Heritage (Eds.) *Structures of social action: Studies in conversation analysis*[C] (pp. 28–52). Cambridge University Press.

Schegloff, Emanuel A. (2007). *Sequence organization in interaction: A primer in conversation analysis I*[M]. CUP.

Scherer, Klaus R. (1987). Toward a dynamic theory of emotion: The component process model of affective states[J]. *Geneva Studies in Emotion and Communication*, *1*, 1–98.

Scherer, Klaus R. (2003). Vocal communication of emotion: A review of research paradigms[J]. *Speech Communication*, *40*, 227–256.

Scherer, Klaus R. (2009). The dynamic architecture of emotion: Evidence for the component process model[J]. *Cognition and Emotion*, *23*, 1307–1351.

Scherer, Klaus R. (2014). Corpus design for studying the expression of emotion in speech[A]. In Tommaso Raso & Heliana Mello (Eds.) *Spoken corpora and linguistic studies*[C] (pp. 210–232). John Benjamins.

Scherer, Klaus R. & Ekman, Paul. (1984). *Approaches to emotion*[M]. Lawrence Erlbaum.

Schmitt, Norbert (Ed.) (2004). *Formulaic sequences: Acquisition, processing and use*[C]. John Benjamins.

Scollon, Ron & Suzie W. Scollon. (2004). *Nexus analysis: Discourse and the emerging internet*[M]. Routledge.

Searle, John R. (1969). *Speech acts: An essay in the philosophy of language*[M]. Cambridge University Press.

Searle, John R. (1971). What is a speech act?[A]. In John R. Searle (Ed.) *The philosophy of language*[C] (pp. 39–53). Oxford University Press.

Searle, John R. (1976). A classification of illocutionary acts[J]. *Language in Society*, *5*(1), 1–23.

Searle, John R. (2001/1979). *Expression and meaning: Studies in the theory of speech acts*[M]. Foreign Langauge Teaching & Research Press.

Searle, John R., Kiefer, Ferenc, & Bierwisch, Manfred. (1980). *Speech act theory and pragmatics*[M]. Reidel.

Searle, John R., Parret, Herman, & Verschueren, Jef. (1992). *On Searle on conversation*[C]. John Benjamins.

Searle, John R. & Vanderveken, Daniel. (1985). *Foundations of illocutionary logic*[M]. Cambridge University Press.

Seilkirk, Elisabeth O. (1984). *Phonology and syntax*[M]. MIT Press.

Selting, Margret. (2010). Prosody in interaction: State of the art[A], In Dagmar Barth-Weingarte, Elisabeth Reber, & Margret Selting (Eds.) *Prosody in interaction*[C] (pp. 3–40). John Benjamins.

Shen, J. (1998). Some preliminary views on the classification of Chinese intonation and its way of marking[J]. *Applied Linguistics*, (1), 102–104.

Shi, P. W. (1980). Intonational changes in four types of sentences[J]. *Language Teaching and Linguistic Studies*, (2), 71–81.

Siegman, Aron W. & Feldstein, Stanley. (1987). *Nonverbal behavior and communication* (2nd ed.)[M]. Lawrence Erlbaum.

Sinclair, John. (1991). *Corpus, concordance, collocation*[M]. Oxford University Press.

Sinclair, John. (1996). The search for units of meaning[J]. *TEXTUS*, IX, 75–106.

Sinclair, John. (2008). Borrowed ideas[A]. In A. Gerbig & O. Mason (Eds.) *Language, people, numbers – Corpus linguistics and society*[C] (pp. 21–42). Rodopi.

Sinclair, John & Coulthard, R. Malcolm. (1975). *Toward an analysis of discourse*[M]. Oxford University Press.

Stavropoulou, Pepi. (2002). *Predicting prosodic phrasing*[D], Master dissertation. Department of Theoretical and Applied Linguistics, University of Edinburgh.

Stefanucci, Jeanine K. & Proffitt, Dennis R. (2005). *Multimodal interfaces improve memory*[R]. Paper presented at the 11th International Conference on Human-Computer Interaction, Las Vegas, NV.

