Synthesis Lectures on
Information Concepts, Retrieval, and Services

SYNTHESIS
COLLECTION OF TECHNOLOGY

Betsy Van der Veer Martens

Keywords In and Out of Context

Springer

Synthesis Lectures on Information Concepts, Retrieval, and Services

Series Editor

Gary Marchionini, School of Information and Library Science, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

This series publishes short books on topics pertaining to information science and applications of technology to information discovery, production, distribution, and management. Potential topics include: data models, indexing theory and algorithms, classification, information architecture, information economics, privacy and identity, scholarly communication, bibliometrics and webometrics, personal information management, human information behavior, digital libraries, archives and preservation, cultural informatics, information retrieval evaluation, data fusion, relevance feedback, recommendation systems, question answering, natural language processing for retrieval, text summarization, multimedia retrieval, multilingual retrieval, and exploratory search.

Betsy Van der Veer Martens

Keywords In and Out of Context

Springer

Betsy Van der Veer Martens
University of Oklahoma
Norman, OK, USA

ISSN 1947-945X ISSN 1947-9468 (electronic)
Synthesis Lectures on Information Concepts, Retrieval, and Services
ISBN 978-3-031-32529-8 ISBN 978-3-031-32530-4 (eBook)
https://doi.org/10.1007/978-3-031-32530-4

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To my husband Jon, with "love" as keyword

Preface

Why this book about the "keyword"? Today, keywords and their multiple uses serve to bridge between the humanities and technology, between librarianship and information science, between the searcher and the web resource, between the search engine and the advertiser, between the social influencer and the audience, between the political strategist and the voter turnout, and between our contemporary world and that of our ancient predecessors. The keyword in its various guises (key word, concept symbol, hashtag, and search term) can point not only to text and other forms of media, but to associated ways of thinking and acting based on specific words that we may consider "key."

This project is an effort to explore the rich history of the keyword from its earliest manifestations (long before it appeared anywhere in Google Trends or library cataloging textbooks) in order to illustrate its implicit and explicit mediation of human cognition and communication processes, from its deictic origins in primate and proto-speech communities, through its semiotic and symbolic instantiation in various physical artifacts and structures, through its development within oral traditions, through its initial appearances in numerous graphical forms, through its workings over time within a variety of indexing traditions and technologies, to its role in search engine optimization and social media strategies, to its potential as an element in the slowly emerging semantic web as well as in multiple voice search applications. The purpose of the book is to synthesize different perspectives on the significance of this often-invisible intermediary, both in and out of the library and information science context, and to understand how it has come to be so embedded in our daily life.

Norman, USA — Betsy Van der Veer Martens

Contents

Representation, Reference, Relevance, and Retention

1

Like many of the words that matter most, that tell us most about our intellectual and material life and about our cognitive and perceptual habits, 'keyword' hides in plain view. —Michael Leja, "Keyword" (2009)

Abstract

The long history of keywords and their predecessors as semiotic, symbolic, and semantic pointers to key concepts over time is introduced. This chapter describes current findings on four sensory specifics that are generally not considered as being aspects of library and information science but that are keywords which ground the discipline both physically and conceptually: that is, vision for representation, voice for reference, hearing for relevance, and memory for retention.

While words in any natural language can serve as symbolic and semantic tools in individual and social cognition, due to their ability to mobilize both abstract and concrete concepts in representation and reference and even to immobilize these in various ways for retention and retrieval over time, some of these verbal tools appear to be particularly useful as keys to communication, maintaining and retaining their significance for larger groups of people and for longer periods of time. However, as Leja (2009) observes above, the presence of these "key" words, especially in their functions as "keywords," is largely taken for granted.

B. V. V. Martens, *Keywords In and Out of Context*, Synthesis Lectures on Information Concepts, Retrieval, and Services, https://doi.org/10.1007/978-3-031-32530-4_1

1.1 Defining the Keyword

The Oxford English Dictionary defines "keyword" as either "a. A word that serves as the key to a cipher or code" or "b. A word or idea that serves as a solution or explanation for something; a word, expression, or concept of particular importance or significance." Rosenberg (2021) in discussing these two definitions suggests that they are best represented today by the prevalence of keyword searching in search engines and by the prevalence of polarizing terms in social discourse. His genealogy of the modern trajectory for both definitions can be traced to 1958, the year in which IBM engineer Hans-Peter Luhn (1958) published his method for automatically extracting and indexing "significant" words from scientific and technological articles and in which cultural theorist Raymond Williams (1958) published his initial analysis of how particular words as used by individual writers were the key to analyzing changes in cultural and social mores. These two trajectories, the technological and the cultural, have coincided in today's information environment, to the point that, as Rosenberg says, "What makes *keyword* such a powerful idea is precisely the ambiguity of the relationship that it mediates between what is informative and what is significant, a conundrum of our time if there ever was one" (Rosenberg 2021, p. 121).

Bernard, in his history of the hashtag, which he calls the latest incarnation of the keyword, dismisses most of these earlier incarnations, saying that "without a doubt the category of 'the keyword' had occupied a rather inconspicuous place prior to the twenty-first century…Today, every Twitter feed and Instagram post provides further testimony to the collective indexing or 'keywording' of the world" (Bernard 2019, p. 2).

In their encyclopedic examination of "keyword" and its related terms (term, index term, free-text term, Uniterm, heading, subject heading, descriptor, concept symbol, tag, word, stopword, N-gram, and keyphrase), Lardera and Hjørland (2021) provide an in-depth intellectual background for the keyword in both library and information science (LIS) theory and practice. Stubbs (2010, p. 25) has done the same for the discipline of linguistics, arguing that "keyness is a textual matter" because "keywords are words which are significantly more frequent in a sample of text than would be expected, given their frequency in a large general reference corpus."

Nevertheless, the keyword has its origins in a much longer history than proposed by these authors, so the rest of this chapter will explore the grounds on which our understanding of "key" words should begin.

1.2 Grounding the Keyword

Clearly, before there can be "keywords," there must be a conceptual and communicative infrastructure in which any such coded or clear reference to or perceived relevance of "something" (or, indeed, "anything") can be meaningful enough to be memorable. Bruner

referred to all of these as "routes to reference" (Bruner 1998). It is well accepted by now that evolutionary approaches to communication and information are necessarily intertwined, as both survival and reproduction require some successful internal and external communication of information at every level of taxa, from the lower bacteriological levels (Lyon 2015) to the higher zoological ones (Hoffecker 2013). In particular, the question of the development of the human "faculty of language" (Hauser et al. 2002) can no longer be considered in isolation from epigenetic (Gokhman et al. 2016), genetic (Graham et al. 2015), neural (Konopka and Roberts 2016), and other environmental (Greenhill 2016) factors.

Changeux and his colleagues (2021) have proposed that seemingly minor changes in the human genome since our fairly recent evolution from nonhuman primates can explain fundamental features of human brain connectivity, especially the tripling in size of the global neural architecture within the original primate brain, resulting in a larger number of neurons and areas and the increased modularity, efficiency, and differentiation of cortical connections. "The combination of these features with the developmental expansion of upper cortical layers, prolonged postnatal brain development, and multiplied nongenetic interactions with the physical, social, and cultural environment gives rise to categorically human-specific cognitive abilities including the recursivity of language. Thus, a small set of genetic regulatory events affecting quantitative gene expression may plausibly account for the origins of human brain connectivity and cognition" (2021, p. 2425).

Worden (2022) suggests that the traditional notion of language evolution through natural selection alone cannot account for the fact that the energy expenditure for our language-enabled brain (roughly 20% of our metabolic requirements) is much too high in comparison with that of a simpler brain with primitive language capabilities and smaller metabolic costs, which would be both more efficient and entirely adequate for our original survival needs. He proposes instead that both natural selection and sexual selection played a role in the evolution of language and intelligence, probably at different times. Specifically, he theorizes that early language was driven by natural selection to facilitate within-group collaboration. Early forms of information exchange, probably developed as sounds and gestures in various hunting and gathering activities over time, along with an emerging theory of mind, began to serve as markers for this superior intelligence, which ultimately played a critical role in sexual selection for early man, as the qualities of empathy and leadership it can embody are attractive to both peers and potential mates, thus precipitating the unique refinement of pragmatics, the development of spoken symbols, and the construction of syntax which was eventually to become modern language. As it does today, language acts as the main display mechanism for intelligence and also determines which "keywords" will become critically important in particular contexts, whether it is the term for a prey animal or the name of a political party.

Within library and information science, this is consistent with today's "informational" turn in which, despite some skepticism, such information-oriented scholars as Bates (2022), Beynon-Davies (2011), Brier (2010), Madden (2004), O'Connor (1996), Shah

(2023), Spink (2010), and Stonier (1997) continue their interdisciplinary investigations of what may quite reasonably be termed the evolution of information research. These initiatives tend to stress the continuities and similarities among different forms of informationally oriented cognitive systems over space and time and to take a much wider perspective than usual. They also tend to support the utility of a broader approach to the information problematic, such as that posed by and through "keywords."

Thinking about keywords as part of that problematic can raise central issues of representation (that is, presentation and organization of data or information) and reference (that is, meaning as intended by the speaker or writer), relevance (that is, meaning as understood by the listener or reader), and retention (that is, the varied forms of internal and external memory that may also be archived in both individual and social forms): all of these are fundamental aspects of LIS. In the absence of words, would any of these exist? Conversely, in their absence, might words still exist? This chapter explores some of the current findings on systems, symbols, and speech that must necessarily (though invisibly) ground any concept of "keywords."

The biological systems necessary for representation, reference, relevance, and retention are usually taken for granted, as human beings are so used to our visual, oral, aural, and retentive processing capabilities that it seldom occurs to us to wonder how these are affecting what we see, say, hear, and remember, other than perhaps in thinking in terms of extending these senses through novel technologies. Nevertheless, a knowledge of some sensory specifics and their embodiment may be helpful in grounding this discussion, especially those related to the so-called "symbol grounding problem," in which the question is how any one thing can be connected to a meaningful interpretation of that thing. This has been neatly expressed both by Searle (1980, p. 424) who observed "Of course the brain is a digital computer. Since everything is a digital computer, brains are too. The point is that the brain's causal capacity to produce intentionality cannot consist in its instantiating a computer program, since for any program you like it is possible for something to instantiate that program and still not have any mental states. Whatever it is that the brain does to produce intentionality, it cannot consist in instantiating a program since no program, by itself, is sufficient for intentionality") and by Harnad (1990, p. 335) who queried "How can the semantic interpretation of a formal symbol system be made *intrinsic* to the system, rather than just parasitic on the meanings in our heads?"

As Barsalou (2016, p. 1129) put it, "To a large extent, grounding concerns itself with the grounding problem raised initially by Searle and Harnad which asks how amodal symbols, specifically, and cognition, more generally, are linked to the modalities, body, and environment. In a review of research on grounding, Barsalou argued that researchers have attempted to ground concepts and cognition by establishing their relations with modality-specific systems, the body, the physical environment, and the social environment. …Thus, at a general level, grounding simply refers to programmatically studying cognition in new ways. Rather than studying cognitive mechanisms in isolation, establish their relations with the contexts in which they are embedded and on which they depend. At more

specific levels, grounding refers to establishing specific accounts of how cognitive processes in the brain utilize the modalities, the body, and the environment. It does *not* mean reducing concepts and cognition to anything, including sensory-motor mechanisms."

1.3 Representation

Vision, for instance, enables us to r*epresent* the world. Pylyshyn (2000, p. 197) explains that "Representations are the basic building blocks of cognitive explanations of human behavior.... [and] function in the same way as descriptions: they use the conceptual resources of the mind to encode properties of the world in much the same way as language uses words... [but] a conceptual description alone (what Bertrand Russell called a 'definite description') is inadequate for encoding certain types of [physical] knowledge… such as finding one's way home. The most primitive contact that the visual system makes with the world (the contact that precedes the encoding of any sensory properties) is a contact with what have been termed visual objects or proto-objects." Relatedly, Ballard and his colleagues (1997) found that very minute eye and hand movements are linked by processes underlying these elemental perceptual events through "deictic coding" to working memory, at time scales of approximately 1/3 of a second, and play an essential role in the brain's symbolic computations of "embodied" representations. Pitcher and Ungerleider (2021) have proposed that on the lateral brain surface in the primate visual cortex, in addition to the ventral visual pathway, which computes the identity of an object and the dorsal visual pathway, which computes the location of an object and actions related to that object, there exists a third visual pathway which computes the actions of moving faces and bodies and is apparently specialized for the dynamic aspects of social perception.

Symbols are represented in our brains at different levels of complexity: at the initial, simplest level, as physical entities, in the corresponding primary and secondary sensory cortices, as conceptual ones. Symbols, however, no matter how simple their surface forms may appear, evoke higher order multifaceted representations that are implemented in distributed neural networks spanning a large portion of the cortex. These internal states that reflect our knowledge of the meaning of symbols are what we call semantic representations (Borghesani and Piazza 2017). Viganò and his colleagues (2021) showed that this categorization within the brain took place in at least three representational stages: first, the sensory regions process the features relevant for categorization, the left angular gyrus integrates the different sensory features into unique object identities, connecting them to the correct name, and the hippocampus encodes the abstract associative rule.

The question of whether there might be a "language of thought" cognitive coding that is separate from natural language, however, is still open, as opined by Mandelbaum and his colleagues (2022): "Recent advances in deep neural networks appear to suggest that there is no need for psychological models beyond ones that posit links between neuron-like nodes. But while Artificial Intelligence (AI) research has moved away from

transparently interpretable, richly structured internal representations, advances in many disparate areas of cognitive science suggest otherwise. Evidence from animal and infant cognition, Bayesian computational cognitive science, unconscious reasoning, and visual cognition suggests that the mind traffics in representations couched in an amodal code with a language-like structure."

1.4 Reference

Like other mammals, our common way of sharing a representation is by making a vocal *reference* to it. (Pointing, the other referential method common in humans, is uncommon among mammals, even among the great apes, though a few have been occasionally observed using whole-hand gesturing to indicate a desired object from a human companion).

The mammalian voice production organ has three subsystems: the pulmonary system, which supplies power through the lungs, a sound generation system, typically the larynx, and a sound modifier system, the (pharyngeal, oral, and/or nasal) vocal tract (Herbst 2016). It was once believed that a descended larynx was uniquely human, but it has now been found in deer, for instance, though the human vocal tract still seems to be similar only to those of the Neanderthals and the Denisovans (Dediu et al. 2021). Regardless of languages and contexts, the amplitude modulation of the speech signal for humans consists of a rhythm that ranges between 3 and 8 Hz, while the vocalizations and facial expressions of monkeys and apes also have this rhythmic structure (Zhang and Ghazanfa 2020). Human infants are attuned to this rhythm even prior to birth, which helps to accelerate their process of language acculturation and accumulation (Ghio et al. 2021).

Although the abundance of sounds found in the world's languages has been thought to have been fixed by biological constraints since the emergence of Homo sapiens, it has recently been proposed that post-Neolithic changes in bite configuration likely caused by diet changes gave rise to a new class of speech sounds, the labio-dentals, produced by positioning the lower lip against the upper teeth (Blasi et al. 2019). In general, vocalization is undergoing intense study at present, especially the questions of communicative exchange (Pika et al. 2018) and vocal learning (Vernes et al. 2021), as it becomes apparent that human vocalization seems to have more in common with other bird and animal sound emissions than earlier researchers believed. However, as Arbib (2021) notes, conveying "aboutness" in general is an apparently uniquely human capability.

1.5 Relevance

Similarly, *relevance*, our ability to make use of articulated connections between the world and others' representations of it, is usually contingent upon the evolution of sound localization in mammals. Hearing relies on existence of an evolutionarily ancient sound recognition mechanism in the brainstem which is capable of implicit learning of sequences of sound, which are integrated and streamed according to the spectrotemporal properties of recognized sound sources, leading to enhanced auditory sensitivity to behaviorally relevant sound sources. As mammals evolved in size, an accompanying increase in head size brought both a larger larynx and an increase in the distance between the ears, which favored evolutionary adaptations that helped to localize those cues indicating relevant sounds or communication calls, as the tympanic ear is particularly adapted to transmit low-frequency sounds via middle-ear bones (McLachlan and Wilson 2017).

Sound comprehension is also fundamental to communication, especially in perceiving voices of members of the same species, and this ability is widely shared across the animal kingdom. Neurophysiological work by Rauschecker and Scott (2009) used findings from both humans and non-human primates to develop a model showing the connections of structures in the temporal, frontal, and parietal lobes linking vocal perception and production. Chief among these were physiological and anatomical studies showing that the primate auditory cortex, across species, shares the patterns of hierarchical structure, topographic mapping, and streams of functional processing that enable speech. Thus, speech appears to be an extreme adaptation of vocalization common to many other species.

Bodin and Belin (2020) suggest the existence of a "voice patch system," a network of interconnected cortical areas providing a common template for the cerebral processing of voices in primates. They argue that the left dorsal pathway is devoted to the processing of complex sounds in monkeys as well in humans, but that the temporo-parietal regions of the right hemisphere may have further evolved in humans to favor multimodal associations and the processing of high-level social information that would support the development of speech.

Speech comprehension in humans is complex, involving multiple stages of neural representations in order to convert sound to meaning. The initial process involves spectrotemporal analysis of the acoustic signal in early auditory cortices, followed by phonetic and phonological processing in the superior temporal lobe, in which continuous acoustic features are projected into categorical representations. Higher levels of speech comprehension transform intermediate speech representations into conceptual and semantic representations in the superior middle temporal lobe, while the posterior dorsal temporal lobe, parietal operculum, and posterior frontal lobe are responsible for translating speech signals into articulatory representations (Martin et al. 2019).

1.6 Retention

Retention, the storage and retrieval of memories within primate cognition, is also complex, as it involves cognitive control, the executive and regulatory processes that allocate attention, metacognitive processes that evaluate available information and seek additional information when required, memory processes that draw on past experiences to plan future behaviors, and self-control processes that deal with distractions and impulses when these are threats to goal achievement (Beran et al. 2016).

Our human memory consists of several types, distinct from one another neurobiologically, functionally, and phenomenologically (Camina and Güell 2017). These include sensory memory, short-term memory, and long-term memory. Sensory memory is very brief (lasting for approximately one second) and includes information gathered through sight (iconic memory), information gathered through hearing (echoic memory), and information gathered through touch (haptic memory). Short term or "working" memory includes a central executive which controls attention, an articulatory loop which retains verbal information, and a visuospatial sketchpad which retains visual and spatial information. Long-term memory includes explicit memory (which, as episodic memory, includes conscious recall of episodic autobiographical memory, or, as semantic memory, includes conscious recall of learned facts, events, or data), implicit memory (subconscious recall of previously learned procedures or subconscious priming through previously experienced emotions or interactions), and prospective memory (formation and implementation of intended future actions).

Evidence from brain imaging research indicates that the human brain's "default network" for working memory is not significantly different from those of other primates (Buckner et al. 2008). As noted above, however, memory has different components, some of which, such as autobiographical memory, are apparently unique to humans (Bjorklund and Sellers II 2013), which foregrounds the human perception of time as opposed to that of other creatures and which assists in supporting social memory within human groups.

1.7 Grounds for Key Concepts

All of these sensory systems, therefore, can provide the "grounds" for concepts. Carruthers (2013, p. 238) says "The global broadcasting of attended sensory representations in the brain is almost certainly an ancient adaptation, designed to coordinate the activity of numerous higher cognitive systems around a common focus. …And we also know that just as concepts are bound into the content of perception and globally broadcast along with it (resulting in a state of seeing something *as* a car or *as* one's mother, say), so are concepts bound into the contents of visual and other imagery (resulting in an image of a car, as such, or of one's mother, as such)".

According to Barsalou (2021), a concept develops by aggregating information from perception, action, and internal states, and then using integration mechanisms to integrate it with background situational knowledge already in memory. Although learning plays central roles in establishing concepts, genetic and epigenetic processes constrain the features that can be represented for a concept, and also their integration in the brain's association areas. Once the conceptual system is in place, it supports virtually all other forms of cognitive activity, both in the current situation and when representing the world in language, memory, and thought. The cortex's activation of associations for concrete and abstract concepts is still under investigation (Henningsen-Schomers and Pulvermüller 2022). Further, the sharing of concepts, whether through coded (Origgi and Sperber 2000) or inferential (Smith 2008) transmission, is far from well understood, though there is increasing interest in the cognitive sciences on shared cognition (Dingemanse et al. 2023), conceptual alignment (Stolk et al. 2016), and joint epistemological engineering (Stolk et al. 2022).

The discovery of mirror neurons (neurons within the brain that fire both when the self performs an action and when the self observes the same action performed by another) in the brain of the macaque monkey created considerable excitement as to how these might encode abstract concepts of actions which may then relate to imitation and other forms of social learning in animals. For humans, brain activity consistent with that of mirror neurons has been found in the premotor cortex, the supplementary motor area, the primary somatosensory cortex, and the inferior parietal cortex, also raising questions regarding a possible innate ability of humans to imitate, learn from, and empathize with others. However, studies involving the abilities of human (Oostenbroek et al. 2016) and monkey (Simpson et al. 2014) newborns indicate that their imitations did not show any patterns that would suggest the existence of innate imitative abilities by the infants of either species, leaving social learning as the most probable explanation for these abilities. Nevertheless, the human infant's initial grasping abilities, then of reaching for more distant objects, and later of showing and pointing as markers of shared attention, are all considered signs of developing deictic capabilities, the ability to distinguish between self, objects, and others, which will help to form the basis for language acquisition (Boundy et al. 2016).

There are now various theories within cognitive science and cognate disciplines attempting to "ground" new research in motor control, perception, and cognition into more "embodied" approaches to thought (Gentsch et al. 2016) and language (García and Ibáñez 2016). Based on new findings that the brain's hippocampal formation, anterior thalamus, and claustrum provide rich, rapid coding of spatial representations, O'Mara and Aggleton (2019) suggest that this may support the notion that the cognitive system uses both the body and the environment as external informational structures in order to ground internal representations. They note that the idea that the brain uses the external environment as a kind of "cognitive surface," or that it supports the structure of cognition has garnered little attention in neuroscience to date but could offer productive areas of inquiry. This

may relate particularly to deixis, which, as Galbraith (2021) defines it, is "the semiotic term for particularized space and time in embodied existence. This ever-present deictic field is both ordinary and unexplainable: how is it that this space and this body exist in this moment? The elemental semiotic function of calling attention to particulars from the perspectival orientation of a bodily self in time and space foregrounds such ineluctable properties as presence, immediacy, and the vulnerability of being, and is a central topic for philosophers, linguists, anthropologists, cognitive scientists, and literary theorists."

In other words, the concepts of representation, reference, relevance, and retention are instantiated in various ways that may be productive in considering these as "key" to the development of all other systems, including biological systems, semiotic systems, and information systems, which are essential to the creation of words, the symbolic key to human understanding.

Even when specific languages have long been forgotten, it is still possible to recover key meanings, as in the case of cuneiform (Robson 2019) and khipus (Hyland 2020). The next chapter will explore the beginnings of semiotics, the study of semiosis (sign processes) in all living beings, focusing on the importance of zoosemiotics and the challenges that researchers face in understanding non-human primate communication.

References

Arbib, M. A. (2021). The aboutness of language and the evolution of the construction-ready brain. In A. Lock, C. Sinha, & N. Gonthier (Eds.), *The Oxford handbook of symbolic evolution*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198813781.013.7

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, *20*(4), 723–742. https://doi.org/10.1017/S0140525X97001611

Barsalou, L. W. (2016). On staying grounded and avoiding quixotic dead ends. *Psychonomic Bulletin & Review*, *23*(4), 1122–1142. https://doi.org/10.3758/s13423-016-1028-3

Barsalou, L. W. (2021). Categories at the interface of cognition and action. In I. Fiorentini & E. Goria (Eds.), *Building categories in interaction: Linguistic resources at work* (pp. 35–72). John Benjamins Publishing.

Bates, M. J. (2022). A proto-paradigm for information science research. *Information Research*, *27*. https://doi.org/10.47989/colis2201

Beran, M. J., Menzel, C. R., Parrish, A. E., Perdue, B. M., Sayers, K., Smith, J. D., & Washburn, D. A. (2016). Primate cognition: Attention, episodic memory, prospective memory, self-control, and metacognition as examples of cognitive control in nonhuman primates. *Wiley Interdisciplinary Review of Cognitive Science*, *7*(5), 294–316. https://doi.org/10.1002/wcs.1397

Bernard, A. (2019). *Theory of the hashtag* (V. A. Pakis, Trans.). Polity Press.

Beynon-Davies, P. (2011). *Significance: Exploring the nature of information, systems and technology*. Palgrave Macmillan.

Bjorklund, D. F., & Sellers II, P. D. (2013). Memory development in evolutionary perspective. In P. J. Bauer & R. Fivush (Eds.), *The Wiley handbook on the development of children's memory* (Vol. 1, pp. 126–156).

Blasi, D. E., Moran, S., Moisik, S. R., Widmer, P., Dediu, D., & Bickel, B. (2019). Human sound systems are shaped by post-Neolithic changes in bite configuration. *Science*, *363*(6432), Article eaav3218. https://doi.org/10.1126/science.aav3218

Bodin, C., & Belin, P. (2020). Exploring the cerebral substrate of voice perception in primate brains. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1789), Article 20180386. https://doi.org/10.1098/rstb.2018.0386

Borghesani, V., & Piazza, M. (2017). The neuro-cognitive representations of symbols: The case of concrete words. *Neuropsychologia*, *105*, 4–17. https://doi.org/10.1016/j.neuropsychologia.2017.06.026

Boundy, L., Cameron-Faulkner, T., & Theakston, A. (2016). Exploring early communicative behaviours: A fine-grained analysis of infant shows and gives. *Infant Behavior and Development*, *44*, 86–97. https://doi.org/10.1016/j.infbeh.2016.06.005

Brier, S. (2010). Cybersemiotics: An evolutionary world view going beyond entropy and information into the question of meaning. *Entropy*, *12*(8), 1902–1920. https://doi.org/10.3390/e12081902

Bruner, J. S. (1998). Routes to reference. *Pragmatics and Cognition*, *6*(1/2), 209–227. https://doi.org/10.1075/pc.6.1-2.11bru

Buckner, R. L., Andrews Hanna, J. R., & Schacter, D. L. (2008). The brain's default network. *Annals of the New York Academy of Sciences*, *1124*(1), 1–38. https://doi.org/10.1196/annals.1440.011

Camina, E., & Güell, F. (2017). The neuroanatomical, neurophysiological, and psychological basis of memory: Current models and their origins. *Frontiers in Pharmacology*, *8*, Article 438. https://doi.org/10.3389/fphar.2017.00438

Carruthers, P. (2013). Animal minds are real, (distinctively) human minds are not. *American Philosophical Quarterly*, *50*(3), 233–248. https://www.jstor.org/stable/24475346

Changeux, J. P., Goulas, C. C., & Hilgetag, C. C. (2021). A connectomic hypothesis for the hominization of the brain. *Cerebral Cortex*, *31*(5), 2425–2449. https://doi.org/10.1093/cercor/bhaa365

Dediu, D., Moisik, S. R., Baetsen, W. A., Bosman, A. M., & Waters-Rist, A. L. (2021). The vocal tract as a time machine: Inferences about past speech and language from the anatomy of the speech organs. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*, Article 20200192. https://doi.org/10.1098/rstb.2020.0192

Dingemanse, M., Liesenfeld, A., Rasenberg, M., Albert, S., Ameka, F. K., Birhane, A., Bolis, D., Cassell, J., Clift, R., & Wiltschko, M. (2023). Beyond single-mindedness: A figure-ground reversal for the cognitive sciences. *Cognitive Science*, *47*, Article e13230. https://doi.org/10.1111/cogs.13230

Galbraith, M. (2021). Deixis. In *Oxford research encyclopedia of literature.* Oxford University Press. https://doi.org/10.1093/acrefore/9780190201098.013.1055

García, A. M., & Ibáñez, A. (2016). A touch with words: Dynamic synergies between manual actions and language. *Neuroscience & Behavioral Reviews*, *68*, 59–95. https://doi.org/10.1016/j.neubiorev.2016.04.022

Gentsch, A., Weber, A., Synofzik, M., Vosgerau, G., & Schütz-Bosbach, S. (2016). Towards a common framework of grounded action cognition: Relating motor control, perception, and cognition. *Cognition*, *146*, 81–89. https://doi.org/10.1016/j.cognition.2015.09.010

Ghio, M., Cara, C., & Tettamanti, M. (2021). The prenatal brain readiness for speech processing: A review on foetal development of auditory and primordial language networks. *Neuroscience & Biobehavioral Reviews*, *128*, 709–719. https://doi.org/10.1016/j.neubiorev.2021.07.009

Gokhman, D., Meshorer, E., & Carmel, L. (2016). Epigenetics: It's getting old. Past meets future in paleoepigenetics. *Trends in Ecology and Evolution*, *31*(4), 290–300. https://doi.org/10.1016/j.tree.2016.01.010

Graham, S. A., Deriziotis, P., & Fisher, S. (2015). Insights into the genetic foundations of human communication. *Neuropsychology Review*, *25*(1), 3–26. https://doi.org/10.1007/s11065-014-9277-2

Greenhill, S. J. (2016). Overview: Debating the effect of environment on language. *Journal of Language Evolution*, *1*(1), 30–32. https://doi.org/10.1093/jole/lzv007

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1–3), 335–346. https://doi.org/10.1016/0167-2789(90)90087-6

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, *298*(5598), 1569–1579. https://doi.org/10.1126/science.298.5598.1569

Henningsen-Schomers, M. R., & Pulvermüller, F. (2022). Modelling concrete and abstract concepts using brain-constrained deep neural networks. *Psychological Research*, *86*(8), 2533–2559. https://doi.org/10.1007/s00426-021-01591-6

Herbst, C. T. (2016). Biophysics of vocal production in mammals. In R. A. Suthers, W. T. Fitch, R. R. Fay, & A. N. Popper (Eds.), *Vertebrate sound production and acoustic communication* (pp. 159–189). Springer.

Hoffecker, J. F. (2013). The information animal and the super-brain. *Journal of Archaeological Method and Theory*, *20*(1), 18–41. https://doi.org/10.1007/s10816-011-9124-1

Hyland, S. (2020). Subject indicators and the decipherment of genre on Andean Khipus. *Anthropological Linguistics*, *62*(2), 137–158. https://doi.org/10.1353/anl.2020.0004

Konopka, G., & Roberts, T. F. (2016). Insights into the neural and genetic basis of vocal communication. *Cell*, *164*(6), 1269–1276. https://doi.org/10.1016/j.cell.2016.02.039

Lardera, M., & Hjørland, B. (2021). Keyword. In B. Hjørland & C. Gnoli (Eds.), *ISKO encyclopedia of knowledge organization.* https://www.isko.org/cyclo/keyword

Leja, M. (2009). Keyword. *American Art*, *23*(1), 34–35.https://doi.org/10.1086/599062

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, *2*(2), 159–165. https://doi.org/10.1147/rd.22.0159

Lyon, P. (2015). The cognitive cell: Bacterial behavior reconsidered. *Evolutionary and Genomic Microbiology, Frontiers in Microbiology*, *6*, Article 264. https://doi.org/10.3389/fmicb.2015.00264

Madden, A. D. (2004). Evolution and information. *Journal of Documentation*, *60*(1), 9–23. https://doi.org/10.1108/00220410410516626

Mandelbaum, E., Dunham, Y., Feiman, R., Firestone, C., Green, E. J., Harris, D., Kibbe, M. M., Kurdi, B., Mylopoulos, M., Shepherd, J., & Wellwood, A. (2022). Problems and mysteries of the many languages of thought. *Cognitive Science*, *46*(12), Article13225. https://doi.org/10.1111/cogs.13225

Martin, S., Millán, J. D. R., Knight, R. T., & Pasley, B. N. (2019). The use of intracranial recordings to decode human language: Challenges and opportunities. *Brain and Language*, *193*, 73–83. https://doi.org/10.1016/j.bandl.2016.06.003

McLachlan, N. M., & Wilson, S. J. (2017). The contribution of brainstem and cerebellar pathways to auditory recognition. *Frontiers in Psychology*, *8*, Article 265. https://doi.org/10.3389/fpsyg.2017.00265

O'Connor, B. C. (1996). *Explorations in indexing and abstracting: Pointing, virtue, and power*. Libraries Unlimited.

O'Mara, S. M., & Aggleton, J. P. (2019). Space and memory (far) beyond the hippocampus: Many subcortical structures also support cognitive mapping and mnemonic processing. *Frontiers in Neural Circuits*, *13*, Article 52. https://doi.org/10.3389/fncir.2019.00052

Oostenbroek, J., Suddendorf, T., Nielsen, M., Redshaw, J., Kennedy-Constantini, S., Davis, J., Clark, S., & Slaughter, V. (2016). Comprehensive longitudinal study challenges the existence of neonatal imitation in humans. *Current Biology*, *26*(10), 1334–1338. https://doi.org/10.1016/j.cub.2016.03.047

Origgi, G., & Sperber, D. (2000). Evolution, communication, and the proper function of language. In P. Carruthers & A. Chamberlain (Eds.), *Evolution and the human mind: Modularity, language, and meta-cognition* (pp. 140–169). Cambridge University Press.

Pika, S., Wilkinson, R., Kendrick, K. H., & Vernes, S. C. (2018). Taking turns: Bridging the gap between human and animal communication. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1880), Article 20180598. https://doi.org/10.1098/rspb.2018.0598

Pitcher, D., & Ungerleider, L. G. (2021). Evidence for a third visual pathway specialized for social perception. *Trends in Cognitive Sciences*, *25*(2), 100–110. https://doi.org/10.1016/j.tics.2020.11.006

Pylyshyn, Z. W. (2000). Situating vision in the world. *Trends in Cognitive Sciences*, *4*(5), 197–207. https://doi.org/10.1016/S1364-6613(00)01477-7

Rauschecker, J. P., & Scott, S. K. (2009). Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. *Nature Neuroscience*, *12*(6), 718–724. https://doi.org/10.1038/nn.2331

Robson, E. (2019). *Ancient knowledge networks: A social geography of cuneiform scholarship in first-millennium Assyria and Babylonia.* UCL Press.

Rosenberg, D. (2021). Keyword. In M. Kennerly, S. Frederick, & J. E. Abel (Eds.), *Information: Keywords* (pp. 121–132). Columbia University Press.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756

Shah, C. (2023). The past, the present, and the future of information and data sciences: A pragmatic view. *Data and Information Management*, 100028. https://doi.org/10.1016/j.dim.2023.100028

Simpson, E. A., Paukner, A., Suomi, S. J., & Ferrari, P. F. (2014). Visual attention during neonatal imitation in newborn macaque monkeys. *Developmental Psychobiology*, *56*(4), 864–870. https://doi.org/10.1002/dev.21146

Smith, A. D. (2008). Protolanguage reconstructed. *Interaction Studies*, *9*(1), 100–116. https://doi.org/10.1075/is.9.1.08smi

Spink, A. (2010). *Information behavior: An evolutionary instinct.* Springer-Verlag.

Stolk, A., Bašnáková, J., & Toni, I. (2022). Joint epistemic engineering: The neglected process in human communication. In A. M. García & A. Ibáñez (Eds.), *The Routledge handbook of semiosis and the brain* (pp. 259–278). Routledge.

Stolk, A., Verhagen, L., & Toni, I. (2016). Conceptual alignment: How brains achieve mutual understanding. *Trends in Cognitive Sciences*, *20*(3), 180–191. https://doi.org/10.1016/j.tics.2015.11.007

Stonier, T. (1997). *Information and meaning: An evolutionary perspective.* Springer.

Stubbs, M. (2010). Three concepts of keywords. In M. Bondi & M. Scott (Eds.), *Keyness in texts* (pp. 21–42). John Benjamins Publishing.

Vernes, S. C., Kriengwatana, B. P., Fischer, J., Tyack, P. L., ten Cate, C., & Janik, V. M. (2021). The multi-dimensional nature of vocal learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*, Article 20200236. https://doi.org/10.1098/rstb.2020.0236

Viganò, S., Borghesani, V., & Piazza, M. (2021). Symbolic categorization of novel multisensory stimuli in the human brain. *Neuroimage*, *235*, Article 118016. https://doi.org/10.1016/j.neuroimage.2021.118016

Williams, R. (1958). *Culture & society: 1780-1950*. Columbia University Press.

Worden, R. (2022). The evolution of language by sexual selection. *Frontiers in Psychology, 13*. https://doi.org/10.3389/fpsyg.2022.1060510

Zhang, Y. S., & Ghazanfa, A. S. (2020). A hierarchy of autonomous systems for vocal production. *Trends in Neurosciences*, *43*(2), 115–126. https://doi.org/10.1016/j.tins.2019.12.006

Signals, Semiotics

2

Abstract

This chapter introduces semiotics, the study of signs, which includes signals as biological communication among animals and words as abstract symbols among humans. It focuses on zoosemiotics, the study of biological signals involving the various sensory apparatus that animals possess, as they send, receive, process, and react to those signals in different ways, emphasizing the communication behavior of non-human primates which may be related to that of humans.

Semiotic communication has a far longer history than anything conveyed via words, speech, or even sound, since the lowest so-called "semiotic threshold zone" begins with bacteria quora employing chemical cues in communication (Van der Veer Martens 2023), so any discussion of the origins of communication should begin with semiotics, the study of signs, which is also inclusive of signals as biological communication (semiosis) among animals and words as abstract symbols among humans. This chapter will focus on zoosemiotics, the study of animal signs and signals.

2.1 Introduction to Semiotics

Semiotics derives from *sêma*, the Greek word for sign, although the word itself did not emerge from its Greek translation as "doctrine of signs" as briefly mentioned in Locke's *An Essay Concerning Human Understanding* until a century later and has appeared in several variants, usually as semiotics or as semiology (Nöth 1995). This study of signs as indicative of various phenomena can be traced back to such early observers as Aristotle, Lucretius, Hippocrates, and Augustine. There are numerous typologies of signs, but the

B. V. V. Martens, *Keywords In and Out of Context*, Synthesis Lectures on Information Concepts, Retrieval, and Services, https://doi.org/10.1007/978-3-031-32530-4_2

two dominant semiotic perspectives today originated in nineteenth century work by Ferdinand de Saussure in Switzerland, and by Charles S. Peirce in the United States, both of whose works are still undergoing interpretation by scholars more than a century later. This chapter will discuss Peircean semiotics while De Saussure's semiotic contributions will be discussed in Chap. 4. Peirce's intellectual legacy to semiotics is both complex and incomplete, as he left behind a massive collection of published and unpublished papers still in the process of being edited and published as a multivolume critical edition (Keeler 2020).

The Peircean perspective is best known for Peirce's introduction of the notion that a sign stands for something or somebody in some respect or context. This three-way relationship among the *object* (a phenomenon of some kind), the *sign* (that represents the phenomenon in some way), and the *interpretant* (that translates into some meaning) is essential to Peircean thinking, as each element is essential to the semiotic process. A later version of his thinking further specified that the "sign" could relate to the object or phenomenon in one of three general ways: it could share some particularly salient and perceptible quality with it, in which case it is called an "icon," or it could be seen as symptomatic in some way of the phenomenon itself, in which case it is called an "index," or, if the conceptual connection could be viewed in some conventional way, it is called a "symbol."

An illustration of this in modern terms is as follows: a photograph taken by a security camera of a person outside a building could be viewed as an "icon" of the particular individual depicted, a large plume of smoke observed coming from inside that building could be viewed as an "index" of a fire within, and the name of someone suspected of setting fires in that area could be viewed as a "symbol" of potential arson. However, it should be noted that all of these interpretations are independent of one another; there are no necessary causal or correlative relationships involved in Peirce's system, though there are ample opportunities for inference and investigation (Eco and Sebeok 1988).

Peirce further developed the notion that these signs can be categorized in several additional ways, although his categorizations eventually became so complex, leading to a total of 66 signs (Restrepo and De Mesa 2021), that only a few scholars draw upon the extended versions. Most users of the Peircean perspective tend to use the less complicated schemes, due to the simplicity of the approach and their applicability to a wide variety of situations. In particular, the original *interpretant* element leaves unspecified any human cognition in the translation process, unless the sign itself is a linguistic one, as famously noted by Dewey (1946), while Millikan (2004) later pointed out that both humans and animals are unable to interpret signs for which they lack any mental representations.

Among those who helped to further the Peircean perspective were English scholars Ogden and Richards, whose graphic depiction of what they called the "symbol," the "referent," and the "reference" in a "triangle of reference" in *The Meaning of Meaning* (1927, p. 11) made semiotics much more widely known. The most popular modification of Peirce's original model is arguably that of American linguist Charles Morris (1971),

which includes (1) the "sign vehicle" (the object or event which functions as a sign), (2) the "designatum" (the kind of object or class of objects that the sign designates), (3) the "interpreter" (the individual for whom the sign vehicle functions as a sign), and (4) the "interpretant" (the disposition of an interpreter to initiate a response sequence as a result of perceiving the sign). This novel "stimulus–response" behaviorist approach also defines a sign as any preparatory stimulus producing a disposition in the interpreter to respond to something that is not an immediate stimulus. According to Morris, all signs may also be divided into symbols, which can be interpreted to signify other signs, and signals, which cannot. Morris's model deviates sharply from that of Peirce but has been influential in extending semiotic study into non-human semiosis.

Among Morris's students was linguist Thomas Sebeok, whose application of aspects of Morris's integrated behavioral and semiotic model to his own investigations into the biological roots of culture (Sebeok 1963) became the foundation for the field of zoosemiotics (Maran 2014), and eventually more broadly for what has become known as biosemiotics, which involves the study of the semiosis of all living beings (Favareau 2009). Sebeok (2001) argued that biosemiotics could be used to analyze the cognitive and communicative capabilities of other species as well as those of human beings, as behavior in all biological organisms is indicative of semiosis, which is, fundamentally, activities based on signs. In particular, analysis of the "signification" of signs (which, however, according to Morris, did not equate to "meaning" as generally understood by humans) could be applied to the study of biological signals involving the various sensory apparatus that animals possess, as they receive, process, and react to those signals in quite different but clearly functional ways.

Sebeok also edited a volume on animal communication to which a fellow linguist, Charles Hockett, contributed an important chapter, listing 16 "design features" of language systems, inspired by the Shannon–Weaver model of information theory, and applied to animal communication (Hockett and Altmann 1968). Hockett noted that all primate communication systems include at least some of the following features: (1) a vocal-auditory channel; (2) broadcast transmission and directional reception of sound; (3) rapid fading of sound; (3) interchangeability of signals (members of the speech community can both transmit and receive linguistic signals); (4) complete feedback (the speaker can hear what is said); (6) specialization (the triggering consequences of signals are biologically important; the energy costs of transmitting them are not; (7) semanticity (linguistic signals have denotations that tie them to perceptions in and of the world); (8) arbitrariness (the relation between a meaningful element in a language and its denotation is independent of any other resemblance between the two); (9) discreteness (possible messages in any language constitute a discrete repertoire rather than a continuous one); (10) displacement (things that are remote in time, space, or both from the site of the communicative transaction can be referenced); (11) openness (new linguistic messages are easily created and, in context, are usually understood); (12) tradition (language conventions are passed down by teaching

and learning, not through heredity); (13) duality of patterning (arbitrary but stable meaningless signal-elements and also patterning in terms of minimum meaningful arrangements of those elements); (14) prevarication (signals can be sent that are intentionally false or without meaning); (15) reflexiveness (communication about the communication system itself is possible); and (16) learnability (outsiders to the speech community can learn the language).

This list was widely influential in the study of animal communication in the 1950s and 1960s, though, as Wacewicz and Żywiczyński (2015) point out, Hockett's set of design features, with its focus on language as a product, emphasizing the code itself rather than the cognitive abilities of its users, is incompatible with today's research emphasis on the study of those suites of sensorimotor, cognitive, and social abilities that may enable the acquisition of language by biological creatures. It is still undetermined to what extent these 16 communicative features are shared by different primate groupings, although it remains an active topic of investigation (Amphaeris et al. 2022).

Zoosemiotics studies the biological signals involving the various sensory apparatus that animals possess, as they send, receive, process, and react to those signals in different ways. Ongstad (2021) lists these biological signals as auditory (such as vocalization, stridulation, vibration, and chest beating) electric (both electrogenic and electro-receptive), olfactory (chemical, scent marking, and scent rubbing), seismic (self-generated vibrations transmitted through external substance), thermal (infrared, thermo-regulation), touch (fighting, mating, social integration, foraging, and huddling), and visual (gestures, facial expressions, gaze following, color change, and bioluminescence). Von Uexküll's (2001) concept of animal *umwelt* (the animal's perception of its environment) has also become highly influential in zoosemiotics.

Zoosemiotics, combined with earlier ethological interest in the logic of animal communication (Marler 1961), interest in evolutionary approaches in the logic of natural selection (Dawkins and Krebs 1978), and methods applied from game theoretic approaches to signaling strategies (Maynard Smith 1982) have converged in the realization that these separate approaches could be fused into a modern synthesis that combine proximate questions about how communication mechanisms such as signals serve to fulfill immediate needs of the individual animals with ultimate questions about these mechanisms may be favored by natural and sexual selection (Owings et al. 1997).

2.2 Studying Signals

Recent examples of biosemiotic research include the study of chemical signaling by fish (Wisenden 2015) and of coloration communication among freshwater turtles (Brejcha and Kleisner 2016). However, most attention has focused on the communicative behavior of non-human primates, as their semiosis can be presumed to be closest to that of our hominid ancestry. The distress calls of both human and non-human primate infants share

a similar acoustic range and are adapted to elicit a similar range of more or less immediate responses by their caretakers (Lingle et al. 2012). Recent studies in the communication behavior of marmoset monkeys have shown that normal vocal development in monkey infants can be influenced by parental feedback, potentially somewhat similar to the vocal interactions between human parents and their infants (Ghazanfar et al. 2019).

Analysis of the much-studied vervet monkey genome indicates that their ancestral branch diverged from that leading to hominid ancestry between 27 and 29 million years ago (Warren et al. 2015) while that of the human–gorilla divergence is estimated to be in the range of 9.4–12.2 million years ago, that of the human–chimpanzee divergence in the range of 6.5–9.3 million years, and the divergence of bonobos from chimpanzees approximately 1 million years ago (Moorjani et al. 2016).

The unusually clear alarm calls of the vervet monkey have been closely analyzed for 50 years (Price et al. 2015) to determine whether these calls represent instinctual responses to a perceived stimuli and whether such calls routinely provoke particular behaviors in other group members. Seyfarth, Cheney, and Marler's work in the wild (Seyfarth et al. 1980) demonstrated that these alarm calls elicited behaviors in vervet group members adaptive to the relevant predators: recipients of leopard calls climbed trees, recipients of eagle calls hid within thick brush, and recipients of snake calls sought higher ground. Experimental recordings of the calls elicited the same behavior by recipients, regardless of the degree of caller arousal as indicated by acoustic variations in call length and volume, showing that it was signal content rather than signal intensity that was important.

Seyfarth and Cheney (1993, p. 208) proposed that their vervet monkey results indicated "some representation of the objects and events denoted by different call types and that [the monkeys] compare and respond to vocalizations on the basis of these representations." Thus, the call both makes reference by the sender to a particular phenomenon, as shown by its specificity, and is perceived as relevant by the call receivers, as shown by their reactions. However, Evans and Clark (2010) noted that more compelling, comparative evidence would be required to support such "representational signaling" as indicating the presence of complex cognitive processes. Accordingly, researchers have tended to use the neutral term "functional reference" to describe these types of communicative behavior as it is evident, whatever the reason, that such calls do function to alert group members to some presumably observed phenomenon, even though they may not be intended to be "informative" in the human sense. In any event, Scarantino (2013, p. 1008) noted that the vervet work "galvanized the field of animal communication. It provided preliminary support for what we may call the referential view of animal signaling, according to which at least some animal signals refer to the external world."

More broadly, Allen and Saidel (1998) relate functional reference to semiotics, suggesting that functional reference may consists of different types: mimetic, proxy, and conceptual reference. Mimetic reference consists of a signal closely resembling the referent so that it is capable of directly causing the same kind of response, equating to an "iconic" sign. Proxy reference consists of a signal functioning as proxy for its referent

in the sense that the signal elicits the same kind of response that the referent would but does so by a different cognitive mechanism, in contrast to mimetic reference which works by stimulating the same sensory/cognitive pathways as the referent would, equating to an "indexical" sign. Conceptual reference consists of signals that may refer to external conditions without it being normal for such uses to elicit the responses that the referent itself would elicit if it were present, equating to a "symbolic" sign. For instance, a mimetic reference signal might be the exaggerated sexual swellings of female baboons during their peak periods of fertility, while proxy reference signals might include the vervet alarm calls that alert to the presence of a particular predator, and conceptual reference signals might include the "play face" that gorillas use to indicate non-aggressive intent during casual physical interactions with their companions rather than the show of fully bared teeth that would precede a serious attack.