Stiles, William B. (1992). *Describing talk: A taxonomy of verbal response modes*[M]. Sage.

Stockley, Sue. (2006). *The development of an analytical tool for automated dialogue act annotation of spoken corpora*[D], PhD dissertation. Lancaster University.

Stolcke, Andreas, Ries, Klaus, Coccaro, Noah, Shriberg, Elizabeth, Bates, Rebecca, Jurafsky, Daniel, Taylor, Paul, Martin, Rachel, Van Ess-Dykema, Carol, & Meteer, Marie. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech[J]. *Computational Linguistics*, *26*(3), 339–371.

Strawson, Peter F. (1964). Intention and convention in speech acts[J]. *The Philosophical Review*, *73*(4), 439–460.

Sun, B. Z. (2009). Review of researches of literacy amount requirement and distribution for pupils[J]. *Shanghai Research on Education*, *266*(10), 71–72.

Taavitsainen, Irma & Jucker, Andreas H. (Eds). (2014). *Diachronic corpus pragmatics*[C]. (Pragmatics & Beyond New Series). John Benjamins.

Tang, X. W. (2006). *Introduction to brain science*[M]. Zhejiang University Press.

Tao, Hongyin. (1996). *Units in mandarin conversation: Prosody, discourse, and grammar*[M]. John Benjamins.

Tao, Hongyin. (2004). Fundamentals in spoken language research[J]. *Linguistic Sciences*, *3*(1), 50–67.

Tao, J. H. & Tan, T. N. (2004). Digitalizing human emotions: Affective computing in a harmonious environment of human-computer interaction[J]. *PC World China*, (1), 29–32.

Tao, J. H., Yang, M. H., Li, Y., Pan, S. F., & Mu, K. H. (2011). Multimodal fusion of interpersonal discourse system[J]. *Communications of the CCF*, *7*(11), 30–39.

Taylor, Martin M., Néel, Françoise, & Bouwhuis, Don. (Eds.) (2000). *The structure of multimodal dialogue II*[C]. John Benjamins.

Tench, Paul. (1996). *The intonation systems of English*[M]. Cassell.

Thomsen, Ole N. (2010). From talking heads to communicating bodies: Cybersemiotics and total communication[J]. *Entropy*, *12*, 390–419.

Thrift, Nigel. (1977). *An introduction to time geography (Concepts and techniques in modern geography no. 13)*[M]. Geo-Abstracts.

Tseng, Shu-Chuan. (Ed.) (2009). *Linguistic patterns in spontaneous speech*[M]. Academia Sinica.

Tsui, Ami. (1994). *English conversation*[M]. Oxford University Press.

Turner, Jonathan H. (2000). *On the origins of human emotions: A sociological inquiry into the evolution of human affect*[M]. Stanford University Press.

Turner, Jonathan H. & Stets, Jane. (2005). *The sociology of emotions*[M]. Cambridge University Press.

Turner, Jonathan H. & Stets, Jane E. (2007). *The sociology of emotions*[M]. (J. C. Sun & J. Wen, Trans.). Shanghai People's Publishing House.

UNESCO. (1953). *World illiteracy at mid-century*[R]. UNESCO.

van Kuppevelt, Jan C. J., Dybkjær, Laila, & Bernsen, Niels Ole. (2005). *Advances in natural multimodal dialogue systems*[C]. Springer.

van Lancker Sidtis, D. (2004). When novel sentences spoken or heard for the first time in the history of the universe are not enough: Toward a dual-process model of language[J]. *International Journal of Language and Communication Disorders*, *39*(1), 1–44.

van Lancker Sidtis, D. (2008). The relation of human language to human emotion[A]. In Brigitte Stemmer & Harry H. Whitaker (Eds.) *Handbook of the neuroscience of language*[C] (pp. 199–208). Academic Press.

van Leeuwen, Theo. (1999). *Speech, music, sound*[M]. Macmillan.

van Leeuwen, Theo. (2005). *Introducing Social Semiotics*[M]. Routledge.

Vanderveken, Daniel. (1994). A complete formulation of a single logic of elementary lllocutionary acts[A]. In Savas L. Tsohatzidis (Ed.) *Foundations of speech act theory*[C]. Routledge.

Vaughan, Elaine & Clancy, Brian. (2013). Small corpora and pragmatics[A]. In Jesús Romero-Trillo (Ed.) *Yearbook of corpus linguistics and pragmatics 2013*[C] (pp. 53–76). Springer.

Verschueren, Jef. (1980). *On speech act verbs*[M]. John Benjamins.

Verschueren, Jef. (1999). *Understanding pragmatics*[M]. Arnold.

Vogel, Irene. (2009). Universals of prosodic structure[A]. In Sergio Scalise, Elisabetta Magni, & Antonietta Bisetto (Eds.) *Universals of language today*[C] (pp. 59–82). Springer.

Wahlster, Wolfgang. (2006). *SmartKom: Foundations of multimodal dialogue systems*[M]. Springer.

Wang, D. C. (1983). *Exploration on rhetoric*[M]. Beijing Publishing Group.

Wang, H. J. (2006). Annotation and criticism on the classifications of "illocutionary act"[J]. *Journal of Jiangxing University*, *18*(1), 84–87.

Wang, H. J. (2008). *Non-linear phonology of Chinese (Rev. ed.)*[M]. Peking University Press.

Wang, J. (2014). An exploration of empathy, sympathy, and fellow-feeling from the perspective of neuroscience and western moral psychology[J]. *Social Sciences in Yunnan*, *197*(1), 50–55.

Wang, M. L. (2003). *Prosodic patterns of Chinese spontaneous speech*[D], Ph.D. dissertation. University of Chinese Academy of Social Sciences, Beijing.

Wang, S. Y. (2006a). Language is a complex adaptive system[J]. *Journal of Tsinghua University (Philosophy and Social Sciences)*, *21*(6), 5–13.

Wang, S. Y. (2006b). Computational modeling in evolutionary linguistics[J]. *Journal of Peking University (Philosophy and Social Sciences)*, *43*(2), 17–22.

Wang, S. Y. & Ke, J. Y. (2001). A preliminary study on language emergence and simulation modeling[J]. *Studies of the Chinese Language*, *282*(3), 195–200, 282.

Wang, S. Y., & Peng, G. (2006). *Language, speech, and technology*[M]. Shanghai Educational Publishing House.

Wang, Z. C., & Wang, Y. (2007). Research evolvement and direction of complex system simulation conceptual model[J]. *Journal of Astronautics*, *28*(4), 779–785.

Wang, Z. Y. (1990). Sidelights of the first seminar of pragmatics in China[J]. *Foreign Language Teaching and Research*, (1), 9–13.

Weisser, Martin. (2003). SPAACy: A semi-automate tool for annotating dialogue acts[J]. *International Journal of Corpus Linguistics*, *8*(1), 63–74.

Weisser, Martin. (2015). Speech act annotation[A]. In Karin Aijmer & Christoph Rühlemann (Eds.) *Corpus pragmatics: A handbook*[C] (pp. 84–110). Cambridge University Press.

Wennerstrom, Ann. (2001). *The music of everyday speech: Prosody and discourse analysis*[M]. Oxford University Press.

Wharton, Tim. (2009). *Pragmatics and non-verbal communication*[M]. Cambridge University Press.

Whitaker, Harry A. (2010). *Concise encyclopedia of brain and language*[K]. Elsevier.

Wichmann, Anne & Blakemore, Diane. (2006). Introduction: The prosody-pragmatics interface[J]. *Journal of Pragmatics*, *38*(10), 1537–1541.

Wierzbicka, Anna. (1987). *English speech act verbs*[M]. Academic Press.

Wierzbicka, Anna. (1999). *Emotions across languages and cultures*[M]. Cambridge University Press.