Functional reference might presumably serve as a ladder of ascent to fully referential communication, as both the behavior and the semiotic elements become more complex. Referential communication is both anomalous in animal communication because it involves context-dependent behaviors designed to influence specific addressees rather than involving stable traits designed by natural selection to influence bystanders, and complex, because these behaviors can carry multiple meanings, and these meanings can be conveyed by a variety of behaviors (Stolk et al. 2015). Manser (2020) identifies the different types of referents in animal vocalizations as those associated with social factors (age, dominance, group membership, sex, etc.), behavioral factors (contactual, emotional, and motivational), and external factors (presence of food, predators, etc.). A growing body of evidence attests to the prevalence of this vocal behavior among different species and the different forms that it may take but the argument continues as to whether this does (Arnold and Bar-On 2020) or does not (Fischer and Price 2017) have broader implications in terms of the evolution of language within the primate lineage.

2.3 Gestural Repertoires

Vocal communication has been the primary focus of much of this research, but gestural communication has been found to be equally important among non-human primates. According to Byrne and Cochet (2017), a biological repertoire consists of specific gestures for specific meanings, and all apes have the innate potential to develop substantial repertoires. Studies show a large overlap in the species-typical gestural repertoires of all great apes, approximately 90% overlap between chimpanzees and bonobos, approximately 80% between chimpanzees and orangutans, and approximately 60% overlap between chimpanzees and gorillas (Graham et al. 2022).This overlap is likely due to their common descent, as mentioned earlier.

This repertoire includes the so-called "deictic" gestures or "pointing" behavior, which is of particular interest due to its potential connection to the evolution of language and

to pre-linguistic behavior by human infants. According to Gillespie et al. (2013), gestural communication among apes tend to involve dyadic (that is, indicating a partner) rather than triadic (indicating a third entity) gesture and often carries an immediate, imperative request such as for grooming or sharing food. However, pointing to objects has been observed (Leavens et al. 2015). While the common human gesture of pointing with the index finger is infrequently observed in ape behavior, whole hand pointing (or indicative reaching) is used by captive apes when they have a receptive audience, though these gestures tend to be imperative as well (Tomasello and Call 2019).

Liebal and Call (2012) found that gesture repertoires among great apes lack the abstract qualities of human gesturing, and that their individual repertoires apparently emerge from action via three potential pathways: genetic predispositions towards certain species-specific ones, ontogenetic ritualizations that appear to pattern certain movements in coordination with those of another conspecific, and social learning that may facilitate the adoption of idiosyncratic or novel gestures.

Cissewski and Boesch (2016) observed that different great ape communities employ similar gestures to convey different meanings. All of these findings regarding gestural communication would seem to provide evidence regarding the development of early human multimodal communication beyond that shown by great ape repertoires (Levinson and Holler 2014). An extensive review of the pointing literature by Krause and his colleagues (2018, p. 339) concludes that "Early conceptions of pointing with the index finger as a human species-specific gesture derived from our unique adaptations for language have been revealed by subsequent research to be both cross-culturally and evolutionarily inadequate to account for the full range of nonverbal referential capacities manifested by a large range of vertebrate species."

Leavens et al. (2009, p. 164) assert that "deixis, the ability to direct the attention of another to a specific locus, is a shared capacity of great apes and humans. Because deixis in great apes cannot ultimately derive from adaptations for bipedalism, profligate encephalization, or neurobiological or cognitive adaptations for speech, then this suggests that our hominin ancestors were pre-adapted for joint attention, that joint attention is a faculty of 'language' in a broad sense (shared by humans and other animals)."

Although Tomasello and Call (2019, p. 467) conclude from these and other considerations that "although great apes do understand about directing the attention of others to things, they do not understand reference in human-like ways because human reference is, in effect, an invitation to share attention … The deepest underlying issue that differentiates great ape gestural communication from human gestural communication, therefore, is cooperation. Great apes are essentially communicating in order to fulfill individualistic goals—and they understand others to be doing this as well—whereas humans are communicating cooperatively in the context of joint goals and joint attention," this may be altered by new research, for example, into the nature of vocal communication during cooperative hunting by wild chimpanzees (Mine et al. 2022) and of declarative referential gesturing by wild chimpanzee pairs (Wilke et al. 2022).

However, the view of Wheeler and Fischer (2012) still appears to be valid: that even the most sophisticated ape signals appear to be at the stage of Dennett's "first order intentionality" in which the signal is intended to alter the immediate behavior of the receiver, though the sender's intent may not be a conscious one, rather than at the second stage, in which the sender intends to alter the receiver's actual knowledge state rather than any immediate behavior. They then suggest that "context-specific signals" would be a more useful term, as the investigation of ecologically based factors surrounding the development and use of various calls in particular species shows that the physical and social environment may have a much more direct impact on these than previously realized. Rendall and colleagues (2009) argue that exerting influence rather than exchanging information may well be the underlying "goal" of animal communication, while Scarantino (2010) believes that the broader term "functional signaling" should continue to be used by animal communications researchers without any implications regarding sender intentionality. Sievers and Gruber (2016) propose instead that researchers turn to actual situations of animal signaling and adopt a pragmatic approach of identifying specific "acts of reference," with a focus on the signaler and whether it intentionally produces a referential signal with the goal of modifying audience behavior. Liebal and her colleagues (2022) lament the fact that multimodal communication research in primate communities continues to be uncommon, despite the fact that analyzing the combination of facial, gestural, and vocal manifestations is a routine method of research in understanding human communication and could well be applied in studying our fellow primates.

Relatedly, researchers continue to disagree to what extent "theory of mind" or "perspective taking" appears among different species of non-human primates in the wild (Royka and Santos 2022). Some research suggests that certain apes do have cognitive abilities that might accurately termed be "beliefs" and "plans" related to their environment. For example, female chacma baboons are known to engage in a variety of proactive strategies to protect their nursing young from potentially infanticidal males in their own and neighboring troops (Palombit 2015), and wild orangutan males communicate their intent to travel by specific calls to conspecifics up to a day in advance of any actual movements (van Schaik et al. 2013). However, research also indicates that non-human primates cannot identify the truth or falsity of a belief held by another if it differs from their own (Martin and Santos 2016). Kano and his colleagues (2020) contest this, saying that the evidence of failure is largely due to experimental design issues, as the so-called false belief test, originally applied to human children, does not adequately control for nonhuman primates' cognitive differences and their experiments indicate that chimpanzees can pass appropriate false belief tests, arguing that much more research on non-human theory of mind is necessary.

Bermúdez (2007) has theorized that while nonlinguistic creatures are unable to apply formal rules of inference defined over propositions, since they lack the public language required to think about thoughts, they are able to make certain rational causal connections between pairs of contrary concepts in a form of proto-inference, though they are obviously

unable to describe the inferential process itself. This is exemplified in the many studies of problem-solving by nonhuman primates and by human infants (Hopper 2010). In the case of human infants, who will eventually achieve linguistic competence, it also supplies the initial scaffolding necessary to ascend to logical thought over time.

These issues are integral to developing a fully human "theory of mind," which is the ability of both speakers and listeners to make attributions about one another's beliefs, knowledge, and other mental states, and is theoretically inherent in all forms of human communication as it is pragmatically what connects speaker reference and listener relevance (Sperber and Wilson 1986). Even though enculturated apes raised in human-created environments that support their learning of symbolic communication systems show some evidence of sophisticated usage of imperative signs (Rivas 2005), deictic signs (Lyn et al. 2014), and even declarative signs (Lyn et al. 2011), they usually confine their usage to requests for particular objects and actions, rather than engaging in the type of exploratory communicative interactions common in human children beyond the age of two or three years. This is thus a less telling source of evidence for the actual evolution of proto-language, given the unnatural nature of the experimental environment as opposed to non-human primate interactions in the wild, which are obviously much richer and unstructured as shown by the observational studies that have been made over time (Leavens et al. 2009).

More broadly, it has been argued that there has been undue emphasis by researchers on defining cognition from a purely human point of view and that there are undoubtedly certain cognitive skills that other animals have evolved to perform better than humans, but these have been both under-valued and under-researched (Bräuer et al. 2020), as are the dimensions of animal consciousness in general (Birch et al. 2020), while the cognitive and communicative potential of non-human primates today remains a matter of both ethical and legal concern (Benvenuti 2016). Obviously, non-human primates have evolved in a different direction than humans in terms of their cognitive and communication capabilities, so the next chapter will focus on the fossil hominin record in order to further explore the development of vocal and gestural deixis into some form of proto-language, the unknown communication system that over millennia was to become modern speech.

References

Allen, C., & Saidel, E. (1998). The evolution of reference. In D. Cummins & C. Allen (Eds.), *The evolution of mind* (pp. 183–203). Oxford University Press.

Amphaeris, J., Shannon, G., & Tenbrink, T. (2022). Overlap not gap: Understanding the relationship between animal communication and language with prototype theory. *Lingua*, *272*, Article 103332. https://doi.org/10.1016/j.lingua.2022.103332

Arnold, K., & Bar-On, D. (2020). Primate pragmatics, expressive behavior, and the evolution of language. *Animal Behavior and Cognition*, *7*(2), 117–130. https://doi.org/10.26451/abc.07.02.06.2020

Benvenuti, A. (2016). Evolutionary continuity and personhood: Legal and therapeutic implications of animal consciousness and human unconsciousness. *International Journal of Law and Psychiatry*, *48*, 43–49. https://doi.org/10.1016/j.ijlp.2016.06.001

Bermúdez, J. L. (2007). *Thinking without words*. Oxford University Press.

Birch, J., Schnell, A. K., & Clayton, N. S. (2020). Dimensions of animal consciousness. *Trends in Cognitive Sciences*, *24*(10), 789–801. https://doi.org/10.1016/j.tics.2020.07.007

Bräuer, J., Hanus, D., Pika, S., Gray, R., & Uomini, N. (2020). Old and new approaches to animal cognition: There is not "one cognition." *Journal of Intelligence*, *8*(3), Article 28. https://doi.org/10.3390/jintelligence8030028

Brejcha, J., & Kleisner, K. (2016). Turtles are not just walking stones: Conspicuous coloration and sexual selection in freshwater turtles. *Biosemiotics*, *9*(2), 247–266. https://doi.org/10.1007/s12304-015-9249-9

Byrne, R. W., & Cochet, H. (2017). Where have all the (ape) gestures gone? *Psychonomic Bulletin & Review*, *24*, 68–71. https://doi.org/10.3758/s13423-016-1071-0

Cissewski, J., & Boesch, C. (2016). Communication without language: How great apes may cover crucial advantages of language without creating a system of symbolic communication. *Gesture*, *15*(2), 224–249. https://doi.org/10.1075/gest.15.2.04cis

Dawkins, R., & Krebs, J. R. (1978). Animal signals: Information or manipulation. In J. R. Krebs & N. B. Davies (Eds.), *Behavioral ecology: An evolutionary approach* (pp. 282–309). Blackwell Scientific.

Dewey, J. (1946). Peirce's theory of linguistic signs, thought, and meaning. *The Journal of Philosophy*, *43*(4), 85–95. https://doi.org/10.2307/2019493

Eco, U., & Sebeok, T. A. (1988). *The sign of three: Dupin, Holmes, Peirce*. Indiana University Press.

Evans, C. S., & Clarke, J. A. (2010). Referential signaling. In M. D. Breed & J. Moore (Eds.), *Encyclopedia of animal behavior* (pp. 27–32). Elsevier.

Favareau, D. (2009). Introduction: An evolutionary history of biosemiotics. In M. Barbieri (Ed.), *Essential readings in biosemiotics: The new synthesis* (pp. 1–67). Springer.

Fischer, J., & Price, T. (2017). Meaning, intention, and inference in primate vocal communication. *Neuroscience & Biobehavioral Reviews*, *82*, 22–31. https://doi.org/10.1016/j.neubiorev.2016.10.014

Ghazanfar, A. A., Liao, D. A., & Takahashi, D. Y. (2019). Volition and learning in primate vocal behaviour. *Animal Behaviour*, *151*, 239–247. https://doi.org/10.1016/j.anbehav.2019.01.021

Gillespie-Lynch, K., Greenfield, P., Feng, Y., Savage-Rumbaugh, S., & Lyn, H. (2013). A cross-species study of gesture and its role in symbolic development: Implications for the gestural theory of language evolution. *Frontiers in Psychology*, *4*, Article 160. https://doi.org/10.3389/fpsyg.2013.00160

Graham, K. E., Badihi, G., Safryghin, A., Grund, C., & Hobaiter, C. (2022). A socio-ecological perspective on the gestural communication of great ape species, individuals, and social units. *Ethology, Ecology, & Evolution*, *34*(3), 235–259. https://doi.org/10.1080/03949370.2021.1988722

Hockett, C. F., & Altmann, S. A. (1968). A note on design features. In T. A. Sebeok (Ed.), *Animal communication: Techniques of study and results of research* (pp. 61–72). Indiana University Press.

Hopper, L. M. (2010). "Ghost" experiments and the dissection of social learning in humans and animals. *Biological Reviews*, *85*(4), 685–701. https://doi.org/10.1111/j.1469-185X.2010.00120.x

Kano, F., Call, J., & Krupenye, C. (2020). Primates pass dynamically social anticipatory-looking false-belief tests. *Trends in Cognitive Sciences*, *24*(10), 777–778. https://doi.org/10.1016/j.tics.2020.07.003

Keeler, M. (2020). The hidden treasure of C. S. Peirce's manuscripts. *Chinese Semiotic Studies*, *16*(1), 155–166.https://doi.org/10.1515/css-2020-0008

Krause, M. A., Udell, M. A., Leavens, D. A., & Skopos, L. (2018). Animal pointing: Changing trends and findings from 30 years of research. *Journal of Comparative Psychology*, *132*(3), 326–345. https://doi.org/10.1037/com0000125

Leavens, D. A., Racine, T. P., & Hopkins, W. D. (2009). The ontogeny and phylogeny of non-verbal deixis. In R. Botha & C. Knight (Eds.), *The prehistory of language* (pp. 141–165). Oxford University Press.

Leavens, D. A., Reamer, L. A., Mareno, M. C., Russell, J. L., Wilson, D., Schapiro, S. J., & Hopkins, W. D. (2015). Distal communication by chimpanzees (Pan troglodytes): Evidence for common ground? *Child Development*, *86*(5), 1623–1638. https://doi.org/10.1111/cdev.12404

Levinson, S. C., & Holler, J. (2014). The origin of human multi-modal communication. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), Article 20130302. https://doi.org/10.1098/rstb.2013.0302

Liebal, K., & Call, J. (2012). The origins of non-human primates' manual gestures. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1585), 118–128. https://doi.org/10.1098/rstb.2011.0044

Liebal, K., Slocombe, K. E., & Waller, B. M. (2022). The language void 10 years on: Multimodal primate communication research is still uncommon. *Ethology, Ecology, & Evolution*, *34*(3), 274–287. https://doi.org/10.1080/03949370.2021.2015453

Lingle, S., Wyman, M. T., Kotrba, R., Teichroeb, L. J., & Romanow, C. A. (2012). What makes a cry a cry? A review of infant distress vocalizations. *Current Zoology*, *58*(5), 698–726. https://doi.org/10.1093/czoolo/58.5.698

Lyn, H., Greenfield, P. M., Savage-Rumbaugh, S., Gillespie-Lynch, K., & Hopkins, W. D. (2011). Nonhuman primates do declare! A comparison of declarative symbol and gesture use in two children, two bonobos, and a chimpanzee. *Language & Communication*, *31*(1), 63–74. https://doi.org/10.1016/j.langcom.2010.11.001

Lyn, H., Russell, J. L., Leavens, D., Bard, K. A., Boysen, S. T., Schaeffer, J. A., & Hopkins, W. D. (2014). Apes communicate about absent and displaced objects: Methodology matters. *Animal Cognition*, *17*(1), 85–94. https://doi.org/10.1007/s10071-013-0640-0

Manser, M. B. (2020). Referents and semantics in animal vocalizations. In M. A. Bee & C. T. Miller (Eds.), *Psychological mechanisms in animal vocalizations* (pp. 222–249). Springer.

Maran, T. (2014). Dimensions of zoosemiotics: Introduction. *Semiotica*, *2014*(198), 1–10. https://doi.org/10.1515/sem-2013-0098

Marler, P. (1961). The logical analysis of animal communication. *Journal of Theoretical Biology*, *1*(3), 295–317. https://doi.org/10.1016/0022-5193(61)90032-7

Martin, A., & Santos, L. R. (2016). What cognitive representations support primate theory of mind? *Trends in Cognitive Sciences*, *20*(5), 375–382. https://doi.org/10.1016/j.tics.2016.03.005

Maynard Smith, J. (1982). *Evolution & the theory of games*. Cambridge University Press.

Millikan, R. G. (2004). On reading signs: Some differences between us and the others. In D. Kimbrough Oller & U. Griebel (Eds.), *Evolution of communication systems: A comparative approach* (pp. 15–29). MIT Press.

Mine, J. G., Slocombe, K. E., Willems, E. P., Gilby, I. C., Yu, M., Thompson, M. E., Muller, M. N., Wrangham, R. W., Townsend, S. W., & Machanda, Z. P. (2022). Vocal signals facilitate cooperative hunting in wild chimpanzees. *Science Advances*, *8*(30). https://doi.org/10.1126/sciadv.abo5553

Moorjani, P., Amorim, C. . E. G., Arndt, P. F., & Przeworski, M. (2016). Variation in the molecular clock of primates. *PNAS*, *113*(38), 10607–10612. https://doi.org/10.1073/pnas.1600374113

Morris, C. W. (1971). *Writings on the general theory of signs*. Mouton.

Nöth, W. (1995). *Handbook of semiotics*. Indiana University Press.

Ogden, C. K., & Richards, I. A. (1927). *The meaning of meaning*. Harcourt, Brace, Jovanovich.

Ongstad, S. (2021). Can animals refer? Meta-positioning studies of animal semantics. *Biosemiotics*, *14*(2), 433–457. https://doi.org/10.1007/s12304-021-09401-7

Owings, D. H., Beecher, M. D., & Thompson, N. S. (1997). Introduction: Perspectives on communication. In D. H. Owings, M. D. Beecher, & N. S. Thompson (Eds.), *Communication: Perspectives in ethology* (Vol. 12, pp. 1–6). Springer.

Palombit, R. A. (2015). Infanticide as sexual conflict: Coevolution of male strategies and female counterstrategies. *Cold Spring Harbor Perspectives in Biology*, *7*(6), Article a017640. https://doi.org/10.1101/cshperspect.a017640

Price, T., Wadewitz, P., Cheney, D. L., Seyfarth, R. M., Hammerschmidt, K., & Fischer, J. (2015). Vervets revisited: A quantitative analysis of alarm call structure and context specificity. *Scientific Reports*, *5*, Article 13220. https://doi.org/10.1038/srep13220

Rendall, D., Owren, M. J., & Ryan, M. J. (2009). What do animal signals mean? *Animal Behaviour*, *78*(2), 233–240. https://doi.org/10.1016/j.anbehav.2009.06.007

Restrepo, J. A. F., & De Mesa, J. A. L. (2021). A new approach to the problem of the order of the ten trichotomies and the classification of sixty-six types of signs in Peirce's late speculative grammar. *Transactions of the Charles S. Peirce Society*, *57*(3), 374–396. https://muse.jhu.edu/pub/3/article/845176

Rivas, E. (2005). Recent use of signs by chimpanzees (Pan troglodytes) in interactions with humans. *Journal of Comparative Psychology*, *119*(4), 404–417. https://doi.org/10.1037/0735-7036.119.4.404

Royka, A., & Santos, L. R. (2022). Theory of mind in the wild. *Current Opinion in Behavioral Sciences*, *45*, Article 101137. https://doi.org/10.1016/j.cobeha.2022.101137

Scarantino, A. (2010). Animal communication between information and influence. *Animal Behaviour*, *79*(6), e1–e5. https://doi.org/10.1016/j.anbehav.2010.03.005

Scarantino, A. (2013). Rethinking functional reference. *Philosophy of Science*, *80*(5), 1006–1018. https://doi.org/10.1086/673900

Sebeok, T. A. (1963). Reviewed works: Communication among social bees; porpoises and sonar; man and dolphin. *Language*, *39*, 448–466. https://doi.org/10.2307/411126

Sebeok, T. A. (2001). Biosemiotics: Its roots, proliferation, and prospects. *Semiotica*, *134*(1/4), 61–78. https://doi.org/10.1515/semi.2001.014

Seyfarth, R. M., & Cheney, D. L. (1993). Meaning, reference, and intentionality in the natural vocalizations of monkeys. In H. L. Roitblat, L. M. Herman, & P. E. Nachtigall (Eds.), *Language and communication: Comparative perspectives* (pp. 1995–219). Psychology Press.

Seyfarth, R. M., Cheney, D. L., & Marler, P. (1980). Vervet monkey alarm calls: Semantic communication in a free-ranging primate. *Animal Behaviour*, *28*(4), 1070–1094. https://doi.org/10.1016/S0003-3472(80)80097-2

Sievers, C., & Gruber, T. (2016). Reference in human and nonhuman primate communication: What does it take to refer? *Animal Cognition*, *19*(4), 758–768. https://doi.org/10.1007/s10071-016-0974-5

Sperber, D., & Wilson, D. (1986). *Relevance: Communication & cognition.* Harvard University Press.

Stolk, A., Blockpoel, M., van Rooij, I., & Toni, I. (2015). On the generation of shared symbols. In R. E. Willsems (Ed.), *Cognitive neuroscience of natural language use* (pp. 201–226). Cambridge University Press.

Tomasello, M., & Call, J. (2019). Thirty years of great ape gestures. *Animal Cognition*, *22*(4), 461–469. https://doi.org/10.1007/s10071-018-1167-1

Van der Veer Martens, B. (2023). On thresholds: Signs, symbols, and significance. *Journal of Documentation.* Advance online publication. https://doi.org/10.1108/JD-08-2022-0168

van Schaik, C. P., Damerius, L., & Isler, K. (2013). Wild orangutan males plan and communicate their travel direction one day in advance. *PLOS ONE*, *8*(9), Article e74896. https://doi.org/10.1371/journal.pone.0074896

Von Uexküll, J. (2001). An introduction to Umwelt. *Semiotica*, *134*(1/4), 107–110. https://doi.org/10.1515/semi.2001.017

Wacewicz, S., & Żywiczyński, P. (2015). Language evolution: Why Hockett's design features are a non-starter. *Biosemiotics*, *8*(1), 29–46. https://doi.org/10.1007/s12304-014-9203-2

Warren, W. C., Jasinska, A. J., García-Pérez, R., Svardal, H., Tomlinson, C., Rocchi, M., Archidiacono, N., Capozzi, O., Minx, P., Montague, M. J., & Kyung, K. (2015). The genome of the vervet (Chlorocebus aethiops sabaeus). *Genome Research*, *25*(12), 1921–1933. https://doi.org/10.1101/gr.192922.115

Wheeler, B. C., & Fischer, J. (2012). Functionally referential signals: A promising paradigm whose time has passed. *Evolutionary Anthropology: Issues, News, and Reviews*, *21*(5), 195–205. https://doi.org/10.1002/evan.21319

Wilke, C., Lahiff, N. J., Sabbi, K. H., Watts, D. P., Townsend, S. W., & Slocombe, K. E. (2022). Declarative referential gesturing in a wild chimpanzee (Pan troglodytes). *PNAS*, *119*(47), Article e2206486119. https://doi.org/10.1073/pnas.2206486119

Wisenden, B. D. (2015). The cue–signal continuum. In P. W. Sorensen & B. D. Wisenden (Eds.), *Fish pheronomes and related cues* (pp. 149–158). Wiley Blackwell.

Proto-Signs, Proto-Words

3

Abstract

This chapter reviews some of the more important scenarios for language evolution, including Bickerton's "power scavenging" hypothesis and Chomsky's "Merge" hypothesis, all of which attempt to theorize the emergence of the earliest "proto-words" and "proto-signs," the potentially multiple communicative variants using speech and/or gestures that may have been developed by the small, scattered human groups existing long before any known proto-language became dominant enough to leave traces in modern language.

Christiansen and Kirby (2003) call the evolution of language "the hardest problem in science" largely because, as has often been said, "language does not fossilize" (Botha and Everaert 2013, p. 1). Any inferences about language evolution must be made from very different sources of evidence, not only the communicative behavior and cognition of living non-human primates and other animals as discussed in the previous chapter, but those genes believed to be involved in human language and speech, and the fossil remains of prehistoric hominins and associated artifacts. There have been a number of different initial evolution of language scenarios offered during the past several decades, but adducing empirical evidence in support of these theories has only recently become prevalent (Wacewicz and Zywiczynski 2017). Fitch (2017, p. 3) comments that "there is an increasing amount of relevant data allowing empirical evaluation of such models.…The key challenge facing the study of language evolution is not a lack of data, but rather a weak commitment to hypothesis-testing approaches and strong inference, exacerbated by the broad and highly interdisciplinary nature of the relevant data."

B. V. V. Martens, *Keywords In and Out of Context*, Synthesis Lectures on Information Concepts, Retrieval, and Services, https://doi.org/10.1007/978-3-031-32530-4_3

3.1 Proto-World

Lameira and Call (2020) explain that language evolution researchers are broadening their focus: "Fifteen million years ago, unlike today, the world was inhabited by multiple and contemporaneous hominid species, of which at least 22 are recognized extinct genera, excluding the direct ancestors who would become genus Homo. Given how improbable fossilization is, numbers were beyond doubt much larger than what the fossil record shows. Besides ourselves, the only surviving traces of this diversity nowadays are the seven (nonhuman) great ape species alive today, three orangutan, two gorilla, one chimpanzee, and one bonobo species, comprising three genera, Pongo, Gorilla, and Pan, respectively. Each extant great ape lineage diverged at different times from the one that eventually gave raise to Homo. Pongo was the first lineage to diverge, Gorilla the second, Pan the last. This is reflected in the level of genetic similarity between each great ape genus and humans, as well as their taxonomy… Acknowledging the diversity and variance of great ape traits (and more generally across animal taxa) that contributed to shape cognitive and linguistic evolution will help make new strides in the effort of reconstructing the evolutionary timeline of language."

This chapter briefly reviews some of the more important scenarios and evidence that connect them to the inferences made by different scenarios about developments that may have led to the earliest "proto-signs" and "proto-words," both of which belong to what has been called "Proto-World," which is alleged to be that period approximately 100,000 years prior to the emergence of what historical linguists, somewhat confusingly, also call "proto-languages," meaning those earliest identifiable languages such as Proto-Afro-Asiatic and Proto-Indo-European which preceded the emergence of languages in their modern forms.

"Proto-World" is, of course, only a metaphor, which covers multiple possibilities: that human language was a unique "saltationist" event that eventually came to affect all human populations, or that multiple communicative variants using speech and/or gestures repeatedly emerged due to various factors but became extinguished among the small human groups that occupied various continents over the millennia before any known historical proto-language became dominant enough to leave traces, or that some still unknown series of events was the catalyst for human language.

Arbib (2005a, 2005b) describes a proto-language as a system of utterances used by a particular hominin species (possibly including Homo sapiens), which could be recognizable as precursor to human language but which is not itself a human language in the modern sense. He defines "proto-language" as a system of communication using "proto-signs" or "proto-words" whose meaning can be established within a community by conventionalization (and in this sense is open-ended) but which has little or no grammar. In contrast, a true language has a large and open-ended stock of words (the lexicon) and a grammar which supports both familiar and novel combination of words into meaningful patterns that comprise phrases and sentences in that language (compositional semantics).

It should be emphasized that most contemporary linguists are dismissive of efforts to hypothesize the initial emergence of language in this way, given their discipline's adherence to a form of uniformitarianism (the principle originally articulated by Lyell in regard to geological processes) that any natural processes must be observable in the present in order to hypothesize that they operated similarly in the past) for the study of language in general. Further, given that the rate of change in language over time is very high, and the risk of accidental similarities in comparing various words from different languages being seen as significant ones is also very high, no traces of a "Proto-World" origin for any word can reasonably be expected to be found in any known language. (See Campbell and Poser (2008) for an extensive discussion of these issues.) Nevertheless, while there may be no direct bridge between proto-words in Proto-World and any known language today, it is likely that certain of our key contemporary concepts, such as family, fire, food, and fornication, would have survived over the millennia, even though their original expressions have not.

Although no specimens of prehominin or early hominin vocalizations exist, it is clear that modern human cognition and communication must have been enabled by the development of particular genetic and physical capabilities to support them, so the genetic and paleoanthropic findings from both extinct and very early modern hominins are useful in assessing different theories regarding these (Lieberman 2016). Similarly, even though genetic analysis indicates that chimpanzees and bonobos have diverged about as much from the common human-chimpanzee ancestor as have humans (Wildman et al. 2003) and even that substantially more genes underwent positive selection in the chimpanzee lineage than in the human lineage, suggesting more adaptive genetic changes during chimpanzee evolution than during human evolution (Balzeau et al. 2014), some inferences can reasonably be made from data regarding their genetic and morphological data as well.

Approximately 99% of human DNA is similar to that of the chimpanzees and bonobos with whom we shared a common ancestry roughly 13 million years ago, and even more with the Neanderthals and Denisovans with whom humans shared ancestry 600,000 years ago, so the genetic findings that the FOXP2 gene, which is required for normal development of speech and language, although present in all lineages, acquired two amino acid substitutions (N303 and S325) after the split of hominids from their common ancestor with chimpanzees, suggesting that the two gene changes might have contributed to the development of hominin linguistic abilities (Mozzi et al. 2016). The version of the FOXP2 transcriptional factor that is shared by humans, Neanderthals, and Denisovans enhanced synaptic connectivity and malleability in these neural circuits. Tiwary (2020) has recently reported findings that the FOXP2 gene may have coevolved with 6 neighboring genes (*DGKQ, KLF10, HSPBP1, RIMS1, TAC3, and PRUNE*) to produce the cognition necessary for human speech.

As Tallerman (2006) points out, though the common ancestor of chimp and human at around 5–6 million years ago had a vocal tract, tongue and lips and glottis, ears, and much else besides that have been co-opted for speaking and listening, speech (production,

perception, and the neural mechanisms supporting it) itself is not language, which must be learned. However, studies in humans and a small number of other species have provided insights into the neural and genetic basis for learned vocal communication and are helping to delineate the roles of brain circuits across the cortex, basal ganglia, and cerebellum in generating vocal behaviors (Konopka and Roberts 2016).

The cerebellum, widely thought to play a critical role in the executive functions of both cognition and motion (Mariën et al. 2014), is also large in ancestral apes, suggesting that the initial impetus for this may have been to improve branch locomotion and route planning, and may have later helped in the development of both technologically and socially complex behaviors. Additional cerebellular specializations may have contributed to planning and comprehension, and may also have laid foundations for syntactical aspects of language (Barton 2012). It is notable also that the very different configurations of tongue, palate, larynx, and dentition in the modern human, Neanderthal, and Denisovan would tend to preclude the development of identical vocalization assemblages, though there is growing consensus that the latter two also had developed language capabilities (Progovac 2016).

3.2 Theories of Language Evolution

Given the above background, various theories that have represented the diversity of thought about language evolution over the past several decades are of particular interest here. These fall into two general categories, focusing primarily on the importance of communication or of culture, though many include elements of both perspectives. Among the communication-centric theories are those of Michael Arbib, Derek Bickerton, Dorothy Cheney and Robert Seyfarth, Noam Chomsky, Michael Studdert-Kennedy and Louis Goldstein, and Jordan Zlatev. The culture-centric theories include those of Terrence Deacon, Merlin Donald, Robin Dunbar, Dean Falk, James Hurford, Steven Mithen, and Thom Scott-Phillips.

3.3 Communication-Centric Theories of Language Evolution

Arguably the best-known communication-centric theory of language evolution is that of Chomsky (2010), whose "Merge" hypothesis about the emergence of language, part of his so-called "Minimalist program," which proposes that grammatical recursion is the minimum characteristic of the faculty of language, and that it was the result of a single human mutation sometime within the past 100,000 to 60,000 years. Rather than suggesting a gradual transition from any form of proto-language, Chomsky's hypothesis continues to posit an unprecedented, immediate leap into modern language.

No trace of this has been found in genetic drift as yet, though Benítez-Burraco and Uriagereka (2015) have suggested the possibility of a link to ancient viral vectors and their potential epigenetic impact on a developing immune system and brain, and Fujita (2009) has proposed that "Merge" evolved from a preexisting capacity for hierarchical and sequential object combination as typically observed in tool using and tool making. Huybregts (2017) offers an alternative suggestion for the existence of "Merge": that the capacity for internalized language ("Merge") existed internally before human populations became separated, but that externalized language (speech) only emerged after that separation. He gives the example of the Khoisan "click" languages that are only found among biologically Khoisan peoples as empirical support for Chomsky's "Merge" hypothesis.

Many non-Minimalists, however, find "Merge" unconvincing in light of growing evidence that the various shifts that eventually enabled language were more gradual and less dramatic than the Minimalist program appears to allow (Botha 2011). Murphy (2019) notes that there exist 1,241 primate-specific genes, 280 of which are human-specific. Fifty-four percent of these human-specific genes are upregulated in the prefrontal cortex, a brain area implicated in higher cognition. Given this level of regulatory complexity, Murphy considers it most likely that the neurocomputational properties required for language emerged after the mutation of multiple regulatory genes acting in concert rather than through a singular mutational event such as "Merge."

Such a more gradual timeline is offered by the "Mirror System Hypothesis," which is founded on the idea that mechanisms which support language in the human brain evolved atop a basic mechanism not originally related to communication (Arbib 2017; Rizzolatti and Arbib 1998). Instead, the mirror neuron system for grasping, with its capacity to generate and recognize a set of actions, provides the evolutionary basis for language parity (that is, that an utterance means roughly the same for both speaker and hearer). In particular, the Broca's area in humans contains, but is not limited to, a mirror system for grasping, which is homologous to that seen in the macaque. Arbib proposes that Homo habilis through to early Homo sapiens had a proto-language based primarily on manual gestures ("proto-sign"), which provided the essential scaffolding for the emergence of a proto-language based primarily on vocal gestures ("proto-speech"), but that the hominin line saw advances in both proto-sign and proto-speech feeding off each other in an expanding spiral so that proto-sign did not attain the status of a full language prior to proto-speech. The claim is that the demands of an increasingly spoken proto-vocabulary then provided the evolutionary pressure that yielded a vocal apparatus and corresponding neural control to support the human ability for rapid production and co-articulation of phonemes that underlie human speech.

Arbib also advances the notion of proto-words as individual holographic units that represent complex propositions, more like whole modern sentences, from which "true" language developed. He supports this by suggesting aspects of "language-readiness" supported by hominin brain mechanisms that evolved prior to the emergence of language.

These aspects include complex action recognition and complex imitation; intended communication; symbolization; parity (that is, what counts for the producer of one or more symbols must count, frequently, as approximately the same for the receiver of those symbols, extending the role of the mirror neurons for grasping and other actions by "lifting" complex imitation from praxis to communication); the perception of hierarchical structure of scenes to determine what actions to execute and when to execute them to achieve their goals; the ability to recall past events or imagine future ones; a prolonged period of infant dependency, especially pronounced in humans, combined with the willingness of adults to act as caregivers and the consequent development of social structures to provide the conditions for complex social learning; symbols becoming words in the modern sense, interchangeable and composable in the expression of meaning; the matching of syntactic to semantic structures growing in complexity, with the nesting of substructures making some form of recursion inevitable; verb tenses or other tools expressing the ability to recall past events or imagine future ones; and, finally, to qualify as a human language, much of the syntax and semantics of a human language must be learnable by most human children.

Criticism of proto-words as individual holographic units centers around the apparent contradiction that the segmentation it depends on can only work if both signals and meanings contain sub-units, yet these holographic units by definition are unitary, unstructured utterances that do not easily lead to the combinatorial complexity required for evolution into modern languages (Tallerman 2007).

Another communication-centric theory, that of "particulate language," is based on the existence of a neuroanatomical system of independently movable parts that can be coordinated to produce rapid sequences of motor actions that can result in articulation (Studdert-Kennedy and Goldstein 2004). Speech is a motor function that draws on ancient mammalian oral capacities, notably sucking, a defining mammalian behavior, which presumably began the neuroanatomical differentiation of the mammalian tongue, culminating in its usage for calls and cries, and eventually for the hominin proto-syllable, precursor of the modern child's unit of babble and the modern adult's unit of rhythm and melody. This theory views the early proto-syllable as a gesture, since it involves the constriction and release of one of the vocal organs. Once these vocal gestures had evolved as discrete, combinable units, the expansion of "particulate language" could have occurred through a process of mutual attunement between individual and eventual communal speakers/hearers without any further genetic change. The "particulate" theory of language, therefore, hypothesizes the eventual precipitation of these gestures and sounds into further processes of actual acculturation.

A theory based on bodily mimetic behavior also focuses on gestural evolution from a cognitive semiotic perspective, suggesting that pantomime formed the original human-specific communication system through the emergence of iconic representation through mimicry, dominated by bodily gestures for enactment and supported by expressive vocalization and facial expressions (Zlatev et al. 2020). This represents a final stage of mimesis,

as earlier stages (proto-mimesis as expressed by neonatal mirroring, dyadic mimesis as expressed by object movement re-enactment, and triadic mimesis as expressed through the imitation of sign use) have all been observed to some extent in apes (Zlatev et al. 2005).

Perniss and Vigliocco (2014) have urged a similarly semiotic framework in which this iconicity in face-to-face communication (spoken and signed) serves as a powerful vehicle for bridging between human sensorimotor experience and language. They claim that iconicity might have played a key role in establishing displacement (the ability of language to refer beyond what is immediately present) and in supporting referentiality (learning to map linguistic labels to objects, events, etc., in the world), and in providing a mechanism to account for how language comes to be embodied (grounded in our sensory and motor systems), which is core to meaningful communication.

3.4 Culture-Centric Theories of Language Evolution

The communication-centric theories, however, do not explain what might have precipitated the necessity for more sophisticated semiotic functioning among early hominins. Bickerton's (1990, 1995, 2014) theory for the origins of proto-language presents such a need, speculating that "power scavenging" was necessitated by a dramatic change in habitat from jungle to savanna during the early African migration period. The ability of these small groups to search for available animal carcasses and effectively communicate their whereabouts to others, then successfully distract larger predators from these kills, and eventually to compete as groups of hunters with other predators for their prey created an ecological niche that, in turn, facilitated the development of a proto-language that emphasized observation, tracking, and coordination of these activities.

Since contemporary chimpanzees also sporadically engage in group hunting activities (Moore et al. 2017), it is likely that early hominins did the same. MacDonald and Roebroeks (2013) have argued that a reduction of information search costs in finding prey animals by this common knowledge of animal behavior, communicated with mimicry, natural gestures, and a small vocabulary would have significant survival advantages.

Bickerton proposed a "compositional" approach: that is, that the proto-language must have contained a small set of easily expressed sounds with simple meanings expressing basic cognitive concepts, which could be called the "ancestors" of modern nouns and verbs. He first proposed proto-language as an intermediate stage between a non-linguistic state and a fully fledged modern language and suggests, moreover, that proto-language still exists in the human brain, only resurfacing when normal linguistic development fails. Each word in this proto-language refers to a single, simple concept, and although these could have been concatenated together in a "slow, clumsy, ad hoc stringing together of symbols" (1995, p. 65), Bickertonian proto-language is always characterized as "a lexicon without syntax" (1995, p. 51).

One weakness of this theory is that it is unlikely that the so-called "displaced reference" (the ability to refer to something that is not immediately visible, which Bickerton identifies as the key driver behind successful power scavenging groups) could have been the only factor at work in the development of a proto-language, given that this does not account for the maternal transmission of general "displaced reference" skills to infants, which is the primary means of such skill development in humans. It is of course possible that, like other, later hunter-gatherer groups, both hunting and foraging survival required the ability to locate and describe especially suitable food sources to other members of the group, which would allow infants to pick up this ability by observation and imitation as well. Szilágyi and colleagues (2023) note that the synergistic fitness advantage provided by these several components of the scavenging-specific hypothesis could have evolved more readily in combination rather than in isolation.

Cheney and Seyfarth's theory regarding the evolution of language, based on their extensive research among non-human primate communities, especially the importance of member knowledge of dominance hierarchy and kinship, proposes that (2010, pp. 294–295) "Gestural communication, including pointing and pantomime, combined with a few vocalizations representing key concepts such as kinship or danger would suffice for much of the proposed protolanguage. We suggest, then, that long before our ancestors spoke in sentences, they had a language of thought in which they represented the world—and the meaning of call sequences—in terms of actors, actions, and those who are acted upon. The linguistic revolution occurred when our ancestors began to express this tacit knowledge, and to use their cognitive skills in speaking as well as listening. The prime mover behind this revolution was a theory of mind that had evolved to the point where its possessors did not just recognize other individuals' goals, intentions, and even knowledge—as monkeys and apes already do—but were also motivated to share their own intentions, beliefs, and knowledge with others... It also provided the selective pressure for the evolution of the physiological adaptations that enabled vocal modifiability. Whatever the selective pressures that prompted this change, the complex suite of skills that we call human speech built upon mental computations that had their origins and foundations in social interactions."

The most obvious criticism of Cheney and Seyfarth's "social interactions" theory is that, without additional evidence, contemporary ape interactions in the wild cannot provide sufficient justification for theories about hominid branches that diverged from a common ancestral ape millions of years ago. Nevertheless, their empirical contributions are clearly relevant and cannot readily be discounted as evidence for or against other theories.

Among the acculturation theories is Deacon's theory about the evolution of language, based on the notion that symbolic thinking triggered a co-evolutionary exchange between languages and brains over two million years of hominin evolution (Deacon 1998). He argues that the first symbolic communication evolved as a way for long-term, sexually exclusive pair bonds to co-exist with cooperative group foraging, which became a critical

factor when hunter-gatherer activities became more prevalent in a changing climate, as shown by the increase in stone tools and fossilized bones of other animals within prehistoric archaeology sites. The theory proposes that the natural instincts of males to ensure that they would have exclusive access to sexually receptive females and the instincts of females to ensure an adequate food supply for their offspring created an effective reproductive strategy in which mating became a symbolic social contract recognized by the group. This symbolic social contract, initially involving a very small symbolic domain, mediated by a more or less typical primate repertoire of non-symbolic vocal, olfactory, and gestural displays of affiliation toward the mate and aggression toward interlopers, was what provided a significant selective advantage through increased offspring survival over time to the majority of those who employed it. The development of kinship systems enabled by this initial level of monogamy may also have helped to promote group survival (Schacht and Kramer 2019).

Another culturally oriented theory is that of Donald (1991) whose argument regarding the evolution of the modern mind is that the so-called "mimetic culture" of early hominins was founded on the ability to provide conscious, self-initiated, representational acts that serve a variety of purposes, from self-reflection, to skill rehearsal, to social communication. Donald makes the provocative suggestion that, in fact, the rudiments of culture must have preceded and laid the foundations for the development of symbolic cognition:

> Although our brains have undoubtedly evolved a capacity for symbolic thought, this capacity is only vaguely defined in the nervous system itself. The brain is not, on its own, a symbolizing organ. The brain depends entirely on culture for the exploitation of its symbolic capacity, and some of its most impressive functions have a purely cultural origin. Symbolizing minds, as we know them, are not self-sufficient neural devices, as are eyes. They are hybrid products of a brain-culture symbiosis. Without cultural programming, they could never become symbolizing organs. They would become something else, very powerful perceptual-motor systems, like those of a superprimate, perhaps, but not truly symbolic. . .Unthinkable as it may seem, we are not even certain that spoken language, as we know it, was part of our primordial profile as a species. We have no firm empirical evidence by which we can dismiss the notion that language itself might be, especially in some of its most esoteric semantic and grammatical features, just another product of our deep symbiosis with culture (2000, pp. 22–25).

Increasing interest in what is being termed "animal culture" lends some support to Donald's suggestion, as research shows that there is a wide continuum of social learning behaviors among various species that can correctly be called some form of cultural transmission that are apparent in such behavioral domains as foraging techniques, mate selection, migratory pathways, nesting sites, preferences for particular prey, tool use, and vocal communication, though these are much less complex than those of human culture (Whiten 2021). Since language clearly is not necessary for these social learning behaviors to occur within animal groups, it is quite possible that such forms of early hominin culture did precede any spoken language, though the evaluation of such evidence is contingent upon one's definition of "culture."

Any use of proto-words by early hominins, therefore, was likely to have begun as subsidiary to gestural communication and to focus the attention of the group on immediate matters of survival, such as the location of seasonal food sources, and associated problem solving, such as crossing difficult terrain. However, this quite probably also offered the opportunity for "bootstrapping" the group's cognitive and communicative abilities over generations. Lock (1999, page 347) observes that "interaction constructs contexts that language can come to symbolize, thereby providing a cognitive technology that bootstraps the increasing discovery of those 'things' that are implied by what has already been symbolised." In particular, this presents the possibility of increasing the group's communicable store of shared memories, which could improve group problem-solving abilities and also represent an advance over what Donald terms "the episodic culture" of non-human primates, which defines their inability to focus attention beyond a very limited span of time, and an approach toward mimetic culture.

Falk's theory of the evolution of language (2007) hypothesizes that the vocal substrates for proto-language had prosodic features similar to contemporary so-called "motherese," which evolved as the enlarging brains in early hominins made birthing more difficult, thus causing a selective shift toward relatively undeveloped newborns and mothers that had genetically based potential for modifying their maternal behavior. These infants, unable to cling like infant apes, required their mothers to adopt new foraging strategies that allowed them to attend vigilantly to their infants by creating and using prosodic and gestural markings to encourage them to behave and follow. These proto-motherese markings eventually became conventionalized utterances that formed "proto-words" for those born into these particular hominin groups. Poliva (2015) has recently offered some tentative neuroanatomical support for the hypothetical development of these infant-mother location calls into the beginnings of speech.

Critics of Falk's theory argue, however, that the substrates of proto-speech/language may not be found in proto-motherese but rather in other primate vocalizations involved in courting, pair bonding, social grooming, or some other related phenomenon, and that her theory requires substantially more work on how contemporary motherese and first-language acquisition on the one hand are interrelated with prelinguistic evolution in early hominins on the other hand (Botha 2008).

Hurford (2007) presents a more abstract and theoretical focus on the interactions of social trust and evolving communication capabilities in the development of shared meaning-making among proto-humans. His theoretical perspective focuses on "proto-humans, who are innately disposed to bond with each other by imitation of meaningful gestures and vocalizations, and who show a high degree of trust and regard for others. Furthermore, their life-history includes a long period of dependency and cognitive plasticity. They have inherited from their ancestors a largely innate, but partly learned, set of calls, of crucial utility for survival and reproduction. The plasticity of the infants allows the genetic component of these calls to diminish while the learned component can maintain the system... In times this leads to an enlarged repertoire of learned conventionalized

meaningful calls, which simultaneously convey practical information and group membership. The calls are taken seriously (e.g., believed and acted upon) by members of the group because their conventionalized form reliably signals group membership....This keeps the group together as a cohesive social unit, which can outcompete less cohesive groups because of the practical utility of its communal conventional communicative code" (2007, pp. 329–330).

One problem with Hurford's theoretical perspective is that it lacks an explanation of how any set of animal natural signals could become conventionalized to the extent to which they can form the basis of more arbitrary proto-words. This may have been addressed by Sievers and Gruber (2020) who posit the existence of a continuum between arbitrary and non-arbitrary signals communicated by humans and non-humans, arguing further that there are at least some partially arbitrary elements in non-human primate communication, especially in the use of specific objects in various social interactions, as well as some non-arbitrary elements in human communication, especially in high arousal situations that can produce, for instance, screams. This also implies that social learning mechanisms for a nonhuman signal that is partially arbitrary do not have to be at the same level of complexity as that of human signal learning (Whiten et al. 2022) and that certain non-arbitrary signals such as screams are still inherent in humans (Arnal et al. 2015). Thus, communication conventions may have emerged over an extremely long time period, making use of existing natural signals.

Dunbar's "Gossip and Grooming" hypothesis (1995, 2017) is also culture-oriented, focusing on the transition from daytime grooming and gossip among very small groups of hominins to the larger groupings made possible by the ability to start and control campfires, which also facilitated nighttime "vocal grooming" activities such as rhythmic drumming and singing that would have helped to form larger group cohesion. The argument is that vocal grooming thus evolved gradually into vocal language through the medium of gossip and other narratives by the group members. However, there is little evidence to show that "vocal grooming" itself could become cognitively complex enough to transition into syntactical speech.

Mithen's (2006) "Hmmmmm" theory is an elaboration of Darwin's original notion of a so-called proto-language originating as proto-music. According to Mithen, early hominin communication was made up of "holistic" phrases, each of which had a unique meaning and which could not be broken down into meaningful constituent parts. Each such phrase also made extensive use of variation in pitch, rhythm, and melody to communicate information, express emotion and induce emotion in other individuals. "Hmmmmm" is said to have had the following characteristics: it was holistic, manipulative, multi-modal, musical, and mimetic.

Mithen argues that Neanderthal skeletal fossils suggest they had the capacity for vocal communication, but there are few traces of linguistically mediated behavior in other, archaeological evidence. His claim is that they had a complex vocal communication system, but it was a type of Hmmmmm rather than language, inherited from ancestral

hominins common to them and to homo sapiens, who employed holistic communication similar to those of non-human primate calls, which have often been described as musical. These vocalizations of Pliocene ancestral Neanderthals would have been similar and eventually also evolved into human language and music during the course of hominin evolution, most likely as part of the process by which modern Homo sapiens originated in Africa.

Critics note that Mithen's Hmmmmm phrases would have been an indivisible unit that had to be learned, uttered, and understood as a single acoustic sequence, similar to the critique of Arbib's homophrastic proto-words (Tallerman 2008). Each utterance would also have to be stored as a single concept in the hominin's mental lexicon and retrieved from storage to be uttered, and today nothing remotely as complex is stored as a single concept in any known languages. A second criticism is that like all linguistic or proto-linguistic communication, holistic utterances must be culturally transmitted; in other words, they are unlike ape calls, which are essentially innate. So, the mechanism by which Neanderthals learned the highly specific meanings that Mithen attributes to his holistic utterances is left unspecified.

Another criticism is that the study of Neanderthal communication is extremely limited by the lack of soft tissue evidence and the existing fossil evidence (Johansson 2015; Lieberman 2015) so it is especially difficult to evaluate his hypotheses regarding these. Nevertheless, there is growing evidence that, regardless of the possible musicality of their communication systems, Neanderthals and Denisovans possessed language capabilities, given that, as stated by Dediu and Levinson (2018, p. 53): "If one considers all of the cultural skills needed to survive in ecologies from the Arctic to game-poor Mediterranean littorals, it is difficult to argue that Neanderthals lacked complex linguistic codes, capable of communicating about spatial locations, hunting and gathering, fauna and flora, social relations, technologies, and so on," as well as the fact that traces of Neanderthal or Denisovan DNA are present in all modern populations (Chen et al. 2020).

The work of Scott-Phillips (Scott-Phillips et al. 2009; Scott-Phillips and Kirby 2010) centers on his view that primate intelligence is to a significant degree explained by the highly social nature of primate life and explains why humans, who are extremely social even by primate standards, would have evolved particularly sophisticated forms of social cognition, which eventually led to ostensive (declarative) communication. His approach is largely predicated on the results of simulations of such evolution through signaling games. He suggests that early instances of ostensive communication would have been halting ones, but a number of cognitive adaptations that make ostensive communication operate smoothly and efficiently would probably have followed. A parallel development would have been the emergence, through repeated interaction and cultural transmission, of conventional ways in which to use ostensive communication, conventions that would in time become ossified into words, grammars, and the other constituent parts of languages.

The criticisms of language evolution study through simulated signaling games, such as Scott-Phillips's work, center around the difference between artificial agents and natural

ones, whether human or otherwise, and the necessity of incorporating those factors into the model (Liu 2014). However, there now seems to be a growing amount of empirical evidence to support the existence of deixic behavior and even ostensive communication by great apes, which would tend to foreground the role of social cognition within the evolution of language (Scott-Phillips and Heintz 2023).

In summary then, the earliest referential communication demonstrated by displaced reference (the use of alarm calls to alert conspecifics to impending threats) and direct deixis (the use of gestures to indicate desired objects and actions) appears to be related to the so-called "E and G" ("embodied" and "grounded") cognition. Such rudimentary cognition and communication capabilities form the grounds on which more sophisticated referential communication may become possible, whether by ape, Homo antecessor, or artificial intelligence (Borghi and Pecher 2011) as well as the evolutionary possibilities leading toward more socially distributed cognition (Gamble et al. 2011).

The continuing development of shared attention, which has been observed repeatedly in non-human primate communities (Leavens, 2011), is another obvious requisite for more extended "reference" and "relevance" exchanges, with or without speech or language. The evolution of human cooperation beyond small kinship groups over time would depend upon the recognition of identity and reciprocity, which would eventually lead both to the emergence of practices for naming individuals and accounting for actions, both of which are prerequisite for truly human culture (Mullins et al. 2013).

Despite the likelihood that we may never be able to identify what Botha (2012) ironically termed "the God particle" of language evolution and the reality that scientists continue to argue vehemently over the possibilities of a seminal proto-language prior to all known proto-languages and modern languages, there continues to be important new work in a variety of disciplines that can illuminate aspects of how words became "key" to human cognition. The next chapter will deal with the issues of human language: its varied philologies, philosophies, and pragmatics.

References

Arbib, M. A. (2005a). From monkey-like action recognition to human language: An evolutionary framework for neurolinguistics. *Behavioral and Brain Sciences*, *28*(2), 105–124. https://doi.org/10.1017/S0140525X05000038

Arbib, M. A. (2005b). Interweaving protosign and protospeech: Further developments beyond the mirror. *Interaction Studies*, *61*(2), 145–171. https://doi.org/10.1075/is.6.2.02arb

Arbib, M. A. (2017). Toward the language-ready brain: Biological evolution and primate comparisons. *Psychonomic Bulletin & Review*, *24*, 142–150. https://doi.org/10.3758/s13423-016-1098-2

Arnal, L. H., Flinker, A., Kleinschmidt, A., Giraud, A. L., & Poeppel, D. (2015). Human screams occupy a privileged niche in the communication soundscape. *Current Biology*, *25*(15), 2051–2056. https://doi.org/10.1016/j.cub.2015.06.043

Balzeau, A., Gilisse, E., Holloway, R. L., Prima, S., & Grimaud-Hervé, D. (2014). Variations in size, shape and asymmetries of the third frontal convolution in hominids: Paleoneurological implications for hominin evolution and the origin of language. *Journal of Human Evolution*, *76*, 116–128. https://doi.org/10.1016/j.jhevol.2014.06.006

Barton, R. A. (2012). Embodied cognitive evolution and the cerebellum. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1599), 2097–2107. https://doi.org/10.1098/rstb.2012.0112

Benítez-Burraco, A., & Uriagereka, J. (2015). The immune syntax revisited: Opening new windows on language evolution. *Frontiers in Molecular Neuroscience*, *8*, Article 84. https://doi.org/10.3389/fnmol.2015.00084

Bickerton, D. (1990). *Language and species*. University of Chicago Press.

Bickerton, D. (1995). *Language and human behavior*. University of Washington Press.

Bickerton, D. (2014). *More than nature needs*. Harvard University Press.

Borghi, A. M., & Pecher, D. (2011). Introduction to the special topic: Embodied and grounded cognition. *Frontiers in Psychology*, *2*, Article 187. https://doi.org/10.3389/fpsyg.2011.00187

Botha, R. (2008). On modelling prelinguistic evolution in early hominins. *Language & Communication*, *28*(3), 258–275. https://doi.org/10.1016/j.langcom.2007.11.001

Botha, R. (2011). Constraining the arbirariness of exaptationist accounts of the evolution of language. *Lingua*, *121*(9), 1552–1563. https://doi.org/10.1016/j.lingua.2011.05.003

Botha, R. (2012). Protolanguage and the "God particle". *Lingua*, *122*(12), 1308–1324. https://doi.org/10.1016/j.lingua.2012.07.005

Botha, R., & Everaert, M. (2013). Introduction: Evidence and inference in the study of language evolution. In R. Botha & M. Everaert (Eds.), *The evolutionary emergence of language: Evidence and inference* (pp. 1–17). Oxford University Press.

Campbell, L., & Poser, W. J. (2008). *Language classification: History and method*. Cambridge University Press.

Chen, L., Wolf, A. B., Fu, W., Li, L., & Akey, J. M. (2020). Identifying and interpreting apparent Neanderthal ancestry in African individuals. *Cell*, *180*(4), 677–686. https://doi.org/10.1016/j.cell.2020.01.012

Cheney, D. L., & Seyfarth, R. M. (2010). Primate communication and human language: Continuities and discontinuities. In P. Kappeler & J. Silk (Eds.), *Mind the gap* (pp. 283–298). Springer.

Chomsky, N. (2010). Some simple evo devo theses: How true might they be for language? In R. K. Larson, V. Deprez, & H. Yamakido (Eds.), *The evolution of human language* (pp. 45–62). Cambridge University Press.

Christiansen, M. H., & Kirby, S. (2003). Language evolution: The hardest problem in science. In M. H. Christiansen & S. Kirby (Eds.), *Language evolution* (pp. 1–15). Oxford University Press.

Deacon, T. W. (1998). *The symbolic species: The co-evolution of language and the brain*. W. W. Norton & Co.

Dediu, D., & Levinson, S. C. (2018). Neanderthal language revisited: Not only us. *Current Opinion in Behavioral Sciences*, *21*, 49–55. https://doi.org/10.1016/j.cobeha.2018.01.001

Donald, M. (1991). *Origins of the modern mind: Three stages in the evolution of culture and cognition*. Harvard University Press.

Donald, M. (2000). The central role of culture in cognitive evolution: A reflection on the myth of the "isolated mind." In L. Nucci, G. B. Saxe, & E. Turiel (Eds.), *Culture, thought, and development* (pp. 19–38). Lawrence Erlbaum Associates.

Dunbar, R. I. (1995). *Grooming, gossip, and the evolution of language*. Harvard University Press.

Dunbar, R. I. (2017). Group size, vocal grooming and the origins of language. *Psychonomic Bulletin & Review*, *24*, 209–212. https://doi.org/10.3758/s13423-016-1122-6

Falk, D. (2007). Delving into the ancient brain. *Nature*, *450*(7166), 31–32. https://doi.org/10.1038/450031a

Fitch, W. T. (2017). Empirical approaches to the study of language evolution. *Psychonomic Bulletin & Review*, *24*, 3–33. https://doi.org/10.3758/s13423-017-1236-5

Fujita, K. (2009). A prospect for evolutionary adequacy: Merge and the evolution and development of human language. *Biolinguistics*, *3*(2/3), 128–153. https://doi.org/10.5964/bioling.8721

Gamble, C., Gowlett, J., & Dunbar, R. I. (2011). The social brain and the shape of the Palaeolithic. *Cambridge Archaeological Journal*, *21*(1), 115–136. https://doi.org/10.1017/S0959774311000072

Hurford, J. R. (2007). *The origins of meaning: Language in the light of evolution.* Oxford University Press.

Huybregts, M. R. (2017). Phonemic clicks and the mapping asymmetry: How language emerged and speech developed. *Neuroscience & Biobehavioral Reviews*, *81*, 279–294. https://doi.org/10.1016/j.neubiorev.2017.01.041

Johansson, S. (2015). Language abilities in Neanderthals. *Annual Review of Linguistics*, *1*(11), 311–332. https://doi.org/10.1146/annurev-linguist-030514-124945

Konopka, G., & Roberts, T. F. (2016). Insights into the neural and genetic basis of vocal communication. *Cell*, *164*(6), 1269–1276. https://doi.org/10.1016/j.cell.2016.02.039

Lameira, A. R., & Call, J. (2020). Understanding language evolution: Beyond pan-centrism. *BioEssays*, *42*(3), Article 1900102. https://doi.org/10.1002/bies.201900102

Leavens, D. A. (2011). Joint attention: Twelve myths. In A. Seemann (Ed.), *Joint attention: New developments in psychology, philosophy of mind, and social neuroscience* (pp. 43–72). MIT Press.

Lieberman, P. (2015). Language did not spring forth 100,000 years ago. *PLoS Biology*, *13*(2), Article e1002064. https://doi.org/10.1371/journal.pbio.1002064

Lieberman, P. (2016). The evolution of language and thought. *Anthropological Sciences*, *94*, 1–20. https://www.isita-org.com/jass/Contents/2016vol94/Lieberman/26963222.pdf

Liu, H. (2014). Language is more a human-driven system than a semiotic system: Comment on "Modelling language evolution: Examples and predictions" by Tao Gong, Shuai Lan, Menghan Zhang. *Physics of Life Reviews*, *11*(2), 309–310.https://doi.org/10.1016/j.plrev.2013.12.008

Lock, A. (1999). On the recent origin of symbolically-mediated language and its implications for psychological science. In M. C. Corballis & S. E. Lea (Eds.), *The descent of mind: Psychological perspectives on hominid evolution* (pp. 324–356). Oxford University Press.

MacDonald, K., & Roebroeks, W. (2013). Neanderthal linguistic abilities: An alternative view. In R. Botha & M. Everaert (Eds.), *The evolutionary emergence of language: Evidence and inference* (pp. 97–117). Oxford University Press.

Mariën, P., Ackermann, H., Adamaszek, M., Barwood, C. H., Beaton, A., Desmond, J., & Leggio, M. (2014). Consensus paper: Language and the cerebellum: An ongoing enigma. *The Cerebellum*, *13*(3), 386–410. https://doi.org/10.1007/s12311-013-0540-5

Mithen, S. J. (2006). *The singing Neanderthals: The origins of music, language, mind, and body*. Harvard University Press.

Moore, J., Black, J., Hernandez-Aguilar, R., Idani, G. I., Piel, A., & Stewart, F. (2017). Chimpanzee vertebrate consumption: Savanna and forest chimpanzees compared. *Journal of Human Evolution*, *112*, 30–40. https://doi.org/10.1016/j.jhevol.2017.09.004

Mozzi, A., Forni, D., Clerici, M., Pozzoli, U., Mascheretti, S., Guerini, F. R., Riva, S., Bresolin, N., Cagliani, R., & Sironi, M. (2016). The evolutionary history of genes involved in spoken and written language: Beyond FOXP2. *Scientific Reports*, *6*. https://doi.org/10.1038/srep22157

Mullins, D. A., Whitehouse, H., & Atkinson, Q. D. (2013). The role of writing and recordkeeping in the cultural evolution of human cooperation. *Journal of Economic Behavior & Organization*, *90*, S141–S151. https://doi.org/10.1016/j.jebo.2012.12.017

Murphy, E. (2019). No country for Oldowan men: Emerging factors in language evolution. *Frontiers in Psychology*, *10*, Article 1448. https://doi.org/10.3389/fpsyg.2019.01448

Perniss, P., & Vigliocco, G. (2014). The bridge of iconicity: From a world of experience to the experience of language. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), Article 20130300. https://doi.org/10.1098/rstb.2013.0300

Poliva, O. (2015). From where to what: A neuroanatomically based evolutionary model of the emergence of speech in humans. *F1000Research*, *4*(67). https://doi.org/10.12688/f1000research.6175.3

Progovac, L. (2016). A gradualist scenario for language evolution: Precise linguistic reconstruction of early human (and Neandertal) grammars. *Frontiers in Psychology*, *7*, Article 1714. https://doi.org/10.3389/fpsyg.2016.01714

Rizzolatti, G., & Arbib, M. A. (1998). Language within our grasp. *Trends in Neurosciences*, *21*(5), 188–194. https://doi.org/10.1016/S0166-2236(98)01260-0

Schacht, R., & Kramer, K. L. (2019). Are we monogamous? A review of the evolution of pair-bonding in humans and its contemporary variation cross-culturally. *Frontiers in Ecology and Evolution*, *7*, Article 230. https://doi.org/10.3389/fevo.2019.00230

Scott-Phillips, T. C., & Heintz, C. (2023). Animal communication in linguistic and cognitive perspective. *Annual Review of Linguistics*, *9*, 93–11. https://doi.org/10.1146/annurev-linguistics-030421-061233

Scott-Phillips, T. C., & Kirby, S. (2010). Language evolution in the laboratory. *Trends in Cognitive Sciences*, *14*(9), 411–417. https://doi.org/10.1016/j.tics.2010.06.006

Scott-Phillips, T. C., Kirby, S., & Ritchie, G. R. (2009). Signalling signalhood and the emergence of communication. *Cognition*, *113*(2), 226–233. https://doi.org/10.1016/j.cognition.2009.08.009

Sievers, C., & Gruber, T. (2020). Can nonhuman primate signals be arbitrarily meaningful like human words? An affective approach. *Animal Behavior and Cognition*, *7*(2), 140–150. https://doi.org/10.26451/abc.07.02.08.2020

Studdert-Kennedy, M., & Goldstein, L. (2004). Launching language: The gestural origin of discrete infinity. In M. H. Christiansen & S. Kirby (Eds.), *The evolution of language* (pp. 235–254). Oxford University Press.

Szilágyi, A., Kovács, V. P., Czárán, T., & Szathmáry, E. (2023). Evolutionary ecology of language origins through confrontational scavenging. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *378*(1872), Article 20210411. https://doi.org/10.1098/rstb.2021.0411

Tallerman, M. (2006). Abracadabra! Early hominin for 'I think my humming's out of tune with the rest of the world!' *Cambridge Archaeological Journal*, *16*(1), 106–107. https://doi.org/10.1017/S0959774306000060

Tallerman, M. (2007). Did our ancestors speak a holistic protolanguage? *Lingua*, *117*(3), 579–604. https://doi.org/10.1016/j.lingua.2005.05.004

Tallerman, M. (2008). Holophrastic protolanguage: Planning, processing, storage, and retrieval. *Interaction Studies*, *9*(1), 84–99. https://doi.org/10.1075/is.9.1.07tal

Tiwary, B. K. (2020). The cognitive and speech genes are jointly shaped by both positive and relaxed selection in the human lineage. *Genomics*, *112*(5), 2922–2927. https://doi.org/10.1016/j.ygeno.2020.05.006

Wacewicz, S., & Zywiczynski, P. (2017). The multimodal origins of linguistic communication. *Language & Communication*, *54*, 1–8. https://doi.org/10.1016/j.langcom.2016.10.001

Whiten, A. (2021). The burgeoning reach of animal culture. *Science*, *372*(6537). https://doi.org/10.1126/science.abe6514

Whiten, A., Harrison, R. A., McGuigan, N., Vale, G. L., & Watson, S. K. (2022). Collective knowledge and the dynamics of culture in chimpanzees. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1843), Article 20200321. https://doi.org/10.1098/rstb.2020.0321

Wildman, D. E., Uddin, M., Liu, G., Grossman, L. I., & Goodman, M. (2003). Implications of natural selection in shaping 99.4% nonsynonymous DNA identity between humans and chimpanzees: Enlarging genus Homo. *PNAS*, *100*(12), 7181–7188. https://doi.org/10.1073/pnas.1232172100

Zlatev, J., Persson, T., & Gärdenfors, P. (2005). Bodily mimesis as "the missing link" in human cognitive evolution. *Lund University Cognitive Studies*, *121*, 1–45. https://lucris.lub.lu.se/ws/files/4680259/625029.pdf

Zlatev, J., Żywiczyński, P., & Wacewicz, S. (2020). Pantomime as the original human-specific communicative system. *Journal of Language Evolution*, *5*(2), 156–174. https://doi.org/10.1093/jole/lzaa006

Philologies, Philosophies, Pragmatics

4

Abstract

This chapter discusses past and present work in philology (the historical study of language) and pragmatics (the naturally occurring social usage of language), as well as some of the more important contributions of philosophers of language and linguists as they may relate to key words.

As explained in Chap. 3, it is still unknown as to whether a single seminal proto-language might have developed in one small group millennia ago and diffused more broadly as its speakers migrated over time, or whether the very first proto-languages were independently or repeatedly invented by several geographically separated small bands of humans, or whether these processes were much more complex and diverse, involving gestural or vocal initiatives toward simple symbolic communication that eventually resulted in a more general "proto-language" that could be understood by contacts among these proto-speech communities. It is, however, a reasonable assumption that the beginning of these processes took place roughly 100,000 years before the resulting development of the first identified proto-languages studied by today's historical linguists (Arbib 2012).

4.1 Philology and Plato

As Goody and Watt (1963) pointed out, in illiterate societies a word necessarily takes its meaning in the context of its usage, as there is no way to accumulate meaning over time except in the individual memories of a speaker and listener. This presumably would apply to an even greater extent to the gradual development over time of a consistent set of proto-words that could eventually comprise such proto-languages.

B. V. V. Martens, *Keywords In and Out of Context*, Synthesis Lectures on Information Concepts, Retrieval, and Services, https://doi.org/10.1007/978-3-031-32530-4_4

Proto-Indo-European, the ancestral form of the Indo-European language family, is the most studied of the known proto-languages and can be traced back to 3500 BC through the evidence of incised clay seals unearthed in Mesopotamia. Proto-Sino-Tibetan, the ancestral form of the largest and most complex group of languages, including Chinese, has not yet been firmly dated, but the large numbers of pottery vessels with incised signs found at the Dawenkou site near the Yellow River in China may indicate that 3500 BC is a reasonable estimate there as well. Proto-Afro-Asiatic, identified through what are believed to be makers marks on pots unearthed at Abydos in Egypt, can be traced back to 3400 BC (Mitchell 1999) but is likely to be earlier as well.

However, Dematté (2010) rightly observes that print-acculturated scholars may tend to conflate the development of language with the development of those writing systems that provide the archaeological evidence for the existence of various languages, but the writing systems themselves were invented for other purposes, such as accounting, astronomy, divination, meteorology, and taxation, and only gradually began to serve more general linguistic purposes. Surviving samples of early writing systems, therefore, cannot provide conclusive evidence for dating the origins of spoken languages that they may document, and many spoken languages have developed and flourished without associated writing systems. In the case of proto-Indo-European, for instance, historical linguistic research has found that the five major Indo-European language subfamilies (Celtic, Germanic, Italic, Balto-Slavic, and Indo-Iranian) were diverging from the Indo-European proto-language as early as 6000 BC (Bouckaert et al. 2012), indicating an even earlier origin for the parent proto-language.

Interest in protolinguistics and the historical etymology of words also dates back several millennia, though the techniques employed in their study were far more similar to folklore and discursive practices intended to facilitate the formation of in-group identity through the creation of myths and genealogies rather than to any modern linguistic methods (Sluiter et al. 2015).

The most celebrated example of this early interest is found in the *Cratylus*, a Platonic dialogue from the fourth century BC in which Greek philosopher Socrates is shown debating the purpose of words in both practical and philosophical terms that still resonate today, and which provides a marked contrast to the much more limited linguistic views of his sophist and stoic philosophical contemporaries. The wide-ranging nature of the dialogue even makes it possible to find in it precursors to philosophical views ranging from that of Quine (1960) on the interdeterminacy of reference to that of Goodman on the irrealism of reference (1975) to that of Davidson (1979) on the inscrutability of reference. Plato's Socrates brings all this to a brilliantly inconclusive and remarkably contemporary conclusion ("And when you have found the truth, come and tell me") as, despite all our frustration with words as tools, it is indisputable that we have almost no others with which to work.

Accordingly, commentators have continued to be fascinated by this dialogue, in which Plato cleverly weaves references to his own literary and philosophical training, attempts

to analyze differences between word etymologies that seem to have philosophical foundations and those that do not, examines the scientific and social constructions of reality made possible by language, and explores the impossibilities of determining meaning with or without words. However, modern scholars have puzzled over the lack of intellectual sophistication shown by the etymological section of this dialogue as compared to the other sections, not realizing that it accurately represents both earlier and current thinking about word origins and meanings in Plato's time, which was very unlike how we think about these today (Ademollo 2011).

Later classical thinkers, the first philologists, attempted to create order from a growing number of texts in different dialects as their culture was emerging from orality into literacy. Their problem was how to choose among word forms: which could be considered true, or correct, or usual, or preferred. They decided upon four principles: *natura* (nature), *consuetudo* (custom or usage), *analogia* (analogy), and *auctoritas* (authority). These principles conveyed the notions that the nature of any word is shown by its meaning, something that is expressed in normal usage, while analogy in language shows the patterns that apply to numerous words of similar types, and authority determines whose word usage is considered more important than that of others (Zetzel 2019). These original principles were, however, applied and ranked in different ways over time and in differing contexts and circumstances through antiquity, into the medieval period, and even into modern times.

4.2 Historical Linguistics

More modern efforts at historical linguistics also demonstrate the particular preoccupations and proclivities of the researcher's own era. Daston and Most (2015, p. 384) note that "well into the nineteenth century in Europe—indeed, especially in the nineteenth century—philology not only counted as a science; it was *the* science, the model of the highest form of knowledge. The discoveries of the philologists, whether concerning the authorship of the *Iliad*, the decipherment of hieroglyphics, or the ancestry of modern languages, counted alongside those of the chemists and the physicists as among the most spectacular of the age and their methods as among the most rigorous. Big Science (including the phrase) began not with the science-based industries of coal tar derivatives or optical glass but with the grand projects of the philologists, such as Theodor Mommsen's *Corpus Inscriptionum Latinarum*."

Building on the ideas of his contemporary Charles de Brosses, French encyclopedist Denis Diderot argued in 1760 that an original "universal core vocabulary" developed due to the commonalities of human physiology that facilitated the creation of certain sounds that became the initial syllables, and only much later became differentiated into early languages (Clark-Evans 1993). In Germany, romanticist Johann Gottfried Herder's 1770 *Treatise on the Origin of Language* laid the foundations for both the field of comparative

linguistics and linguistic nationalism, arguing that every people's language and culture became so naturally interwoven as to constitute a unique folk identity (Wolff 2016).

In Britain, there was an intense intellectual interest in antiquity and its antecedents, often from a religious perspective (Hiebert 2007). Sharon Turner published several papers in the first issue of the *Transactions of the Royal Society of Literature* comparing specific words from a wide variety of global languages, ranging from Hebrew to Huron, and concluded that humanity's original (or "antediluvian") words included terms for "mother," "father," and "water," but that the many obvious variations in their subsequent linguistic forms could be attributed to postdiluvian separation into different populations after the destruction of the Tower of Babel (Turner 1827). His etymological investigation was, of course, only one of many addressing these popular questions during this period of intense intellectual interest in language and its antique precedents by figures such as Condillac in France, Grimm in Germany, and Locke in England (Eddy 2011).

Investigations into the origin of language became more controversial in the aftermath of the publication of *The Origin of Species* in 1859, with the dawning of the realization that those evolutionary processes described in Darwin's work might extend beyond the book's contents to include its readers as well. Not coincidentally, the *Société de Linguistique de Paris* in France banned any discussion of the topic of the origins of language in 1864 (Andresen 2013). At the 1870 meeting of the American Philological Association in Rochester, New York, a similar ban was discussed, and linguistic luminary W. D. Whitney of Yale, editor of the *Century Dictionary*, despite defending the topic as a legitimate one for science, would nevertheless continue to confine his attention to more tractable word derivation investigations than a truly scientific search for a "first word" would entail (2013, p. 1).

Darwin himself propounded in *The Descent of Man* (1871, pp. 59–61) that "The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously the same. But we can trace the origin of many words further back than in the case of species, for we can perceive that they have arisen from the imitation of various sounds, as in alliterative poetry. …The survival or preservation of certain favoured words in the struggle for existence is natural selection." Even the London Philological Society banned all such discussions in 1873, despite or perhaps because of the fact that their massive New English Dictionary on Historical Principles project to collect and define all English words was underway and would eventually become the *Oxford English Dictionary*. German linguist August Schleicher was one of the few who continued his empirical efforts in philology (Taub 1993). However, while proposing that contemporary languages had inherited traits which could be traced back to their original proto-languages, he proposed that these were subject to a sort of Hegelian dialectic: after a period of development and growth, language repeatedly encounters various conceptual challenges that are antithetical to the original linguistic "thesis," and results in newly "synthesized" languages over time. Although this idea was

never widely adopted, Schleicher's efforts to trace existing languages to their proto-Indo-European roots did lead to a more systematized genealogical approach than those used by earlier philologists.

Besides wide recognition as a founder of structural linguistics, Ferdinand de Saussure is also celebrated for his sharp differentiation between the diachronic approach (which compares language at two different points in time and is used for comparative and historical linguistics) and the synchronic approach (which focuses on language at one particular point in time and is used for modern linguistics) and which he expressed as "the opposition between the two viewpoints—synchronic and diachronic—is absolute and allows no compromise," (de Saussure 2011, p. 83). Until very recently, his dictum discouraged most linguists from exploring other forms of symbolic communication not expressed in modern language. Interestingly, however, de Saussure's own thinking may not be fully reflected in his posthumously edited and published *Course in General Linguistics*, as his extensive surviving manuscript materials show that he was also well aware that the theoretical language system he modeled was only one among many important sign systems (Stawarska 2015). Indeed, some of his own historical work on the Hittite language, reconstructing laryngeal consonants missing in certain words, disparaged at the time by other linguists due to the fact that these phonemes were unknown in other surviving Indo-European languages, was only vindicated some 50 years later with the discovery of Hittite texts preserved on cuneiform tablets in Anatolia (Sturtevant 1952).

Although historical and comparative linguistics have had numerous successes since de Saussure's time in building empirically based genealogies of languages (Jager and Wichmann 2016), approximately 10,000 years seems to represent the maximum extent of that time horizon. Newer efforts focus on correlational studies (Ladd et al. 2015) and on computational linguistics, both of which rely largely on statistical methods and processing immense volumes of data, and attempt more ambitious investigations into the past.

For example, Pagel (2017) and his collaborators (Pagel et al. 2013) have taken an evolutionary, statistical linguistic approach to the search for what might be considered original "key words." Their analysis of the World Etymology Project corpus has proposed that a very small group of reconstructed proto-words and their cognates in each of the Altaic, Chukchi-Kamchatkan, Dravidian, Eskimo, Indo-European, Kartvelian, and Uralic language families, unlike the vast majority of words that have emerged or evolved over time, could be traced back 15,000 years to an unknown ancestral Eurasian language. Their list of 23 "ultraconserved words" (Thou, I, Not, That, We, To give, Who, This, What, Man/male, Ye, Old, Mother, To hear, Hand, Fire, To pull, Black, To flow, Bark, Ashes, To spit, Worm) has attracted criticism on the grounds that such shorter words tend to appear related more frequently by chance alone (Mahowald and Gibson 2013). There has not been general acceptance of this finding by mainstream linguists, who also tend to look askance at other statistically oriented models not supported by empirical linguistic data. These critics, such as Perslsvaig and Lewis (2015), argue that the so-called reinvention of historical linguistics through the use of phylogenetic and phylogeographical

analysis, treating cognates like genes and conceptualizing the spread of languages in terms of the diffusion of viruses, produces incoherent results, contradicted by the empirical record, since clearly languages do not evolve like biological species and do not spread like viruses.

4.3 Reference and Relevance

There are two primary problems for spoken language: the philosophical problem of reference, which is the question of how words relate to objects in the world, while the practical problem of relevance is the question of how speakers and listeners come to understand what one another may mean.

As noted above, the problem of reference was identified by Plato in the *Cratylus* and has continued to fascinate many other philosophers, from Russell (1910) to Wittgenstein (1958) to Searle (1969) to Putnam (1975): how is it possible for our shared verbal expressions to refer to objects in the world as well as to objects that do not exist? Moreover, how is it possible that these speech acts have become so taken for granted that even very small children are expected to learn to perform them with spontaneity and rapidity? Bruner aptly referred to this process as the "routes to reference" (1998). Reference (or, in the case of non-verbal adults or pre-verbal infants, declarative pointing) is such a fundamental intellectual skill for human beings that inability to perform it is considered a serious learning disability.

In addition to de Saussure in Switzerland, those philosophers of the late nineteenth and early twentieth centuries who possessed a marked scientific interest in logic and linguistics, notably Mill and Peirce in the United States, Ogden, Quine, Richards, and Russell in Great Britain, and Wittgenstein in Germany, all tended to focus on developing more abstract systems, and individual words themselves were valued primarily as compositional elements within the contexts of these theoretical models.

Ferdinand De Sassure's legacy to linguistics, which he termed *semiology*, is best known for its formal approach to the *linguistic sign*, which is comprised of the signifier (the *signifier* (*signifiant*), and the *signified* (*signifié*). Unlike Peircean semiotics discussed in Chap. 2, which can also be applied to a non-linguistic environment, de Sassurean semiology's arbitrary relation between words and their meanings is almost totally constrained by language conventions within a particular speech community. According to de Saussure, language is a meaningful system of social discourse that expresses concepts through a dyadic relation to particular chosen sound patterns at any given time, and though there is often a *referent* beyond the dyad, the linguist need not take that into account, nor are the individual utterances of a speaker (*parole*) of more importance than their overall contribution to the system of discourse (*langue*). Semiology, therefore, is well suited to the study of a closed system of signs, and de Saussure's seminal *Course in General Linguistics* has had an enduring impact on linguistic research in general (Seuren 2016).

Early theorists focused on the so-called "formal language" in which logic is studied, while later theorists became more interested in the "natural language" in which human beings converse. For instance, in his seminal philosophical article, "On Denoting" (1905), Russell concluded that some words within a sentence, while contributing to its meaning, may not have meaning in isolation, similar to symbols that are meaningful only in specific contexts in logic.

In his *Theory of Language*, the German linguist Bühler (1934; 2011) developed his "Organon" model, positing a communicative triangle (a state of affairs or objects that are symbolically represented through language by the expressions of some speaker and the resulting significance to some listener) model, which may be considered as both a successor to Peirce's semiotic model and a precursor to Shannon–Weaver's decoding model.

The Soviet linguist *Vološinov* challenged both Peirce and de Saussure as he focused attention on words as ideological signs: that is, the relationship of language and ideologies as these are constructed through dialogic communication. His *Marxism and the Philosophy of Language* (1973), although published in Russian in the 1930s was not translated into English until 40 years after his death.

Much better known in the Western world, Wittgenstein worked within the tradition of formal language and logic in his early works, but his interests later shifted to natural language, with its "language games" in which the problems of determining meaning become central in his *Philosophical Investigations* (1958).

As the field of contemporary linguistics began to separate from that of the philosophy of language after the Second World War, linguists such as Firth (1957, p. 11) began to take a more analytic approach, with his aphorism "you shall know a word by the company it keeps," indicating that the immediate situational context of the word is critical to determining meaning. This analytic approach, both quantitative and contextual, was foundational to the emerging field of corpus linguistics, in which large collections of texts in specific languages are mined to discover how specific words were used in context. Other important work in contemporary linguistics include that of Empson on word ambiguity (1947), that of Sinclair on word collocation (1991), and that of Stubbs on lexical semantics (2001).

The practical problem of relevance (how speakers and listeners come to understand each other's meanings) has been addressed by Grice (1975), who proposed that the analysis of the meaning of signs (that is, words) should incorporate methods to deal with statements regarding what a particular speaker or writer meant by a particular sign on a particular occasion, which might well vary from the standard meaning of that sign. He also identified certain cooperative principles that are often unconsciously employed by speakers during conversational exchanges. These include optimal principles for the quantity of information, quality of information, relevance of information, and presentation of information.

These Gricean "maxims" regarding listener expectations regarding speaker content and communication were further developed by Sperber and Wilson (1986, 2012) in their work on relevance theory, which focused attention on the effects of context on comprehension. This becomes especially critical in conversational exchanges, as mental lexicons are necessarily unique to each individual, and words may accordingly be associated with differing concepts as well. Word meanings, therefore, are "underspecified" until additional context is provided. Words also differ in terms of the amount of encoded conceptual or procedural content they may carry within sentences: "dog" and "red" are perceived as content words while "but" and "also" are perceived as discourse connectives (Blakemore 1987). However, Carston (2016) has suggested that that most "content" words encode not only a concept but also a procedure for ad hoc concept construction. Relevance theorists have argued for some time that the interpretation of every "content" word is fine-tuned in context so it is possible that the processing of a "content" word might automatically activate such a fine-tuning procedure.

Research into conversation analysis has shown that word meanings are also very often negotiated and renegotiated during the course of an exchange, as the participants attempt to understand one another. This is a complicated interaction, even when the conversational partners are members of the same speech community, particularly as there are more concepts than words to express them, and words often contain multiple concepts that may need additional words to express the speaker's intent (Sperber and Wilson 2012). A related phenomenon is that of lexical entrainment, identified by Brennan and Clark (1996) as the process of developing "conceptual pacts" during an exchange during which one participant adopts the same wording as the other participant in order to facilitate the attainment of conceptual clarity.

This problem of meaning is exacerbated when the communicative "partnership" becomes more taxing: for example, when translation is necessary (He et al. 2016) or when the exchange of communicative acts is separated by space or time, as in a long-distance phone call over a faulty connection (Schoenenberg et al. 2014). However, research on the well-known "cocktail party" phenomenon of divided attention among multiple ongoing conversations first identified by Cherry (1953) and later investigated by Wood and Cowan (1995) has shown that words perceived as "key" (such as hearing one's name spoken in an adjacent conversation) tend to shift one's attention to that away from the immediate conversation.

Understanding the problem of relevance also requires the empirical investigation of naturally occurring social interaction (Moore 2013), which is the focus of the field of language pragmatics. Quine (1987, p. 5) famously said, "Each of us learns his language by observing other people's verbal behavior and having his own faltering verbal behavior observed and reinforced or corrected by others. We depend strictly on overt behavior in observable situations. As long as our command of our language fits all external checkpoints, where our utterance or our reaction to someone's utterance can be appraised in

the light of some shared situation, so long all is well. Our mental life between checkpoints is indifferent to our rating as a master of the language. There is nothing in linguistic meaning, then, beyond what is to be gleaned from overt behavior in observable circumstances."

4.4 The Problems of Pragmatics

Nevertheless, despite Quine, there must be some unobservables at work in the acquisition of words. The cognitive theory of language acquisition was proposed by Jean Piaget in 1926 (2001). He believed that language is another manifestation of the symbolic functions that children between the ages of two and three begin to develop as part of their general cognitive development. Thus, he proposed that children's use of language reflects how they think about the world, and their vocabulary expands accordingly as they acquire new concepts.

Chomsky (1965) claimed that there is a "Language Acquisition Device" that children possess at birth, providing them with a set of universal principles of language that are triggered by minimal language experience. Slobin and Bowerman (2007) later modified this position, arguing that rather than possessing substantive rules of language at birth, children are born with a set of procedures and inferences that allow them to learn any language in the world. However, there is increasing evidence that what allows rapid language learning in children is a set of generalized learning competencies that can be applied to a variety of domains (Dąbrowska 2015). The interactionist or social theories of language suggest that language exists for the purpose of communication and can therefore be learned only in the context of interpersonal interaction (Lytle and Kuhl 2017). According to this theoretical viewpoint, the language behaviors of adults, especially the infant-directed speech often termed "motherese," are critical to this process.

Wierzbicka's (2021) work on the so-called "semantic primitives" lends some support to several of these theories, as her comparison of multiple languages has found that they contain common conceptual frameworks (such as up/down) but that these frameworks find expression in highly diverse grammars and vocabularies that are specific to each language or language family, indicating that while certain modes of perception may be innately human, shared modes of expression are culturally specific.

Xu and Tenenbaum (2007) have proposed a Bayesian model of word learning that claims both children and adults quickly infer the meaning of a word based not on a process of deductive hypothesis elimination or by associative learning, but rather by integrating prior knowledge about plausible word meanings with the statistical structure of the observed examples.

Regardless of the specific underlying explanation, the initial stage of this process is prenatal. The near-term fetus has a mature auditory system that can detect and respond to

sound (Voegtline et al. 2013). Fetuses actively attend to their mothers' voices during pregnancy, which provides a continuous source of information about the mother's language. A newborn can recognize the voice of its mother and prefers this to unfamiliar voices and languages. Prenatal exposure to language can also enable the newborn to recognize the basic structures of that language (Hepper 2015).

Communicating with a newborn usually involves special infant-directed speech ("motherese") that employs a singsong vocal style in a higher than usual register, exaggerated phrasing, short, simple words, and modified word order, coupled with intimate, intensive interactions with the child (Parish-Morris et al. 2013). This development of shared attention, involving gaze direction, demonstrative and deictic gesturing, and continuing vocal exchanges between caregiver and infant, helps to orient the child toward recognition of the relationships among different objects in the world and the public words for these objects. Even during the prelinguistic states of vocal production, termed "babbling," this feedback is critical in helping the infant acquire preliminary skills necessary for eventual word production (Goldstein and Schwade 2008; Gros-Louis et al. 2008). One of the very earliest words that infants reliably recognize is the sound patterns of their own names (Mandel et al. 1995).

A cross-cultural study has shown that the first five words in an infant's vocabulary tend to be these other "key" terms: Mommy, Daddy, hi, bye, and uhoh (Tardif et al. 2008). Further cross-cultural research (Waxman et al. 2013) suggests that infants find it easier to identify and name objects (nouns) than actions (verbs) regardless of whether the specific language environment is noun-friendly (English, French) or verb-friendly (Japanese, Korean).

As Saint-Georges et al. (2013) note, clearly "motherese" is helpful in the development of the child's cognition and communication skills, given that 18-month-olds are slower to interpret target words (i.e., familiar object names) in isolation, and putting target words in sentence-final positions may help them segment the linguistic stream. Children themselves help in this early process by both pointing to and naming objects to clarify and modify their meaning (Cartmill et al. 2014).

By age 5, children can recognize 10,000 words in spoken language, though they may not be able to produce all of these correctly themselves (Law et al. 2016). This process is particularly important because words learned earlier in life have been shown to be easier to access than those learned later (Elsherif et al. 2023).This so-called "age-of-acquisition effect" will also have life-long impact in terms of word production, recognition, and retention (Brysbaert and Ellis 2015; Cuetos et al. 2017).

By adulthood, the average lexicon size is estimated to be approximately 50,000 words (Aitchison 2012). Almost all of these words are acquired through hearing or seeing their use in context, either informally through conversation or more formally through education (Sternberg 1987). An associated process is the development of the individual's mental grammar, which contains the linguistic rules by which the lexemes are to be combined to make understandable sentences.

Each individual's mental lexicon is also made up of two integrated, interacting components: an internal "dictionary" of lexemes and their formal meanings, and the internal "encyclopedia" of the connections between formal meanings and the expressive, expansive network of concepts that associate these with reality (Begby 2016). The importance of this dual processing is that the "dictionary" allows fast retrieval of specific words and their denotations, while the "encyclopedia" allows more gradual comprehension of any additional connotations and implications. Finding the "right words" for any given situation, therefore, is a complex though opaque mental operation, though in routine interactions it is performed with great efficiency. Early communicative practices also continue to influence adult use of words. Moore (2013) found that when people do not know the correct word for an object, they tend to utilize the same "referential practices" as do young children: they try other possible names, they point if the object is present, or they attempt to create some kind of iconic gesture or depiction if the object is not present.

4.5 Expanding the World of Words

The "right words" are also key to much of LIS research. Taylor's seminal question-negotiation model (1968) is predicated on a "visceral need" which can only be satisfied by finding the right words in which to couch the request for information. The Belkin et al. anomalous state of knowledge (ASK) model (1982) requires "linguistic, pragmatic transformations" into a statement expressing a request for information, while the "exploration" stage of Kuhltau's Information Search Process (ISP) model (1993) similarly demands identification of the right words in order to move on to the more satisfying stages of query formulation and information collection.

All of this seminal work is founded on what is generally known as the conduit metaphor (Reddy 1979), which is a figurative expression often used in describing communication as a process in which speakers or writers insert their ideas, feelings, concepts, etc. into containers such as words, sentences, poems, books, etc. whose contents are then extracted by listeners or readers. Day (2000) observes that, while loosely related to the information processing paradigm of cognitive science, the conduit metaphor is not theoretically or empirically grounded in contemporary cognitive research and it has been utilized in LIS and many other fields without a continuing examination of its origins and outcomes. It can, however, be truthfully said that the conduit metaphor itself would not be able to function without the use of words.

On an even broader level, "key" words can serve to express attention-focusing concepts within different cultures. Much work in anthropology over the past century has been devoted to investigating whether and to what extent language may constrain thought. According to Lucy (1997), empirical approaches to this issue of the so-called "linguistic relativity" are classified into three types: structure-centered approaches that begin with language differences and ask about their implications for thought; domain-centered

approaches that begin with experienced reality and ask how different languages encode it; and behavior-centered approaches that begin with some practical concern and seek an explanation in language.

Among the seminal structure-centered approaches are those of anthropologist Franz Boas, credited with the seminal insight, based on his work with Native American languages, that, contrary to general belief among linguists at the time, there is no "ideal" language toward which all languages are in various stages of evolving, and that underestimation of non-Western languages was due to the imposition by Western linguists and anthropologists of their own categories and concepts on these rather than an empirical investigation of the languages themselves; Edward Sapir, who emphasized the importance of the categories that comprise the formal structure underlying any linguistic system, and argued that these varying formal structures influence the meaning value of their linguistic elements, leading the language speakers to interpret events within the systematically coherent construal of reality imposed by their language; and that of Benjamin Whorf, who recognized that large-scale structural meanings interrelate semantically, requiring the study of the specific grammar as a whole (Lucy 2016, pp. 488–489).

Ironically, the linguistic contribution for which Boas, Sapir, and Whorf are best known, the so-called "Whorf-Sapir hypothesis" that language constrains the thoughts available to speakers of any particular language, exemplified by the supposed multiple Inuit terms for "snow," was not in fact developed by any of the three; language behaviorists Lennehen and Brown synthesized some of the elements in their predecessors' writings in order to come up with a version they believe could be tested empirically (Pavlenko 2016). The "Whorf-Sapir hypothesis" has since been criticized for its implicit emphasis on monolingualism as a cultural norm, which is neither historically nor ethnographically sound, for its explicit emphasis on general laboratory tests of concepts such as color rather than fieldwork with specific languages, and for the various misconceptions its popularization has created about particular languages (Martin 1986).

There has also been substantial work on prejudice in language, which has been found to be nearly universal as native speakers either consciously or unconsciously denigrate those perceived to be outsiders with xenophobic terms (Green 1996).

An important international extension of the work of cultural theorist Raymond Williams (1976) on "keywords" in English (which was itself an expansion of his seminal 1958 book on the cultural importance of specific key words) and which continues to enjoy new revisions as well as new applications (Hart et al. 2005), and to inspire new approaches to analyzing English, technological (Scott 1997) as well as theoretical (Highmore 2022), is that of Polish linguist Anna Wierzbicka (1997, 2014) who warns against the hazards of theorizing from a single dominant language such as English in her documentation of the variety of "key words" as perceived by a variety of linguistic communities, including Japanese and Russian, emphasizing the dangers of monolingualism, while stressing the cross-cultural understanding of core values that can be achieved by better understanding the "key concepts" embedded in such "key words."

Risager (2020) dissects the concept of what has been called both linguaculture and languaculture, which is comprised of three complex and highly inter-related cultural/linguistic dimensions: the semantic-pragmatic dimension, the poetic dimension, and the identity dimension. This area of study analyzes the mutual influence of culture on language and language on culture, identifying four critical elements (linguistic flows, which are codes such as English or Swahili; linguacultura flows (meanings related to first language use of particular languages); discursive flows (meanings not necessarily related to particular languages); and other cultural flows (non-language meanings, including semiotic meanings in art and music). As global migration affects the language composition in various geographic settings, an understanding of how people learn to interact within different linguacultures becomes increasingly important.

Recent investigations have shown that multilingualism can destabilize entrenched habits of thought through the availability of new concepts through language learning (Evans 2010). Although it is now clear through recent research in brain scanning that cognition is individually embodied, so-called social cognition is necessarily reliant on communication with others through concepts shared with words, so the brain is not nearly as constrained by language as it is freed by it.

Clearly "key words" have a much richer sociocultural history than commonly recognized, and their ongoing influence is still not well understood. The next chapter will review some of the earliest history of keywords as preserved in rituals and religions, some of the methods for their communication, and some of their ongoing impacts.

References

Ademollo, F. (2011). *Plato's Cratylus: A commentary*. Cambridge University Press.

Aitchison, J. (2012). *Words in the mind: An introduction to the mental lexicon*. John Wiley & Sons.

Andresen, J. T. (2013). *Linguistics and evolution: A developmental approach*. Cambridge University Press.

Arbib, M. A. (2012). Evolutionary parallels between language and tool use. In T. C. Scott-Phillips, M. Tamariz, E. A. Cartmill, & J. R. Hurford (Eds.), *The evolution of language* (pp. 3–10). World Scientific. https://doi.org/10.1142/8449

Begby, E. (2016). Deranging the mental lexicon. *Inquiry*, *59*(1), 33–33. https://doi.org/10.1080/0020174X.2015.1115276

Belkin, N. J., Oddy, R. N., & Brooks, H. M. (1982). ASK for information retrieval: Part I. Background and theory. *Journal of Documentation*, *38*(2), 61–71. https://doi.org/10.1108/eb026722

Blakemore, D. (1987). *Semantic constraints on relevance*. Blackwell Publishing.