Wittenburg, Peter. (2008). Preprocessing multimodal corpora[A], In Anke Lüdeling & Merja Kytöl (Eds.) *Corpus linguistics: An international handbook*[C] Vol. 1 (pp. 664–684). Walter de Gruyter.

Wu, J. M. & Lü, S. N. (2011). A discussion on the prosodic characteristics of the expressive intonation[J]. *Studies of the Chinese Language*, *345*(6), 540–549.

Wu, J. M. & Zhu, H. D. (2001). *Chinese prosody* (pp. 395–403)[M]. Language & Culture Press.

Wu, Z. J. (2004). Intonation rules in Chinese[A]. *Collected essays on linguistics of Wu Zongji*[C]. The Commercial Press.

Xiang, Ji. (2007). *Semantic prosody and role of promises in commissives: Corpus-based study of perlocutionary effect of commissives*[D], Master dissertation. Shanghai Jiao Tong University, Shanghai.

Xu, M. F. (2008). The default rules of speech act verbs[J]. *Contemporary Rhetoric*, *145*(1), 10–18.

Yang, L. X. & Fu, X. L. (2001). Development in the study of syntactic ambiguity resolution: The constraint-satisfaction model of ambiguity resolution[J]. *Psychological Science*, *24*(4), 465–467.

Yang, X. Z. & Xin, Z. Y. (2010). Reviews of multimodal studies[A]. In G. W. Huang & C. G. Chang (Eds.) *Annual review of functional linguistics*[C]. Higher Education Press: 1, 23–34

Yi, W. & Lu, S. Y. (2013). The psychological reality of formulaic language[J]. *Advances in Psychological Science*, *21*(12), 2110–2117.

Yin, J. & Wang, Y. N. (2016). The applicability of complex systems paradigms in social sciences[J]. *Social Sciences in China*, *243*(3), 62–79.

Yu, J. Y. (2014). Philosophy and architecture of Qian Xuesen's system science[J]. *Scientific Decision Making*, *209*(12), 2–22.

Zaefferer, Dietmar. (1981). On a formal treatment of illocutionary force indicators[A]. In Herman Parret, Marina Sbisa, & Jef Verschueren (Eds.) *Possibilities and limitations of pragmatics*[C], Proceedings of the Conference on Pragmatics, Urbino, July 8–14, 1979. John Benjamins.

Zhang, De-Lu. (2009). Multimodal discourse theory and its application to foreign language teaching with modern media technology[J]. *Foreign Language Education*, *30*(4), 15–20.

Zhang, De-Lu. (2012). Exploration of multimodal learning ability training model[J]. *Foreign Languages Research*, *132*(2), 9–14.

Zhang, De-Lu & Wang, Lu. (2010). The synergy of different modes in multimodal discourse and their realization in foreign language teaching[J]. *Foreign Languages Research*, *153*(2), 97–102.

Zhang, L. X. (2014). *A study of sentential accent in Chinese discourse from the perspective of pragmatics*[M]. World Publishing Corporation, Guangdong branch.

Zhang, Y. (2008). A review of studies on prosodic features of modal particles[J]. *Language Teaching and Linguistic Studies*, (2), 53–59.

Zhang, Z. C. & Chen, Y. M. (2011). A critical survey of three approaches to multimodal discourse[J]. *English Education in China*, *31*(1), 1–11.

Zhao, X. F. (2013). On multimodal cognitive poetics: A new direction of cognitive poetics[J]. *Journal of Sichuan International Studies University*, *29*(6), 43–51.

Zhong, S. M. (2008). Paraphrase of speech act verbs and its interrelated study[J]. *Foreign Language Education*, *29*(5), 13–17.

Zuo, Y. (1999). An elementary analysis of a few concepts in prosodic studies[J]. *Journal of Peking University (Humanities and Social Sciences)*, *36*(S1), 85–88.

Index

Note: **Bold** page numbers refer to tables, *Italic* page numbers refer to figures and page number followed by “n” refer to end notes.