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., Gray, R. D., Suchard, M. A., & Atkinson, Q. D. (2012). Mapping the origins and expansion of the Indo-European language family. *Science*, *337*(6907), 957–960. https://doi.org/10.1126/science.1219669

Brennan, S., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482–1493. https://doi.org/10.1037/0278-7393.22.6.1482

Bruner, J. S. (1998). Routes to reference. *Pragmatics and Cognition*, *6*(1/2), 209–227. https://doi.org/10.1075/pc.6.1-2.11bru

Brysbaert, M., & Ellis, A. W. (2015). Aphasia and age of acquisition: Are early-learned words more resilient? *Aphasiology*, *30*(11), 1240–1263. https://doi.org/10.1080/02687038.2015.1106439

Bühler, K. (2011). *Theory of language: The representational function of language* (D. F. Goodwin, Trans.). John Benjamins Publishing.

Carston, R. (2016). The heterogeneity of procedural meaning. *Lingua*, *175–176*, 154–166. https://doi.org/10.1016/j.lingua.2015.12.010

Cartmill, E. A., Hunsicker, D., & Goldin-Meadow, S. (2014). Pointing and naming are not redundant: Children use gesture to modify nouns before they modify nouns in speech. *Developmental Psychology*, *50*(6), 1660–1666. https://doi.org/10.1037/a0036003

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, *25*(5), 975–979. https://doi.org/10.1121/1.1907229

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Clark-Evans, C. (1993). Charles de Brosses and Diderot: Eighteenth-century arguments concerning primitive language, particular natural languages and a national language. *History of European Ideas*, *16*(1–3), 183–188. https://doi.org/10.1016/S0191-6599(05)80117-2

Cuetos, F., Arce, N., Martinez, C., & Ellis, A. W. (2017). Word recognition in Alzheimer's disease: Effects of semantic degeneration. *Journal of Neuropsychology*, *11*(1), 26–39. https://doi.org/10.1111/jnp.12077

Dąbrowska, E. (2015). What exactly is Universal Grammar, and has anyone seen it? *Frontiers in Psychology*, 6, Article 852. https://doi.org/10.3389/fpsyg.2015.00852

Darwin, C. (1871). *The descent of man, and selection in relation to sex*. John Murray.

Daston, L., & Most, G. W. (2015). History of science and history of philologies. *Isis*, *106*(2), 378–390. https://doi.org/10.1086/681980

Davidson, D. (1979). The inscrutability of reference. *Southwestern Journal of Philosophy*, *10*(2), 7–19. http://www.jstor.org/stable/43155338

Day, R. E. (2000). The "conduit metaphor" and the nature and politics of information studies. *Journal of the American Society for Information Science and Technology*, *51*(9), 805–811. https://doi.org/10.1002/(SICI)1097-4571(2000)51:9<805::AID-ASI30>3.0.CO;2-C

de Saussure, F. (2011). *Course in general linguistics* (P. Meisel & H. Saussy, Eds.; W. Baskin, Trans.). Columbia University Press.

Dematté, P. (2010). The origins of Chinese writing: The Neolithic evidence. *Cambridge Archaeological Journal*, *20*(2), 211–228. https://doi.org/10.1017/S0959774310000247

Eddy, M. D. (2011). The line of reason: Hugh Blair, spatiality and the progressive structure of language. *Notes and Records of the Royal Society*, *65*(1), 9–24. https://doi.org/10.1098/rsnr.2010.0098

Elsherif, M. M., Preece, E., & Catling, J. C. (2023). Age-of-acquisition effects: A literature review. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* Advance online publication. https://doi.org/10.1037/xlm0001215

Empson, W. (1947). *Seven types of ambiguity*. New Directions Publishing.

Evans, N. (2010). *Dying words: Endangered languages and what they have to tell us*. Wiley-Blackwell.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. In J. R. Firth (Ed.), *Studies in linguistic analysis* (pp. 1–32). Basil Blackwell.

Goldstein, M. H., & Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. *Psychological Science*, *19*(5), 515–523. https://doi.org/10.1111/j.1467-9280.2008.02117.x

Goodman, N. (1975). Words, works, worlds. *Erkenntnis*, *9*(1), 57–73. https://www.jstor.org/stable/20010456

Goody, J., & Watt, I. (1963). The consequences of literacy. *Comparative Studies in Society and History*, *5*(3), 304–345. https://www.jstor.org/stable/177651

Green, J. (1996). *Words apart: The language of prejudice.* Kyle Cathie Ltd.

Grice, H. P. (1975). Logic and conversation. In D. J. Levitin (Ed.), *Speech acts* (pp. 41–58). Brill.

Gros-Louis, J., West, M. J., Goldstein, M. H., & King, A. P. (2008). Mothers provide differential feedback to infants' prelinguistic vocalizations. International Journal of Behavioral Development, 30(6), 509–516. https://doi.org/10.1177/0165025406071914

Hart, J. P., Jarvis, S. E., Jennings, W. P., & Smith-Howell, D. (2005). *Political keywords: Using language that uses us.* Oxford University Press.

He, H., Boyd-Graber, J., & Daumé III, H. (2016). Interpretese vs. translationese: The uniqueness of human strategies in simultaneous interpretation. In *Proceedings of the 2017 conference of the North American Association for Computational Linguistics: Human language technologies* (pp. 971–976). Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1111

Hepper, P. (2015). Behavior during the prenatal period: Adaptive for development and survival. *Child Development Perspectives*, *9*(1), 38–43. https://doi.org/10.1111/cdep.12104

Hiebert, T. (2007). The Tower of Babel and the origin of the world's cultures. *Journal of Biblical Literature*, *126*(1), 29–58. https://doi.org/10.2307/27638419

Highmore, B. (2022). Keywords and keywording. *Cultural Studies*, *36*(6), 875–898. https://doi.org/10.1080/09502386.2021.1947336

Jager, G., & Wichmann, S. (2016). Inferring the world tree of languages from word lists. In S. G. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Fehér, & T. Verhoef (Eds.), *The evolution of language: Proceedings of the 11th international conference (EVOLANGX11).* http://evolang.org/neworleans/papers/147.html

Kuhlthau, C. C. (1993). *Seeking meaning: A process approach to library and information services.* Ablex.

Ladd, D. R., Roberts, S. G., & Dediu, D. (2015). Correlational studies in typological and historical linguistics. *Annual Review of Linguistics*, *1*, 221–241. https://doi.org/10.1146/annurev-linguist-030514-124819

Law II, F., Mahr, T., Schneeberg, A., & Edwards, J. (2016). Vocabulary size and auditory word recognition in preschool children. *Applied Psycholinguistics*, *38*(1), 89–125. https://doi.org/10.1017/S0142716416000126

Lucy, J. A. (1997). Linguistic relativity. *Annual Review of Anthropology*, *26*, 291–312. https://doi.org/10.1146/annurev.anthro.26.1.291

Lucy, J. A. (2016). Recent advances in the study of linguistic relativity in historical context: A critical review. *Language Learning*, *66*(3), 487–515. https://doi.org/10.1111/lang.12195

Lytle, S. R., & Kuhl, P. K. (2017). Social interaction and language acquisition: Toward a neurobiological view. In E. M. Fernández & H. S. Cairns (Eds.), *Handbook of psycholinguistics* (pp. 615–634). John Wiley & Sons.

Mahowald, K., & Gibson, E. (2013). Short, frequent words are more likely to appear genetically related by chance. *PNAS*, *110*(35), Article E3253. https://doi.org/10.1073/pnas.1308822110

Mandel, D. R., Jusczyk, P. W., & Pisoni, D. B. (1995). Infants' recognition of the sound patterns of their own names. *Psychological Science*, *6*(5), 314–317. https://doi.org/10.1111/j.1467-9280.1995.tb00517.x

Martin, L. (1986). "Eskimo words for snow": A case study in the genesis and decay of an anthropological example. *American Anthropologist*, *88*(2), 418–423. https://doi.org/10.1525/aa.1986.88.2.02a00080

Mitchell, L. (1999). Earliest Egyptian glyphs. *Archaeology*, *52*(2). https://archive.archaeology.org/9903/newsbriefs/egypt.html

Moore, R. J. (2013). When names fail: Referential practice in face-to-face service encounters. *Language in Society*, *37*(3), 385–413. https://doi.org/10.1017/S004740450808055X

Pagel, M. (2017). Darwinian perspectives on the evolution of human languages. *Psychonomic Bulletin & Review*, *24*, 151–157. https://doi.org/10.3758/s13423-016-1072-z

Pagel, M., Atkinson, Q. D., Calude, A. S., & Meade, A. (2013). Ultraconserved words point to deep language ancestry across Eurasia. *PNAS*, *110*, 8471–8476. https://doi.org/10.1073/pnas.1218726110

Parish-Morris, J., Golinkoff, R. M., & Hirsh-Pasek, K. (2013). From coo to code: Language acquisition in early childhood. In P. Zelazo (Ed.), *Oxford handbook of developmental psychology* (Vol. 1, pp. 868–909). Oxford University Press.

Pavlenko, A. (2016). Whorf's lost argument: Multilingual awareness. *Language Learning*, *66*(3), 581–607. https://doi.org/10.1111/lang.12185

Perslsvaig, A., & Lewis, M. W. (2015). *The Indo-European controversy.* Cambridge University Press.

Piaget, J. (2001). *The language and thought of the child* (M. Gabain & R. Gabain, Trans.). Routledge.

Putnam, H. (1975). The meaning of "meaning." In K. Gunderson (Ed.), *Language, mind, and knowledge* (pp. 131–193). University of Minnesota Press.

Quine, W. V. O. (1960). *Word and object*. MIT Press.

Quine, W. V. O. (1987). Indeterminacy of translation again. *Journal of Philosophy*, *84*(1), 5–10. https://doi.org/10.2307/2027132

Reddy, M. (1979). The conduit metaphor: A case of frame conflict in our language about language. In A. Ortony (Ed.), *Metaphor and thought* (pp. 284–324). Cambridge University Press.

Risager, K. (2020). Linguaculture and transnationality: The cultural dimensions of language. In J. Jackson (Ed.), *Routledge handbook of language and intercultural communication* (2nd ed., pp. 109–123). Routledge.

Russell, B. (1905). II.-On denoting. *Mind*, *14*(4), 479–493. https://doi.org/10.1093/mind/XIV.4.479

Russell, B. (1910). Knowledge by acquaintance and knowledge by description. *Proceedings of the Aristotelian Society*, *11*, 108–128. http://www.jstor.org/stable/4543805

Saint-Georges, C., Chetouani, M., Cassel, R., Apicella, F., Mahdhaoui, A., Muratori, F., Laznik, M.-C., & Cohen, D. (2013). Motherese in interaction: At the cross-road of emotion and cognition? *PLoS ONE*, *8*(10), Article e78103. https://doi.org/10.1371/journal.pone.0078103

Schoenenberg, K., Raake, A., & Koeppe, J. (2014). Why are you so slow? Misattribution of transmission delay to attributes of the conversation partner at the far-end. *International Journal of Human-Computer Studies*, *72*(5), 477–487. https://doi.org/10.1016/j.ijhcs.2014.02.004

Scott, M. (1997). PC analysis of key words—And key key words. *System*, *25*(2), 233–245. https://doi.org/10.1016/S0346-251X(97)00011-0

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language.* Cambridge University Press.

Seuren, P. A. M. (2016). Saussure and his intellectual environment. *History of European Ideas*, *42*(6), 819–847. https://doi.org/10.1080/01916599.2016.1154398

Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford University Press.

Slobin, D. I., & Bowerman, M. (2007). Interfaces between linguistic typology and child language research. *Linguistic Typology*, *11*(1), 213–226. https://doi.org/10.1515/LINGTY.2007.015

Sluiter, I., Matthaios, S., & Rengakos, A. (2015). Ancient etymology: A tool for thinking. In F. Montanari, S. Matthaios, & A. Rengakos (Eds.), *Brill's companion to ancient Greek scholarship* (pp. 896–922). Brill.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication & cognition.* Harvard University Press.

Sperber, D., & Wilson, D. (2012). The mapping between the mental and the public lexicon. In D. Wilson & D. Sperber (Eds.), *Meaning and relevance* (pp. 31–46). Cambridge University Press.
Stawarska, B. (2015). *Saussure's philosophy of language as phenomenology: Undoing the doctrine of the course in general linguistics.* Oxford University Press.
Sternberg, R. J. (1987). Most vocabulary is learned from context. In M. McKeown & M. E. Curtis (Eds.), *The nature of vocabulary acquisition* (pp. 89–105). Erlbaum.
Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics.* Blackwell Publishing.
Sturtevant, E. H. (1952). The prehistory of Indo-European: A summary. *Language*, *28*(2), 177–181. https://doi.org/10.2307/410510
Tardif, T., Flectcher, P., Liang, W., Zhang, Z., Kacaroti, N., & Marchman, V. (2008). Baby's first 10 words. *Developmental Psychology*, *44*(4), 929–939. https://doi.org/10.1037/0012-1649.44.4.929
Taub, L. (1993). Evolutionary ideas and "empirical" methods: The analogy between language and species in the works of Lyell and Schleicher. *British Journal for the History of Science*, *26*(2), 171–193. https://www.jstor.org/stable/4027096
Taylor, R. S. (1968). Question-negotiation and information seeking in libraries. *College & Research Libraries*, *29*(3), 178–194. https://doi.org/10.5860/crl_29_03_178
Turner, S. (1827). Turner on language: On the affinities and diversities of the world, and their primeval cause; on the elementary and compound terms, and their classifications and affinities for the numeral two; on the terms used by the different nations to express "mother"; on the classification and affinities for "father"; further illustrations of the primeval cause of the affinities and diversities of language; further illustrations of the preceding principles. *Transactions of the Royal Society of Literature*, *1*(1), 17–106.
Voegtline, K. M., Costigan, K. A., Pater, H. A., & DiPietro, J. A. (2013). Near-term fetal response to maternal spoken voice. *Infant Behavior and Development*, *36*(4), 526–533. https://doi.org/10.1016/j.infbeh.2013.05.002
Vološinov, V. N. (1973). *Marxism and the philosophy of language* (L. Matejka & I. R. Titunik, Trans.). Harvard University Press.
Waxman, S., Fu, X., Arunachalam, S., Leddon, E., Geraghty, K., & Song, H. (2013). Are nouns learned before verbs? Infants provide insight into a long-standing debate. *Child Development Perspectives*, *7*(3), 155–159. https://doi.org/10.1111/cdep.12032
Wierzbicka, A. (1997). *Understanding cultures through their key words: English, Russian, Polish, German, and Japanese.* Oxford University Press.
Wierzbicka, A. (2014). *Imprisoned in English: The hazards of English as a default language.* Oxford University Press.
Wierzbicka, A. (2021). "Semantic Primitives" fifty years later. *Russian Journal of Linguistics*, *25*(2), 317–342. https://doi.org/10.22363/2687-0088-2021-25-2-317-342
Williams, R. (1976). *Keywords: A vocabulary of culture and society.* Oxford University Press.
Wilson, D., & Sperber, D. (2012). *Meaning and relevance.* Cambridge University Press.
Wittgenstein, L. (1958). *Philosophical investigations* (G. E. M. Anscombe, Trans.). Basil Blackwell.
Wolff, T. (2016). Arbitrary, natural, other: J. G. Herder and ideologies of linguistic will. *European Romantic Review*, *27*(2), 259–280. https://doi.org/10.1080/10509585.2016.1140044
Wood, N., & Cowan, N. (1995). The cocktail party phenomenon revisited: How frequent are attention shifts to one's name in an irrelevant auditory channel? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(1), 255–260. https://doi.org/10.1037/0278-7393.21.1.255
Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272. https://doi.org/10.1037/0033-295X.114.2.245
Zetzel, J. (2019). Natural law and natural language in the first century BCE. In G. Pezzini & B. Taylor (Eds.), *Language and nature in the classical Roman world* (pp. 191–211). Cambridge University Press.

Rites, Religions

5

Abstract

This chapter describes the possible origins of rites and ritual communication through the study of embedded mnemonics in carvings, landscape modification, and monuments as well as through the creation and memorization of "key symbols" such as the names of gods and magical invocations that were to become foundational to early religions as they began to add written narratives to their oral traditions. It also describes three primary categories of religious keywords that are still pertinent today: the imperative (commands), the performative (rituals), and the interpretive (explanations).

While symbolic communication is generally accepted as the beginning of truly human culture, early hominin fossils and associated artifacts are usually scattered and fragmentary, and any material objects found with these tend to be difficult to interpret, especially when the site has been compromised by natural processes or other disturbances over time. Even identifying the remains of possible communicative or cognitive activities is a challenge, since the discipline of paleoanthropology has traditionally followed the larger discipline of archaeology in using a "ladder of inference" (Hawkes 1954) in which symbolic interpretation of the material evidence is usually considered irrelevant or impossible (Robb 1998). Malafouris (2013, pp. 89–118) notes that a second problem is that of conflating symbolic representation with language semiology from the Saussurian perspective rather than with material semiotics in the Peircean perspective, which he believes is much more appropriate to the study of such material signs. A third problem currently receiving some attention is that cognitive archaeology is not immune to the same biases prevalent among Western-centric research in the other human sciences, in that there is seldom much consideration of sample diversity and that species-wide inferences and generalizations are

B. V. V. Martens, *Keywords In and Out of Context*, Synthesis Lectures on Information Concepts, Retrieval, and Services, https://doi.org/10.1007/978-3-031-32530-4_5

being made without noting potential differences between modern populations, populations of ancient humans, their hominin ancestors, or cousin lineages (Killin and Pain 2022).

5.1 Origins of Symbols

Nevertheless, the very early indications of possibly symbolic communication are now being considered of more interest (Kissel and Fuentes 2016, 2017). For instance, the enigmatic stalagmite structures within Bruniquel Cave in southwest France from roughly 176,000 years ago are now tentatively being attributed to Neanderthal builders and may represent one of the earliest sites of ritual significance (Jaubert et al. 2016).

The earliest widely accepted evidence for reference to symbolic concepts includes the figurative rock art at Sulawesi in Indonesia and the cave paintings at Cueva de El Castillo in Spain, both dating to between 35,000 and 40,000 years ago, and the set of 32 geometric symbols found in caves throughout Europe from the same period (Von Petzinger 2016). These markings, while known since the discovery of the caves, were often dismissed as inexplicable or incidental, especially when compared to the representational art found within. Some markings, such as the meandering lines that von Petzinger spotted at a site in Portugal's Côa Valley region, may have been maplike representations of a river or other landscape features. Other signs, such as the lines inscribed on a deer-tooth necklace, could have served as memory aids for ceremonialists presiding over important rituals or recounting a tribe's origin stories, which would indicate an oral tradition. Such markings, says von Petzinger, seem to be a way of storing information externally—a form of graphic communication that may have eventually led to writing.

These substantially predate the stone megaliths that indicate the remains of a ritualized culture at Göbekli Tepe in Turkey as early as the ninth millennium BC, the incised clay tokens from Mureybet in Syria that recorded specific contributions of cereal grains to feasts and other ritual ceremonies from the eighth millennium BC (Schmandt-Besserat 1982) or the incised tortoise shells possibly used for shamanic purposes found at Jihau in China (Li et al. 2003). Bradley (2012) has studied the prevalence of certain architectural archetypes, notably circular barrows and stone henges, across prehistoric Europe, which he believes indicates a common orientation toward space and time and an increase in shared symbolic behavior. More broadly, Kelly (2015) proposes that certain prehistoric societies invented sophisticated "oral technologies," devoting significant physical and mental efforts to extend beyond their original oral traditions by embedding their cultural knowledge through carvings, landscape, and monumental mnemonics that served as knowledge and memory "spaces" for rituals and other important activities, and thus could transmit this essential societal information more accurately through generations than could oral accounts alone. She argues (p. 9) that this "Method of Loci" was originally based on physical sites used by far earlier cultures than even those of pre-literate Greece and greatly antedates their celebrated "memory palace" technique utilized well

into the European Renaissance. However, such "oral technologies" themselves may eventually lose their last interpreters, and archaeologists must extrapolate the uses of labyrinths and megaliths, for example, from whatever similarities these may bear to evidence from other sources.

Language obviously has played an essential role in what Edwards and Middleton (1987) term the "transcendence of the serial chain of behavior" over time as the social basis of "symbolic remembering," which is the use of language to construct joint understandings that are accessible and communicable via a common code (i.e., conventional symbols, grammatical rules, and rules of interpretation). The "shared reference" to culture and custom becomes immeasurably more complex than was possible through gesture or other forms of communication prior to the development of a spoken language (Finnegan 2015).

In accordance with Cosmides and Tooby's (2013) work on evolutionary psychology, which considers the human mind as a computer populated by a large number of adaptive specializations, each equipped with content-rich representations, concepts, inference systems, and regulatory variables that are functionally organized to solve the complex problems of survival and reproduction encountered by the ancestral hunter-gatherers, and for which one of the greatest challenges is the problem of coordinated and cooperative group action, Watson-Jones and Legare (2016) argue that the capacity to engage in ritual is a psychologically prepared, culturally inherited behavior geared toward facilitating social group dynamics, group cohesion, and group survival. The invention of new forms of physical and verbal deixis also allowed "pointing" to presumed natural or supernatural forces, providing potential explanations for otherwise inexplicable phenomena such as lightning, drought, or disease and offering possible ways to deal with these.

The development of "key symbols" is integral to religious and ritual practices of all kinds, as noted by Geertz (1957), Turner (1970), and Ortner (1973). Henderson (1999) found that divination rites can be considered as the earliest form of exegesis of such supernatural sources. However, most contemporary research on rituals has tended to focus on their discursive aspect as preserved within religious traditions and texts, while neglecting these embodied and iconic aspects, which likely emerged first (Kitts 2017). Interestingly, though, Watts (2005) found that written texts tend to be used as sacred objects in rituals rather than completely replacing them. For instance, after the destruction of the Temple in Jerusalem, Hebrew rituals began to become increasingly textualized (Schmid 2012).

5.2 Oral Traditions

The development of such integral social innovations as religion is incumbent on the ability of a group to cohere through particular ritual behaviors (Rossano 2012) and to remember certain symbolic concepts expressed by "key words" that help to define their social existence (Kirkpatrick and Rossano 2021). The names of gods and goddesses served this

purpose in many cultures. Moreover, an oral culture tends to favor the development of rhetorical, memorization, and improvisational skills to a greater extent than does a print culture, so key words play an essential role in defining that society, as shown by the early emergence of both orators and actors in classical Greece, for example (Yates 1966). Although, as Havelock (1982, p. 8) pointed out, while there was no specific term for the abstract concept of "word" in antiquity, and the discrete sound of individual words are normally elided in vocal communication, it can be taken for granted that the use of names and naming demonstrates deliberate verbal deixis in oral societies as well. The names of gods, kings, and heroes served as social glue for these communities.

In the Mesopotamian epics of Sumer, the lengthy recitation of kings lists, whether or not these were historically accurate, served to provide "heroic" legitimacy for the ruler and his heirs (Noegel 2005). The belief in the power of specific words in ritual charms was well established in Hittite fertility practices by 1650 BC (Collins 2002) and in Egyptian medicine by the time of the Papyrus Ebers in 1550 BC (Graf 2002) and undoubtedly has much earlier roots. Drout (2011, p. 456) finds that "traditional referentiality is generated by the combination of repetition with associative memory," making certain key words and phrases more likely to be transmitted down generations within specific formulaic genres.

Archaic Greek culture had no universally recognized sacred texts but relied heavily on a complex oral tradition of polytheism as well as sacrificial rites and other rituals passed on by ancestral custom (Henrich 2003). Havelock (1963) emphasized the importance of the oral transmission of traditional knowledge through didactic poetry such as that of Hesiod and epic poetry such as that of Homer. Havelock (1986) famously argued that these were so entrenched in the mind-set of classical Greek culture that Plato's real purpose in banning poets from his imaginary *Republic* was to prevent such traditional lore from being equally influential in his proposed ideal society, preventing the growth of his "forms" of new knowledge. Although Havelock may have overemphasized the pragmatic educational value of epic poetry, clearly its role in providing a grand narrative of presumed historical value to Greek society is clear.

While it is no longer widely accepted that the works of Plato mark the boundary between oral and written communication, and media theorists now agree that there must have been a longer and more gradual transition over time (Gibson 2005), Collins's magisterial work on the sociology of philosophy (1998) also posits that much of the cognitive advances in the field stemmed from the spoken and remembered exchanges over time by seminal thinkers such as Plato and the short chains of students who conveyed those thoughts into future generations through their own teaching.

Most of the scholarly attention to Walter Ong's classic trilogy (1967, 1977, 1982) focused on the gradual development and impact of literacy, but the importance of oral communication itself is now well established. As Ong noted, the influence of a technologized print culture has affected almost all traditional performance-based oral cultures, though certain traces of it still exist in even the most print-acculturated societies, notably in areas such as law, literature, and religion. There is also a growing realization that oral

and print cultures co-exist, forming a continuum rather than the sharp dichotomy hypothesized earlier (Goody 2000). The early presence and prevalence of pseudepigraphy (that is, those manuscripts deliberately and incorrectly attributed to ancient authors) testify to this continuum, in that many of these falsified works were purportedly "authored" in light of ancient oral originals in order to enhance their legitimacy (Lange 2010).

As scholars such as Lord (1969), Finnegan (1970), and Foley (1988) discovered, traces of ancient oral traditions also remain in narratives from a variety of cultures, ranging from the Homeric epics of Greece (Ready 2019), to India's Mahabharata (Blackburn et al. 1989) to Africa's Ntsomi performances (Scheub 1970). These epic tales have been found to focus on crucial events and personages, and, although usually narrated with a particular order and emphasis, rely on the absorbing presentation of communal cultural concepts and characterizations accepted by the audience rather than on either rote recitation or novel interpretation.

Foley (1992, p. 178) argued that, in a truly oral culture, this traditional "word-power" also engages contexts and mediated communication during what he termed "the enabling event—performance—and the enabling referent—tradition—that give meaning to word-power. Like the coinage created to carry this metonymic burden, the investigation has two interlocking primary responsibilities: the problem of defining what a 'word' is in this communicative medium and the coordinate task of locating and explaining its peculiar 'power'." Words in an oral culture are not defined by marks on a page but rather by their particular performance in particular places. He noted also that the power of these words varies in different contexts and cultures. In the bardic recitations of Homeric epics, for instance, the bard's ability to draw on an audience's existing stock of knowledge by making references, often couched in the form of familiar epithets such as "grey-eyed Athena," produced a much richer experience than is possible for the present-day reader encountering the text without such socialization. Nidich (2016), in her examination of early Hebrew oral traditions that influenced the creation of seminal biblical texts, further suggests that an awareness of these referential traditions and their reception is important both in understanding them and avoiding the imposition of modern normative ideas upon these narratives.

Speech-act theory (Austin 1962; Searle 1969), with its focus on performative speech, can be viewed as an archaeology of the oral tradition, in which verbal promises preceded written contracts both in time and in tenure. One example is the Hippocratic oath in its earliest form as recorded in the fifth century BC, which included certain key words that endure almost intact in its modern form ("do no harm") while others ("treat the children of the one who taught me this art as my own brothers") have fallen into disuse, showing how the performance of the medical practitioner over time has evolved from an apprenticeship into a profession.

The early Anglo-Saxons used charms as "word magic" to invoke supernatural assistance in their attempts to exert control over their environments (Arnovick 2006, p. 30). A lesser known form of this presumed performative power of words is that of curses,

a surprising number of which have survived from ancient times in written form (Gager 1992). Healing words too as part of early medical treatments played a larger part in the development of medicine than is generally recognized (Bishop 2007). Similarly, the use of saints' names in both prayers for intercession and in the taking of oaths was a common feature of medieval societies, showing the importance attached to particular names perceived as holy (Bartlett 2013). In Islam, the use of the Durood Sharif blessing upon any invocation of the Prophet Mohammad's name continues to be widespread today.

The importance of "word power" and rote memorization in these oral cultures is underscored by its survival in religious teaching. Even in modern societies, children are often indoctrinated at a very young age in particular belief systems by instruction in traditional formulaic recitations, such as the Vedic mantras in Hinduism, the catechism questions and responses of Catholicism, the Kiddush blessings of Judaism, or the obligatory Salat prayers of Islam. These belief systems offer evidence of the importance of the oral traditions in establishing and maintaining thought regimes throughout history, by focusing on a divine "word" as transmitted through elders, prophets, priests, or other religious authorities, with direct effects on the maintenance of social order, the creation of laws, and the development of hierarchies.

Even presumably divinely delivered words, whatever their origins, have been subjected to interpretation, whether this is exemplified by ritual poetry about the Vedic hymns (Jackson 2005) or by the *Abhidhamma Piṭaka*, which remains the classic commentary on Buddhist sutras. As these interpretations are conveyed over time, distinctions still continue to be made between the original oral and subsequent written channels of diffusion, usually favoring the oral ones, such as the preference for the so-called oral Torah above the written Torah for legal decisions within the rabbinate (Jaffee 1997), the tradition of the superiority of the spoken Qur'an in its original language to all written transcriptions and translations (Small 2011), and the continuance of otherwise unused early languages in the forms of ecclesiastical Latin in the Latin liturgical rites of the Catholic Church and Koine Greek in the ceremonies of the Greek Orthodox Church.

5.3 Ritual Writings

Despite the value placed on the spoken word, the growth and sustainability of religious and other traditional authority, however, eventually required the evolution of additional cognitive tools to supplement even the most sophisticated of memorization and rhetorical practices. The development of cuneiform in Mesopotamia, the development of hieroglyphics in Egypt, and the development of Shang script in China all represented more sophisticated logographic writing systems, in which each picture was an ideogram that could be considered "key" to its particular content, whether it was sag (head), pictured in cuneiform, ankh (life) pictured in hieroglyphics, or fu (husband) in Shang script. Such an extensive "key word" vocabulary necessitated the presence of a scribal or priestly class

capable of memorizing the hundreds (and, for Shang, thousands) of individual words depicted in these systems, which were largely used for ceremonial purposes.

One ingenious cognitive tool still in use today is the I-Ching (Book of Changes) used in ancient Chinese divination practices, which used "trigrams," comprised of eight different three-line symbols (each comprised of patterns of solid and broken lines) to represent natural phenomena. These trigrams were then doubled to form 64 individual "hexagrams" to denote particular phenomena that were used in predictions. Thus, the trigrams served as an index to the hexagrams, as the meanings of each symbol (Humphreys 2011).

Yet another cognitive tool in widespread religious use is the alphabet. The Kabbalah, with its mystical approach to the Hebrew alphabet, represents an example of how linguistic formulae and ritual could be viewed as manipulating reality (Harari 2019, p. 73). Notable among the acrostic psalms is the long Psalm 119, which typically is printed in subsections named after the 22 letters of the Hebrew alphabet, each section consisting of 8 verses, each of which begins with the same letter of the alphabet and the entire psalm consisting of $22 \times 8 = 176$ verses, the longest "chapter" in the Hebrew Bible. This psalm also famously references the Torah (or "Law") in each of its verses and is firmly embedded within Jewish tradition (Berger 2006). Within Christian tradition, the alphabet sublimates "the beginning and the end" as symbolized by the Alpha and Omega (the first and last letters of the Greek alphabet) in 21:6 of the Revelations of John. Despite this, in medieval times alphabetical order was considered contrary to divine order, which would, for instance, inevitably place Adam before Abel, because biblically correct chronologies took precedence over merely alphabetical order, considered appropriate mostly for lesser purposes, such as in reference to the lower orders of creation (Carruthers 1990; Clanchy 1993).

Experimentation over time with various writing surfaces, such as animal hides, bamboo, bone, bronze, clay, silk, stone, wax, and wood, and the invention of writing technologies, such as the stylus and ink, also led to the development of new forms of religious deixis, such as making the name of a god or king appear larger and more legible. Early forms of highlighting "key words" included the use of special formats for important words, such as the opening words of ritual spells in the Egyptian Book of the Dead, which were inscribed in colored pigment rather than black (Blumel 2009) and painting a red background for the names of demons in a papyrus fragment from the Tebtunis Temple library (Ryhol 2018).

Rubrication, in which colored pigments (usually red) were also used to indicate particularly important words and other textual elements in early Christian manuscripts, also became more common (Smith 2010). These "nomina sacra" in early Christian texts of the second and third century also used special Greek abbreviations for a few particular words, such as "Jerusalem" and "Jesus," in place of the more usual variants (Hurtado 2006). This practice may have begun as a variation of the Jewish practice of using the Tetragrammaton (a four-letter transliteration of the Hebrew name for Jehovah) to satisfy the religious

prohibition against speaking or viewing the holy name directly, but in the Christian context it instead became a way to quickly identify and emphasize specific sanctified words, particularly in public readings.

Carruthers (1990) explains that the extensive use of illustration in medieval Christian manuscripts also became a way for Church authorities to signpost significant passages for the devout, regardless of their level of literacy, similar to the use of stained glass windows and cathedral design to support liturgical rites. She argues that medieval literacy differed from modern literacy in that a medieval reader did not distinguish between personal experience and experience gained through reading in the same way that a modern reader does, precisely because of the extensive meditation and memorization required in the consumption of devotional texts (1990, p. 160). Preaching for the edification of the faithful involved both rhetorical and intellectual skills in elucidating various biblical passages through allegory, metaphor, and literal interpretation for both educated and non-educated listeners according to accepted doctrine (Menache and Horowitz 1996). According to Burke (2000), the ultimate goal of education during the medieval period was to make orthodox doctrine the foundation of all knowledge. Although this educational goal is clearly no longer relevant to much of the modern world, it has not disappeared, as its similarities are apparent in educational endeavors in different forms of contemporary religious environments, such as the Hasidic yeshiva, the Islamic madrasa, and the Christian fundamentalist school.

The creation of concordances to various "divine words" also has a long history: Jewish scholars in Jerusalem were initially entrusted with the accurate oral transmission of the Hebrew Biblical text and, by the fourth century were also responsible for its standardized written transmission. The Hebrew Bible then accumulated numerous thematic word-lists (the so-called Mesocratic notes), which often consisted of a key Bible word and the phrase in which it occurred, arranged in the canonical order of the 24 books of the Hebrew Bible, and then by the section in which the word appears (Weinberg 1997, 2001). This remained a satisfactory solution to most Jewish scholarship until a true concordance to the Hebrew Bible was compiled by 1445 by Isaac Nathan, primarily as an aid to countering Christian argumentation rather than as a real necessity of Jewish scholarship.

The first concordance in Latin to the Vulgate Bible was compiled by 1244 by Dominican monks as an aid to Christian preachers in need of specific Biblical verses, while the first concordance to the English language Tyndale Bible was viewed as heretical and its compiler, John Marbeck, narrowly avoided being burnt at the stake in 1544, though his concordance was eventually published in 1550. The first concordance to the English language King James Bible was created single-handedly by bookseller Alexander Cruden in 1734 and is still in use today (Keay 2005).

The Qur'an, however, in accordance with its believers' strict reliance on accurate memorization, the intricate relationships among its various verses, and complexities of the Arabic language did not prompt the production of written concordances until the end

of the nineteenth century, and these originally existed only as unpublished manuscripts (Hairutdinov 2016; Kokabi 2011).

Of course, the Jewish, Christian, and Islamic traditions all possess extensive bodies of exegesis (that is, critical interpretation or explanation of a specific text) originating in rabbinic midrash and peshat on the Torah (Benin 2003) and Talmudic scholarship (Samely 2007), early patristic exegesis of the Christian Gospels (Simonetti 1994), medieval scholasticism (Dufal 2020), and tafsir commentaries on Qur'anic verses (Burge 2015) respectively. The practice of exegesis itself is continually undergoing change, as new interpretative methodologies are developed and refined.

The meaning of religious keywords continues to inspire debate, as demonstrated, for instance, by the multiple definitions of particular characters in the Confucian *Analects*, which have undergone generations of commentary in the past two millennia as to how the definition might impact the specifics of a particular saying by the sage (Xiao 2007) or by the doctrine of "abrogation" (whether certain verses of the Qur'an can be considered to have been superseded by later ones), which has been debated by Islamic commentators for centuries (McAuliffe 2006).

We may view religious keywords, therefore, as falling into three main categories: the imperative (commands), performative (rituals), and the interpretive (explanations). Imperative keywords mandate general cognitive requirements for members of a specific religious group, whether these requirements are, for instance, to reflect on daily actions or inactions in relation to the requirements, or to comprehend that no boundaries exist between physical and spiritual worlds, or to accept that each individual is capable of enlightenment. Performative keywords mandate specific behavioral requirements, whether these involve dietary practices, holy wars, or marriage ceremonies. Interpretive keywords explain the connections among all of these in accordance with the particular belief system, sometimes by means of traditional legends, sometimes by cryptic aphorisms, and sometimes by elaborate exegesis.

Examples of imperative keywords from one of the earliest religions are those of the Jain dharma, which mandate *ahiṃsā* (non-violence), *satya* (truth), *asteya* (not stealing), *brahmacharya* (sexual continence), and *aparigraha* (non-possessiveness) as a way of life. An example of performative keywords from the medieval Christian period are the many European sermons that invoked the importance of "crusades" against perceived "infidels" or "heretics" as a way of achieving divine favor (Gaposchkin 2017; Valenin 2019). An example of interpretive keywords from the early modern African period is "Ubuntu," which describes the significance of group solidarity on the basis of survival as the essence of being African, transcending beyond a narrow Western view of individualism to a holistic African outlook that recognizes four basic attributes of human beings: human dignity and equality, universal brotherhood, sacredness of life, and "Being" as the most desirable state of life (Fagunwa 2019).

The importance of religious keywords has not diminished in today's globalized communication societies, as they still serve to arouse emotion and to mobilize both social

and political action within Christian (Burge and Williams 2019), Hindu (Thomas 2021), Islamic (Rifat et al. 2022), Jewish (Osterbur and Kiel 2021), and other religious communities, even as much of society has become increasingly secularized and globalized in its communications, as will be explored in the next chapter.

References

Arnovick, L. (2006). *Written reliquaries: The resonance of orality in medieval English texts*. John Benjamins Publishing.

Austin, J. L. (1962). *How to do things with words*. Harvard University Press.

Bartlett, R. (2013). *Why can the dead do such great things? Saints and worshippers from the martyrs to the Reformation*. Princeton University Press.

Benin, S. (2003). The search for truth in sacred scripture: Jews, Christians and the authority to interpret. In J. D. McAuiliffe, B. D. Walfish, & J. W. Goering (Eds.), *With reverence for the word: Medieval scriptural exegesis in Judaism, Christianity, and Islam* (pp. 13–32). Oxford University Press.

Berger, J. (2006). Indexation, memory, power and representations at the beginning of the 12th century: The rediscovery of pages from the tables to the Liber de Honoribus, the first Cartulary of the Collegiate Church of St Julian of Auvergne (Brioude). *The Indexer*, *25*(2), 95–99. https://doi.org/10.3828/indexer.2006.26

Bishop, L. M. (2007). *Words, stones, and herbs*. Syracuse University Press.

Blackburn, S. H., Claus, P. J., Flueckinger, J. B., & Wadley, S. S. (Eds.). (1989). *Oral epics in India*. University of California Press.

Blumel, L. H. (2009). Report of proceedings in red ink from late second century AD Oxyrhynchus. *Bulletin of the American Society of Papyrologists*, *46*, 23–30.

Bradley, R. (2012). *The idea of order: The circular archetype in prehistoric Europe*. Oxford University Press.

Burge, S. R. (2015). Introduction: Words, hermeneutics and the construction of meaning. In S. R. Burge (Ed.), *The meaning of the word: Lexicology and the Qur'anic exegesis* (pp. 1–40). Oxford University Press.

Burge, R. P., & Williams, M. D. (2019). Is social media a digital pulpit? How evangelical leaders use Twitter to encourage the faithful and publicize their work. *Journal of Religion, Media and Digital Culture*, *8*(3), 309–339. https://doi.org/10.1163/21659214-00803004

Burke, P. (2000). *A social history of knowledge: From Gutenberg to Diderot*. Polity Press.

Carruthers, M. (1990). *The book of memory: A study of memory in medieval culture*. Cambridge University Press.

Clanchy, M. (1993). *From memory to written record: England 1066–1307*. Blackwell Publishing.

Collins, R. (1998). *The sociology of philosophies: A global theory of intellectual change*. Harvard University Press.

Collins, B. J. (2002). Necromancy, fertility and the dark earth: The use of ritual pits in the Hittite cult. In P. Miriki & M. Meyer (Eds.), *Magic and ritual in the ancient world* (pp. 224–241). Brill.

Cosmides, L., & Tooby, J. (2013). Evolutionary psychology: New perspectives on cognition and motivation. *Annual Review of Psychology*, *64*, 201–229. https://doi.org/10.1146/annurev.psych.121208.131628

Drout, M. D. C. (2011). Variation within limits: An evolutionary approach to the structure and dynamics of the multiform. *Oral Tradition*, *26*(2), 447–474.

Dufal, B. (2020). The fathers of scholasticism: Authorities as totems. In A. Fitzpatrick & J. Sabapathy (Eds.), *Individuals and institutions in medieval scholasticism* (pp. 53–69). University of London Press.

Edwards, D., & Middleton, D. (1987). Conversation and remembering: Bartlett revisited. *Applied Cognitive Psychology*, *1*(2), 77–92. https://doi.org/10.1002/acp.2350010202

Fagunwa, T. (2019). Ubuntu: Revisiting a endangered African philosophy in quest of a Pan-Africanist revolutionary ideology. *Genealogy*, *3*(3), Article 45. https://doi.org/10.3390/genealogy3030045

Finnegan, R. H. (1970). *Oral literature in Africa.* Clarendon Press.

Finnegan, R. H. (2015). *Where is language?* Routledge.

Foley, J. M. (1992). Word-power, performance, and tradition. *The Journal of American Folklore*, *105*(417), 275–301. https://doi.org/10.2307/541757

Foley, J. M. (1988). *The theory of oral composition: History and methodology*. Indiana University Press.

Gager, J. G. (1992). *Curse tablets and binding spells from the ancient world*. Oxford University Press.

Gaposchkin, M. C. (2017). *Invisible weapons: Liturgy and the making of crusade ideology*. Cornell University Press.

Geertz, C. (1957). Ethos, world-view and the analysis of sacred symbols. *The Antioch Review*, *17*(4), 421–437. https://doi.org/10.7723/antiochreview.74.3.0622

Gibson, T. (2005). Epilogue to Plato: The bias of literacy. *Proceedings of the Media Ecology Association*, *6*. https://www.media-ecology.org/resources/Documents/Proceedings/v6/v6-05-Gibson.pdf

Goody, J. (2000). *The power of the written tradition.* Smithsonian Press.

Graf, F. (2002). Theories of magic in antiquity. In P. Miriki & M. Meyer (Eds.), *Magic and ritual in the ancient world*. Brill.

Hairutdinov, A. G. (2016). "Miftah al-Qur'an" by Kayum Nasyri: The first Quran concordance in the Muslim world. *Islam in the Modern World*, *12*(1), 57–76. https://doi.org/10.20536/2074-1529-2016-12-1-57-76

Harari, Y. (2019). "Practical Kabbalah" and the Jewish tradition of magic. *Aries*, *19*(1), 38–82. https://doi.org/10.1163/15700593-01901003

Havelock, E. A. (1963). *Preface to Plato.* Princeton University Press.

Havelock, E. A. (1982). *The literate revolution in Greece and its cultural consequences*. Princeton University Press.

Havelock, E. A. (1986). *The muse learns to write: Reflection on orality and literacy from antiquity to the present*. Yale University Press.

Hawkes, C. (1954). Archaeological theory and method: Some suggestions from the Old World. *American Anthropologist*, *56*(2), 155–168. https://doi.org/10.1525/aa.1954.56.2.02a00660

Henderson, J. B. (1999). Divination and Confucian exegesis. *Extrême-Orient, Extrême-Occident*, *21*, 78–89.

Henrich, A. (2003). Writing religion: Inscribed texts, ritual authority and the religious discourse of the Polis. In: H. Yunis (Ed.), *Written texts and the rise of literate culture in ancient Greece* (pp. 38–58). Cambridge University Press. https://doi.org/10.1017/CBO9780511497803.004

Humphreys, N. K. (2011). The world's oldest profession: Indexing? *The Indexer*, *29*(4), 161–165. https://doi.org/10.3828/indexer.2011.44

Hurtado, L. (2006). *The earliest Christian artifacts.* Eerdmans.

Jackson, P. (2005). Retracing the path: Gesture, memory, and the exegesis of tradition. *History of Religions*, *45*(1), 1–28. https://doi.org/10.1086/497814

Jaffee, M. S. (1997). A rabbinic ontology of the written and spoken word: On discipleship, transformative knowledge, and living tests of oral Torah. *Journal of the American Academy of Religion*, *65*(3), 525–549.

Jaubert, J., Verheyden, S., Genty, D., Soulier, M., Cheng, H., Blamart, D., Burlet, C., Camus, H., Delaby, S., Deldicque, D., Edwards, R. L., Ferrier, C., Lacrampe-Cuyaubère, F., Lévêque, F., Maksud, F., Mora, P., Muth, X., Régnier, É., & Frédéric, S. (2016). Early Neanderthal constructions deep in Bruniquel Cave in southwestern France. *Nature*, *534*, 111–114. https://doi.org/10.1038/nature18291

Keay, J. (2005). *Alexander the corrector: The tormented genius whose Cruden's concordance unwrote the Bible*. Overlook Press.

Kelly, L. (2015). *Knowledge and power in prehistoric society: Orality, memory, and the transmission of culture*. Cambridge University Press.

Killin, A., & Pain, R. (2022). How WEIRD is Cognitive Archaeology? Engaging with the challenge of cultural variation and sample diversity. *Review of Philosophy & Psychology*, Advance online publication. https://doi.org/10.1007/s13164-021-00611-z

Kirkpatrick, L. A., & Rossano, M. J. (2021). The advent of religion. In T. Henley & M. J. Rossano (Eds.), *Psychology & cognitive archaeology: An interdisciplinary approach to the study of the human mind* (pp. 126–139). Routledge.

Kissel, M., & Fuentes, A. (2016). From hominid to human: The role of human wisdom and distinctiveness in the evolution of modern humans. *Philosophy, Theology, and the Sciences*, *3*(2), 217–244. https://doi.org/10.1628/219597716X14696202742253

Kissel, M., & Fuentes, A. (2017). Semiosis in the Pleistocene. *Cambridge Archaeological Journal*, *27*(3), 397–412. https://doi.org/10.1017/S0959774317000014

Kitts, M. (2017). Discursive, iconic, and somatic perspectives on ritual. *Journal of Ritual Studies*, *31*(1), 11–26.

Kokabi, M. (2011). A survey on the concordance of I'timad-Assaltanih and other information retrieval tools in a copy of Koran dated 1313 Lunar Hegira (1892 AD). *The International Journal of the Book*, *8*(3), 129–145. https://doi.org/10.18848/1447-9516/CGP/v08i03/36881

Lange, A. (2010). In the second degree: Jewish paratextual literature in the context of Graeco-Roman and ancient Near Eastern literature. In P. Alexander, A. Lange, & R. Pillinger (Eds.), *In the second degree: Paratextual literature in ancient Near Eastern and ancient Mediterranean culture and its reflections in medieval literature*. Brill.

Li, X., Harbottle, G., Zhang, J., & Wang, C. (2003). The earliest writing? Sign use in the seventh millennium BC at Jiahu, Henan Province, China. *Antiquity*, *77*(295), 31–41. https://doi.org/10.1017/S0003598X00061329

Lord, A. B. (1969). *The singer of tales*. Harvard University Press.

Malafouris, L. (2013). *How things shape the mind: A theory of material engagement*. MIT Press.

McAuliffe, J. D. (2006). Exegetical sciences. In A. Rippin (Ed.), *The Wiley Blackwell companion to the Qur'an* (pp. 403–419). John Wiley & Sons.

Menache, S., & Horowitz, J. (1996). Rhetoric and its practice in medieval sermons. *Historical Reflections/Réflexions Historiques*, *22*(2), 321–350.

Nidich, S. (2016). Folkore and Israelite tradition: Appreciation and application. In S. Nidich (Ed.), *The Wiley Blackwell companion to early Israel* (pp. 87–102). John Wiley & Sons.

Noegel, S. B. (2005). Mesopotamian epic. In J. M. Foley (Ed.), *A companion to ancient epic*. Blackwell Publishing.

Ong, W. J. (1967). *The presence of the word: Some prolegomena for cultural and religious history*. Yale University Press.

Ong, W. J. (1977). *The interfaces of the word: Studies in the evolution of consciousness and culture*. Cornell University Press.

Ong, W. J. (1982). *Orality and literacy: The technologizing of the word*. Methuen & Co.
Ortner, S. B. (1973). On key symbols. *American Anthropologist*, *75*(5), 1338–1346.
Osterbur, M., & Kiel, C. (2021). Tweeting in echo chambers? Analyzing Twitter discourse between American Jewish interest groups. *Journal of Information Technology & Politics*, *18*(2), 194–213. https://doi.org/10.1080/19331681.2020.1838396
Ready, J. L. (2019). *Orality, textuality, and the Homeric epics: An interdisciplinary study of oral texts, dictated texts, and wild texts*. Oxford University Press.
Rifat, M. R., Amin, M. R., & Ahmed, S. I. (2022). Situating public speaking: The politics and poetics of the digital Islamic sermons in Bangladesh. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW1), Article No. 64. https://doi.org/10.1145/3512911
Robb, J. E. (1998). The archaeology of symbols. *Annual Review of Anthropology*, *27*(1), 329–346. https://doi.org/10.1146/annurev.anthro.27.1.329
Rossano, M. J. (2012). The essential role of ritual in the transmission and reinforcement of social norms. *Psychological Bulletin*, *138*(3), 529–549. https://doi.org/10.1037/a0027038
Ryhol, K. (2018). Scribal habits at the Tebtunis temple library: On materiality, formal features, and palaeography. In J. Cromley & E. Grossman (Eds.), *Scribal repositories in Egypt from the New Kingdom to the Early Islamic period* (pp. 153–183). Oxford University Press.
Samely, A. (2007). *Forms of rabbinic literature and thought: An introduction*. Oxford University Press.
Scheub, H. (1970). The technique of expansible image in Xhosa Nitsomi-performances. *Research in African Literatures*, *1*(2), 119–146.
Schmandt-Besserat, D. (1982). The emergence of recording. *American Anthropologist*, *84*(4), 871–878.
Schmid, K. (2012). The canon and the cult: The emergence of book religion in ancient Israel and the gradual sublimation of the temple cult. *Journal of Biblical Literature*, *131*(2), 289–305. https://doi.org/10.2307/23488226
Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge University Press.
Simonetti, M. (1994). *Biblical interpretation in the early church: An historical introduction to patristic exegesis*. Continuum Publishing.
Small, K. E. (2011). *Textual criticism and Qur'an manuscripts*. Lexington Books.
Smith, M. M. (2010). Red as a textual element during the transition from manuscript to print. In O. Da Rold & E. Treharne (Eds.), *Textual cultures: Cultural texts*. D. S. Brewer.
Thomas, P. (2021). Religion, media and culture in India: Hindutva and Hinduism. In S. M. Hoover & N. Echchaibi (Eds.), *Media and religion: The global view* (Vol. 74, pp. 205–218). Walter de Gruyter. https://doi.org/10.1515/9783110497878-012
Turner, V. (1970). *The forest of symbols*. Cornell University Press.
Valenin, P. (2019). God wills it! Supplementary divine purposes for the crusades according to crusade propaganda. *The Journal of Ecclesiastical History*, *70*(3), 472–486. https://doi.org/10.1017/S0022046918002610
Von Petzinger, G. (2016). *The first signs: Unlocking the mysteries of the world's first symbols*. Atria Books.
Watson-Jones, R. E., & Legare, C. H. (2016). The social functions of group rituals. *Current Directions in Psychological Science*, *25*(1), 42–46. https://doi.org/10.1177/0963721415618486
Watts, J. W. (2005). Ritual legitimacy and scriptural authority. *Journal of Biblical Literature*, *124*(3), 401–417. https://doi.org/10.2307/30041032
Weinberg, B. H. (1997). The earliest Hebrew citation indexes. *Journal of the American Society for Information Science*, *48*(4), 318–330. https://doi.org/10.1002/(SICI)1097-4571(199704)48:4<318::AID-ASI5>3.0.CO;2-Z

Weinberg, B. H. (2001). Index structures in early Hebrew biblical word lists. *The Indexer*, *22*(4), 178–186. https://doi.org/10.3828/indexer.2001.22.4.5

Xiao, Y. (2007). How Confucius does things with words: Two hermeneutic paradigms in the "Analects" and its exegeses. *The Journal of Asian Studies*, *66*(2), 497–532. https://doi.org/10.1017/S0021911807000897

Yates, F. (1966). *The art of memory*. University of Chicago Press.

Writing, Indexing

6

Abstract

This chapter describes the development of writing, the diffusion of alphabetic technologies, the creation of administrative, legal, and theocratic structures dependent on written communication, and the growing importance of textual interpretation and information retrieval methods with the expansion of knowledge made possible by the invention of printing, the increased international transmission of manuscripts, books, and news, and a gradual increase in literacy leading into the early modern era.

The primary preservation of keywords has been through writing: "a system of more or less permanent marks used to represent an utterance in such a way that it can be recovered more or less exactly without the intervention of the utterer" (Daniels 2007). This chapter discusses the evolution of writing and the various ways in which key concepts have been distinguished as key words over the millennia.

6.1 The Rise of Writing

The prototype of all dictionaries today was the lists of Sumerian words (up to 1400) with their Akkadian equivalents, written in cuneiform script on clay tablets roughly 4700 years ago (Klein 2015). Based on archaeological evidence from the first half of the second millennium BCE, Egyptian hierographic and hieratic logographic systems seemed to have influenced the creation of the first phonemic system, known as the proto-Canaanite, by providing a few ideograms, such as "ox" ("aleph") and "water" ("net") that were borrowed as the initial sounds for the letters that eventually became A and M, resulting in a powerful communication technology that diffused across the Near East in numerous local variations

B. V. V. Martens, *Keywords In and Out of Context*, Synthesis Lectures on Information Concepts, Retrieval, and Services, https://doi.org/10.1007/978-3-031-32530-4_6

of an alphabet. Although some variations died out over time, the resulting Arabic, Coptic, Hebrew, Greek, and Latin alphabets all exist today in more modern forms.

Similarly, the characters of Chinese writing evolved over millennia from Shang dynasty pottery inscriptions and diffused in various forms to other Asian languages, resulting in more flexible forms of writing. Chinese character organization involves a logic which groups the roughly 214 "radicals" which are free-standing graphical character-words, by the number of strokes, and which can then be combined with other characters to create other words.

Alphabetical order in Western written culture was also established fairly early though not nearly as strictly enforced as in the modern sense, often confined only to the first letter of a word (Daly 1967). The Stoic philosopher Quintius Sextus originally organized his *Pythagorean Sentences* by the so-called "catch words" from each sentence, though later editors eventually rearranged these into alphabetical order by the first word of each sentence, demonstrating the gradual shift between mnemonic recital and visual retrieval, which would eventually culminate in the critical importance of "known order" in indexing (Witty 1973).

The Greeks inherited the order of the letters when they adopted and adapted an alphabet for their own language. Abcedaries for the use of children existed at least as early as the eighth century B.C. (West 2015). Goody (2000) argued that the alphabet was instrumental in the development of Greek philosophy, as writing can preserve unchanged opinions throughout time, and individuals can thus take a critical stance toward them, separating themselves from the inherited discourse. The earliest example of this can be found in Aristotle's *Metaphysics*, which reports the opinions of his predecessors as well as his own. The very rediscovery of the works of Aristotle in the twelfth century A.D. by the Christian intellectual world was only made possible by their written preservation in the Arabic world.

Like early Greek ritual, Greek law was originally oral in nature (Gagarin 2003). With the development of writing and the growth of Greek city-states, legal regulations gradually became distinguished from other unwritten rules and traditions by being permanently inscribed in stone and prominently displayed in public (2003, pp. 70–71). Similarly, Roman law underwent a transition from a performative set of legal actions in which the repetition of exact words and gestures, similar in nature to ritual religious practices, was essential for legality, to the creation and display of the Twelve Tablets stating the laws as they applied to all inhabitants of Rome (Meyer 2004). This was particularly important, in that most Graeco-Roman writing was done on wax tablets, which were easily erased and reused, and not generally intended for long-term storage (Small 1997).

Butler (2002) notes that the Roman public "spectacle" continued to include a variety of deliberate semiotic as well as semantic elements, making the concept of ancient "literacy" far more complex than commonly understood today. What MacMullen (1982) termed "the sense of audience" refers to the intended effect of Roman statues and monumental inscriptions in public places for both literate and illiterate spectators. The public

display by Mark Antony in the Forum of the severed head and hands of his assassinated rival Cicero, serving as graphic rebuttal of Cicero's spoken and written works on the Roman republic, was clearly intended to have a similarly dramatic effect on his audience. Ironically, Cicero's influence on the future, exerted through his consideration of both the epistemological and rhetorical role of "evidence" in his book, *Orator,* was to play an even greater role, as his writings became central to the classical foundations of grammar, logic, and rhetoric in medieval learning (Lancaster & Raiswell, 2018). Indeed, his letters to his friend Atticus also provided the first use of the term "scholia" specifically in reference to a gloss or comment contained within a particular manuscript (Reeve, 2016).

Authors and copyists of other early manuscripts began to organize their contents through the emergence of abstracts and tables of contents, though these are often omitted from their modern editions (Gibson, 2014). Witty (1973) dates the origins of abstracting from the hypotheses (plot summaries) of classical Greek plays as preserved in papyri at the Library of Alexandria. Pliny the Elder's *Natural History*, an early encyclopedic work of science, offered a "summarium" or table of contents as its Book I, but, as Doody (2010) notes, its purpose was more generally introductory to natural history than indexical to specific facts. However, Pliny's contemporary, the Greek military physician Dioscorides, did produce an herbarium, the *De Materia Medica*, which was specifically devoted to the pharmaceutical qualities of various plants, and was used as a medical reference source well into the medieval period. The only surviving complete manuscript (known as the *Codex Vindobonensis*) was produced in the sixth century and featured an alphabetical index to the plant names for ready reference (Discorides 2000).

This more specific indexical approach was further developed as the codex replaced the scroll across the Roman empire (Marchese 2012). Although there are various theories concerning this gradual replacement of the scroll by the codex, many having to do with its several advantages to the growing Christian textual community, which rapidly adopted and diffused the new format (Roberts and Skeat 1987; Skeat 1994) or with reasons for its more gradual acceptance by the elite in the larger population (Hartnett 2018), or with the artisanal innovativeness shown by reworking existing storage technologies, such as woven baskets, into containers for the earliest codex forms (Boudali 2018), there is general consensus that the "book" provided novel access points to its contents, important to those who needed to point quickly to particular, authoritative passages, ultimately giving it an advantage over other formats. As Kloppenborg (2014) notes, while certainly possible, repeatedly locating any particular section of interest in a scroll was far more difficult. A codex, however, allowed primitive versions of random access information retrieval through the use of bookmarks and marked book edges, for instance (Marchese 2012).

Similarly, in China, the use of writing on bamboo slats bound together into rolls, had become routine by the Warring States period (481–256 B.C.), and facilitated the voluminous works of the Han dynasty (206 B.C.–A.D. 220). Paper was invented there by the first century B.C. and facilitated production of a broad range of manuscripts. By

the eighth century (at the latest), the Chinese had invented xylography, the technology that reproduced text from characters cut in relief on wooden blocks. Developed first for the production of Buddhist texts, the technology gradually turned to the production of dictionaries, medical texts, almanacs, astrological, divination and geomancy manuals, and then to government printing, which established standard editions of the Confucian classics in the tenth century. By the end of that century, with the establishment of a text-centered civil service examination system as the primary means of official recruitment, literacy and mastery of classic texts became routinized within the state bureaucracy (Brokaw 2007, p. 254).

6.2 The Importance of Interpretation

Although Latin was not the only language in use as the Roman empire dissolved and was succeeded by the early medieval period, it remained the lingua franca of scholarship, providing a common framework for communication across Europe (McKitterick 1990), even as what was to become the vernacular English language transitioned through Norse, Anglo-Saxon, Old English, Norman French, Middle English, and modern English iterations, adopting and adapting words and grammatical constructions along the way (Nevalainen et al. 2020). Ecclesiastical Latin dominated medieval thinking, however, and the manuscript copying workshops (scriptoria) of the monastic houses, especially the Dominicans, were the major sources of Biblical texts and theological manuscripts.

The thirteenth-century Dominican John Balbi constructed the *Catholicon,* which helped to interpret Biblical words as well as Latin grammar for education within these monastic houses, while the role of "word-lists" defining specific terms became essential for those not otherwise schooled in Latin.

The thirteenth-century Dominican Vincent of Beauvais is credited with the encyclopedic work *Speculum Maius,* intended as a compendium of universal knowledge. De Beauvais's pioneering effort included the attachment of a list of topics in alphabetical order (tabulae) to each of the four parts of the *Speculum*, providing a "facilitatis inveniendi" that mapped particular book and chapter locations to each topic, as even the best-trained memory would be unable to retain that much knowledge, although, as Rouse and Rouse (1991, pp. 191–255) observe, a well-laid-out manuscript coupled with a well-trained memory remained essential for medieval scholars through the thirteenth century.

The use of "glosses" to explain or interpret an authoritative text, such as the Bible or the legal Code of Justinian, became more prevalent as well. These glosses were initially placed between the lines or in the margins of the original texts, but eventually some took a more independent form. Such collections as Lombard's *Four Books of Sentences*, which compiled authoritative statements on theologically important biblical passages and which marked the beginnings of systematic theology (Rouse and Rouse 1991, pp. 191–219),

and Gratian's *Decretum,* which compiled significant legal precedents from both canon and Roman law (Winroth 2021), the so-called *sedes materiae* (the seat of the matter), became foundational to medieval scholasticism in an attempt to harmonize the various authorities of their own tradition and to reconcile Christian theology with the philosophers of classical and late antiquity. The ultimate aim was to determine divine truth and human obligation as evidenced through these texts, which included the Bible, the writings of the Church Fathers, the works of certain philosophers such as Plato and Aristotle, and the two *corpora* of Roman and canon law.

Within medieval law, the so-called "Glossators" developed a fairly consistent and uniform system, based on the belief that numbers were too hard to remember and not suitable for the oral disputations in which lawyers frequently engaged. The earliest uniform practice of the Glossators was to give first a *siglum* (symbol) representing the work: after the *siglum*, the first few words of the *rubric* (a subject heading usually written in red ink) of the title were given, followed by the *initium* (the opening words) of the lex and, if necessary, of the paragraph. Numbers were only occasionally used. This early method, though with greater use of numbers, was also adopted by canon lawyers, especially after the compilation of the decretals by Gratian around 1140. One of the earliest methods to refer to a legal document was to quote the opening words or *incipit*, much as the Glossators used the initia. Writs often took names from their opening words, and this was one of the earliest methods of referring to statutes (Cooper 1982).

For other scholarly purposes, Parkes (1976) notes that the growing sophistication of the later medieval book can be traced to the rise of scholasticism in the medieval universities. Scholastic logic required an apparatus of thesis statements along with explanations of terms and the arguments involved, resulting in the creation of a more organized structure for books, including an accepted order of hierarchy for individual topics and their subtopics, and an increasing emphasis on both original texts and a growth of compilations and commentaries. The proliferation of compilations also promoted the development of the so-called "tabula," alphabetical indices to their contents, which were organized by selected and defined catchwords from the topics and subtopics. These tabula could refer to a single work or to multiple related works in the context of related arguments (1976, pp. 131–132). Medieval "authorship" of such compilations was also "editorial": it involved separating the subject matter hierarchically into *quaestiones* and *distinctiones*, but also involved the design of each page, using visual cues to mark divisions. Ideally, the handwritten book served as a physical mirror of the internal structure of its contents and the scholar's memory accordingly reflected what was read (1976, p. 121). The printing of the first Gutenberg Bible in 1441 using movable metal type was gradually to revolutionize such scholarly practices as well (Eisenstein 1983).

Alphabetical indices themselves were somewhat disdained by many scholars, presumably because they represented a more mechanical aid to memory than it was felt someone with truly scholarly mastery of an entire subject should require (Zedelmaier 2007). Such

retention through memory alone, however, became increasingly difficult with the "renaissance" of new knowledge into the West from the East in the form of new translations from ancient sources and novel works in such fields as mathematics and medicine (Nelles 2009). By the fourteenth century, if a manuscript was not supplied with this type of organized structure, scholarly readers were beginning to add some of the organizational apparatus specific to an individual copy, but, by the fifteenth century, even secular works often enjoyed an analytical table of contents, text disposed into books, chapters, and paragraphs, accompanied by footnotes and an index that often contained "see also" notations (Parkes 1976, pp. 134–135).

The growth of scholarship, facilitated by the so-called "Republic of Letters" representing an international correspondence network (which continued to be conducted mostly in Latin) among learned and aristocratic individuals in the West, is exemplified by the Dutch theologian Erasmus, who first popularized the term in his 1495 *Antibarbari* (*Book Against the Barbarians*), an argument for the embrace of classical languages and learning in order to enhance Christian intellectual endeavors (Burke 1999). This ongoing exchange of letters and manuscripts helped to bring new translations and new ideas for the humanities, mathematics, philosophy, and science into circulation across Europe.

In 1545, Conrad Gessner, the "father of bibliography," listed, annotated, and evaluated 10,000 works in his *Bibliotheca Universalis* (Nelles 2009, p. 150). The *Bibliotheca* contained two alphabetical lists of its authors and an index. The second volume, the *Pandectae* (1548), contains a section with a list of keywords, ordered thematically to the first volume, as well the first known instructions on index construction ("De indicibus librorum") intended to help scholars process textual material. Gessner recommended that, during the reading process, anything that appeared important or useful was to be written on a piece of paper as it was encountered, later to be cut into individual slips, and arranged in any desired order, after which they could be copied, or pasted onto another sheet, or inserted into a special binding for purposes of retention (Zedelmaier 2007). Krajewski (2011) points out that these cards can be considered as forming an early "paper machine" that facilitated random information retrieval from multiple books, similar to but more powerful than the index within a single book, but not as powerful as the computing machines to come.

6.3 The Literacy Revolution

Meanwhile, vernacular languages began to replace Latin on the continent as well as in England, necessitating more vernacular words to express various concepts, often called "hard words" because they were both novel and complex. The phrase "hard words" would appear to have been used for the first time in the title of John Day's glossary *A gatheryng of certayne harde wordes in the newe Testament*, *with their exposicion* (1551), translated from the expression *mot difficiles* used in the original French title of this work, which was

one of the study aids produced in advance of the famous Geneva Bible of 1560, featuring an extensive apparatus of indices and commentaries for the Protestant reader. The use of the phrase "hard words" later came to carry a political connotation, indicating their borrowed rather than naturalized status as English words, indicating to some observers that these were terms used by the elite and therefore should be adopted and to others that they were foreign and therefore suspect. The rapid increase in the English lexicon during the sixteenth and seventeenth centuries, due to the adoption of foreign terms, the adaptation of foreign terms to create their English equivalents, and the need for new terms to indicate new ideas in religion, science, geography, medicine, and literature eventually generated a term of its own: the "inkhorn" controversy over the use of such words (Durkin 2014, pp. 317–319).

While there continued to be marked differences in education between the landed (Levy 1982) and the laboring (Laqueur 1976) classes, the rudiments of literacy became increasingly available to more people over time. Accounts of witchcraft trials, for instance, were aimed at both literate and semi-literate audiences (Suhr 2012). Written handbooks for the use of magic and witchcraft and for their discovery and prevention became available, though the grimoires (a term used for occult handbooks) were usually handwritten and carefully concealed from other readers (Orgel 2017) while the witch-finders' handbooks such as the celebrated *Hammer of Witches* became increasingly popular as they found their way into print (Broedel 2010). It should be remembered that expressing disbelief in demonology and witchcraft even during the early modern era might also provide cause for accusations of heresy (Clark 1997). Church inquisitors were notoriously adept at using their extensive records of interrogations in order to entrap victims (Sherwood 2012).

While heterodox as well as various heretical beliefs gained access to written as well as oral channels of communication, and more of the population became literate, both the content and the indices of works also began to provide different perspectives on what might be considered knowledge (Biller 1994). Censorship by both church and state continued to pose an ongoing danger to many forms of innovative thought, ranging from medicine (Marcus 2020) to politics (Loades 1974) to science (Marcus and Findlen 2019) and, of course, to theology (Watson 1995). Even translators of such works ran the risk of imprisonment and execution. Although some have argued that that the invention of the printing press in the early fifteenth century was not the unique genesis of the so-called "print culture" (Love 2003), printing dramatically impacted the transmission of knowledge, also making it easier to censor certain works such as the Talmud and rabbinical scholarship (Morsel-Eisenberg 2021) and to broaden the circulation of others, such as Arthurian legends and classic myths (Wang 2004). It also helped to facilitate the domination of non-Western cultures through colonialism, as the printed book became associated with the privileging of Western forms of literacy over other forms of writing (Mignolo 1995).

Printing allowed the alphabetical index to assume additional importance, in that what was originally intended as an aid to personal memory for an individual using a particular manuscript became a contribution to social memory, as the format of printed books became more uniform than their handwritten predecessors. Cevolini (2014) analyzes the evolution of indexing itself and finds it among the most important knowledge management tools of the typographic age, showing how both cognitive performance and social media can be shaped by communication media.

Cevolini comments that "The transition from manuscript to printed texts triggered a functional change that worked as a selective force on indexing procedures, turning a mnemotechnical aid into a search engine of virtual memories" (2014, p. 51). Now a more standardized, less idiosyncratic index could allow any reader not only to compensate for limitations imposed by a failure to remember where in the text a specific word or passage of note occurred, but also offered an opportunity to innovate through the discovery of particular words or passages not reached through typical linear reading. In other words, the evolution of the index was from a memory tool relevant to a pre-selected compilation of authorized texts to a finding tool for novel ideas. The art of memory gradually gave way to a science of information retrieval, as the amount of potential knowledge became, in Blair's phrase "too much to know" (Blair 2010) for any one individual, and card indices, archives, and libraries all became more prevalent over time.

These technologies of deixis served as black boxes of exteriorized, virtual memory for those with both access and appropriate knowledge management skills. The "art of semantic association as practised by orators was now replaced by the somewhat mechanical procedure of the alphabetical register, leaving users to take their bearings from the terms from which the index was constructed through semantic paths. Information retrieval took place in a 'methodical' manner, through clear and simple rules that allowed for a progressive increase in knowledge without weighing down the mind" (Cevolini 2014, p. 54). The evolution of these technologies of deixis also gradually created what we might term a "new world order" in which printing only gradually displaced handwritten manuscripts, borrowing many of their existing structures while standardizing and lending more precision to such formats as page numbering, and revolutionizing the social production and consumption of knowledge, as the "key words" from the variety of chapters and titles that formed the infrastructure of both book indices and library catalogs allowed more freedom in forming and informing new thoughts and different ways of thinking (Dover 2021; Eisenstein 1983; McKitterick 2003).

While records administration practices by church and state varied widely by region, they too were generally becoming more sophisticated, as documents were perceived as increasingly essential to legal and political operations. Registries were set up to track the flow of documents related to particular areas of government, letter-books were created to contain copies of the documents with notes regarding their contents and any actions taken in response, chronological order was maintained by binding these letter-books into codices to improve access, separate inventories were developed for additional security

against forgery and theft, and special cabinets were built to house these in various official repositories (Head 2016).

The slow shift toward both intellectual and political freedom also manifested itself in a growing interest in news, as traditional oral news transmission methods (village gossip, town criers announcing royal edicts, sung ballads about notorious events, etc.) became supplemented by written communications, such as the coranto broadsheets that began to appear around 1620 providing updates on European battles and politics and which were the predecessors of newspapers a century later. Increasing literacy fueled this growth as well, provoking an increase in censorship on printed materials of all types, though these attempts at news censorship were often unsuccessful (Espejo 2011). Pamphlets and newspapers attempted to stress their novelty, credibility, and timeliness to prospective audiences through title page and paragraph design, and eventually through the creation of newspaper headlines, while their audiences began to respond more actively to what they were learning about events, governance, and society (Barker 2014).

Such different ways of thinking gradually came to include more scientific habits of mind and more politically charged forms of expression, though both were deeply rooted in earlier cognitive endeavors. For individuals, the keeping of the so-called "commonplace books" did indeed become commonplace, as the process was gradually transformed from an earlier scholastic endeavor in which pupils were rigorously drilled in the still-dominant classic traditions by transcribing selected Greek and Latin extracts into elegant English prose in order to learn the principles of rhetoric and argument, into a much more individualized one in which both men and women transcribed various literary, poetic, or religious quotations of particular interest and the thoughts that these engendered into their personal notebooks for future reflection. Erasmus wrote *De Copia* in 1512, which suggested extensive reading, note-taking, and careful thought would facilitate this new type of learning from a variety of printed sources. Philosopher John Locke's 1706 *A New Method of Making Common-Place Books* also urged the use of consistent order in constructing these, with the goals of better of understanding principles of organization and better retention of the contents. The compilation of commonplace books continued to provide structure for students as they absorbed familiar quotations and useful pieces of knowledge, though without the early emphasis on classic authors and rhetorical arts as the ultimate model for thinking (Moss 1996). Mass literacy required different modes of teaching and learning than did earlier forms of elite education (Vincent 2000).

As Lancaster and Raiswell (2018, p. 3) point out, "What is distinct about the early modern period is that the cultural, social, and intellectual spaces in which knowledge was generated and the ends to which it was brought were changing as new groups of people developed new methods to solve a different set of problems. As these people came to assert increasingly more control of the intellectual agenda in venues outside those traditionally associated with knowledge production in the Middle Ages and Renaissance, they formulated new conceptions of what constituted acceptable proof, tested it in new social and intellectual contexts, and found new roles for it in their arguments and worldviews.

… In this way, the empirical, mathematical, geographical, natural philosophical, and historical mixed with the authority of classical antiquity, Scripture, and analogous textual traditions, along with the hermetic, legal, and medical, to beget new conceptions of evidence. … Although the early modern period witnessed the notion of evidence as it would later come to be understood in the modern sciences emerge, this process was far from direct, and the form it took was anything but inevitable."

For instance, the "science" of demonology that dominated the early modern period "was a composite subject consisting of discussions about the workings of nature, the processes of history, the maintenance of religious purity, and the nature of political authority and order" (Clark 1997, p. viii). Caciola and Sluhovsky argue that earlier processes of "discernment" of the preternatural world through the investigation of such supposed phenomena as "witches' marks" eventually paved the way for a more empirical approach to the natural world by both physicians and scientists, while Daston (1998, p. 72) notes "The most impressive machinery of proof and legitimation in early modern Europe was to be found not in science but in religion, in the elaborate and refined procedures used by ecclesiastical authorities to investigate miracles. It is not an accident that Hume developed his rules for the evaluation of the evidence of testimony and things in the context of an inquiry into miracles." This new focus on scientific enquiry as a method of establishing credible evidence, moreover, helped to create appreciation for more accurate sources of information of all kinds within the emerging "public sphere" in Western societies (Dooley 2001).

6.4 The Language of Science and the Science of Language

In his magisterial overview of language changes during the early modern era, Wootton (2015) notes that The Royal Society of London for Improving Natural Knowledge quickly disclaimed any interest in investigating such things as demonic possession or miraculous powers and its formation in 1660 is considered to mark the start of organized scientific inquiry. However, even existing scientific language could be considered suspect: for instance, alchemy provided much of the foundation for chemistry in the words that were used to describe chemical processes (Nummedal 2011). The Society's first secretary, John Wilkes, embarked on a project to provide a new classificatory language for science and technology, which was intended to provide a means of universal communication for scientists and scholars that would be superior to Latin and all other languages in that it would be based on universal laws that account for the natural world, and the "things and notions" that are represented by them (Schulte-Albert 1979). Despite having intrigued such luminaries as Locke, this project never came to full fruition and, instead, in 1704, the Royal Society's other secretary, John Harris, compiled the first English-language encyclopedia, the *Lexicon technicum*, which also served as a dictionary of mathematics, mechanical arts, and natural sciences (Lonati 2007).

The gradual development of scientific and technical vocabularies by working scientists continued throughout the early modern era, shown by examples such as Lavoiser's early attempt to systematize chemical nomenclature (Wloch, 2015), Linnaeus's binomial nomenclature for botany (Müller-Wille and Charmantier 2012), and Messier's catalog of astronomical objects (Dick 2013). The development of a scientific vocabulary and associated classificatory terms is generally considered the hallmark of an emerging scientific field as they facilitate the exchange of knowledge, even as both vocabularies and classifications are subject to change as the particular science advances in knowledge. Starting in 1665, learned journals such as the Royal Society's *Philosophical Transactions* in England and the *Journal des Scavans* in France provided important channels of communication for this flood of new scientific information.

The modern encyclopedic genre dates from the 1700s, with the stated goal of containing an encapsulated summary of all human knowledge as a substitute for multiple books within a library, demonstrated by examples such as Chambers' 1728 *Cyclopedia*, Diderot's 1751 *Encyclopédie*, and Smellie's 1768 *Encyclopedia Britannica*. These were all commercial multi-volume publications aimed at a wide public readership, providing an ambitious range of subjects that went beyond traditional academic disciplines, and usually offered an alphabetical arrangement of topics rather than the traditional systems imposed by academic or theological hierarchies (Yeo 2007). These efforts were massive: for instance, while the *Lexicon Technicum* contained roughly a million words, the *Cyclopedia* contained 3 million, the *Encyclopedia Britannica* contained 10 million, and the *Encyclopédie* 19 million (Loveland 2012).

An obvious problem for the encyclopedia's user was the difficulty of locating specific topics within a more general alphabetized list of subjects. Chambers addressed the problem of alphabetized headings and terms by providing a preliminary "View of Knowledge" diagram, based on commonplace book organizational schemes with which encyclopedia readers could be presumed to be familiar. Each subject shown in the diagram was linked to a footnote containing a list of the terms belonging to it, so that, with cross-references, the reader could reconstitute a science that had been scattered alphabetically. The diagram showed 47 headings, numbered simply according to their position on the diagram. The notes attached to each art or science show the cognate terms belonging to it in order to help the reader locate a specific topic within various headings, and also a listing of the suggested order in which these topics should be read. Roughly half of the articles also contained cross-references to other relevant articles (Yeo 2003). An alternative solution was found by the editors of the *Britannica*, which included individual indices to its lengthier subject treatments for its second edition.

The boundaries between encyclopedias and dictionaries have often been viewed as porous but the development of lexicography as a separate field helped to further separate the two (Hancher 2019), even while the nature of "evidence" and the importance of "organization" also clearly impacted the emerging study of words and their usage (Mugglestone 2015). While medieval word-lists, such as the lists of Latin words with their

vernacular equivalents for the use of students in monasteries and later in the universities, had been common, the decline of Latin as a universal language prompted new lexicographic endeavors for vernacular languages, beginning with John Florio's 1598 *Worlde of Words*, which attempted to singlehandedly if unsuccessfully undertake for the English language what the contemporaneous scholarly groups of the Academie Francaise and the Accademia della Crusca were attempting to legislate for the French and Italian languages respectively (Considine 2014), expanded both by Thomas Blount's exhaustive 1656 *Glossographia* and by John Kersey's 1702 emphasis on "common" rather than "hard" words (Lancashire 2005), and refined by Samuel Johnson's 1755 *Dictionary*, which proposed to identify both common and important words as defined by distinguished writers and acknowledged specialists (Osselton 2006).

The lexicographers who compiled the dictionaries of the seventeenth and eighteenth century believed that the oldest meaning of a word was the most literal, which informed much of their efforts to trace and record older uses of each word and to arrange them in a coherent order. Although this belief was incorrect, it did help to systematize the practice of dictionary-making (Hanks 2020). Roget's *Thesaurus* represents the best known in a series of earlier attempts to depict the connections among various words through the organization of synonyms rather than through alphabetization (Hüllen 2003).

Kreidler (1998) points out that another dominant viewpoint among lexicographers was that language diversity was undesirable, and that attempts should be made to "fix" language into a norm to which all educated individuals should conform. In the United States, Noah Webster subscribed to these ideas, hoping the diffusion of correct usage would gradually destroy the differences of dialect which had been brought from the British Isles. Deeply influenced by the evangelical Christianity of the day, his 1828 dictionary's preface incorporated the idea that that all languages are descended from the biblical "Chaldee" and form three groups, Hamitic, Semitic, and Japhetic, to which English belonged.

As dictionaries evolved, it became clear that they faced continual challenges as words became obsolete, changed meaning, or were replaced or joined by others reflecting new concepts. The usefulness of dictionaries, as well as their prevalence, became vested not only in their authoritativeness but in their comprehensiveness. The *Oxford English Dictionary*, work on which began in 1857, is well known for its empirical approach, soliciting input from a wide variety of contributors rather than relying on linguists and grammarians (Winchester 2003) while still enforcing its own editorial discretion in adjudicating among words and their definitions (Mugglestone 2005).

The diffusion of modern technologies can be tracked by the initial dates in the dictionary for the published use of such neologisms as newspaper (1667), bureaucracy (1759), telegraph (1793), steamship (1819), railroad (1822), analytical engine (1843), telephone (1864), and typewriter (1868), all of which helped to accelerate the global growth of communication. For instance, newspaper headlines increasingly served as "relevance optimizers" in which to mobilize public opinion and actions (Dor 2003) while the telegraph accelerated involvement in international affairs (Yeh 2015). Such key words from the

nineteenth century such as "colonization," "Dred Scott decision," and "l'affaire Dreyfus" continue to resonate today, just as today's keywords may do for future students and researchers in those dictionaries and other textual corpora currently being compiled.

Lexicography itself, while retaining many of its traditional methods even as it has transitioned to an almost completely electronic environment, is gradually being reconceptualized as a result of new lexicographic theories, massively increased availability of textual corpora and new ways of analyzing these, and the influx of ideas from other disciplines (Leroyer and Simonsen 2020).

The next chapter will describe the role of information, particularly scientific and technological information, as it contributed to commercial competition and the growth of nation-states.

References

Barker, H. (2014). *Newspapers, politics and English society, 1695–1855*. Routledge.

Biller, P. (1994). Heresy and literacy: Earlier history of the theme. In P. Biller & A. Hudson (Eds.), *Heresy and literacy 1000-1530* (pp. 1–17). Cambridge University Press.

Blair, A. M. (2010). *Too much to know: Managing scholarly information before the modern age*. Yale University Press.

Boudali, G. (2018). *The codex and crafts in late antiquity*. University of Chicago Press.

Broedel, H. P. (2010). *The Malleus Maleficarum and the construction of witchcraft: Theology and popular belief*. Manchester University Press.

Brokaw, C. J. (2007). Book history in premodern China: The state of the discipline I. *Book History*, *10*(1), 253–290. https://www.jstor.org/stable/30227406.

Burke, P. (1999). Erasmus and the Republic of Letters. *European Review*, *7*(1), 5–17. https://doi.org/10.1017/S1062798700003690.

Butler, S. (2002). *The hand of Cicero*. Routledge.

Cevolini, A. (2014). Indexing as preadaptive advance: A socio-evolutionary perspective. *The Indexer*, *32*(2), 50–57. https://doi.org/10.3828/indexer.2014.19.

Clark, S. (1997). *Thinking with demons: The idea of witchcraft in early modern Europe*. Oxford University Press.

Considine, J. (2014). *Academy dictionaries 1600–1800*. Cambridge University Press.

Cooper, B. D. (1982). Anglo-American legal citation: Historical development and library implications. *Law Library Journal*, *75*(3), 3–33. https://www.repository.law.indiana.edu/facpub/1745.

Daly, L. W. (1967). *Contributions to a history of alphabetization in antiquity and the Middle Ages*. Latomus.

Daniels, P. T. (2007). Littera ex occidente: Toward a functional history of writing. In C. L. Miller (Ed.), *Studies in Semitic and Afroasiatic linguistics presented to Gene B. Gragg* (pp. 53–68). Oriental Institute of the University of Chicago.

Daston, L. (1998). The nature of nature in early modern Europe. *Configurations*, *6*(2), 149–172. https://doi.org/10.1353/con.1998.0014.

Dick, S. J. (2013). *Discovery and classification in astronomy: Controversy and consensus*. Cambridge University Press.

Discorides. (2000). *De materia medica: Five books in one volume* (T. A. Osbaldeston, Trans.). IBIDIS Press.

Doody, A. (2010). *Pliny's Encyclopedia: The reception of the Natural History*. Cambridge University Press.

Dooley, B. (2001). News and doubt in early modern culture: Or, are we having a public sphere yet? In B. Dooley & S. A. Baron (Eds.), *The politics of information in early modern Europe* (pp. 275–290). Routledge.

Dor, D. (2003). On newspaper headlines as relevance optimizers. *Journal of Pragmatics*, *35*(5), 695–721. https://doi.org/10.1016/S0378-2166(02)00134-0.

Dover, P. M. (2021). *The information revolution in early modern Europe*. Cambridge University Press.

Durkin, P. (2014). *Borrowed words: A history of loanwords in English*. Oxford University Press.

Eisenstein, E. (1983). *The printing revolution in early modern Europe*. Cambridge University Press.

Espejo, C. (2011). European communication networks in the early modern age: A new framework of interpretation for the birth of journalism. *Media History*, *17*(2), 189–202. https://doi.org/10.1080/13688804.2011.554730.

Gagarin, M. (2003). Letters of the law: Written text in archaic Greek law. In H. Yunis (Ed.), *Written texts and the rise of literate culture in ancient Greece* (pp. 59–96). Cambridge University Press.

Gibson, R. (2014). Starting with the index in Pliny. In L. Jansen (Ed.), *The Roman paratext: Frames, texts, readers* (pp. 33–55). Cambridge University Press.

Goody, J. (2000). *The power of the written tradition*. Smithsonian Press.

Hancher, M. (2019). Dictionary vs. encyclopedia, then and now. *Dictionaries: Journal of the Dictionary Society of North America*, *40*(1), 113–138. https://doi.org/10.1353/dic.2019.0003.

Hanks, P. (2020). The unmasking of English dictionaries by R. M. W. Dixon (review). *Language*, *96*(2), 460–462. https://www.linguisticsociety.org/sites/default/files/12_96.2Hanks.pdf.

Hartnett, B. (2018). The diffusion of the codex. *Classical Antiquity*, *36*(2), 183–235. https://doi.org/10.1525/ca.2017.36.2.183.

Head, R. C. (2016). Configuring European archives: Spaces, materials and practices in the differentiation of repositories from the late Middle Ages to 1700. *European History Quarterly*, *46*(3), 498–518. https://doi.org/10.1177/0265691416648530.

Hüllen, W. (2003). *A history of Roget's Thesaurus: Origins, development, and design*. Oxford University Press.

Klein, W. (2015). Lexicology and lexicography. In *International encyclopedia of the social & behavioral sciences* (2nd ed., Vol. 13, pp. 938–942). Elsevier.

Kloppenborg, J. S. (2014). Literate media in early Christ groups: The creation of a Christian book culture. *Journal of Early Christian Studies*, *22*(1), 21–59. https://doi.org/10.1353/earl.2014.0004.

Krajewski, M. (2011). *Paper machines: About cards & catalogs, 1548-1929*. MIT Press.

Kreidler, C. W. (1998). Noah Webster's linguistic influences. *Language & Communication*, *18*(2), 101–110. https://doi.org/10.1016/S0271-5309(98)00004-4.

Lancashire, I. (2005). Johnson and seventeenth-century English glossographers. *International Journal of Lexicography*, *18*(2), 157–171. https://doi.org/10.1093/ijl/eci018.

Lancaster, J. A. T., & Raiswell, R. (2018). Evidence before science. In J. A. T. Lancaster & R. Raiswell (Eds.), *Evidence in the age of the new sciences* (pp. 1–29). Springer.

Laqueur, T. (1976). The cultural origins of popular literacy in England 1500-1850. *Oxford Review of Education*, *2*(3), 255–275. https://doi.org/10.1080/0305498760020304.

Leroyer, P., & Simonsen, H. K. (2020). Reconceptualizing lexicography: The broad understanding. In Z. Gavriilidou, M. Mitsiaki, & A. Fliatouras (Eds.), *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion* (Vol. 1, pp. 183–192). Ljubljana University Press.

Levy, F. J. (1982). How information spread among the gentry, 1550-1640. *Journal of British Studies*, *21*(2), 11–34. https://doi.org/10.1086/385788.

Loades, D. M. (1974). The theory and practice of censorship in sixteenth-century England. *Transactions of the Royal Historical Society*, *24*, 141–157. https://doi.org/10.2307/3678936.

Lonati, E. (2007). Blancardus' Lexicon medicum in Harris's Lexicon technicum: A lexicographic and lexicological study. In J. Considine & G. Iamartino (Eds.), *Words and dictionaries from the British Isles in historical perspective* (pp. 91–108). Cambridge Scholars Publishing.

Love, H. (2003). Early modern print culture: Assessing the models. *Parergon*, *20*(1), 45–64. https://doi.org/10.1353/pgn.2003.0071.

Loveland, J. (2012). Why encyclopedias got bigger ... And smaller. *Information & Culture*, *47*(2), 233–254. https://doi.org/10.1353/lac.2012.0012.

MacMullen, R. (1982). The epigraphic habit in the Roman Empire. *American Journal of Philology*, *103*, 233–246. https://doi.org/10.2307/294470.

Marchese, F. T. (2012). Tables and early information visualization. In F. T. Marchese & E. Banissi (Eds.), *Knowledge visualization currents: From text to art to culture* (pp. 35–61). Springer-Verlag.

Marcus, H. (2020). *Forbidden knowledge: Medicine, science, and censorship in early modern Italy*. University of Chicago Press.

Marcus, H., & Findlen, P. (2019). Deciphering Galileo: Communication and secrecy before and after the trial. *Renaissance Quarterly*, *72*(3), 953–995. https://doi.org/10.1017/rqx.2019.253.

McKitterick, D. (2003). *Print, manuscript and the search for order, 1450-1830*. Cambridge University Press.

McKitterick, R. (Ed.). (1990). *The uses of literacy in early medieval Europe*. Cambridge University Press.

Meyer, E. A. (2004). *Legitimacy and law in the Roman world: Tabulae in Roman belief and practice*. Cambridge University Press.

Mignolo, W. D. (1995). *The darker side of the Renaissance: Literacy, territoriality, and colonization*. University of Michigan Press.

Morsel-Eisenberg, T. (2021). Anxieties of transmission: Rabbinic responsa and early modern "print culture." *Journal of the History of Ideas*, *82*(3), 377–404. https://doi.org/10.1353/jhi.2021.0019.

Moss, A. (1996). *Printed commonplace-books and the structuring of Renaissance thought*. Oxford University Press.

Mugglestone, L. (2005). *Lost for words: The hidden history of the Oxford English Dictionary*. Yale University Press.

Mugglestone, L. (2015). *Samuel Johnson and the journey into words*. Oxford University Press.

Müller-Wille, S., & Charmantier, I. (2012). Natural history and information overload: The case of Linnaeus. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 4–15. https://doi.org/10.1016/j.shpsc.2011.10.021.

Nelles, P. (2009). Reading and memory in the universal library: Conrad Gessner and the Renaissance book. In D. Beecher & G. Williams (Eds.), *Ars Reminiscendi: Mind and memory in Renaissance culture* (pp. 147–169). Centre for Reformation and Renaissance Studies, ITER Press.

Nevalainen, T., Säily, T., Vartiainen, T., Liimatta, A., & Lijffit, J. (2020). History of English as punctuated equilibria? A meta-analysis of the rate of linguistic change in Middle English. *Journal of Historical Sociolinguistics*, *6*(2), 20190008. https://doi.org/10.1515/jhsl-2019-0008.

Nummedal, T. E. (2011). Words and works in the history of alchemy. *Isis*, *102*(2), 330–337. https://doi.org/10.1086/660142.

Orgel, S. (2017). Secret arts and public spectacles: The parameters of Elizabethan magic. *Shakespeare Quarterly*, *68*(1), 80–91. https://doi.org/10.1353/shq.2017.0004.

Osselton, N. E. (2006). Usage guidance in early dictionaries of English. *International Journal of Lexicography*, *19*(1), 99–105. https://doi.org/10.1093/ijl/eci053.

Parkes, M. B. (1976). The influence of the concepts of ordinatio and compilatio on the development of the book. In J. J. G. Alexander & M. T. Gibson (Eds.), *Medieval learning & literature: Essays in honor of Richard William Hunt* (pp. 115–141). Clarendon Press.

Reeve, M. D. (2016). Scholia. In T. Whitmarsh (Ed.), *Oxford classical dictionary*. Oxford University Press. https://doi.org/10.1093/acrefore/9780199381135.013.5738.

Roberts, C. H., & Skeat, T. C. (1987). *Birth of the codex*. Oxford University Press/British Academy.

Rouse, M. A., & Rouse, R. H. (1991). *Authentic witnesses: Approaches to medieval texts and manuscripts*. University of Notre Dame Press.

Schulte-Albert, H. G. (1979). Classificatory thinking from Kinner to Wilkins: Classification and thesaurus construction, 1645-1668. *Library Quarterly*, *49*(1), 42–64. https://doi.org/10.1086/600900.

Sherwood, J. (2012). The inquisitor as archivist, or surprise, fear, and ruthless efficiency in the archives. *The American Archivist*, *75*(1), 56–80. https://doi.org/10.17723/aarc.75.1.a2712l7ur075j10h.

Skeat, T. C. (1994). The origin of the Christian codex. *Zeitschrift für Papyrologie und Epigraphik*, *102*, 263–268. https://www.jstor.org/stable/20189148.

Small, J. P. (1997). *Wax tablets of the mind: Cognitive studies of memory and literacy in classical antiquity*. Routledge.

Suhr, C. (2012). Publishing for the masses: Early modern English witchcraft pamphlets. *Neuphilologische Mitteilungen*, *113*(1), 118–121.

Vincent, D. (2000). *The rise of mass literacy: Reading and writing in modern Europe*. Blackwell Publishing.

Wang, Y.-C. (2004). Caxton's romances and their early Tudor readers. *Huntington Library Quarterly*, *67*(2), 173–188. https://www.jstor.org/stable/https://doi.org/10.1525/hlq.2004.67.2.173.

Watson, N. (1995). Censorship and cultural change in late-medieval England: Vernacular theology, the Oxford translation debate, and Arundel's Constitutions of 1409. *Speculum*, *70*(4), 822–864. https://doi.org/10.2307/2865345.

West, W. C. (2015). Learning the alphabet: Abecedaria and the early schools in Greece. *Greek, Roman, and Byzantine Studies*, *55*(1), 52–71. https://grbs.library.duke.edu/article/view/15231/6549.

Winchester, S. (2003). *The meaning of everything: The story of the Oxford English Dictionary*. Oxford University Press.

Winroth, A. (2021). Gratian and his book: How a medieval teacher changed European law and religion. *Oxford Journal of Law and Religion*, *10*(1), 1–15. https://doi.org/10.1093/ojlr/rwab003.

Witty, F. J. (1973). The beginnings of indexing and abstracting: Some notes towards a history of indexing and abstracting in antiquity and the Middle Ages. *The Indexer*, *8*(4), 193–198. https://doi.org/10.3828/indexer.1973.8.4.1.

Wloch, J. (2015). The language of chemistry: A study of English chemical vocabulary. *Beyond Philology*, *12*, 77–106. https://fil.ug.edu.pl/sites/default/files/_nodes/strona-filologiczny/33797/files/beyond_philology_no_12_2015.pdf#page=77.

Wootton, D. (2015). *The invention of science: A new history of the scientific revolution*. Harper Collins.

Yeh, P. E. (2015). The role of the Zimmermann telegram in spurring America's entry into the First World War. *American Intelligence Journal*, *32*(1), 61–64. https://www.jstor.org/stable/26202105.

Yeo, R. R. (2003). A solution to the multitude of books: Ephraim Chambers's Cyclopaedia (1728) as "the best book in the universe." *Journal of the History of Ideas*, *64*(1), 61–72. https://doi.org/10.1353/jhi.2003.0020.

Yeo, R. R. (2007). Lost encyclopedias: Before and after the Enlightenment. *Book History*, *10*(1), 47–68. https://www.jstor.org/stable/30227399.

Zedelmaier, H. (2007). Facilitas inveniendi: The alphabetical index as a knowledge management tool. *The Indexer*, *25*(4), 235–242. https://doi.org/10.3828/indexer.2007.29.

Progress, Public

7

Abstract

This chapter describes the emphasis on organization and efficiency that resulted from the rise of nation-states and international commercial competition during the latter part of the nineteenth century and how this impacted the growth of science and technical knowledge and the operations of libraries, especially advances in bibliographic cooperation and indexing technologies.

While the rise and dominance of the factual in modern times have often been separately explained as a result of economic advances (Poovey 1998), educational structures (Vincent 2000), political pressures (Buzan and Lawson 2015), scientific progress (Kitcher 1993), and societal conditioning (Latour 2012), all of these clearly played a part in the specific directions taken in the creation, collection, and dissemination of public knowledge.

7.1 The Growth of Governments and the Push Toward Progress

The growth of governments throughout the latter part of the nineteenth century increasingly required the utilization of sets of standardized and aggregated facts such as national censuses (Scott 1998). Census methodologies had developed over time, shifting from local, to regional, to national levels, from simple counts of all residents to nominal listings of household heads with numeric information about other household residents, then to census lists with individual names and data for all persons. The national information requirements for military recruitment, imposition of taxes, civic activities, demographic trends, and provision of social services grew more sophisticated, with modernized census techniques being widely employed by the 1840s, though attempts at global standardization

B. V. V. Martens, *Keywords In and Out of Context*, Synthesis Lectures on Information Concepts, Retrieval, and Services, https://doi.org/10.1007/978-3-031-32530-4_7

of census measurements through such efforts as the International Statistical Conferences, held from 1853 through 1872, failed due to resistance by individual governments to interference with their own procedures (Emigh et al. 2016). Krajewski (2011, pp. 27–32) makes an intriguing analogy between the era's emphasis on the improvement of government censuses and the improvement of library collections in terms of the somewhat similar problems in identifying and locating individual persons and in identifying and locating individual books as both populations were rapidly growing.

Book and periodical indexing were becoming professionalized with the increase in commercial publishing, the growth of national and specialized libraries, and the formation of professional library associations with their emphasis on bibliographic description and cataloging. Subject access in particular had been problematic, as the organization of any particular library's contents tended to be local and often idiosyncratic. Even those so-called "classified catalogs" in printed format that listed a library's contents in a hierarchical arrangement by subject used very broad categories, making the identification of books on specific subjects increasingly difficult as the collection grew, further highlighting issues of inefficiency within library processes.

In 1858, British publisher Samson Low released a retrospective catalog of all British books printed between 1837 and 1857. This included several indices, one based on title-term indexing, which identified the subject matter of a book by catchwords extracted from its title. His assistant, Andrea Crestadoro, described this method in an 1856 pamphlet, *The Art of Making Catalogues of Libraries: Or, a Method to Obtain in a Short Time a Most Perfect, Complete, and Satisfactory Printed Catalog of the British Museum Library*, although it was never in fact adopted by the British Museum Library, which was cataloging its collection at the time under library director Antonio Panizzi's *91 Rules for Compilation of the Catalogue*: a project which lasted a staggering 40 years.

Although title-term indexing was clearly a precursor to keyword indexing, the method was quickly overshadowed by advances in other forms of subject classification for libraries, notably American Charles Cutter's 1876 *Rules for a Dictionary Catalog*, which solved one of the major problems in title-term indexing, which was the lack of a way to identify synonyms for specific title terms. Cutter's development of a subject authority file was clearly superior to title-term indexing in regard to finding linkages to important variations in terminology and eventually formed the basis for the Library of Congress subject headings classification system.

However, as Griffiths (2015, p. 135) points out, "There is an inherent tension between the power of an expansive index to coordinate a wide range of materials and the limitations of the embedded procedures that institutions use to decide what to index and how. Expansions in scale bring this tension into crisis, as the increasing quantity of sources demands qualitatively different procedures to draw them together. This problem of the scalar transformation of print materials gained particular focus in the nineteenth century, as institutions including the British Museum, the French Bibliothèque Nationale, and the

rapidly growing American Library of Congress built collections that surpassed a million items by the close of the century."

A start toward global bibliographic cooperation was made at the two international Conferences of Librarians, the first held in New York in September of 1877 and the second in London in October of that same year. Topics of discussion included the recent appearance of such professional library publications as Cutter's *Rules*, Dewey's 1876 *A Classification and Subject Index for Cataloguing and Arranging the Books and Pamphlets of a Library*, and the first issue of *Library Journal*, with its call for international cooperation through shared cataloging and indexing and the development of standardized formats, as expressed both through British librarian J.A. Cross's call for "A Universal Index of Subjects" to be compiled from all existing published subject bibliographies, plus the subject cards or lists from all British libraries, and American librarian William Poole's call for both American and English libraries to contribute individual indices from their own journal holdings to be incorporated into his *Index to Periodical Literature* (Gambee 1967). Neither suggestion was successfully implemented at the time, though innovations such as H.B. Wheatley's 1878 guide to indexing, *What is an Index*? (Bell 2001), John Shaw Billings's 1879 *Index Medicus*, a regularly published index of important recent medical books and journal articles (Marson 1969), and John Butler Johnson's 1884 *Engineering Index* abstracts (Landau 1984) soon began to appear.

The so-called search for order in the late nineteenth century and early twentieth century was also driven by new pragmatic and progressive ideas in politics that actively promoted the use of scientific findings and technological advances intended to improve social well-being (Wiebe 1967). Informed by these ideals, American librarian Henry Bliss developed his Bliss Bibliographic Classification as a socially conscious alternative to the Dewey Decimal System (Broughton 2008) though it was not widely adopted in the United States. However, the now-widely used term "knowledge organization" was also first used by Bliss in his two seminal books, *The Organization of Knowledge and the System of the Sciences* (1929) and *The Organization of Knowledge in Libraries* (1933). All of this led to a corresponding growth in the amounts of information being collected, archived, exchanged, and exploited, with a consequent increased need for filing and retrieval at almost all levels of governmental, commercial, and educational institutions (Muddiman 2008).

Traditional forms of government document handling, such as the use of registers for logging individual documents, the use of string-bound volumes in which to preserve them, with minimal inventorying and indexing for the retrieval of these documents, began to be replaced by the newly invented filing cabinets and associated forms of indexing that allowed easier access (Robertson 2021). For instance, patent filings for new inventions increased dramatically during the late nineteenth century and became a global driver of technological change (Pretel 2018). Words also accrued economic value during this period, as trademarks and brand names were developed and legitimized through legal protection (Mercer 2010) while, as intangible assets essential to a firm's identity, they facilitated the growth of the modern corporation (Wilkins 1992).

As many American librarians were both influenced by and influential within the Progressive movement, libraries began to adopt and adapt these managerial techniques and business-oriented vocabularies (Nardini 2001) while at the same time, certain library techniques and technologies to improve efficiency, such as indexing and filing systems and equipment (Casey 1981) began to be widely adopted by other organizations (Flanzraich 1993). Vertical files replaced earlier filing technologies as the preferred form of records storage and card files became the preferred form of locating specific information (Yates 1993).

7.2 Cooperation and Competition

Nevertheless, limitations of even these novel library practices were becoming ever more evident, as the imposition of international copyright restrictions, funding shortfalls faced by expanding national libraries, the increasing necessity of translation among major European languages, and even the constraints imposed by the physicality of books and journals all increased the complexity of the librarian's task. This complexity was viewed by some visionaries as obstacles to universal shared knowledge, which should be overcome by various types of world-wide initiatives, although this ideal of open communication was far from universally accepted (Crawford 1988). Adding to the complexity of the situation, such pernicious theories as eugenics were prominent in the scientific communication of the period (Kevles 2016) while theories of imperialism and colonialism were predominant in political communications (Richards 1993).

This period prior to the First World War can also be said to mark the origins of the division between "documentation" (a key word made popular by the work of Belgian lawyer Paul Otlet, which focuses on documents or "information" within documents as discrete units of interest) and "libraries" (a long-established key word which focuses on the books and periodicals in which these are contained). The use of index cards as, in Krajewski's phrase, "paper machines" (2011) were to play a bridging role between these different perspectives. Index cards offered a surprising amount of flexibility in organizing and finding for special purposes: in addition to the standardized card catalog files maintained in alphabetical order, cards themselves could be modified through trimming edges or punching holes, connecting these to specified codes, words, or topics, allowing easy retrieval from a card deck through use of a sorting spindle or, later, through "punch card" machines designed to automate the process. This "coordinate" form of indexing through individual terms contrasts with "subordinate" indexing, the form of library classification in which terms are subordinated to subject headings. "Pre-coordinate" refers to the original assignment of terms in indexing; "post-coordinate" refers to the subsequent selection of terms in searching. Theorizing these processes would later become fundamental to library and information science.

The best-known effort to advance global knowledge sharing was launched by Paul Otlet and Henri La Fontaine in Belgium in 1895, with their Universal Bibliographic Repertory (UBR), a collection of index cards with an eventual 15 million entries summarizing social facts from books and catalogs, systematized through the Universal Decimal Classification (UDC), an approved modification of the Dewey Decimal Classification that included special symbols (Rayward 1997).The project was based on Otlet's vision of a "Universal Book," an encyclopediac endeavor to supplement traditional libraries with a free, global source of factual knowledge available to all.

A closely related endeavor also begun in 1895 was that of American zoologist Herbert Field, who began the Concilium Bibliographicum in Zurich, Switzerland, with the goal of providing scientists with coverage of the entire world's output of zoological literature in installments to be supplied every two weeks in the form of 75×125 mm catalog cards using UDC notation (Burke and Buckland 2016). While the UDC still exists, neither of these two projects long survived the death of their founder.

A third globally oriented scientific information and communication effort was that of Nobel Prize-winning Latvian chemist, Wilhelm Ostwald, who used funds from his award to found The Bridge, International Institute for the Organization of Intellectual Work in 1911, with the goal of making scientific communication more efficient through the use of Esperanto as a universal scientific language, through the centralization of specific fields of science into global institutes that would oversee publication processes, and through the standardization of scientific paper publication into a single "World Format." The First World War interrupted The Bridge's activities, which were never re-established (Hapke 1999; Holt 1977).

Finally, a little-known attempt by two other Nobel Laureates to create an inclusive "World Library" was that of Indian poet Rabindranath Tagore and French novelist Romain Rolland, who in 1923 proposed a global collection of the most important literary works from both the East and West. Although some translation rights were obtained, the project lacked substantial financial support and was abandoned in the 1930s (Mani 2016). What might perhaps be viewed as its successor is today's Project Gutenberg, the freely available online collection of works of world literature in the public domain, which was begun in 1971 by University of Illinois graduate student Michael Hart on the ARPANET computer network, and which now includes the works of both Rolland and Tagore, as well as thousands of other authors.

Billings, editor of the *Index Medicus* and an accomplished military physician in the US Surgeon General's Office who masterminded what would eventually become the National Library of Medicine, was keenly interested in vital statistics and worked closely with the US Census Bureau. He is credited with suggesting to Bureau employee Herman Hollerith the potential benefits to the rapidly growing annual census of some type of mechanical system to tabulate the census records. Hollerith, a gifted engineer familiar with the automated jacquard looms whose silk fabrics were made through the use of designs controlled by punch cards (Essinger 2004, pp. 160–161), followed up on this suggestion by

inventing his Hollerith Electric Tabulating System, which converted the data conveyed by series of holes punched into special "Hollerith cards" into electrical impulses, which would then activate mechanical counters. His system was successfully used for the 1890 census, leading Hollerith to start his Tabulating Machine Company, which was eventually to become the foundation of the International Business Machines corporation. Tabulator punch cards (31/4 by 73/8 inches or 82.550 by 187.325 mm) and library catalog cards ($71/2 \times 121.2$ cm size no. 33 catalog) were early perceived as being related: in the spring of 1896, Hollerith and Melvil Dewey agreed on a three-year partnership under which Dewey's Library Bureau would supply standardized punch cards and cabinets to commercial and government customers who used Hollerith's tabulating equipment, just as it supplied catalog cards and catalog cabinets to libraries, though this particular arrangement lasted only three years (Wright 2014).

Such tabulators and cards became increasingly common: for instance, during the 1920s the Copeland-Chatterson firm in London also began production of edge notched punch cards, the so-called Cope-Chat cards, in their new Paramount Sorting System, which were used for a wide variety of scientific and technological records. Kilgour (1997) describes a number of successful indexing systems employing Cope-Chat cards. One form of punch card system refined by W.E. Batten of the Imperial Chemical Industries in Britain in 1944 involved the so-called "peek-a-boo" or optical coincidence: each Batten card was given a subject heading from an approved list of terms and was marked off in tiny squares with assigned numbers that corresponded to specific files. If the specific file related to the subject heading, that square on the subject heading file was punched, indicating that the collection contained items relevant to that heading. When multiple subject cards were stacked together, those punched holes that penetrated throughout the entire stack identified the items relevant to all the selected subjects, using Boolean AND logic (Kilgour 1997). The efficiencies involved in such an inverted file system for specialized purposes were obvious, serving as a precursor to today's online ontologies, though its limitations in terms of size and scope were obvious as well.

Efficiency like this was also central to British science fiction author H. G. Wells's (1938) book *World Brain,* which offered a monolithic notion of global informativeness, in which a permanent world encyclopedia staffed by editors, abstractors, and indexers would accumulate files from libraries, museums, universities, and other research organizations around the world on an ongoing basis and serve as a "sort of mental clearing house for the mind, a depot where knowledge and ideas are received, sorted, summarized, digested, clarified, compared" (Wells 1938, p. 69) as well as a centralized organization "for the collection, indexing, summarizing and release of knowledge." It would be "a synthesis of bibliography and documentation with the indexed archives of the world" (Wells 1938, p. 85) culminating in information products for worldwide distribution via microfilm, with the ultimate goal of creating global consensus based on expert, scientific reasoning for actions to be taken toward world progress. This initiative, largely adapted from Otlet's earlier UBR project, which had been largely abandoned by the time of Otlet's death in

1940, never evolved in the more grandiose way that Wells envisioned, so the potential problems inherent in his vision failed to materialize as well (Rayward 2008, p. 232).

Another widely read British book of the period, *The Social Function of Science* by the controversial molecular biologist J. D. Bernal (1939), highlighted the contributions that science could make to human well-being and also urged the importance of what would eventually be known as current awareness, noting that for each scientific laboratory "there should be someone deputed to watch the whole of current literature for items which might be relevant to the work of the laboratory, and to be able to indicate without loss of time where such items are likely to be found" (1939, p. 272). Bernal also played an active role in the international scientific communications community, with such provocative proposals as abolishing scientific journals and making all scientific papers available to libraries through a government-run distribution agency (Muddiman 2003).

Soon to become equally well known was another scientist, Vannevar Bush, first head of the National Science Foundation in the United States, whose article, "As We May Think" (1945) in the popular magazine *The Atlantic*, imagined the "Memex" as a machine with two adjacent desktop surfaces, where microfilm reels of documents, notes, and images could be examined. It allowed users to link these within the machine's archive by topic or idea, so they could organize and retrieve files according to their personal indexing systems, managing both the items and the connections they made among various items. Although it was not published until 1945, Bush actually wrote this piece during the 1930s (Nyce and Kahn 1989) so it can easily be considered in conjunction with the "World Brain" vision of H. G. Wells, as both relied on the emerging technology of microfilm. Wells focused on organizational inputs and Bush focused on individual outputs, but both can be considered as elements in what might be considered a global information feedback loop. Although Bush's work has been widely appreciated as that of one of the early pioneers of information science, his "associative trail" approach has been critiqued for its lack of awareness of indexing and retrieval specifics and also for his own neglect of predecessors such as Emmanuel Goldberg, who had actually invented a machine similar to the Memex (Buckland 1992), but it should be noted also that the Memex also represents one of the first illustrations of "situational relevance," essential from a user's perspective.

Another book project with a scientific orientation was the *International Encyclopedia of Unified Science,* headed by noted European scholars Otto Neurath and Rudolph Carnap, and the American linguist Charles W. Morris, a series widely discussed during the 1930s and 1940s. Among the last of the few volumes actually published in this encyclopedic series was Thomas Kuhn's *On The Structure of Scientific Revolutions* (1962), which became celebrated for its statement that working scientists generally had little use for most previous scientific findings except those that directly impacted their own research, a finding about the nature of scientific competition that was amply borne out by subsequent research in the sociology of science (Hollinger 2011).

The lifting of various governmental restrictions on the dissemination of both US and foreign technological and science information after the end of the Second World War

(Vogel 2021) unleashed a flood of technical and scientific reports that required new knowledge organization and finding techniques beyond the traditional ones of library catalogs and abstract journals (Cleverdon 1991). The Scientific Information Conference of the Royal Society held in London in 1947 viewed indexing as one of the major problems in scientific communication and duly noted four fundamental methods of subject indexing: (1) indexing the names of subjects in alphabetical order, (2) classifying the subjects themselves under symbols which serve to pinpoint their positions in a logically constructed map of knowledge, (3) coding the subjects under symbols which can be coded and then stored in random order and run through a machine, when required, which automatically picks out those of them which have been coded in any particular way, and (4) coding the shapes of certain classes of objects which can then be retrieved by use of that code (Holmstrom 1948, p. 284).

The use of the third technique, statistical in nature, would mark the beginning of a change in era from traditional librarianship to what would become known as information science. Also highly influential in this shift was the introduction of Shannon's information theory in 1948 and Cherry's 1951 adaptation of it to communication theory (Pierce 1973). The term "information retrieval" was coined by MIT researcher Calvin Mooers in 1950, to denote the concept of selective rediscovery of information from some form of storage (Mooers 1950) and also would become foundational to the field.

Supporting Kuhn's conclusions about the perceived value of contemporary competitive scientific information, the volume of scientific and technological knowledge was rapidly increasing and could even be quantified. Economist Fritz Machlup's (1962) study, *The Production and Distribution of Knowledge in the United States*, introduced the notion of a "knowledge-based economy" and estimated that in 1959 almost 30% of the US's gross national product had been produced in knowledge industries. There was growing consensus that this increase in scientific and technical information was of vital importance to economic progress and that the improvement of information retrieval was essential to maintaining that progress (Leslie 2020).

Central to this focus on information science was an increased emphasis on discovery and novelty, especially in the growth of scientific and technical vocabularies for various disciplines with highly specific terms and nomenclatures, as well as an increasing emphasis on identifying statistical patterns in these vocabularies and the resulting literatures. The paradigmatic case is that of chemistry, which underwent rapid industrialization starting in the 1860s, establishing the earliest research laboratories, dominating international markets, and generating fierce competition among not only firms but nations with the development of new chemical compounds used for explosives, fertilizers, fibers, fuels, pharmaceuticals, poisons, synthetic rubbers, and other products used for both civilian and military purposes (Lesch 2000).

The rapid growth in the chemical literature during the nineteenth century led to a recognition of the critical need for comprehensive abstracting and indexing services for the chemical sciences, especially as industrial chemicals were being weaponized (Schofield

1999). The American Chemical Society was founded in 1876, starting the Chemical Abstracts Service in 1907, launching a journal devoted to chemical documentation in 1960, and adding the CAS Registry for small organic and inorganic molecules in 1965 to provide access to substance information, eventually leading to the new field of chemoinformatics (Lynch 1966). Several of the pioneers in information science had a background in chemistry, such as Charles Bernier, Eugene Garfield, and James Perry.

However, there were numerous debates during this post-war period sparked by the variations in the bibliographic perspectives held by those working in archives and documentation centers, libraries, and special libraries, exemplified by important publications such as the 1945 founding of the *Journal of Documentation* and Bradford's (1948) textbook *Documentation* in England, Briet's (1951) pamphlet *Qu'est-ce que la documentation?* in France, and the 1950 founding of the journal *American Documentation* and Egan and Shera's (1952) article "Foundations of a theory of bibliography" in the United States.

Robertson (2008, p. 439) explains that "the core of the argument was generally not empirical, but philosophical. Library classification schemes tend to carry with them entire philosophical world-views, concerning the nature of human knowledge, and to some extent of its representation in documents. But the nature of language as such was somewhat separate and peripheral – in some sense the object of a formal classification or indexing system is to avoid all the vagaries and pitfalls of natural language. Of course one has to describe and define the concepts or categories of the scheme in natural language, but the function of this description might be regarded as pedagogic – to help the librarian or user towards an understanding of what the concept or category really is, and to see underneath the surface of language. In constructing such a scheme, one might appeal to literary warrant, but that would not absolve one of the responsibility of understanding the concept. In the context of these arguments, empiricism (let alone a formal scientific experiment) was a radical notion. There was resistance to a strictly functional view of such schemes, quite apart from the difficulties of first formulating the functionality and then operationalizing an experimental framework."

An anonymous editorial in the April 1955 *American Documentation* enthused over the possibilities offered by these new forms of experimentation: "classification unencumbered by the arbitrary delineation of knowledge into rigid compartmentation is now possible through the use of automatic, or semi-automatic mechanisms – machines which give to classification a new flexibility and elasticity through the ease with which entirely new categories, classes, or composite groupings may be generated whenever a particular situation or need so demands. Selection based on combinations, any combinations, of characteristic attributes is a fundamental property of documentation systems utilizing automatic or semi-automatic mechanisms (mechanical aids). Thus with properly designed equipment the full range of combinations permitted by the logical theory of class definitions becomes possible in making selections."

Jesse Shera, dean of the library school at Western Reserve University in Cleveland, established the Center for Documentation and Communication Research in 1955, the

first computerized research center associated with a library school, which was headed by information scientists Allen Kent and James Perry. The center's legacy to LIS (in particular, furthering development of the terms recall and precision in the evaluation of library information retrieval) has been widely acknowledged. Perry (Ranganathan and Perry 1951), Kent (1959), and Shera (1961) were all prolific contributors to the emergent literature of information science which, as key theoretical insights resulting from empirical developments in information retrieval, growing interest in symbolic logic, and emerging applications to knowledge organization, would transform the field within the next 50 years. The next chapter will describe this transformation, as library discovery systems and more general information discovery systems, notably the Internet, even though retaining many commonalities, became increasingly differentiated in their uses.

References

Bell, H. (2001). *Indexers and indexes in fact and fiction.* University of Toronto Press.

Bernal, J. D. (1939). *The social function of science.* George Routledge & Sons.

Bliss, H. E. (1929). *The organization of knowledge and the system of the sciences.* Henry Holt & Co.

Bliss, H. E. (1933). *The organization of knowledge in libraries and the subject approach to books.* Wilson.

Bradford, S. C. (1948). *Documentation.* Crosby Lockwood and Son.

Briet, S. (1951). *Qu'est-ce que la documentation?* (R. E. Day, L. Martinet, & H. G. B. Anghelescu, Trans.). Scarecrow Press.

Broughton, V. (2008). Henry Evelyn Bliss—The other immortal, or a prophet without honour? *Journal of Librarianship and Information Science*, *40*(1), 45–58. https://doi.org/10.1177/0961000607086620

Buckland, M. K. (1992). Emanuel Goldberg, electronic document retrieval, and Vannevar Bush's Memex. *Journal of the American Society for Information Science*, *43*(4), 284–294. https://doi.org/10.1002/(SICI)1097-4571(199205)43:4<284::AID-ASI3>3.0.CO;2-0

Burke, C. B., & Buckland, M. K. (2016). Herbert Haviland Field (1868-1921): Bibliographer of zoology. *Bulletin of the Association for Information Science and Technology*, *42*(6), 10–14. https://doi.org/10.1002/bul2.2016.1720420605

Bush, V. (1945, July). As we may think. *The Atlantic*, 101–108. https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/

Buzan, B., & Lawson, G. (2015). *The global transformation: History, modernity and the making of international relations* (Vol. 135). Cambridge University Press.

Casey, M. (1981). Efficiency, Taylorism, and libraries in progressive America. *The Journal of Library History*, *16*(2), 265–279. https://www.jstor.org/stable/25541194

Cleverdon, C. W. (1991). The significance of the Cranfield tests on index languages. In *SIGIR '91: Proceedings of the 14th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 3–12). Association for Computing Machinery. https://doi.org/10.1145/122860.122861

Crawford, E. (1988). Internationalism in science as a casualty of the First World War: Relations between German and Allied scientists as reflected in nominations for the Nobel prizes in physics and chemistry. *Social Science Information*, *27*(2), 163–201. https://doi.org/10.1177/053901888027002001

Egan, M. E., & Shera, J. H. (1952). Foundations of a theory of bibliography. *Library Quarterly*, *22*(2), 125–137. https://doi.org/10.1086/617874

Emigh, R. J., Riley, D., & Ahmed, P. (2016). *Antecedents of censuses from medieval to nation states: How societies and states count*. Palgrave Macmillan.

Essinger, J. (2004). *Jacquard's web*. Oxford University Press.

Flanzraich, G. (1993). The Library Bureau and office technology. *Libraries & Culture*, *28*(4), 403–429. https://www.jstor.org/stable/25542593

Gambee, B. (1967). The great junket: American participation in the Conference of Librarians, London, 1877. *The Journal of Library History*, *2*(1), 9–44. https://www.jstor.org/stable/25540016

Griffiths, D. (2015). The radical's catalogue: Antonio Panizzi, Virginia Woolf, and the British Museum Library's "Catalogue of Printed Books." *Book History*, *18*, 134–165. https://www.jstor.org/stable/43956371

Hapke, T. (1999). Wilhelm Ostwald, the "Brücke" (Bridge), and connections to other bibliographic activities at the beginning of the twentieth century. In M. E. Bowden, T. B. Hahn, & R. V. Williams (Eds.), *Proceedings of the 1998 Conference on the History and Heritage of Science Information Systems* (pp. 139–147). Information Today.

Hollinger, D. A. (2011). The unity of knowledge and the diversity of knowers: Science as an agent of cultural integration in the United States between the two World Wars. *Pacific Historical Review*, *80*(2), 211–230. https://doi.org/10.1525/phr.2011.80.2.211

Holmstrom, J. E. (1948). The Royal Society Scientific Information Conference: Section 3, indexing and other library services. *Nature*, *162*(4112), 284–286. https://doi.org/10.1038/162279a0

Holt, N. R. (1977). Wilhelm Ostwald's "The Bridge." *The British Journal for the History of Science*, *10*(2), 146–150. https://doi.org/10.1017/S0007087400015399

Kent, A. (1959, April). A machine that does research. *Harper's Magazine*, 1–6.

Kevles, D. J. (2016). The history of eugenics. *Issues in Science and Technology*, *32*(3), 45–49. https://issues.org/the-history-of-eugenics/

Kilgour, F. G. (1997). Origins of coordinate searching. *Journal of the American Society for Information Science*, *48*(4), 340–348. https://doi.org/10.1002/(SICI)1097-4571(199704)48:4<340::AID-ASI7>3.0.CO;2-W

Kitcher, P. (1993). *The advancement of science: Science without legend, objectivity without illusions*. Oxford University Press.

Krajewski, M. (2011). *Paper machines: About cards & catalogs, 1548-1929*. MIT Press.

Kuhn, T. (1962). *The structure of scientific revolutions*. University of Chicago Press.

Landau, H. B. (1984). Engineering Index 1884–1984: Its history and its service to special libraries. *Special Libraries*, *75*(4), 312–318. https://scholarworks.sjsu.edu/sla_sl_1984/4/

Latour, B. (2012). *We have never been modern*. Harvard University Press.

Lesch, J. E. (Ed.). (2000). *The German chemical industry in the twentieth century* (Vol. 18). Springer Science+Business Media.

Leslie, C. (2020). As we could have thought: Deploying historical narratives of the Memex in support of innovation. *Technology and Culture*, *61*(2), 480–511. https://doi.org/10.1353/tech.2020.0050

Lynch, M. F. (1966). Subject indexes and automatic document retrieval: The structure of entries in Chemical Abstracts subject indexes. *Journal of Documentation*, *22*(3), 167–185. https://doi.org/10.1108/eb026399

Machlup, F. (1962). *The production and distribution of knowledge in the United States*. Princeton University Press.

Mani, B. V. (2016). *Recoding world literature: Libraries, print culture, and Germany's pact with books*. Fordham University Press.

Marson, J. (1969). John Shaw Billings as a bibliographer. *Bulletin of the Medical Library Association*, *57*(4), 379–390. https://europepmc.org/backend/ptpmcrender.fcgi?accid=PMC200868&blobtype=pdf

Mercer, J. (2010). A mark of distinction: Branding and trade mark law in the UK from the 1860s. *Business History*, *52*(1), 17–42. https://doi.org/10.1080/00076790903281033

Mooers, C. N. (1950). *The theory of digital handling of non-numerical information and its implications to machine economics.* (Zator Technical Bulletin No. 4). Zator Company.

Muddiman, D. (2003). Red information scientist: The information career of J.D. Bernal. *Journal of Documentation*, *59*(4), 387–409. https://doi.org/10.1108/00220410310485677

Muddiman, D. (2008). Public science in Britain and the origins of documentation and Information science, 1890–1950. In W. B. Rayward (Ed.), *European modernism and the information society*. Ashgate Publishing.

Nardini, R. F. (2001). A search for meaning: American library metaphors, 1876-1926. *Library Quarterly*, *71*(2), 111–140. https://doi.org/10.1086/603259

Nyce, J. M., & Kahn, P. (1989). Innovation, pragmaticism, and technological continuity: Vannevar Bush's Memex. *Journal of the American Society for Information Science*, *40*(3), 214–220. https://doi.org/10.1002/(SICI)1097-4571(198905)40:3<214::AID-ASI12>3.0.CO;2-K

Pierce, J. (1973). The early days of information theory. *IEEE Transactions on Information Theory*, *19*(1), 3–8. https://doi.org/10.1109/TIT.1973.1054955

Poovey, M. (1998). *A history of the modern fact: Problems of knowledge in the sciences of wealth and society*. University of Chicago Press.

Pretel, D. (2018). The global rise of patent expertise during the late nineteenth century. In D. Pretel & Camprubi, L. (Eds.), *Technology and globalization* (pp. 129–157). Palgrave Macmillan. https://doi.org/10.1007/978-3-319-75450-5_6

Ranganathan, S. R., & Perry, J. W. (1951). External memory and research. *Journal of Documentation*, *7*(1), 10–14. https://doi.org/10.1108/eb026165

Rayward, W. B. (1997). The origins of information science and the International Institute of Bibliography/International Federation for Information and Documentation (FID). *Journal of the American Society for Information Science*, *48*(4), 289–399. https://doi.org/10.1002/(SICI)1097-4571(199704)48:4<289::AID-ASI2>3.0.CO;2-S

Rayward, W. B. (2008). *European modernism and the information society*. Ashgate Publishing.

Richards, T. (1993). *The imperial archive: Knowledge and the fantasy of empire.* Verso.

Robertson, C. (2021). *The filing cabinet*. University of Minnesota Press.

Robertson, S. (2008). On the history of evaluation in IR. *Journal of Information Science*, *34*(4), 439–456. https://doi.org/10.1177/0165551507086989

Schofield, H. (1999). The evolution of the secondary literature in chemistry. In M. E. Bowden, T. B. Hahn, & R. V. Williams (Eds.), *Proceedings of the 1998 Conference on the History and Heritage of Science Information Systems* (pp. 94–106). Information Today.

Scott, J. C. (1998). *Seeing like a state.* Yale University Press.

Shera, J. H. (1961). Automation without fear. *ALA Bulletin*, *55*(9), 787–794.

Vincent, D. (2000). *The rise of mass literacy: Reading and writing in modern Europe*. Blackwell Publishing.

Vogel, W. F. (2021). *"The mighty microbe can go to war": Scientists, secrecy, and American biological weapons research, 1941–1969* [Doctoral dissertation]. University of Minnesota.

Wells, H. G. (1938). *World brain.* Doubleday, Doran & Co.

Wiebe, R. H. (1967). *The search for order, 1877–1920*. Hill and Wang.

Wilkins, M. (1992). The neglected intangible asset: The influence of the trade mark on the rise of the modern corporation. *Business History*, *34*(1), 66–95. https://doi.org/10.1080/00076799200000004

Wright, A. (2014). *Cataloging the world: Paul Otlet and the birth of the information age*. Oxford University Press.

Yates, J. (1993). *Control through communication: The rise of system in American management*. John Hopkins University Press.

Discovery, Retrieval

8

Abstract

This chapter describes the emergence of the differentiation between traditional librarianship with its focus on collections, which has become incorporated into today's library discovery systems, and information science with its focus on retrieval, which has become incorporated into algorithms and ontologies for today's global Internet, which together comprise the contemporary discipline of library and information science.

Keywords are integral to library and information science. Saracevic (2008) identified the emphasis on search and relevance as a defining characteristic of the emerging field, with its focus on information retrieval. Search processes and relevance to users, mostly implicit within the traditional reference functions that promoted and supported the use of library collections, were to become explicit in the new information retrieval functions, most of which involved utilizing the statistical and semantic characteristics of individual words. While the study of so-called "indexing languages" (that is, the controlled vocabulary and formalized structure used to organize materials) was also becoming much more sophisticated (Coates 1970; Soergel 1974) and eventually resulted in today's conceptual theories of knowledge organization (Machado et al. 2019), both LIS endeavors continued on diverging and occasionally converging paths: one toward today's library discovery systems and the other toward search algorithms and ontologies for the global Internet.

8.1 Indexing Essentials

Indexing remained an essential function for all types of knowledge discovery, however, even as the differences between pre-correlative (that is, an indexer's correlation of multiple headings and terms to position an individual document within a syndetic system such as a

B. V. V. Martens, *Keywords In and Out of Context*, Synthesis Lectures on Information Concepts, Retrieval, and Services, https://doi.org/10.1007/978-3-031-32530-4_8

library subject catalog) and post-correlative (that is, a searcher's subsequent correlation of multiple individual terms to locate an individual document) indexing became more marked with the development of new technologies intended to automate indexing (Kilgour 1997). It should be noted that even computerized information retrieval systems normally required an intermediary, such as a librarian or online specialist, to assist the user during the first few decades of their existence.

The current ANSI/NISO Standard Z39.4-2021 Criteria for Indexes (2021) differentiates types of indexes by (1) periodicity (one-time, closed indexes; continuing, open indexes); (2) method of document analysis (human intellectual analysis to identify topics and concepts; computer algorithms to identify useful terms, phrases, or features; combination of algorithmic and human analysis); (3) type or extent of indexable matter (full text of documents; abstracts; titles only; first lines only); (4) arrangement of entries (designed for human scanning, including alphabetical or alphanumeric; classified by relationships among concepts; based on enumerative or faceted classification schemes; alphabetico-classed: broad headings arranged alphabetically; narrower headings are grouped under broad headings and arranged alphanumerically or relationally based on hierarchy, inclusion, chronology, or other association; not designed for human scanning); and (5) type of entries (titles of documents; creators and contributors; subjects including topics or features such as genres, format, methodological approach; identifiers such as numbers that can be used to locate documents; physical descriptions such as terms that describe the type of content, and other types of single entry–type indexes, including place name, company name, law, case name, medication, and industrial classification codes). Importantly, all of these index types are found both in library and in general informational settings.

8.2 Library Discovery Systems

Today's library discovery systems can trace their history through the first computerized circulation systems in the early 1960s, followed by applications to library acquisitions, serials control, and cataloging (Kilgour 1970). The development of the MARC (Machine Readable Cataloging) standard for bibliographic information allowed the distribution of the Library of Congress's cataloging data through magnetic tape to participating libraries in 1969. Cooperative cataloging using computerized networks such as the Ohio College Library Center (now known as OCLC), which was started in 1970, and the Research Libraries Information Network, which was started in 1974, remain among the best known of the various cataloging consortia.

Also essential to the evolution from library catalog to library discovery were the multiple changes in library cataloging standards, starting with the Anglo-American Cataloging Rules (AACR), first published in 1967 and going through several editions until 2005, when it was abandoned in favor of a new model to be based on the 1998 Functional

Requirements for Bibliographic Records (FRBR) created by the International Federation of Library Associations and Institutions (IFLA), which was developed into the Resource Description and Access (RDA) framework in 2010, considered to be much more adapted to use in contemporary digital environments. The Library of Congress began work in 2011 toward updating its original MARC format into its new Bibliographic Framework (BIBFRAME) format for similar reasons (Kalita and Deka 2021).

Individual access to computerized library systems gradually began to be made available to academic and public library users, with the Ohio State University Libraries' Open Public Access Catalog (OPAC) being the first in 1975. This OPAC was launched as an outgrowth of its computerized circulation system (Su 1994, p. 136) and introduced a subject search capability in 1978. Within a few years, OPACs were being widely incorporated into the so-called integrated library systems (ILS), which were initially designed by individual libraries and later offered as turnkey systems by vendors, many of whom were the database providers as well. These developments are often referred to as first-generation, usually replicating the information from physical catalog cards, second-generation, often offering adjacency, proximity, wild card, and keyword searching, as well as Boolean logic, while third-generation OPACs offered graphical user interfaces and the ability to query on language, publication date, and other specifications (Wells 2021).

In the 1990s, library systems began to incorporate link resolvers to connect between their catalog records and external database content, as full text became increasingly available, and to add federated search capabilities for searching simultaneously across multiple external databases. By the 2000s, library search had begun to be influenced by the example of web browsers (Peters and Bell 2006; Schneider 2006a, b, c), first with the creation of the so-called "next-generation catalogs" and then by "discovery layers" which were separated from the library's own ILS in order to be compatible with any vendor database, especially those with full journal contents. The term "discovery system" began to be used by 2010 in order to emphasize the move from internal library holdings to the discovery of multiple external resources. Another major change was the wide-spread implementation of "one-box" searching to emphasize the similarity to general-purpose search engines like Google and the incorporation of non-transparent relevance algorithms and faceted browsing (Wells 2021).

While these systems became increasingly popular with users, much of the LIS research on academic libraries continues to focus on the inability of these users to understand the use of Boolean logic, to identify relevant keywords, to employ suitable search strategies for subject headings, or even to devote sufficient time to scanning results (Cribbs and Gardner 2022), although over time these research efforts have also expanded into much broader investigations of information literacy (Hicks et al. 2022).

8.3 Information Discovery Systems

Book indexing had always focused on individual words within specific content rather than on subject headings or other forms of library classification, but individual words were now also perceived as key entry points to larger, specialized forms of discourse such as science. Quantitative linguistics (Reed 1949) and specifically, the work of Harvard linguist George Kingsley Zipf (1945) on word frequency, with the most common words (the, and, a) showing distinctive differences from less common words drew particular attention, as the statistical pattern Zipf described seemed related to a similar distribution identified by Bradford (1948, pp. 106–121) showing the productivity of scientific journals in publishing in areas of specialized research, in which the scientific publications of most impact in any particular field comprised a very small subset of all the journals with occasional articles about that field, a distribution which Vickery (1948) also found approximated to the publications that scientists he surveyed in particular research areas actually read. It was soon realized that application of statistical regularities within document terminology, especially in conjunction with the advances in mechanized searching, might increase the efficiency of the document discovery process (Brookes 1968).

Although the study of scientific terminology was in its infancy, Kent wrote, "To take a word as having a precise meaning kills it as an instrument of scientific progress. The great scientists have recognized this in their practice, if not always in what they have said about their practice. Words don't file things into pigeon holes, to stay there, but are creative. Words help create, among other things, new and better meanings for themselves. If science is viewed statically, words seem to be precise labels. But if science is viewed dynamically, as the moving thing it actually is, words are imprecise but constantly improving tools" (1958, p. 186). Scientific terms are constantly evolving through such processes as semantic change, morphophonemic change, functional change, compositional change, neologizing, and borrowing (Caso 1980), increasing their rarity in comparison with more common words and their significance thus becoming easier to identify.

While the meanings of scientific terms were being debated by philosophers of science (Achinstein 1964) especially in reference to their "indeterminacy" (that is, how their interpretation might be impacted over time by shifts in associated concepts and theories), contemporary discussions of scientific terms by information scientists focused primarily on their role as document surrogates in indexing systems, viewing ambiguities as "semantic noise" and seeking quantitative approaches to resolve these (Tinker 1966). This was presumably because of the pragmatic nature of information organization and retrieval work, in which "indeterminacy" between term selection by indexer and searcher is viewed as more important than "indeterminacy" in the terms themselves (Blair 1986).

IBM engineer Hans-Peter Luhn observed (1957) that the statistical structures of different categories of information affected their ease of retrieval. For instance, "ready reference" systems such as dictionaries and parts catalogs were easiest, followed by narrowly defined and repetitive systems of fixed categories such as personnel lists and

medical case histories, followed by inventories of uniquely definable structures such as those found in chemistry, followed by highly disciplined intellectual systems such as those found in law, logic, and mathematics, followed by evolving systems involving phenomena and objects in applied sciences and technology, followed by more loosely associated systems dealing with various concepts in the human sciences and the humanities.

Another way in which the nascent information science field distinguished itself from traditional forms of librarianship was by emphasizing its commercial aspects for business and government information and by employing terms such as "coding" rather than "indexing" and "classification" rather than "cataloging." Calvin Mooers, who had originated the phrase "information retrieval," also founded the first commercial information retrieval company, the Zator Company, shortly thereafter. His "Zator" system for information retrieval was based on superimposed codes (Zatocoding) on marginal punched cards (Mooers 1951). He envisioned his zatocoding technique as part of a more sophisticated information retrieval machine, the DOKEN (Document Engine), which he claimed "would be able to search the entire Library of Congress in ten seconds, select all the information on a given subject and print selected abstracts at the rate of ten a minute" (Mooers 1950), but this highly conceptual invention was never implemented (Aitchison and Cleverdon 1963).

Mooers was also dismissive of existing inventions such as the "Rapid Selector" initially developed under Vannevar Bush's auspices at MIT and subsequently implemented by Ralph Shaw at the US Department of Agriculture's library in 1949. The Rapid Selector used 35-mm film, on each reel of which were stored the contents of almost 500,000 conventional library cards, each of which also had a pre-determined code pattern, consisting of black and white squares, indicating the subject. A master card could be inserted into the machine to select all the cards for a desired subject. The Selector was capable of being coded for 10 million subjects, but Shaw discovered that even in a fairly structured field such as agriculture, the lack of an adequate indexing system made the Selector impractical for regular use and it was abandoned by 1952 (Varlejs et al. 1999).

In his analysis of conceptual articles appearing in the journal *American Documentation* during the 1950s and 1960s, Smiraglia (2014) identified the most prevalent concepts among these: unsurprisingly, "indexing" appeared frequently as the practical aspect of document representation. These included coordinate indexing (Taube et al. 1952), contextual indexing (Luhn 1960) citation indexing (Garfield and Sher 1963), and associative indexing (Doyle 1962).

Coordinate indexing was popularized by Mortimer Taube (1961), whose Uniterm system was first used for searching the massive collection of documents at the U.S. Armed Services Technical Information Agency. Although certainly not the first to use a selection of individual terms for retrieval, a technique previously well-known as "pre-correlative indexing" (Bernier 1956), Taube was arguably the first to unify and codify the process of both indexing and retrieving specific terms, calling his approach "coordinate indexing,"

branding his terms as "Uniterms" and commercializing his projects under the corporate name "Documentation, Inc." His system assigned single words or simple phrases (Uniterms) representing individual concepts found within a particular document as "key words." The selection of a particular Uniterm would then retrieve the accession numbers of all documents indexed by that Uniterm. Advantages of the system were seen as the depth of information possible from the use of multiple Uniterms and the speed of retrieval through use of an inverted indexing system. Disadvantages were seen as loss of information through problems such as multiple word meanings or synonyms, false retrievals, difficulties in generic or specific description through the sole use of Uniterms, and lack of a thesaurus to guide the assignment of Uniterms (Costello 1961). To automate the search process, Taube also developed the IBM 9900 Special Index Analyzer or COMAC (**Co**ntinuous **M**ultiple **A**ccess **C**ollator), a punched card collator that tracked logical relationships among subject terms, and which used the IBM 305 RAMAC to print out the results.

Comparative testing to determine the relative efficacy of various indexing and retrieval systems was also triggered by the availability of Taube's system. The first known test was an experiment in using Taube's Uniterm system to index 15,000 documents on aeronautics in the Armed Services Technical Information Agency's collection, which were also indexed using an existing subject heading system there. Test queries were submitted to both indexing systems, but the experts judging the results were unable to agree on whether numerous retrieved documents were relevant to the search, so further testing was cancelled due to the lack of consensus (Gull 1956).

A second and better known set of tests using the Taube Uniterm system were the Cranfield tests, which pioneered the use of test collections in order to objectively determine relative success in precision and recall by various indexing systems (Voorhees 2001). The first effort (Cranfield 1) employed a test collection of 18,000 papers in the field of aeronautical engineering, which were indexed separately with the use of Taube's Uniterm system, the Universal Decimal Classification, an alphabetical subject catalog, and a facet classification. Evaluation with 1,200 different queries with the goal of identifying known source documents found that the Uniterm approach of selecting key words from the document appeared to be equally successful in finding relevant documents as the more traditional methods (Cleverdon 1960). A second experiment, using the metallurgical literature indices compiled at Cranfield and at Western Reserve University, indicated that the more exhaustive indexing system at WRU performed no better in terms of retrieval than did the simpler Cranfield system (Aitchison and Cleverdon, 1963), leading WRU's Allan Kent to observe later that the real value of this particular test was that it was the first in which the inverse relationship of recall and precision was experimentally demonstrated (Kent 1987). A third experiment using 221 queries on a test collection of 1,400 documents on high speed aerodynamics and aircraft structures which had been indexed with a facet classification system found again that the use of single terms produced the most relevant results (Cleverdon

et al. 1966). There was considerable criticism about the conclusions drawn from the Cranfield tests, arguing that these were too sweeping based on the actual evidence (Swanson 1965; Taube 1965), and that other important factors were ignored (Coates 1976). The use of test collections instituted by the Cranfield tests, however, became a hallmark of information retrieval research.

Contextual or KWIC (Keyword-in Context) indexing was pioneered by Hans-Peter Luhn's development of a way to automate the process of manual indexing of literature by scanning the titles of documents, eliminating common ("stop") words, and placing the remaining "content" or "key" words in a fixed position for quick identification, with a fixed number of letters on either side of the keywords in order to establish "context" (Luhn 1960). This rough approximation of topicality relied on document authors' supplying the most pertinent words in their titles, but Luhn's automatic document indexing program also used a document's word frequency and distribution to compute a "measure of significance" for both words and sentences, which then formed the "auto-abstract" of the most significant sentences in the document (Luhn 1958b).

He also prototyped a novel selector using punch cards, light, and photocells, with a search speed of 600 cards per minute in 1951, but this was quickly made obsolete by his next prototype, which used punch cards with an IBM mainframe computer (Burton 1981). Luhn's work continually stressed the commercial importance of information retrieval in business intelligence (Luhn 1958a), especially the business value of a quick machine-generated "dissemination" index of recent publications in a particular field as opposed to the more traditional "retrieval" index of all publications in that field (Bernard 2019). One of the first U.S. government agencies to employ Luhn's selector system during the 1950s was the Central Intelligence Agency (Drell 1957).

Both Luhn's KWIC system and a similar one invented separately by Herbert Ohlman of the Systems Development Corporation with a system called "permutation indexing" also using IBM equipment were introduced at the 1958 International Conference on Scientific Information (ICSI) in Washington, DC. The KWIC product, however, quickly achieved wider distribution under IBM's sponsorship, especially after its adoption by the American Chemical Society (Williams 2010). Variations on Luhn's KWIC concept followed, such as Jordan's Keyword-Out-of-Context indexing method, which entailed a non-permutated keyword index, in which each alphabetized block of relevant keywords preceded a block of those associated titles to the right (Jordan and Watkins 1968), the Key-Letter-in-Context method, which involves a lexicographic ordering of all terms in a data base by each character (alpha, numeric or special) in the term or character string, a permuted arrangement sorted by character with the balance of the term wrapped around it and is used to identify letter combinations that are highly specific and would therefore be discriminating search terms (Williams 1972), and the double-KWIC system, in which the first significant word in a title is extracted as the main index term and replaced by an asterisk to indicate its position in the title, while the remaining words in the title are then rotated, permitting each

significant word to appear as the first word of a wrap-around subordinate entry under the main index term (Petrarca and Lay 1969).

Beginning in 1949, IBM selectors were also being used for literary concordances: the best known of which was Catholic priest Roberto Busa's project to develop concordances to the massive body of work of Thomas Aquinas (Arun 2021; Rockwell and Sinclair 2020). By 1958, however, IBM provided access to its IBM 705, an electronic computer, for Busa's concordance to the Dead Sea Scrolls, which greatly facilitated Busa's concordance programs (Tasman 1957). IBM computers were also used for Josephine Miles's 1957 concordance to John Dryden's poetry (Buurma and Hefferman 2018) and Stephen Parrish's 1959 concordance to the poetry of Matthew Arnold (Parrish 1962). These projects represent the earliest beginnings of today's digital humanities, in that they allowed scholars much easier access than the many manually compiled and printed concordances that followed Cruden's 1749 biblical concordance for work in textual research, linguistic analysis, and philological comparisons.

Citation indexing, originally developed by chemist Eugene Garfield during the 1950s for chemical literature, quickly expanded into other scientific disciplines through his Science Citation Index, which collected all the references found in articles published in specific scientific journals and continuously compiled and published an index providing bibliographic information about each reference and listing all subsequent articles which cited these (Garfield 1957). Citation indexing offered an alternative approach to search, as each listed reference is presumably relevant to the citing paper and therefore potentially relevant to the reader as well, similar to the process used in searching legal precedents in decided cases ("shepardizing"), as it was dubbed by lawyers searching *Shepard's Citation Service* in preparation for current cases. Scientists following the work of other laboratories often found Garfield's index invaluable because it only required looking up the names of scientists whose work was of particular interest. The Science Citation Index was also foundational to the emerging field of bibliometrics as it provided an accessible method of statistical analysis and network mapping through the connections that could be seen through the analysis of authors, co-authors, documents, institutions, keywords, and subsequent citations (De Bellis 2009).

"Associative indexing" encompassed several different methods, all referring to the identification of those words whose frequency, proximity, or synonymy made them likely candidates for retrieval, and whose probable relevance to a particular topic or query could thus be evaluated. Guiliano (1965) described three ways of describing a body of natural language text and the three ways to view association measures computed with respect to that body as: (1) a text as a closed formal system which represents only itself, in in which case the computed association measures are descriptive rather than predictive statistics; (2) a body of text representing a much larger corpus of text, in which case some descriptive statistics for that sample will be predictive of patterns likely to be present in the larger population; and (3) a text as a representation of the encoding of concepts

and of conceptual relationships important to some area of discourse, in which case computed associations are correlates of actual relationships which exist among the concepts expressed by language. Guiliano further proposed that the practical utility of computed associations is dependent upon the third kind of interpretation, and that it is possible to obtain at least two types of measurements from text which are under certain conditions interpretable as applying to relationships among the designata of words, namely association through contiguity or synonymy, based upon counting procedures applied to words and word pairs found within text.

One highly influential model proposed by Maron and Kuhns (1960) differentiated between the semantical (grounded in the meaning of a particular term) and statistical (grounded in the frequency of a particular term as relative to other terms) aspects of words, ignoring any linguistic interpretation of the association measures of terms and focusing on the entirely statistical analysis of the co-occurrence of terms in the "index space" to provide an initial relevance ranking for the user. The goal was to assign a probability to the likelihood of a match between the description of a particular document and the description of a particular user query.

Stiles (1961) operationalized this model at the U.S. Department of Defense by using term association mapping in conjunction with Taube's Uniterm system. This approach involved the preparation of a profile for each request term, consisting of terms that had been used with the request term that have an association factor greater than 1; the comparison of these profiles and the identification of the so-called "first-generation terms," those terms which appear in all or in a given number of profiles; the use of these first generation terms as request terms, repeating the first and second steps above to identify the so-called "second generation terms"; the compilation of a table of association factors for the expanded list of request terms, with the sum of the association factors for each term being called its "wezgM," a weight which indicates the degree of association between that term and the complete request; a comparison of the list of expanded request terms with the index terms of each document in the collection and the addition of the weights (wezgM) of the terms that matched, with the sum of the weights (wezgMs) being called the "document relevance number"; and, finally, the document relevance number being used to present the documents to the requester in the order of their probable relevance to the request.

A second probabilistic approach was developed by Sparck-Jones (1972) who experimented with statistically determining the effects both of exhaustivity (the number of terms used to index an individual document) and specificity (the number of documents to which an individual term pertains) on retrieval. Her work introduced the idea that the frequency of occurrence of a word in a document collection was inversely proportional to its significance in retrieval: less common words tended to refer to more specific concepts, which were more important in retrieval. A second experiment involved weighting the terms by a Zipfian term distribution curve for the entire document collection, finding that retrieval performance was significantly improved by using this collection frequency

measure. Sparck-Jones and Robertson (1976) expanded on this by developing a relevance-weighting model in which a random set of relevant documents is used to estimate the probability that any relevant document would have a particular property, or a combination of properties.

Salton and Yang compared the collection frequency measure, which they called "inverse document frequency" or IDF, with the document term frequency weighting method and with their own term discrimination measure. They compared the removal of terms with high document frequency (IDF) to the removal of terms weights, multiplied by the existing document term frequency, finding that retrieval performance was most improved by using the second method, producing the now famous tf*idf (term frequency * inverse document frequency) weighting.

During this period, Salton also introduced the earliest version of what was to become his Vector Space Model, a mathematical alternative to Taube's Boolean and Maron's probabilistic models. He employed matrix and vector notations to present the frequencies of extracted words (the term-document "incidence" matrix) using the vector representation to describe similarities computed using both extracted words and citation data, and recommended the cosine of angles between vectors as a measure of association (Salton 1963, 1973) as well as the pilot of his SMART system for evaluating the results of various forms of associative indexing (Salton 1965). One of the best known of his SMART test results, involving the National Library of Medicine's prestigious MEDLARS system, a leading exemplar of manual indexing in the critical field of medical research, was that machine indexing produced comparable results to expert, extensive indexing, at substantially less cost in money and time (Salton 1972).

Another useful technique that improved retrieval was stemming, the process of matching words to their lexical variants. The Lovins stemmer was introduced in 1968 (Lovins 1968) and was eventually improved upon by its better-known successor, the Porter stemmer (Porter 1980), which significantly reduced the complexity of the rules for suffix removal and applied a single, unified approach to the handling of context.

Probability-oriented approaches to information retrieval became increasingly popular during the 1970s. Robertson (1977) defined the so-called probability ranking principle, which determined how to optimally rank documents based on probabilistic measures with respect to pre-defined evaluation measures. This approach was refined by Van Rijsbergen (1977) with the incorporation of term dependency into ranked retrieval, and this line of research eventually resulted in a family of scoring functions with slightly different components and parameters. Most impactful was the ranking function BM25, a bag-of-words retrieval function that ranks a set of documents based on the query terms appearing in each document, regardless of their proximity within the document.

Advances were also made in the 1980s on Salton's basic Vector Space model, the best known of which is latent semantic indexing (LSI), where the dimensionality of the vector space of a document collection was reduced through the use of singular value decomposition, a mathematical technique that can be used to identify patterns in the

relationships between the terms and concepts contained in an unstructured collection of text (Deerwester et al. 1990).

Network representations also became important in information retrieval during the 1980s for uses in browsing (Thompson and Croft, 1989), document clustering (Croft 1980), spreading activation search (Cohen and Kjeldsen 1987), support for multiple search strategies (Croft and Thompson 1987), the representation of user knowledge (Oddy et al. 1986), and document content (Tong and Shapiro 1985). These advances, particularly the invention of vector representation for semantic models, which define a word as a sequence of numbers that indicate how often it appears near other words, providing a nearly unlimited space of interconnected meanings in which concepts and topics may cluster, along with the staggering increase in available computer power, would shortly transform textual information search capabilities almost beyond recognition.

8.4 The TREC Conferences

Starting in 1990, the U.S. National Institute of Standards &Technology and the U.S. Department of Defense sponsored the TREC (Text Retrieval Conference) series of workshops (Voorhees and Harman, 2005) designed to provide larger test collections than those previously available in order to evaluate how various ranking and weighting approaches were impacted by collection composition and size. Earlier ranking functions were created through experimentation and manually tuned; however, the growing availability of training data based on all queries of a particular collection began to produce superior results (Fuhr 1986, 1989). One important finding was that the automatic construction of queries based on natural language query statements appeared to produce results equivalent to those based on vector or probabilistic approaches. TREC workshops have grown in the number of research groups and research tracks, focusing in various years on such topics as COVID 19 information and misinformation, cross-language retrieval, crowdsourcing, genomic data, legal documents, news stories, medical informatics, patents, patient medical records, social media, and video streams. Evaluation also remained a constant topic in information retrieval through the present, with a variety of new considerations emerging, such as that of fairness (Gao and Shah 2022).

Importantly, the TREC workshops themselves are non-commercial in nature, so specifically marketing-oriented search engine research is not conducted there, though TREC algorithms and applications often migrate into the commercial search sector, which is the topic of the next chapter.

References

Achinstein, P. (1964). On the meaning of scientific terms. *The Journal of Philosophy*, *61*(17), 497–509. https://doi.org/10.2307/2022847

Aitchison, J., & Cleverdon, C. W. (1963). *ASLIB Cranfield Research Project—A report on a test of the index of metallurgical literature of Western Reserve University*. https://dspace.lib.cranfield.ac.uk/handle/1826/860

Arun, J. (2021). Punching holes in the International Busa Machine narrative. In A. Koh & D. Kim (Eds.), *Alternative historiographies of the digital humanities* (pp. 121–144). Punctum Books.

Bernard, A. (2019). *Theory of the hashtag* (V. A. Pakis, Trans.). Polity Press.

Bernier, C. L. (1956). Correlative indexes I. Alphabetical correlative indexes. *American Documentation*, *7*(5), 283–288. https://doi.org/10.1002/asi.5090070404

Blair, D. C. (1986). Indeterminacy in the subject access to documents. *Information Processing and Management*, *22*(3), 229–241. https://doi.org/10.1016/0306-4573(86)90055-5

Bradford, S. C. (1948). *Documentation*. Crosby Lockwood and Son.

Brookes, B. C. (1968). The derivation and application of the Bradford-Zipf distribution. *Journal of Documentation*, *24*(4), 247–259. https://doi.org/10.1108/eb026457

Burton, D. M. (1981). Automated concordances and word indexes: The fifties. *Computers and the Humanities*, *15*(1), 1–14. https://www.jstor.org/stable/30207364

Buurma, R. S., & Hefferman, L. (2018). Search and replace: Josephine Miles and the origins of distance reading. *Modernism/Modernity*, *3*(1). https://modernismmodernity.org/forums/posts/search-and-replace

Caso, A. L. (1980). The production of new scientific terms. *American Speech*, *55*(2), 101–111. https://doi.org/10.2307/3050500

Cleverdon, C. W. (1960). *ASLIB Cranfield Research Project—Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems*. https://dspace.lib.cranfield.ac.uk/handle/1826/836

Cleverdon, C. W., Mills, J., & Keen, M. (1966). *ASLIB Cranfield Research Project—Factors determining the performance of indexing systems; Volume 1, Design; Part 1, Text and Part 2, Appendices*. Cranfield University College of Aeronautics.

Coates, E. J. (1970). Switching languages for indexing. *Journal of Documentation*, *26*(2), 102–110. https://doi.org/10.1108/eb026489

Coates, E. J. (1976). Card indexes or printed pages—Physical substrates in index evaluation. *The Indexer*, *10*(2), 60–68. https://doi.org/10.3828/indexer.1976.10.2.4

Cohen, P. R., & Kjeldsen, R. (1987). Information retrieval by constrained spreading activation in semantic networks. *Information Processing & Management*, *23*(2), 255–268. https://doi.org/10.1016/0306-4573(87)90017-3

Costello, J. C. (1961). Some solutions to operational problems in concept coordination. *American Documentation*, *12*(3), 191–197. https://doi.org/10.1002/asi.5090120305

Cribbs, H. L., & Gardner, G. J. (2022). To pre-filter, or not to pre-filter, that is the query: A multi-campus big data study. *Journal of Librarianship and Information Science*. Advance online publication. https://doi.org/10.1177/09610006221124609

Croft, W. B. (1980). A model of cluster searching based on classification. *Information Systems*, *5*(3), 189–195. https://doi.org/10.1002/(SICI)1097-4571(198711)38:6<389::AID-ASI1>3.0.CO;2-4

Croft, W. B., & Thompson, R. H. (1987). I^3R: A new approach to the design of document retrieval systems. *Journal of the American Society for Information Science*, *38*(6), 389–404. https://doi.org/10.1002/(SICI)1097-4571(198711)38:6<389::AID-ASI1>3.0.CO;2-4

De Bellis, N. (2009). *Bibliometrics and citation analysis: From the Science Citation Index to cybermetrics*. Scarecrow Press.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Doyle, L. B. (1962). Indexing and abstracting by association. *American Documentation*, *13*(4), 378–390. https://doi.org/10.1002/asi.5090130404

Drell, B. (1957). Intelligence research: Some suggested approaches. *Studies in Intelligence*, *1*(4), 79–95. https://www.cia.gov/resources/csi/studies-in-intelligence/archives/vol-1-no-4/intelligence-research-some-suggested-approaches/

Fuhr, N. (1986). Two models of retrieval with probabilistic indexing. In *SIGIR '86: Proceedings of the 1986 ACM Conference on Research and Development in Information Retrieval* (pp. 249–257). Association for Computing Machinery. https://doi.org/10.1145/253168.253221

Fuhr, N. (1989). Models for retrieval with probabilistic indexing. *Information Processing & Management*, *25*(1), 55–72. https://doi.org/10.1016/0306-4573(89)90091-5

Gao, R., & Shah, C. (2022). FAIR: Fairness-aware information retrieval evaluation. *Journal of the Association for Information Science and Technology*, *73*(10), 1461–1473. https://doi.org/10.1002/asi.24648

Garfield, E. (1957). Breaking the subject index barrier—A citation index for chemical patents. *Journal of the Patent Office Society*, *39*(8), 583–593. http://garfield.library.upenn.edu/essays/v6p472y1983.pdf

Garfield, E., & Sher, I. H. (1963). New factors in evaluation of scientific literature through citation indexing. *American Documentation*, *14*(3), 195–201. https://doi.org/10.1002/asi.5090140304

Giuliano, V. E. (1965). The interpretation of word associations. In M. E. Stevens, V. E. Guiliano, & L. B. Heilprin (Eds.), *Statistical association methods for mechanized documentation: Symposium proceedings*. National Bureau of Standards.

Gull, C. D. (1956). Seven years of work on the organization of materials in the special library. *American Documentation*, *7*(4), 320–329. https://doi.org/10.1002/asi.5090070408

Hicks, A., McKinney, P., Inskip, C., Walton, G., & Lloyd, A. (2022). Leveraging information literacy: Mapping the conceptual influence and appropriation of information literacy in other disciplinary landscapes. *Journal of Librarianship and Information Science*. https://doi.org/10.1177/09610006221090677

Jordan, J. R., & Watkins, W. J. (1968). *KWOC index as an automatic by-product of SDI*. Ames Laboratory, USAEC. https://doi.org/10.2172/12647267

Kalita, D., & Deka, D. (2021). Searching the great metadata timeline: A review of library metadata standards from linear cataloguing rules to ontology inspired metadata standards. *Library Hi Tech*, *39*(1), 190–204. https://doi.org/10.1108/LHT-08-2019-0168

Kent, A. (1987). Cyril Cleverdon. *Journal of the American Society for Information Science*, *38*(3), 152–155. https://doi.org/10.1002/(SICI)1097-4571(198705)38:3<152::AID-ASI3>3.0.CO%3B2-%23

Kent, W. (1958). Scientific naming. *Philosophy of Science*, *25*(3), 185–193. https://doi.org/10.1086/287598

Kilgour, F. G. (1970). History of library computerization. *Journal of Library Automation*, *3*(3), 218–229. https://ejournals.bc.edu/index.php/ital/article/view/5256/4731

Kilgour, F. G. (1997). Origins of coordinate searching. *Journal of the American Society for Information Science*, *48*(4), 340–348. https://doi.org/10.1002/(SICI)1097-4571(199704)48:4<340::AID-ASI7>3.0.CO;2-W

Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, *11*(1 and 2), 22–31. https://aclanthology.org/www.mt-archive.info/MT-1968-Lovins.pdf

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literacy information. *IBM Journal of Research and Development*, *1*, 309–317. https://doi.org/10.1147/rd.14.0309

Luhn, H. P. (1958a). A business intelligence system. *IBM Journal of Research and Development*, *2*(4), 314–319. https://doi.org/10.1147/rd.24.0314

Luhn, H. P. (1958b). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, *2*(2), 159–165. https://doi.org/10.1147/rd.22.0159

Luhn, H. P. (1960). Keyword-in-context index for technical literature (KWIC index). *American Documentation*, *11*(4), 288–295. https://doi.org/10.1002/asi.5090110403

Machado, L. M. O., Martínez-Ávila, D., & de Melo Simões, M. D. G. (2019). Concept theory in library and information science: An epistemological analysis. *Journal of Documentation*, *75*(4), 876–891. https://doi.org/10.1108/JD-11-2018-0195

Maron, M. E., & Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the ACM*, *7*(3), 216–244. https://doi.org/10.1145/321033.321035

Mooers, C. N. (1950). *Making information retrieval pay* (Zator Bulletin No. 55). Zator Company.

Mooers, C. N. (1951). Zatocoding applied to mechanical organization of knowledge. *American Documentation*, *2*(1), 20–32. https://doi.org/10.1002/asi.5090020107

National Information Standards Organization. (2021). *ANSI/NISO Z39.4–2021 Criteria for indexes*. https://doi.org/10.3789/ansi.niso.z39.4-2021

Oddy, R. N., Palmquist, R. A., & Crawford, M. A. (1986). Representation of anomalous states of knowledge in information retrieval. In J. M. Hurd (Ed.), *ASIS '86: Proceedings of the 49th Annnual Meeting of the American Society for Information Science* (pp. 248–254). Learned Information.

Parrish, S. M. (1962). Problems in the making of computer concordances. *Studies in Bibliography*, *15*, 1–14. https://www.jstor.org/stable/40371322

Peters, T., & Bell, L. (2006). Making OPACs blow. *Computers in Libraries*, *26*(8), 20–22.

Petrarca, A. E., & Lay, W. M. (1969). The double-KWIC coordinate index. A new approach for preparation of high-quality printed index by automatic indexing techniques. *Journal of Chemical Documentation*, *9*(4), 256–261. https://doi.org/10.1021/c160035a018

Porter, M. F. (1980). An algorithm for suffix stripping. *Program: Electronic Library and Information Systems*, *14*(3), 130–137. https://doi.org/10.1108/eb046814

Reed, D. W. (1949). A statistical approach to quantitative linguistic analysis. *WORD*, *5*(3), 235–247. https://doi.org/10.1080/00437956.1949.11659355

Robertson, S. E. (1977). The probability ranking principle in IR. *Journal of Documentation*, *33*(4), 294–304. https://doi.org/10.1108/eb026647

Robertson, S. E., & Sparck Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, *27*(3), 129–146. https://doi.org/10.1002/asi.4630270302

Rockwell, G., & Sinclair, S. (2020). Tremendous mechanical labor: Father Busa's algorithm. *DHQ: Digital Humanities Quarterly*, *14*(3). http://www.digitalhumanities.org/dhq/vol/14/3/000456/000456.html

Salton, G. (1963). Associative document retrieval techniques using bibliographic information. *Journal of the ACM*, *10*(4), 440–457. https://doi.org/10.1145/321186.321188

Salton, G. (1965). An evaluation program for associative indexing. In M. E. Stevens, V. E. Giuliano, & L. B. Heilprin (Eds.), *Statistical association methods for mechanized documentation: Symposium proceedings*. National Bureau of Standards.

Salton, G. (1972). A new comparison between conventional indexing (MEDLARS) and automatic text processing (SMART). *Journal of the American Society for Information Science*, *23*(2), 75–84. https://doi.org/10.1002/asi.4630230202

Salton, G., & Yang, C. S. (1973). On the specification of term values in automatic indexing. *Journal of Documentation*, *29*(4), 351–372. https://doi.org/10.1108/eb026562

Saracevic, T. (2008). Effects of inconsistent relevance judgments on information retrieval test results: A historical perspective. *Library Trends*, *56*(4), 763–783. https://doi.org/10.1353/lib.0.0000

Schneider, K. G. (2006a, March 13). How OPACs suck, part 1: Relevance rank (or the lack of it). *ALATechSource*. https://www.ala.org/tools/article/ala-techsource/how-opacs-suck-part-1-relevance-rank-or-lack-it

Schneider, K. G. (2006b, April 3). How OPACs suck, part 2: The checklist of shame. *ALATechSource*. https://www.ala.org/tools/article/ala-techsource/how-opacs-suck-part-2-checklist-shame

Schneider, K. G. (2006c, May 20). How OPACs suck, part 3: The big picture. *ALATechSource*.

Smiraglia, R. (2014). The concept of concepts: A case study from *American Documentation*. *Proceedings of the Annual Conference of CAIS/Actes Du Congrès Annuel de l'ACSI*. https://doi.org/10.29173/cais870

Soergel, D. (1974). *Indexing languages and thesauri: Construction and maintenance*. Melville Publishing Co.

Sparck-Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*(1), 11–21. https://doi.org/10.1108/eb026526

Stiles, H. E. (1961). The association factor in information retrieval. *Journal of the ACM*, *8*(2), 271–279. https://doi.org/10.1145/321062.321074

Su, S.-F. (1994). Dialogue with an Opac: How visionary was Swanson in 1964? *Library Quarterly*, *64*(2), 130–161. https://www.jstor.org/stable/4308919

Swanson, D. R. (1965). The evidence underlying the Cranfield results. *Library Quarterly*, *35*(1), 1–20. https://www.jstor.org/stable/4305528

Tasman, P. (1957). Literary data processing. *IBM Journal of Research and Development*, *1*(3), 249–256. https://doi.org/10.1147/rd.13.0249

Taube, M. (1961). Documentation, information retrieval, and other new techniques. *Library Quarterly*, *31*(1), 378–390. https://www.jstor.org/stable/4305078

Taube, M. (1965). A note on the pseudo-mathematics of relevance. *American Documentation*, *16*(2), 69–72. https://doi.org/10.1002/asi.5090160204

Taube, M., Gull, C. D., & Wachtel, I. S. (1952). Unit terms in coordinate indexing. *American Documentation*, *3*(4), 213–218. https://doi.org/10.1002/asi.5090030404

Thompson, R. H., & Croft, W. B. (1989). Support for browsing in an intelligent text retrieval system. *International Journal of Man-Machine Studies*, *30*(6), 639–668. https://doi.org/10.1016/S0020-7373(89)80014-8

Tinker, J. F. (1966). Imprecision in meaning measured by inconsistency of indexing. *American Documentation*, *17*(2), 96–102. https://doi.org/10.1002/asi.5090170207

Tong, R. M., & Shapiro, D. G. (1985). Experimental investigations of uncertainty in a rule-based system for information retrieval. *International Journal of Man-Machine Studies*, *22*(3), 265–282. https://doi.org/10.1016/S0020-7373(85)80003-1

Van Rijsbergen, C. J. (1977). A theoretical basis for the use of co-occurence data in information retrieval. *Journal of Documentation*, *33*(2), 106–119. https://doi.org/10.1108/eb026637

Varlejs, J., Bowden, M. E., Hahn, T. B., & Williams, R. V. (1999). Ralph Shaw and the rapid selector. In M. E. Bowden, T. B. Hahn, & R. V. Williams (Eds.), *Proceedings of the 1998 Conference on the History and Heritage of Science Information Systems* (pp. 148–155). Information Today Inc.

Vickery, B. C. (1948). Bradford's law of scattering. *Journal of Documentation*, *4*(3), 188–203. https://doi.org/10.1108/eb026133

Voorhees, E. M. (2001). The philosophy of information retrieval evaluation. In M. Braschler, J. Gonzalo, & M. Kluck (Eds.), *Evaluation of cross-language information retrieval systems. CLEF 2001. Lecture notes in computer science 2406* (Vol. 2406, pp. 355–370). Springer-Verlag. https://doi.org/10.1007/3-540-45691-0_34

Voorhees, E. M., & Harman, D. K. (Eds.). (2005). *TREC: Experiment and evaluation in information retrieval*. MIT Press.

Wells, D. (2021). Online public access catalogues and library discovery systems. *Knowledge Organization*, *48*(6), 457–466. https://doi.org/10.5771/0943-7444-2021-6-457

Williams, M. E. (1972). Experiences of IIT Research Institute in operating a computerized retrieval system for searching a variety of data bases. *Information Storage and Retrieval*, *8*(2), 57–78. https://doi.org/10.1016/0020-0271(72)90008-3

Williams, R. V. (2010). Hans Peter Luhn and Herbert M. Ohman: Their roles in the origins of keyword-in-context/permutation automatic indexing. *Journal of the American Society for Information Science and Technology*, *61*(4), 835–849. https://doi.org/10.1002/asi.21265

Zipf, G. K. (1945). The meaning-frequency relationship of words. *The Journal of General Psychology*, *33*(2), 251–256. https://doi.org/10.1080/00221309.1945.10544509

Databases, Search Engines

9

Abstract

This chapter describes the history of commercial databases, the rise of search engines, and the range in responses of libraries and LIS researchers to these developments during the past 50 years.

Clearly, the connections between libraries and commercial information suppliers can be dated to far before Gutenberg's print workshop in Germany in 1440, as woodblock printers in Japan were producing thousands of copies of Buddhist sutras for public distribution as early as 770, but it is only recently that many of these relationships have evolved into what might well be described as antagonistic, as "information" has become an increasingly precious resource to some, such as academic libraries, and an increasingly valuable commodity to others, such as global information analytics corporations.

9.1 Commercial Databases

As mentioned in Chap. 8, libraries became accustomed to commercial database searches as these first emerged during the 1960s, with systems such as Lockheed's Dialog, initially developed under contract with NASA, and Systems Development Corporation's ORBIT, initially developed under contract as ELHILL with the National Library of Medicine. These systems began as subscription-based dial-up services through time-sharing mainframes in order to access specific bibliographic databases and initially required specially trained search intermediaries familiar with the products in order to run searches for users. Dialog provided specialized inverted and linear indexes for Boolean and proximity searches, which could be refined on an iterative basis. ORBIT also used inverted files and did proximity and truncation searching, but was limited by its nonrecursive, non-Boolean

B. V. V. Martens, *Keywords In and Out of Context*, Synthesis Lectures on Information Concepts, Retrieval, and Services, https://doi.org/10.1007/978-3-031-32530-4_9

searching (Bourne 1980). Lexis, a legal information search service introduced in 1972 by the Mead Data Corporation, was the first commercial search system available directly to its users in law offices and is now LexisNexis, acquired in 1994 by Reed Elsevier, now RELX. After undergoing several changes in ownership, Dialog was bought in 2021 by Clarivate Analytics, which now also owns what was originally the Science Citation Index. Similarly, Orbit, after several changes in ownership, was bought in 1994 by Questel, an online company based in France, and became Questel-Orbit.

In general, there has been a reduction in the number of commercial database search firms over the years, due in part to multiple acquisitions and mergers and due in part to the influence of the Internet and the growth of general-purpose search engines. However, firms such as Clarivate Analytics and RELX continue to enjoy a dominant position in the scholarly publishing ecosystem, as they publish many of the leading academic journals and provide powerful analytic data based on these.

9.2 Search Engines

Although the first use of the term "search engine" according to the *Oxford English Dictionary* seems to be found in an April 10, 1978 article in the computer industry publication, *Computerworld* ("Primary keys will handle production problems related to data while secondary keys called 'search engines' will speed the location of files within data bases"), it did not come into common usage until the 1990s, when the development of the Internet as a global network of networks marked another transformation of the library and information science environment, impacting it much more than did previous financial support provided by various U.S. government agencies in information research areas of national importance.

While much of the early development of its predecessor, the ARPANET, was done under the auspices of the U.S. Department of Defense, since the goal of a robust and resilient system with multiple nodes that would be less vulnerable to attack entailed basic research in advanced technology rather than the command-and-control systems then in use by the U.S. military, the research community involved often had academic ties (Abbate, 1999). Even though the ARPANET is universally recognized as the predecessor of the Internet, Campbell-Kelly and Garcia-Swartz (2013) correctly point out that, without the multiple commercial and international efforts that built the personal and institutional computing environments, the TCP/IP protocols that became the Internet's foundation would have had little of general interest to connect.

The evolution of the Internet through its various protocols has always included some element of search, whether simply through the GREP (Global Regular Expression Print) command line used for searching and matching text files on the UNIX computers that made up the majority of early Internet servers (Rualthanzauva 2014), through sockets identification for host-host protocols (Fidler and Acker 2014) through the Wide Area

Information System for client systems on the TC/ICP protocol which allowed users to search for and retrieve documents whose text contained specified words; through the Archie system which let users utilize GREP to search a list of public anonymous FTP sites using the Telnet protocol, through the Veronica and Jughead systems for the Gopher protocol which let users select topics from hierarchically organized menus from information providers instead of having to know and type in the names of computers and files, through webcrawlers for the HTTP protocol (McMurdo 1995), which allowed web pages to be indexed, through WHOIS to search for registrations on the Domain Name System for the TCP/IP protocol (Bradshaw and DeNardis 2019), through the hashtag identifier # which originated on Internet Relay Chat to search for group topics but which subsequently attained much greater visibility on Twitter (Pandell 2017), or through the Ahmia search engine for the TOR protocol, which allows users to access the so-called "dark web" (Gehl 2018).

A major change was the hypertextual development of the World Wide Web in 1994 with the creation of the HTTP protocol, its associated mark-up language (HTML) with its anchor tag coding for hyperlinks, and its uniform resource locator (URL) to denote webpage locations. Hypertext at its simplest is the ability to create documents which link to one another, and it sparked the ensuing explosion of "content" beyond what had been previously available through distribution by libraries and print publishers. Web pages proliferated, providing an extraordinary variety of material from both authoritative sources and others. The development and distribution of the free Web browsers Mosaic and Netscape eased navigation between web pages but didn't assist directly in the search process, giving rise to web page "lists of links" to popular resources. A more technologically sophisticated solution was the invention of robot crawlers to gather and compile information from individual web pages. Web crawlers developed selection policies to determine which pages to download, revisit policies to indicate when to check for changes, politeness policies to avoid overloading sites by too frequent visits, and parallelization policies to coordinate distributed crawling (Pinkerton 2000).

Early search engine structure included a crawler module, a page repository, an indexing module, indexes, a query module, and a ranking module (Langville and Meyer 2011). They used inverted indexes that encoded term frequencies, term positions, document structure information, and document metadata that let users query the system with keywords, key phrases, and other advanced search operators. Other index-related innovations were the use of anchor text on web pages for keyword matching and hyperlink connection analysis, which were considerably more complex in the web environment, given the diversity of textual context and users, the rapid growth of the web, and the possibilities of deliberate intervention and manipulation.

According to Reider (2020), the novel ranking strategies HITS (Hypertext Induced Topic Search) by Kleinberg (1999) and PageRank by Brin and Page (1998) revolutionized search engine conceptualization by complementing the traditional notion of a document repository as a flat corpus in which relevance is only determined by a specific information

need with the more sociometrically influenced ideas of status, authority, and influence. This change conceptualized the web corpus as a graph network of documents linked by various attributes. Kleinberg's HITS model retrieved documents matching a specific query and then utilized eigenvector-based evaluative metrics ("hubness" and "authoritativeness") of the links to these found pages to rank them in order of their presumed importance within the domain of the result set rather than of the entire web corpus.

Brin and Page's PageRank model instead conceived the hyperlink network between web pages in terms of a rational attribution of their relative importance by document authors, similar to the concept of scholarly citations, becoming "a singular, universal network of 'authority' that the search system can combine with traditional ranking mechanisms to calculate a final score during the search process" (Reider 2020, p. 286).

Both algorithms represented slightly different approaches to algorithmic evaluation of Internet content, but it was soon realized that algorithms could not remain static in this rapidly evolving and expanding environment. Besides document information (location and frequency of keywords on a webpage) and link analysis (number and quality of hyperlinks) information, the availability of massive amounts of user log information, including queries, query adaptation, click patterns, and abandonment rates became important in training ranking algorithms in search engines (Joachims 2002).

Other important advances on earlier IR techniques included the introduction of probabilistic approaches using natural language models (Hiemstra 1998), improving ranking functionality by the use of term dependence with proximity operators (Metzler and Croft 2005), and incorporating relevance feedback through maximal marginal relevance ranking (Carbonell and Goldstein 1998).

Search engines also improved on Salton's vector model to create denser vector-based indexes with richer semantic representations to enhance recall (Zobel and Moffat 2006). One of the most notable of these advances has been Word2Vec, a group of related models created by Google researchers in 2013, which can ingest a large corpus of text to produce a multi-dimensional vector space. Each unique word in the corpus is positioned within a word vector (which consists of a numerical list) with words that are semantically and/or syntactically similar being located closer to one another, while dissimilar words are located farther away (Mikolov et al. 2013).

The availability of massive amount of data from previous search sessions, vastly enhanced computing power, and the development of artificial intelligence neural network processing capabilities have resulted in a number of new algorithms in recent years. These machine learning systems arbitrarily assign numerical values to words and then calculate the statistical likelihood that one particular word will follow another, based on the analysis of large samples of text with which the system was trained. This analysis is done by a neural network, each layer of which analyzes a different aspect of the samples it is provided: definitions of words, relations of words, phrase, or sentence structures, etc.

Another stage critical to search engine evolution has been the conceptualization of the web as an implementation of knowledge graphs via linked data and the so-called semantic

web. This added to the original PageRank perspective of a network of authoritative sources by envisioning "a web of concepts by starting with the current view of the web (comprised of hyperlinked pages, or documents, each seen as a bag of words), extracting concept-centric metadata, and stitching it together to create a semantically rich aggregate view of all the information available on the web for each concept instance" (Dalvi et al. 2009).

Arguably the best-known example of this was Google's introduction of its "Knowledge Graph" (Singhal 2012), an innovation intended to aggregate factual information from a wide variety of authoritative and accessible resources such as Wikidata to respond to queries, all presented on a single search results page. Other search engines such as Microsoft's Bing also adopted knowledge graph techniques and technologies, the overall goal being to ingest and organize available data on entities and their interrelations and organize these by use of class definitions and entity relationships within a schema, as well as to adapt for potentially interrelating arbitrary entities with each other (Heist et al. 2020).

However, as these knowledge graphs are generally proprietary and are seldom transparent as to their sources, there has been considerable critique of their impartiality and their influence (Juel Vang, 2013; Kraft and Usbeck 2022). Google in particular, with its ability to analyze massive numbers of queries as demonstrated through its "Google Trends" page showing the most popular of these over time, and its massive U.S. government lobbying expenditures on avoiding antitrust measures, is a frequent target for such criticism. Google's implementation of the HTTPS (HyperText Transfer Protocol Secure) protocol in 2012 to increase overall online security meant that website owners would no longer be able to access information on Google queries about their own sites, making Google the sole owner of this knowledge as well (Craver 2013).

The recently developed transformer language models currently dominate the search market (Vaswani et al. 2017). These employ deep learning models using algorithms which can process massive amounts input at once, differentially weighting the significance of each part of the input data by "self-attention," a mechanism which provides context for any position in the input sequence in order to generate every word's embedding representation. For example, in the phrases "he is flourishing the cane" and "the fields of cane are flourishing," the words "flourishing" and "cane" would have very different embedding representations for each.

The transformer language model BERT (Bidirectional Encoder Representations from Transformers), developed by Devlin et al. (2018) at Google and pre-trained with large language datasets, such as the Wikipedia corpus, forms the current backbone of the Google search engine, while GPT3 (Generative Pre-trained Transformer), developed by Radford et al. (2019) at OpenAI and pre-trained with the Common Crawl dataset, forms the current backbone of the Bing search engine.

New variants of transformer language models are continuously being developed for a wide variety of purposes, including error correction, summarization, and translation. Nevertheless, today's systems continue to conform to the earliest information retrieval system

blueprint: (a) building an efficient queryable index for each document in the corpus, (b) retrieving a set of candidates for a given query, and (c) computing a relevance score for each candidate (Metzler et al. 2021, p. 2). Informed speculation is that Google currently uses at least 200 factors in its ranking relevance for queries (Dean 2021, p. 2).

9.3 Effects on LIS Practice and Research

The ongoing technologization of information resources has had unexpected effects on libraries of different types, including school and public libraries. For instance, the U.S. Children's Internet Protection Act of 2000, passed with the express purpose of protecting those under 17 years of age from harmful visual depictions on the Internet, requires school and public libraries to install Internet filters on all computers to block certain content, usually defined by specific keywords or URLs, in order to receive the federally funded Universal Service Discount (E-rate) for their Internet technology (Harris 2019).

Although these filters are intended for young users, all Internet-accessible computers in institutions receiving such funding must have these filters installed, though they may be removed for adult users upon request. Despite a paucity of research on Internet filtering, current findings indicate that many libraries block far more categories than necessary to fulfill their legal obligations, that each library's specific configuration of filtering varies significantly since it is based on individual administrator decisions, that internet filters in general are well-known to underblock and overblock various sites, and that Internet filtering companies are almost exclusively privately run, making those algorithms used to manage the filtering processes proprietary trade secrets unavailable to actual users. As noted, there is a general lack of data regarding these issues (Oltmann et al. 2021). The political implications of this may even be larger, especially given that keywords related to formerly acceptable retrieval queries and websites such as those concerning human reproductive choices may no longer be legal in numerous states (Cardillo 2022).

Technology-driven changes in journal publication and library acquisition practices have also had a lasting effect on academic research and academic libraries, as rapidly rising journal prices largely propagated by what law librarian Sarah Lamdan (2022) has called the "data cartels" in global publishing made libraries and researchers turn to alternatives such as increased reliance on interlibrary loan networks, various forms of institutional repositories, and open access publishing.

For librarians, the emphasis on keywords and keyword matching as facilitated by commercial databases and search engines has always been somewhat problematic, given the traditional emphasis on the educated use of library classification systems and the desire to provide a much broader and richer research experience, often facilitated by librarians and a wide variety of user resources. While the use of "free text" or unstructured vocabularies by users for online searching had been a topic of research since long before the advent of

the Internet (Markey 2007a, b; Markey et al. 1980), and the utility of user-supplied keywords has been well-recognized for decades (Tillotson 1995), as search engines began to replace library catalogs as the accepted starting point for information searching, librarians bemoaned the unstructured messiness of the Internet itself in comparison with the wealth of knowledge available in libraries (Taylor 1995).

In their analysis of over a million public queries to an early search engine (Excite), LIS researchers found that these queries differed significantly from user queries to more traditional library databases in that they were shorter and simpler, employing a limited number of distinctive terms with very high frequency, and seemingly had the potential to change the nature of information retrieval research as a whole (Jansen et al. 2000; Spink et al. 2001).

This wealth of newly available resources, however heterogenous in nature, increasingly led librarians to reconsider their traditional bibliographic data practices within a larger, or "metadata," context (Borbinha 2004) while others suggested that such practices no longer suffice in a "resource discovery" era (Han, 2012). Tagging, or social folksonomies, was adapted for users of library systems to identify items of particular interest to them shortly after user-created labels known as "tags" or "hashtags" became prevalent among users on such popular Internet platforms as Flickr and Twitter (Rafferty 2018). More importantly, academic libraries began turning toward linked open data in such open Creative Commons-licensed repositories as Wikidata (Tharani 2021) in order to make progress toward the promise of what Berners-Lee named the Semantic Web for sharing and publicizing library resources, though that progress has been slow compared to Google's and Microsoft's much more efficient exploitation of Wikidata for their own proprietary knowledge graphs. Today, academic libraries in particular are turning to "semantic search" as a way of incorporating new linguistic advances into their discovery technologies (Eller 2022).

LIS's research shift in knowledge organization systems topics has also been gradual. Si et al. (2023) found that the dominant knowledge organization topics in LIS research publications from 1993 through 2007 were cataloguing and classification, then from 2008 through 2013 classification schemes, thesauri, cataloguing and bibliographic control, application in specific domains, and ontology predominated, while after 2013, the dominant topics have been ontology applications, domain ontology development, ontology evaluation, linked applications, and linked publishing.

When dealing with Internet-related topics, the focus of much of this LIS research, with some exceptions such as Jansen (2011) and Lewandowski (2011), has been on so-called "organic" results (that is, keyword results from those websites that haven't paid the search engine for their inclusion) as those are most similar to the results of library collection searches.

There has been considerable critique from LIS of the evolving search engine industry, ranging from Jeanneney's (2007) scathing denunciation of Google's unsystematic digitization of the holdings of a large number of academic libraries for its Google Books

initiative, to Hjørland's (2012) considered opinion on the enduring value of knowledge organization as a professional rather than commercial practice, to Palfrey's (2015) defense of libraries as the last strongholds of civic literacy and engagement in an increasingly virtual and commercialized world, to Sauvayre's (2022) in-depth analysis of the prevalence of citation errors in Google Scholar data. But, in general, there appears to be wide acceptance in LIS of the dominant position of search engines in the modern information landscape, and, accordingly, the necessity of finding ways to increase digital information literacy, even as Google, for instance, touts its own efforts to make this critical skill easier for its users (Nayak 2022).

Today search engines such as Google, Bing, and DuckDuckGo have become the search platform of choice for all but the most limited of in-collection inquiries, and librarians themselves are routinely expected to learn the basics of search engine marketing and search engine optimization for their own library websites, and to adapt and adopt many search engine modifications for their own systems. With the increasing importance of the so-called "web scale discovery services," (Raieli 2022) in which external web content is brought together with internal library resources to form an aggregated core collection, library search systems are no longer insulated from the Internet and much library instruction today focuses on "information literacy" intended to help users differentiate between accurate and inaccurate sources of information in any kind of search activity (Head et al. 2022).

Considered as a search "platform" (Gillespie 2010), the library discovery system has always been expected to display a professionally circumspect and deliberately circumscribed approach to directing users toward any specific search results on a list of all possible matches to their queries, though there has been increasing evidence that implicit bias is endemic to such systems as well (Reidsma 2019). A major competitor to library discovery systems, the Google Scholar search platform, has been argued to have disruptive effects that impact the scholarly community in opaque, highly quantitative ways that weaken academic autonomy and disciplinary functioning (Goldenfein and Griffin 2022).

Recently, Google researchers also have proposed to "re-think search" by moving from the conventional keyword search model to a question and answer model by utilizing pretrained language models to respond to natural language queries from online users as a way of promoting user engagement by simulating advice from domain experts (Metzler et al. 2021). The improvement in question answering technologies, especially for factually based queries, in recent years (Zhang et al. 2023) makes this hardly an unexpected development.

However, this proposed advance has prompted criticism that Google's stated vision of speed, convenience, and increased user engagement relegates information searching as "a socially and contextually situated activity with diverse sets of goals and needs for support" to an opaque "combination of text matching and text generating algorithms" and that a preferable vision should include support for a wide variety of information search intentions and information seeking strategies as well as transitions between them,

a clear way for the user to carry interactions with the system with iterations of request-response that carry the knowledge from previous interactions to the next, support for these interactions through various modalities and modes of communication, including different types of devices, interfaces, languages, expression of information need (keywords-based queries, questions, gestures, etc.), support for increasing users' information literacy, the provision of sufficient transparency about the sources where the information objects are coming from, as well as the process through which they are either ranked or consolidated and presented, and the system should be free of economic structures that support and even incentivize the monetization of concepts (such as identity terms) and allow commercial interests to masquerade as 'objective' information" (Shah and Bender 2022).

Transparency about algorithms and data sources alone, however, are unlikely to provide the level of accountability that would make the operations of a search engine such as Google, with its continuous cycle of proprietary algorithm updates, understandable to its external stakeholders such as government and users (Ananny and Crawford 2018). Although more transparent alternatives to commercial search engines have recently been proposed, such as Zhitomirsky-Geffet's open network of inter-linked knowledge organization systems with multiple validity scopes (Zhitomirsky-Geffet 2019; Zhitomirsky-Geffet and Hajibayova 2020), it seems likely that the contents of such a system, if developed, would quickly be ingested by commercial search engines, as is currently the case with WikiData, and incorporated into new algorithms, with uncertain results.

Sundin et al. (2022) approach the problem from a different aspect: that of viewing search engines as "multi-relevance machines" which, they suggest, would benefit from incorporating "societal relevance" into search engine results (as they believe that Google has already done in the special case of medical information during the COVID-19 pandemic). "Societal relevance" would join the common relevance categories of system (unpersonalized) and individual (personalized) relevance, especially in instances where societal interests could be considered paramount, such as disaster situations, in which current, correct, and manually curated information would be vital. They note that this type of "relevance" would pose its own dilemmas both for search engines and users, highlighting issues such as absence of data, political implications, and strategic ignorance, and urge that LIS researchers need to study commercial search engine practices, processes, and results much more actively than they have done to date.

Hjørland (2021) has drawn from the philosophy of science in his current proposal that both information retrieval and knowledge organization need to be rethought in order to emphasize the importance of understanding that "the literature (or 'the universe of recorded knowledge') in which we perform IR cannot just be considered bits of true knowledge in which each bit is as good as any other. On the contrary, the literature must be understood as a mixture of different voices, some of which conflict with each other. The relevance of a given search set of documents (or a given relevance ranking of documents) is therefore a hypothesis, the answer to which concerns the conflict between

different paradigms in the subject area", which he argues can be addressed by understanding these paradigms and by developing "pragmatic and critical theories acknowledging the role of goals, interests, and consequences in knowledge." This expansion of Hjørland's earlier seminal work on relevance and domain discourse into "the universe of recorded knowledge" brings new problems, since neither information science nor librarianship traditionally engages in the broader type of social epistemology that such activities would inevitably entail (Van der Veer Martens 2015, pp. 320–324).

This is not to say that such a widening in scope for LIS would be undesirable, given that "the universe of recorded knowledge" is clearly contiguous to if not already subsumed within the rapidly expanding infosphere. The next chapter will explore the various ways in which keywords are being used to attract, influence, and retain today's search engine users.

References

Abbate, J. (1999). *Inventing the Internet*. MIT Press.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency idea and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989. https://doi.org/10.1177/1461444816676645.

Borbinha, J. (2004). Authority control in the world of metadata. *Cataloging and Classification Quarterly*, *38*(3–4), 105–116. https://doi.org/10.1300/J104v38n03_10.

Bourne, C. P. (1980). On-line systems: History, technology, and economics. *Journal of the American Society for Information Science*, *31*(3), 155–160. https://doi.org/10.1002/asi.4630310307.

Bradshaw, S., & DeNardis, L. (2019). Privacy by infrastructure: The unresolved case of the Domain Name System. *Policy & Internet*, *11*(1), 16–36. https://doi.org/10.1002/poi3.195.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, *30*(1–7), 107–117. https://doi.org/10.1016/S0169-7552(98)00110-X.

Campbell-Kelly, M., & Garcia-Swartz, D. D. (2013). The history of the Internet: The missing narratives. *Journal of Information Technology, 28*(1), 18–33. https://doi.org/10.1057/jit.2013.4

Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 335–336). Association for Computing Machinery. https://doi.org/10.1145/290941.291025.

Cardillo, T. (2022, November 8). In a post-Roe v. Wade America, should librarians be worried about helping patrons find abortion information? *Intellectual Freedom Blog*. https://www.oif.ala.org/in-a-post-roe-v-wade-america-should-librarians-be-worried-about-helping-patron-find-abortion-information/.

Craver, T. (2013, September 23). Goodbye, keyword data: Google moves entirely to secure search. *Search Engine Watch*. https://www.searchenginewatch.com/2013/09/23/goodbye-keyword-data-google-moves-entirely-to-secure-search/.

Dalvi, N., Kumar, R., Pang, B., Ramakrishnan, R., Tomkins, A., Bohannon, P., Keerthi, S., & Merugu, S. (2009). A web of concepts. In *PODS '09: Proceedings of the 28th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* (pp. 1–12). Association for Computing Machinery. https://doi.org/10.1145/1559795.1559797.

Dean, B. (2021, October 10). Google's 200 ranking factors: The complete list (2022). *Backlinko.* https://backlinko.com/google-ranking-factors.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, L. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding.* arXiv: preprint: https://arxiv.org/abs/1810.04805

Eller, D. W. (2022). Transparency and the future of semantic searching in academic libraries. *Information Services & Use*, *42*(3–4), 453–461. https://doi.org/10.3233/ISU-220175.

Fidler, B., & Acker, A. (2014). Metadata and infrastructure in Internet history: Sockets in the Arpanet host-host protocol. *Proceedings of the American Society for Information Science and Technology*, *5*(1), 1–8. https://doi.org/10.1002/meet.2014.14505101054.

Gehl, R. W. (2018). *Weaving the dark web: Legitimacy on Freenet, Tor, and I2P*. MIT Press.

Gillespie, T. (2010). The politics of "platforms." *New Media & Society*, *12*(3), 347–364. https://doi.org/10.1177/1461444809342738.

Goldenfein, J., & Griffin, D. (2022). Google Scholar: Platforming the scholarly economy. *Internet Policy Review*, *11*(3), 1–34. https://doi.org/10.14763/2022.3.1671.

Harris, F. K. (2019). CIPA/Internet Filtering. In R. Hobbs & Milhailidis (Eds.), *International encyclopedia of media literacy*. John Wiley & Sons. https://doi.org/10.1002/9781118978238.iiem l0038.

Head, A., Hertling, S., Geofrey, S., & MacMillan, M. (2022). *The Project Information Literacy Retrospective, 2008–2022*. Project Information Research Institute. https://projectinfolit.org/publications/retrospective/.

Heist, N., Hertling, S., Ringler, D., & Paulheim. H. (2020). Knowledge graphs on the Web—An overview. In I. Tiddi, F. Lécué, & P. Hitzler (Eds.), *Knowledge graphs for explainable artificial intelligence: Foundations, applications, and challenges* (pp. 3–22). IOS Press.

Hiemstra, D. (1998). A linguistically motivated probabilistic model of information retrieval. In C. Nikolaou & C. Stephanidis (Eds.), *Research and advanced technology for digital libraries. ECDL 1998. Lecture notes in computer science: Vol. 1513* (pp. 569–584). Springer. https://doi.org/10.1007/3-540-49653-X_34.

Hjørland, B. (2012). Is classification necessary after Google? *Journal of Documentation*, *68*(3), 299–317. https://doi.org/10.1108/00220411211225557.

Hjørland, B. (2021). Information retrieval and knowledge organization: A perspective from the history of science. *Information*, *12*(3), Article 135. https://doi.org/10.3990/Info12030135.

Jansen, J. (2011). *Understanding sponsored search: Core elements of keyword advertising*. Cambridge University Press.

Jansen, M. B. J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the Web. *Information Processing & Management*, *36*(2), 207–227. https://doi.org/10.1016/S0306-4573(99)00056-4.

Jeanneneney, J.-N. (2007). *Google and the myth of universal knowledge* (T. L. Fagan, Trans.). Chicago University Press.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *KDD '02: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 133–142). Association for Computing Machinery.

Juel Vang, K. (2013). Ethics of Google's knowledge graph: Some considerations. *Journal of Information, Communication and Ethics in Society*, *11*(4), 245–260. https://doi.org/10.1108/JICES-08-2013-0028.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, *46*(5), 604–632. https://doi.org/10.1145/324133.324140.

Kraft, A., & Usbeck, R. (2022). The lifecycle of "facts": A survey of social bias in knowledge graphs. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computation Linguistics and the 12th International Conference on Natural Language Processing* (Vol.

1, pp. 639–652). Association for Computational Linguistics. https://aclanthology.org/2022.aacl-main.49.pdf.

Lamdan, S. (2022). *Data cartels*. Stanford University Press.

Langville, A. N., & Meyer, C. D. (2011). *Google's PageRank and beyond*. Princeton University Press.

Lewandowski, D. (2011). The influence of commercial intent of search results on their perceived relevance. *iConference '11: Proceedings of the 2011 iConference*, 452–458. Association for Computing Machinery.

Markey, K. (2007). Twenty-five years of end-user searching, Part 1: Research findings. *Journal of the American Society for Information Science and Technology*, *58*(8), 1071–1081. https://doi.org/10.1002/asi.20462.

Markey, K. (2007). Twenty-five years of end-user searching, Part 2: Future research directions. *Journal of the American Society for Information Science and Technology*, *58*(8), 1123–1130. https://doi.org/10.1002/asi.20601.

Markey, K., Atherton, P., & Newton, C. (1980). An analysis of controlled vocabulary and free text search statements in online searches. *Online Information Review*, *4*(3), 225–236. https://doi.org/10.1108/eb024031.

McMurdo, G. (1995). How the Internet was indexed. *Journal of Information Science*, *21*(6), 479–489. https://doi.org/10.1177/016555159502100607.

Metzler, D., & Croft, W. B. (2005). A Markov random field model for term dependencies. In *SIGIR '05: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 472–479). Association for Computing Machinery. https://doi.org/10.1145/1076034.1076115.

Metzler, D., Tay, Y., Bahri, D., & Najork, M. (2021). Rethinking search: Making domain experts out of dilettantes. *ACM SIGIR Forum*, *55*(1), 1–27. https://doi.org/10.1145/3476415.3476428.

Mikolov, T., Yih, W. T., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies* (pp. 746–751). Association for Computational Linguistics. https://aclanthology.org/N13-1090.

Nayak, P. (2022, August 11). New ways we're helping you to find high-quality information. *The Keyword*. https://blog.google/products/search/information-literacy/.

Oltmann, S. M., Knox, E. J. M., & Peterson, C. (2021). The significance of access to information—And the challenges it faces in librarianship. *Library Philosophy and Practice*, *5094*. https://digitalcommons.unl.edu/libphilprac/5094.

Palfrey, J. (2015). *Bibliotech: Why libraries matter more than ever in the age of Google*. Basic Books.

Pandell, L. (2017). An oral history of the hashtag. *Wired*. https://www.wired.com/2017/05/oral-history-hashtag/.

Pinkerton, B. (2000). *WebCrawler: Finding what people want* [Doctoral dissertation]. University of Washington.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. OpenAI. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Rafferty, P. (2018). Tagging. *Knowledge Organization*, *45*(6), 500–516. https://doi.org/10.5771/0943-7444-2018-6-500.

Raieli, R. (2022). *Web-scale discovery services: Principles, applications, discovery tools and development hypotheses*. Chandos Publishing. https://doi.org/10.1016/C2020-0-03523-1.

Reider, B. (2020). *Engines of order: A mechanology of algorithmic techniques*. Amsterdam University Press.

Reidsma, M. (2019). *Masked by trust: Bias in library discovery*. Litwin Books.

Rualthanzauva, B. (2014, February 5). The history of grep, the 40 years old Unix command. *Medium*. https://medium.com/@rualthanzauva/grep-was-a-private-command-of-mine-for-quite-a-while-before-i-made-it-public-ken-thompson-a40e24a5ef48.

Sauvayre, R. (2022). Types of errors hiding in Google Scholar data. *Journal of Medical Internet Research*, *24*(5), Article 28354. https://doi.org/10.2196/28354.

Shah, C., & Bender, E. M. (2022). Situating search. In *CHIIR '22: Proceedings of the 2022 SIGIR Conference on Human Interaction and Retrieval* (pp. 221–232). Association for Computing Machinery. https://doi.org/10.1145/988672.988675.

Si, L., He, Y. and Liu, L. (2023). Topics and changing characteristics of knowledge organization research in the 21st century: A content analysis. *Journal of Documentation, 79*(2), 487–508. https://doi.org/10.1108/JD-05-2022-0101

Singhal, A. (2012). Introducing the knowledge graph: Things, not strings. *The Keyword*. https://blog.google/products/search/introducing-knowledge-graph-things-not/.

Spink, A., Wolfram, D., Jansen, M. B. J., & Saracevic, T. (2001). Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, *52*(3), 226–234. https://doi.org/10.1002/1097-4571(2000)9999:9999<::AID-ASI1591>3.0.CO;2-R.

Sundin, O., Lewandowski, D., & Haider, J. (2022). Whose relevance? Web search engines as multisided relevance machines. *Journal of the Association for Information Science and Technology*, *73*(5), 637–642. https://doi.org/10.1002/asi.24570.

Taylor, A. G. (1995). On the subject of subjects. *The Journal of Academic Librarianship*, *21*(6), 484–491. https://doi.org/10.1016/0099-1333(95)90097-7.

Tharani, K. (2021). Much more than a mere technology: A systematic review of Wikidata in libraries. *Journal of Academic Librarianship*, *47*(2), Article 102326. https://doi.org/10.1016/j.acalib.2021.102326.

Tillotson, J. (1995). Is keyword searching the answer? *College & Research Libraries*, *56*(3), 199–206. https://doi.org/10.5860/crl_56_03_199.

Van der Veer Martens, B. (2015). An illustrated introduction to the infosphere. *Library Trends*, *63*(3), 317–361. https://doi.org/10.1353/lib.2015.0006.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, & R. Fergus (Eds.), *NIPS '17: Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 5999–6009). Curran Associates.

Zhang, L., Zhang, J., Ke, X., Li, H., Huang, X., Shao, Z., Cao, S., & Lv, X. (2023). A survey on complex factual question answering. *AI Open*, *4*, 1–12. https://doi.org/10.1016/j.aiopen.2022.12.003.

Zhitomirsky-Geffet, M. (2019). Towards a diversified knowledge organization system: An open network of inter-linked subsystems with multiple validity scopes. *Journal of Documentation*, *75*(5), 1124–1138. https://doi.org/10.1108/JD-10-2018-0163.

Zhitomirsky-Geffet, M., & Hajibayova, L. (2020). A new framework for ethical creation and evaluation of multi-perspective knowledge organization systems. *Journal of Documentation*, *76*(6), 1459–1471. https://doi.org/10.1108/JD-04-2020-0053.

Zobel, J., & Moffat, A. (2006). Inverted files for text search engines. *ACM Computing Surveys*, *38*(2), 1–56. https://doi.org/10.1145/1132956.1132959.

Algorithms, Users

10

Abstract

This chapter explores the evolution and economics of search engine results pages, the rise of computational advertising, search engines and the public sphere, and the potential future of search engines as supplemented and/or supplanted by artificial intelligence chatbots.

Online search activities have become so pervasive today that they are considered almost invisible (Sundin et al. 2017). Users employ major search engines such as Bing and Google for everyday exploratory, informational, navigational, and transactional queries (Broder 2002; Palagi et al. 2017) or various combinations of each (Li et al. 2022) with the expectation of reasonably relevant (Abualsaud and Smucker 2022) and credible (Peterson et al. 2022) results. This chapter discusses how these expectations relate to the realities of search engines today and their potential futures.

10.1 Search Engine Results Pages

Commercial search engines have substantial incentives to facilitate the search process through both privatizing and publicizing the search platform on which it takes place (Airoldi and Rokka 2022). The earliest versions of so-called "sponsored search" (the use of keywords in a search engine to point to online advertising sites) matched user-input keywords with relevant advertisements using standard information retrieval algorithms (bag-of-words features like tf-idf) in which each ad was treated as a document consisting of textual information such as the advertising copy and the bid keyword. Since then, Google has devoted massive research attention to natural language query linguistics, notably its recent Multitask Unified Model (MUM), which is said to be 1,000 times

B. V. V. Martens, *Keywords In and Out of Context*, Synthesis Lectures on Information Concepts, Retrieval, and Services, https://doi.org/10.1007/978-3-031-32530-4_10

more powerful than BERT, since it is trained across 75 different languages and has the capacity to understand both images and text (Nayak 2021). Bing has devoted more attention to its coverage of websites, developing IndexNow, an open source protocol which allows website owners to immediately notify Bing of changes in their sites rather than waiting for search engine crawlers to discover these changes (Schwartz 2021), with the goal of providing the most up-to-date site information for users.

Crane (2011, p. 113) observed that "The technological affordances of search engine interfaces and the attending user experiences materially anchor the discursively constructed myth of disintermediation in design and practice. …Results give the impression of neutrality because they are organized according to ranking systems rationalized by the scientistic logic of computation." Visible on every search engine results page (SERP), however, there is evidence of the continuing tension among what users want (accessible, affordable, and reasonably accurate answers to their questions), what organizations want (cost-effective, immediate, and presumably profitable linkage to their proprietary content), and what commercial search engine providers want (comprehensive control over this "matching" process, regardless of the specifics of the individual queries or content).

Hargittai (2000, p. 235) correctly identified the generic source of this tension as being the result of the relative abundance of available information and the relative scarcity of attention for that information in general, but neglected its converse in the particular: the relative scarcity of certain specific information and the relative abundance of attention for that information, much of which tends to be financial in nature. As Segev (2010) noted, this tension marks some of the most disputed territories within the global search engine society. The increase in so-called "search engine law" is indicative of the extent of its societal impact (Grimmelmann 2007; Volkoh and Falk 2012) in protecting both proprietary content and the public interest in an environment in which technological advances routinely outpace regulatory responses.

Various specific issues of information access, both legal and illegal, are constantly circulating around search engines. The analysis of Google search trends, as a predictor of global issues such as public health (Lin et al. 2020) or as a monitor of the diffusion of conspiracy theories (Ballatore 2015), the detection of online child sexual abuse (Lee et al. 2020), increases in cybercrime (Moneva et al. 2022), changing societal mores (Mejova et al. 2022), and the prevalence of deliberate disinformation campaigns in both the commercial and political sectors (Bradshaw 2019; Zade et al. 2022) are receiving continual research attention. Keywords are the surrogates for this tension, as they provide the data around which the entire search engine industry has been built. It is important to note that advertisers do not pay for the mere display of their online ads but rather when users click on these.

Google dominates the online keyword ecology in both organic (non-paid) and sponsored (paid) keywords, with its ad tech stack, an extensive series of coordinated advertising programs that allow advertisers to bid on keywords either within the Google SERP itself or in other sites that either partner with Google or which incorporate

ads via Google for additional monetization of their own content (Morozovaite 2021). Lewandowski et al. (2018) classified the SERP into four sections: organic results, which are results generated from the search engine's index of web pages; sponsored results, which are ads tied to the relevant keywords; universal results, which are generated from vertical search engine indices such as those for news or images, either similar to organic results or from a particular collection of pre-defined sources, and knowledge-graph results, which are answers or facts directly presented in response to the query. Google claims to maintain a strict separation between its organic and advertising businesses (Google, n.d.a) but it has become increasingly difficult to distinguish between the various types of results on a SERP over the years (Farley 2022). Commercial search engines have little incentive to address this issue, given the financial stakes involved (Lewandowski et al. 2020).

The stakes for user attention are clearly very high: the number one organic result on a SERP can expect to receive a 27.1% click-through rate (CTR) on average, which is ten times the rate of the 10th result on the SERP (Dean 2022). The CTR for a branded keyword in this number one position, on the other hand, is a much higher 60.4%, most likely due to the fact that these are often navigational searches for specific brands. It is also important to note that Google's classification algorithms group semantically similar "search terms" input by users in order to group these as "keywords" on which advertisers can bid, with specifications such as an exact match, a broad match, a negative match, or a phrase match (Google, n.d.b).

Research has repeatedly found that many users still cannot distinguish between a search engine's organic (non-paid) and sponsored (paid) results (Lewandowski and Schultheiß 2022), are unaware that there are various ways to improve search results by using search operators and search settings (Liu et al. 2019) or that there may be dramatic differences among the top-ranked results from queries to different search engines (Steiner et al. 2022) or that search engine results on debated topics biased toward a particular viewpoint can favorably influence user attitudes toward that viewpoint in the so-called "search engine manipulation effect" (Draws et al. 2021) or that there is a highly sophisticated search engine optimization industry among advertisers intended to affect these results (Schultheiß and Lewandowski 2020).

From an LIS point of view, these are often considered "information literacy" or "media literacy" issues suggesting the need for more user education or librarian advocacy (Tripodi et al. 2023). From an advertising or sponsorship point of view, such user engagement with their online content is actually an indicator of the strength of their "consumer signals" as part of their "customer journey" and is clearly beneficial (Schweidel et al. 2022). From a psychological point of view, however, the search engine's increasingly sophisticated combination of user personalization within a dynamic "choice environment," real-time adaptive feedback to user inputs and actions, its maximization of predictive capacity through optimization of resources, and the opacity of the algorithms and technologies involved creates the possibility of "hypernudges" that can lead the user in particular directions, whether or not these are initially or ultimately desired by the user (Mills 2022).

10.2 Computational Advertising

As search engine algorithms become more sophisticated, even to the point of being able to separate human-written content from that of AI-written content in order to assess variation in quality (Bahri et al. 2021), so-called "computational advertising," which involves granular-level data collection, mining, and aggregation, resulting in highly individualized and pervasive ad serving, has been developed in response (Helberger et al. 2020).

In their review of the literature on keyword decision-making process by digital advertisers, Yang and Li (2023) describe this process as involving keyword generation at the domain level, keyword targeting at the market level, keyword assignment and grouping at the campaign and ad-group level, and dynamic keyword adjustment, often involving multiple search engines, social media, and other online advertising venues. Long-term considerations can also come into play, especially in the establishment of a brand position, as well as expertise in various consumer click actions within the so-called "purchase funnel" (Erdmann et al. 2022).

Keyword generation generally starts with "seed" keywords, which represent the domain knowledge of the target market, which are then expanded as comprehensively as possible, based either statistically, as shown by co-occurrence relationships among keywords, through data-mining previous queries, metadata, and other ads, or ontologically, based on concept hierarchies. The cost of the specific keyword, whether it refers to a brand or trademark, whether or not the brand or trademark is strongly defended online, and the presence or absence of competitors' bids, all impact the decision process (Simonov et al. 2018). The possibility of "search engine poisoning," that is, the dilution of a particular keyword through competitors' deliberate linkage to misspelled variants (Joslin et al. 2019) or the deliberate diversion of keyword traffic by cybercriminals from a legitimate result to an illegal site (Fittler et al. 2022) must also be taken into account.

The frequent modifications to search engine algorithms and protocols by Google that impact advertising make this very challenging, since these keyword decisions necessarily interact with many other marketing and advertising decisions, particularly the optimal budget allocation and bid determination in order to maximize the performance of campaigns. Keyword auction bidding today by major firms is generally handled by specialized intermediaries, who also use their own algorithms in order to determine the best bidding strategy for campaigns across various online venues (Decarolis and Rovigatti 2021).

The 100 most popular keyword searches for 2022 (excluding pornography searches) in order of frequency were Facebook, YouTube, Amazon, weather, Walmart, Google, Wordle, Gmail, Target, Home Depot, Google Translate, Yahoo Mail, Yahoo, Costco, Fox News, Starbucks, food near me, translate, Instagram, Google Maps, Walgreens, Best Buy, NBA, McDonalds, restaurants near me, NFL, Amazon Prime, CNN, traductor, weather tomorrow, ESPN, Lowes, Chick-Fil-A, news, food, Zillow, Craigslist, CVS, eBay, Twitter,

Wells Fargo, USPS tracking, Bank of America, calculator, Indeed, NFL scores, Google Docs, Etsy, Netflix, Taco Bell, Shein, astronaut, Macys, Kohls, YouTube TV, Dollar Tree, gas station, coffee, NBA scores, Roblox, restaurants, AutoZone, Pizza Hut, USP, Gmail login, Dominos, Chipotle, Google Classroom, tiempo, Hotmail, AOL Mail, Burger King, Facebook login, Google flights, SQM Club, maps, Subway, Dow Jones, Sam's Club, motel, breakfast, English to Spanish, gas, FedEx, Walmart near me, Old Navy, FedEx tracking, Southwest Airlines, Ikea, LinkedIn, Airbnb, Omegle, Planet Fitness, pizza, Spanish to English, Google Drive, MSN, Dunkin Donuts, Capital One, and Dollar General (Hudgens 2022). Locational searches (either geographic or virtual) for services and shopping clearly predominated, with weather and sports news a distant second.

In 2022, the 25 most expensive (cost per click) Google AdSense sponsored keyword phrases were, with some exceptions, not brand specific: Best 18 wheeler accident lawyer, Offshore accident attorney, Construction truck accident lawyer, Best motorcycle attorney, Spectrum business fiber, Flagstar wholesale, Meso law firm, Mesothelioma compensation for family members, Facebook estimated ad recall, Liberty Life structured settlements, Verizon business wifi, Birth injuries law firm, Neonatal intensive care unit, Accredited online colleges for business management, Best business phone line providers, WGU cybersecurity, Google cloud VPS hosting, Donate my car, Sell my home, Telemarketing dialer system, SEO marketing company, Online slots list, Top alcohol rehab centers, Compare car insurance, and Online marketing degree. Competition among law firms for lucrative lawsuit opportunities dominated most of this list, with other business-related searches a distant second (Schrobo 2022).

Although Google is considered the main focus of the online advertising market, social media platforms such as Facebook and ecommerce platforms such as Amazon also utilize keyword bidding auctions for their advertisers. Facebook's marketing algorithms focus on audience demographics, interests, and behaviors within the Facebook environment in order to determine which of the relevant sponsored ads will be shown to individual users. Amazon employs a product relevance ranking algorithm in addition to its keyword bidding auction.

With 1.2 trillion searches per year worldwide, Google currently has roughly 83% of the global search market, while Bing has about 9%, and Yahoo approximately 2.5%. The Baidu search engine, however, has 86% of all search engine users in China, while the Yandex search engine has over 59% of all search engine users in Russa (Bianchi 2023).

Online advertising-related search revenues in the US for 2021 have been estimated at $78.3 billion (Internet Advertising Bureau 2022), while online advertising revenue for Google alone totaled $209.49 billion, which is approximately 80% of its parent company Alphabet's total revenue (Bianchi 2022b).

10.3 Search Engines and the Public Sphere

Shackell (2021) argues that existing trademark law, originally intended to reduce search costs for consumers, has been exploited by Google and other search engines in that trademarks such as "Google" have become "super-generic" because they are both common in everyday language about search and, more crucially, structurally entailed in search. He notes that "a small number of firms now hold trademarks on functions whose pre-Internet equivalents were generic actions such as 'seeking' or 'looking around.' The social costs and hidden externalities created are so ubiquitous and complex that they may take generations to fully perceive and understand" (2021, p. 19). While U.S. trademark law does not currently prohibit the use of competitors' trademarks in search engine advertising or hold search engines liable for any infringement (Franklyn and Hyman 2013), India for example has enacted regulations that do make search engines liable (Pandey 2021). Ironically, Google, like other major technology firms, very vigorously defends its own trademarks (Pritchard 2022).

Further, the larger search engines such as Google and Baidu operate as quasi-monopolies (Si et al. 2022). The U.S. Justice Department filed an anti-trust lawsuit against Google in 2020, claiming that Google exerts monopolistic power over general Internet search services in the United States, search advertising in the United States, and general search text advertising in the United States and citing specific harms to consumers and competitors due to Google's extensive contractual arrangements with leading manufacturers to be the default search engine provided in their products (Bet et al. 2022). Google Chief Economist Hal Varian argues to the contrary, stating that actual economic data for Google shows that such claims are without merit (Varian 2021).

Regardless of the outcome of this particular case, Google's growing power over information access troubles many observers. Google has been the target of various lawsuits claiming that their First Amendment rights had been curtailed by the search engine's algorithms and policies, although most of these have been decided in Google's favor as a result of its constant defense that its practices are "editorial" in nature, similar to that of a newspaper editor selecting specific news stories to run, and that such practices are protected by free speech doctrines. Petersen (2022, p. 12) points out that this analogy is misleading at best: "declaring search engines akin to editors evades complexity in the hope that the companies will either act as trustees of the public interest, diagnosing what informational policies best serve the public interest and enacting them, or that the companies' interests will align with those of the users or public through the alchemy of market mechanisms."

Search engines are also criticized for their aggregation of news items, which are said to have had deleterious effects on traditional news channels (Chyi et al. 2016) and also for a lack of transparency regarding how news sources are algorithmically selected for dissemination of their news items (Nechushtai and Lewis 2019).

Of particular concern is the role of search engines in the dissemination of political information, with their unchecked ability to both amplify and distort perspectives on various issues, as a result of the ongoing technological transformation of the public arena during the twenty-first century (Jungherr and Schroeder 2021). Kawakami et al. (2020) identified 56 news sources that contributed two thirds of the so-called "Top Stories" (a panel of recent news stories shown at the top of the SERP) for 30 political candidates running in the primaries for the 2020 U.S. presidential election and then surveyed voters to see how familiar they were with these sources, finding that some of the most frequent outlets were not familiar to all voters or were not particularly trusted by voters of any political persuasion, leading the researchers to the conclusion that Google was attempting to sample news articles from sources with different political learnings to offer balanced coverage. The researchers note that this is reminiscent of the "fairness doctrine" policy required of broadcast media up until 1997 to air contrasting views about controversial matters. However, because there are more centrist and left-leaning web publications than right-leaning ones, this sampling provided hyper-partisan far-right news sources with more visibility than other, more familiar, and trusted sources received.

Hu et al. (2019) analyzed Google search snippets from numerous webpages about American political issues and discovered that these brief top of page summaries amplified their partisan viewpoints rather than providing objective summaries of the webpages, most likely due to Google's use of extractive rather than abstractive summarization in providing the snippets. In the first large-scale study of publicly available platform disclosures of political advertising during the 2020 U.S. presidential election, researchers found that these fell woefully short of any promised transparency (Papakyriakopoulos et al. 2022).

Fukuyama et al. (2021) have suggested that one possible solution to the current information dominance of search engines such as Google would be for government to mandate a new layer of independent, competitive middleware services with transparent algorithms to serve as gatekeepers between users and the search engines, which would allow users to better determine how information is curated, filtered, and presented to them. Another proposal is to create a separate index on top of a public infrastructure open to everyone, which could form the basis for myriad search engines and other services utilizing Web data (Lewandowski 2019).

Meanwhile, Google's ongoing expansion into mobile technologies has helped to make it almost omnipresent in daily life, as 61% of all Google users accessed it through their mobile phones in 2020 and Google currently has a 93.25% share of the U.S. mobile search market (Bianchi 2022a). Accordingly, Google now employs mobile-first indexing, which means that the mobile version of a website is preferentially indexed and ranked and that having such a mobile-friendly version is strongly recommended as a best practice for all websites. Google Maps' dominance in the online mapping applications category is clearly evident in this as well (Ceci 2023).

Zuboff (2015) argues that Google in particular has profited from what she terms "surveillance capitalism" in which user data (queries, metadata, and associated information) are offered in exchange for free use of the search engine's services and are then stored, mined, aggregated, processed, and utilized for a wide variety of business purposes which are both opaque to users and highly profitable to Google. Zuboff points out as well that Google's search operations are so sophisticated that U.S. government intelligence agencies have used it as a model for their own surveillance systems.

Velkova and Kaun (2021, p. 526) suggest, however, that users and algorithms can and do engage in co-creation and co-curation of this data, offering multiple opportunities for both resistance and repair: "the dominant politics through which algorithmic power functions is an attention politics that defines dominant meanings and representations of objects, people and events. At the same time, what counts as worth being promoted or demoted by algorithms is shaped in a complex interplay between them and users. Algorithms are dependent on users for the queries that make them work and for the generation of data. User-generated data must be produced and aggregated so that algorithms can start to 'define which information is to be included in an analysis; [and] envision, plan for, and execute data trans-formations' (Ananny 2016, p. 98). On the other hand, users are not passive observers in this process. They increasingly recognize the role that they play in shaping the workings of algorithms, and they have begun to strategically intervene in political, commercial, or playful ways in the algorithmic politics of attention." Some of the ways in which consumers resist are algorithm hacking (Muhl 2020), ad avoidance and blocking (Todri 2022), actively complaining and critiquing (Griffin and Lurie 2022), critical ignoring (Kozyreva et al. 2022), and the self-imposing of filter bubbles to avoid particular types of content (Ekström et al. 2022).

10.4 The Future of Search

The recent release of the artificial intelligence conversational agent, ChatGPT by AI Lab, however, is said to have the potential to disrupt the current search engine environment with the agent's ability to provide engaging question and answer responses to natural language queries (Grant and Metz 2022). ChatGPT is a product of the GPT-3 (Generative Pre-trained Transformer 3) transformer neural network model, which was trained on roughly 45 terabytes of text (equivalent to 300 billion words) run through 175 billion parameters. Microsoft is reported to be readying to insert ChatGPT capabilities into its Bing search engine (Bass 2023), however, the beta release has been filled with factual, technical, ethical, and even existential problems (Warren 2023; Willison 2023). The Chinese search engine Baidu is similarly engaged in launching a AI-based chatbot named Ernie Bot (Soo 2023).

Although Google already has a similar AI-based generative language model in development called LaMDA (Language Model for Dialog Applications), the challenge is clearly

not merely a technological one, though Google has expressed concerns that such models are not yet suitable for wide popular release, given their known biases and inaccuracies. And, indeed, Google's promotional launch of its artificial intelligence conversation agent, Bard, was marred by an embarrassing factual mistake that Bard produced in Google's initial advertisement, which was intended to demonstrate the product's capabilities. Bard claimed that the James Webb Space Telescope was the first to take photographs of exoplanets beyond our solar system, when in fact it was the European Southern Observatory's Very Large Telescope that did so in 2004: an egregious error that cost Google's parent company Alphabet $100 billion as its share price fell in response to the ensuing ridicule (Olson 2023).

However, the real economic challenge is that an efficient and effective response to a natural language query, either textual or verbal, does not provide much room for the advertising and associated information that underpins most current search engines' business models. For instance, on January 2, 2023, the current Google Knowledge Panel provided in response to the query "How old is planet Earth?" specified "4.534 billion years, plus or minus about 50 million years" at the top, with a link to the *National Geographic*'s Age of Earth Collection, as well as offering numerous additional links on the rest of the SERP. The first 10 of these were: *National Geographic*, Space.com, Answersin-Genesis.org, the American Museum of Natural History, Wikipedia, HowStuffWorks.com, *New Scientist*, *Smithsonian* magazine, EarthSky.org, and the *Encyclopedia Britannica.* Despite the fact that these results are clearly organic, almost all of them benefit in some measure from the publicity offered by proximity to what can be perceived as the definitive scientific answer to a common query. The same query placed on the Bing search engine resulted in the same answer, supported by the *National Geographic*'s Age of Earth Collection, followed by an "Explore Further (Recommended to You Based on What's Popular)" list of links to NatureNoon.com, Christianity.com, NationalGeographic.com, ScienceTrends.com, and BibleStudy.com, with a similarly varied assortment of additional links following these. It is unclear how a dialogue with a search engine conversational agent could successfully incorporate more information than the basic response.

However, the reliance of search engines on their sponsored search revenue may be the deciding factor in the next wave of keyword innovation, which could be a more multimodal culture of interactive communication, as shown by the popularity of such voice assistants as Google's Siri (Moriuchi 2019) and Baidu's DuerOS (Shen et al. 2022) and the potential for memory-prompting by these devices as well as for ordinary search activities (Atkinson and Barker 2021). All this could well represent an updated version of Vannevar Bush's celebrated concept of an individual researcher's "associative trails" while using the hypothetical "Memex" that could instead be termed an individual user's "accumulative trails" while using those search engines available today. What Bush could not envision was the uses to which governments, commercial entities, and even criminal enterprises could put the keywords found in such individual and aggregated trails on the

Internet. At this point, however, we *can*: though what can or should be done is a topic of continued debate that is likely to generate many new keywords in future.

References

Abualsaud, M., & Smucker, M. (2022). The dark side of relevance: The effect of non-relevant results on search behavior. In *CHIIR '22: Proceedings of the 2022 ACM SIGIR Conference on Human Information Interaction and Retrieval* (pp. 1–11). Association for Computing Machinery. https://doi.org/10.1145/3498366.3505770.

Airoldi, M., & Rokka, J. (2022). Algorithmic consumer culture. *Consumption Markets & Culture*, *25*(5), 411–428. https://doi.org/10.1080/10253866.2022.2084726.

Ananny, M. (2016). Toward an ethics of algorithms: Convening, observation, probability, and timeliness. *Science, Technology, & Human Values*, *41*(1), 93–117. https://doi.org/10.1177/0162243915606523.

Atkinson, P., & Barker, R. (2021). Hey, Alexa, what did I forget? Networked devices, Internet search and the delegation of human memory. *Convergence*, *27*(1), 52–65. https://doi.org/10.1177/1354856520925740.

Bahri, D., Tay, Y., Che, Z., Metzler, D., Brunk, C., & Tomkins, A. (2021). Generative models are unsupervised predictors of page quality: A colossal-scale study. In *WSDM '21: Proceedings of the 14th ACM International Conference on Web Search and Data Mining* (pp. 301–309). Association for Computing Machinery. https://doi.org/10.1145/3437963.3441809.

Ballatore, A. (2015). Google chemtrails: A methodology to analyze topic representation in search engine results. *First Monday*, *20*(7). https://doi.org/10.5210/fm.v20i7.5597.

Bass, D. (2023, January 3). Microsoft hopes OpenAI's chatbot will make Bing smarter. *Bloomberg News*. https://www.bloomberg.com/news/articles/2023-01-04/microsoft-hopes-openai-s-chatbot-will-make-bing-smarter.

Bet, G., Blair, R. D., & Donna, J. D. (2022). The economic rationale of United States v. Google. *The Antitrust Bulletin*, *67*(1), 23–39.https://doi.org/10.1177/0003603X211067116.

Bianchi, T. (2022a, December 1). U. S. market share of mobile search engines 2012–2022a. *Statistica.* https://www.statista.com/statistics/511358/market-share-mobile-search-usa/.

Bianchi, T. (2022b, December 2). Google: Global annual revenue 2002–2021. *Statistica.* https://www.statista.com/statistics/266206/googles-annual-global-revenue/.

Bianchi, T. (2023, January 6). Global desktop market share of search engines 2015–2022. *Statistica.* https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/.

Bradshaw, S. (2019). Disinformation optimised: Gaming search engine algorithms to amplify junk news. *Internet Policy Review*, *8*(4), 1–24. https://doi.org/10.14763/2019.4.1442.

Broder, A. (2002). A taxonomy of web search. *SIGIR Forum*, *36*(2), 3–10. https://doi.org/10.1145/792550.792552.

Ceci, L. (2023, February 7). Most popular navigation apps in the U.S. 2022, by downloads. *Statistica.* https://www.statista.com/statistics/865413/most-popular-us-mapping-apps-ranked-by-audience/.

Chyi, H. I., Lewis, S. C., & Zheng, N. (2016). Parasite or partner? Coverage of Google News in an era of news aggregation. *Journalism & Mass Communication Quarterly*, *93*(4), 789–815. https://doi.org/10.1177/1077699016629370.

Crane, M. (2011). The cultural logic of search and the myth of disintermediation. In C. McCarthy, H. Greenhalgh-Spencer, & R. Mejia (Eds.), *New times: Making sense of critical/cultural theory in a digital age* (pp. 107–128). Peter Lang.

Dean, B. (2022, October 14). We analyzed 4 million Google Search results. *Backlinko*. https://backlinko.com/google-ctr-stats.

Decarolis, F., & Rovigatti, G. (2021). From mad men to maths men: Concentration and buying power in online advertising. *American Economic Review*, *111*(10), 3299–3327. https://doi.org/10.1257/aer.20190811.

Draws, T., Tintarev, N., Gadiraju, U., Bozzon, A., & Timmermans, B. (2021). This is not what we ordered: Why biased search result rankings affect user attitudes on debated topics. In *SIGIR '21: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 295–305). Association for Computing Machinery. https://doi.org/10.1145/3404835.3462851.

Ekström, A. G., Niehorster, D. C., & Olsson, E. J. (2022). Self-imposed filter bubbles: Selective attention and exposure in online search. *Computers in Human Behavior*, *7*, Article 100226. https://doi.org/10.1016/j.chbr.2022.100226.

Erdmann, A., Arilla, R., & Ponzoa, J. M. (2022). Search engine optimization: The long-term strategy of keyword choice. *Journal of Business Research*, *144*, 650–662. https://doi.org/10.1016/j.jbusres.2022.01.065.

Farley, N. (2022, December 17). Is Google Search getting worse? *Search Engine Land*. https://searchengineland.com/is-google-search-getting-worse-389658.

Fittler, A., Paczolai, P., Ashraf, A. R., Pourhashemi, A., & Iványi, P. (2022). Prevalence of poisoned Google search results of erectile dysfunction medications redirecting to illegal Internet pharmacies: Data analysis study. *Journal of Medical Internet Research*, *24*(11), Article e38957. https://doi.org/10.2196/38957.

Franklyn, D. J., & Hyman, D. A. (2013). Trademarks as search engine keywords: Much ado about something? *Harvard Journal of Law & Technology*, *26*(2), 481–543.

Fukuyama, F., Richman, B., & Goel, A. (2021). How to save democracy from technology: Ending big tech's information monopoly. *Foreign Affairs*, *100*(1). https://www.foreignaffairs.com/articles/united-states/2020-11-24/fukuyama-how-save-democracy-technology.

Google. (n.d.a). About measuring paid & organic search results. *Google Ads Help*. https://support.google.com/google-ads/answer/3097241.

Google. (n.d.b). About the search terms report. *Google Ads Help*. https://support.google.com/google-ads/answer/2472708?hl=en.

Grant, N., & Metz, C. (2022, December 21). A new chatbot is a 'Code Red' for Google's Search business. *New York Times*. https://www.nytimes.com/2022/12/21/technology/ai-chatgpt-google-search.html.

Griffin, D., & Lurie, E. (2022). Search quality complaints and imaginary repair: Control in articulations of Google Search. *New Media & Society*. Advance online publication. https://doi.org/10.1177/14614448221136505.

Grimmelmann, J. (2007). The structure of search engine law. *Iowa Law Review*, *93*(1), 3–63. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=979568.

Hargittai, E. (2000). Open portals or closed gates: Channeling content of the World Wide Web. *Poetics*, *27*(4), 233–253. https://doi.org/10.1016/S0304-422X(00)00006-1.

Helberger, N., Huh, J., Milne, G., Strycharz, J., & Sundaram, H. (2020). Macro and exogenous factors in computational advertising: Key issues and new research directions. *Journal of Advertising*, *49*(4), 377–393. https://doi.org/10.1080/00913367.2020.1811179.

Hu, D., Jiang, S., Robertson, R. E., & Wilson, C. (2019). Auditing the partisanship of Google Search snippets. In L. Liu & R. White (Eds.), *Proceedings of the World Wide Web Conference WWW 2019* (pp. 693–704). Association for Computing Machinery. https://doi.org/10.1145/3308558.3313654.

Hudgens, R. (2022, December 22). The 100 most popular Google keywords to inspire 2023. *Siege-Media*. https://www.siegemedia.com/seo/most-popular-keywords.

Internet Advertising Bureau. (2022). *Internet advertising revenue report: Full year 2021 results*. https://www.iab.com/insights/internet-advertising-revenue-report-full-year-2021/.

Joslin, M., Li, N., Hao, S., Xue, M., & Zhu, H. (2019). Measuring and analyzing search engine poisoning of linguistic collisions. In *Proceedings: 2019 IEEE Symposium on Security and Privacy SP 2019* (pp. 1311–1325). IEEE. https://doi.org/10.1109/SP.2019.00025.

Jungherr, A., & Schroeder, R. (2021). Disinformation and the structural transformations of the public arena: Addressing the actual challenges to democracy. *Social Media + Society*, *7*(1). https://doi.org/10.1177/2056305121988928.

Kawakami, A., Kawakami, K., Huang, D., & Kawakami, E. (2020). The "Fairness Doctrine" lives on? Theorizing about the algorithmic news curation of Google's top stories. In *HT '20: Proceedings of the 31st ACM Conference on Hypertext and Social Media* (pp. 59–68). ACM. https://doi.org/10.1145/3372923.3404794.

Kozyreva, A., Wineburg, S., Lewandowsky, S., & Hertwig, R. (2022). Critical ignoring as a core competence for digital citizens. *Current Directions in Psychological Science*, *32*(1), 81–88. https://doi.org/10.1177/09637214221121570.

Lee, H. E., Ermakova, T., Ververis, V., & Fabian, B. (2020). Detecting child sexual abuse material: A comprehensive survey. *Forensic Science International: Digital Investigation*, *34*, Article 301022. https://doi.org/10.1016/j.fsidi.2020.301022.

Lewandowski, D. (2019). The Web is missing an essential part of infrastructure: An open Web index. *Communications of the ACM*, *62*(4), 24–24. https://doi.org/10.1145/3312479.

Lewandowski, D., Kerkmann, F., Rümmele, S., & Sünkler, S. (2018). An empirical investigation on search engine disclosure. *Journal of the Association for Information Science and Technology*, *69*(3), 420–437. https://doi.org/10.1002/asi.23963.

Lewandowski, D., & Schultheiß, S. (2022). Public awareness and attitudes towards search engine optimization.*Behavior & Information Technology*. Advance online publication.https://doi.org/10.1080/0144929X.2022.2056507.

Lewandowski, D., Sünkler, S., & Schultheiß, S. (2020). Studies on search: Designing meaningful IIR studies on commercial search engines. *Datenbank-Spektrum*, *20*(1), 5–15. https://doi.org/10.1007/s13222-020-00331-1.

Li, P., Zhang, Y., & Zhang, B. (2022). Understanding query combination behavior in exploratory searches. *Applied Sciences*, *12*(2), Article 706. https://doi.org/10.3390/app12020706.

Lin, Y.-H., Liu, C.-H., & Chiu, Y.-C. (2020). Google searches for the keywords of "wash hands" predict the speed of national spread of COVID-19 outbreak among 21 countries. *Brain, Behavior and Immunity*, *87*, 30–32. https://doi.org/10.1016/j.bbi.2020.04.020.

Liu, J., Wang, Y., Mandal, S., & Shah, C. (2019). Exploring the immediate and short-term effects of peer advice and cognitive authority on Web search behavior. *Information Processing & Management*, *56*(3), 1010–1025. https://doi.org/10.1016/j.ipm.2019.02.011.

Mejova, Y., Gracyk, T., & Robertson, R. E. (2022). Googling for abortion: Search engine mediation of abortion accessibility in the United States. *Journal of Quantitative Description: Digital Media*, *2*. https://doi.org/10.51685/jqd.2022.007.

Mills, S. (2022). Finding the "nudge" in hypernudge. *Technology in Society*, *71*, Article 102177. https://doi.org/10.1016/j.techsoc.2022.102117.

Moneva, A., Leukfeldt, E. R., Van De Weijer, S. G., & Miró-Llinares, F. (2022). Repeat victimization by website defacement: An empirical test of premises from an environmental criminology perspective. *Computers in Human Behavior*, *126*, Article 106984. https://doi.org/10.1016/j.chb.2021.106984.

Moriuchi, E. (2019). Okay, Google!: An empirical study on voice assistants on consumer engagement and loyalty. *Psychology & Marketing*, *36*(5), 489–501. https://doi.org/10.1002/mar.21192.

Morozovaite, V. (2021). Two sides of the digital advertising coin: Putting hypernudging into perspective. *Market and Competition Law Review*, *5*(2), 105–145. https://doi.org/10.34632/mclawreview.2021.5.2.

Muhl, K. (2020, January 9). The rise of "algorithm hacking" and how it may be leading marketers astray. *MarketingTech.* https://www.marketingtechnews.net/news/2020/jan/09/rise-algorithm-hacking-and-how-it-may-be-leading-marketers-astray/.

Nayak, P. (2021, May 18). MUM: A new AI milestone for understanding information. *The Keyword*. https://blog.google/products/search/introducing-mum/.

Nechushtai, E., & Lewis, S. C. (2019). What kind of news gatekeepers do we want machines to be? Filter bubbles, fragmentation, and the normative dimensions of algorithmic recommendations. *Computers in Human Behavior*, *90*, 298–307. https://doi.org/10.1016/j.chb.2018.07.043.

Olson, E. (2023, February 9). Google shares drop $100 billion after its new AI chatbot makes a mistake. *NPR Technology*. https://www.npr.org/2023/02/09/1155650909/google-chatbot--error-bard-shares.

Palagi, E., Gandon, F., Giboin, A., & Trouncy, R. (2017). A survey of definitions and models of exploratory search. In *ESIDA '17: Proceedings of the 2017 ACM Workshop on Exploratory Search and Interactive Data Analytics* (pp. 3–8). Association for Computing Machinery. https://doi.org/10.1145/3038462.3038465.

Pandey, T. (2021). Google AdWords liability for trademark infringement. *International Journal of Law, Management and Humanities*, *4*(3), 319–323. https://doi.org/10.10000/IJLMH.11456.

Papakyriakopoulos, O., Tessono, C., Narayanan, A., & Kshirsagar, M. (2022). How algorithms shape the distribution of political advertising: Case studies of Facebook, Google, and TikTok. In *AIES '22: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 523–546). ACM. https://doi.org/10.1145/3514094.3534166.

Petersen, J. (2022). Search engines and free speech: A historical analysis of editorial analogies and the position of media companies and users in US free speech discourse. *Television & New Media.* Advance online publication. https://doi.org/10.1177/15274764221137245.

Peterson, A. M., High, A. C., Maragh-Lloyd, R., Stoldt, R., & Ekdale, B. (2022). Trust in online search results during uncertain times. *Journal of Broadcasting & Electronic Media*, *66*(5), 751–771. https://doi.org/10.1080/08838151.2022.2141242.

Pritchard, D. (2022). Big tech trademarks: Trademark law empowers big tech to maintain market dominance. *Hastings Science and Technology Law Journal*, *13*(2), 101–126. https://repository.uchastings.edu/hastings_science_technology_law_journal/vol13/iss2/3/.

Schrobo, M. (2022, March 2). The most expensive keywords for 2022. *Hero Blog*. https://www.ppchero.com/the-most-expensive-keywords-for-2022/.

Schultheiß, S., & Lewandowski, D. (2020). "Outside the industry, nobody knows what we do": SEO as seen by search engine optimizers and content providers. *Journal of Documentation*, *77*(2), 543–557. https://doi.org/10.1108/JD-07-2020-0127.

Schwartz, B. (2021, October 18). IndexNow—New initiative by Microsoft and Yandex to push content to search engines. *Search Engine Land*. https://searchengineland.com/indexnow-new-initiative-by-microsoft-and-yandex-to-push-content-to-search-engines-375247.

Schweidel, D. A., Bart, Y., Inman, J. J., Stephen, A. T., Libai, B., Andrews, M., Babić Rosario, A., Chae, I., Chen, Z., Kupor, D., Longoni, C., & Thomaz, F. (2022). How consumer digital signals are reshaping the customer journey. *Journal of the Academy of Marketing Science*, *50*, 1257–1276. https://doi.org/10.1007/s11747-022-00839-w.

Segev, E. (2010). *Google and the digital divide: The bias of online knowledge*. Chandos Publishing.

Shackell, C. (2021). Is genericness still adequately defined? Internet search firms and the economic rationale for trademarks. *Science, Technology, & Human Values*. Advance online publication. https://doi.org/10.1177/01622439211055482.

Shen, W., He, X., Zhang, C., Zhang, X., & Xie, J. (2022). A transformer-based user satisfaction prediction for proactive interaction mechanism in DuerOS. In *CIKM '22: Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. Association for Computing Machinery. https://doi.org/10.1145/3498366.3505816.

Si, M., Si, Z., & Ye, Y. (2022). The analysis of Google's monopoly in the search engine industry. In *Proceedings of the 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022)* (pp. 3168–2173). Atlantis Press. https://doi.org/10.2991/aebmr.k.220307.356.

Simonov, A., Nosko, C., & Rao, J. M. (2018). Competition and crowd-out for brand keywords in sponsored search. *Marketing Science*, *37*(2), 200–215. https://doi.org/10.1287/mksc.2017.1065.

Soo, Z. (2023, February 22). *Baidu to implement ChatGPT-like Ernie Bot chatbot from March*. https://apnews.com/article/technology-science-baidu-inc-china-artificial-intelligence-427838070f962108cabed1e553fc3d12.

Steiner, M., Magin, M., Stark, B., & Geiß, S. (2022). Seek and you shall find? A content analysis on the diversity of five search engines' results on political queries. *Information, Communication & Society*, *25*(2), 217–241. https://doi.org/10.1080/1369118X.2020.1776367.

Sundin, O., Haider, J., Andersson, C., Carlsson, H., & Kjelberg, S. (2017). The search-ification of everyday life and mundane-ification of search. *Journal of Documentation*, *73*(2), 224–243. https://doi.org/10.1108/JD-06-2016-0081.

Todri, V. (2022). Frontiers: The impact of ad-blockers on online consumer behavior. *Marketing Science*, *41*(1), 7–18. https://doi.org/10.1287/mksc.2021.1309.

Tripodi, F. B., Stevenson, J. A., Slama, R., & Reich, J. (2023). Libraries combating disinformation: From the front line to the long game. *Library Quarterly*, *93*(1), 48–71. https://doi.org/10.1086/722551.

Varian, H. R. (2021). Seven deadly sins of tech? *Information Economics and Policy*, *54*, Article 100893. https://doi.org/10.1016/j.infoecopol.2020.100893.

Velkova, J., & Kaun, A. (2021). Algorithmic resistance: Media practices and the politics of repair. *Information, Communication & Society*, *24*(4), 523–540. https://doi.org/10.1080/1369118X.2019.1657162.

Volkoh, E., & Falk, D. (2012). *First amendment protection for search engine results: White paper commissioned by Google* (Research Paper No. 12–22). UCLA School of Law. https://doi.org/10.2139/ssrn.2055364.

Warren, T. (2023, February 14). These are Microsoft's Bing AI secret rules and why it says it's named Sydney. *The Verge*. https://www.theverge.com/23599441/microsoft-bing-ai-sydney-secret-rules.

Willison, S. (2023, February 15). Bing: "I will not harm you unless you harm me first." *Simon Willison's Weblog*. https://simonwillison.net/2023/Feb/15/bing/.

Yang, Y., & Li, H. (2023). Keyword decisions in sponsored search advertising: A literature review and research agenda. *Information Processing & Management*, *60*(1), Article 103142. https://doi.org/10.1016/j.ipm.2022.103142.

Zade, H., Wack, M., Zhang, Y., Starbird, K., Calo, R., Young, J., & West, J. D. (2022). Auditing Google's search headlines as a potential gateway to misleading content: Evidence from the 2020 US election. *Journal of Online Trust and Safety*, *1*(4). https://doi.org/10.54501/jots.v1i4.72.

Zuboff, S. (2015). Big other: Surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology*, *30*(1), 75–89. https://doi.org/10.1057/jit.2015.5.