Cognition 153 (2016) 182–195

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Original Articles

The length of words reflects their conceptual complexity

Molly L. Lewis *, Michael C. Frank

Department of Psychology, Stanford University, United States

ARTICLE INFO

Article history:
Received 28 August 2015
Revised 31 March 2016
Accepted 5 April 2016

Keywords:
Communication
Pragmatics
Information theory
Language evolution

ABSTRACT

Are the forms of words systematically related to their meaning? The arbitrariness of the sign has long been a foundational part of our understanding of human language. Theories of communication predict a relationship between length and meaning, however: Longer descriptions should be more conceptually complex. Here we show that both the lexicons of human languages and individual speakers encode the relationship between linguistic and conceptual complexity. Experimentally, participants mapped longer words to more complex objects in comprehension and production tasks and across a range of stimuli. Explicit judgments of conceptual complexity were also highly correlated with implicit measures of study time in a memory task, suggesting that complexity is directly related to basic cognitive processes. Observationally, judgments of conceptual complexity for a sample of real words correlate highly with their length across 80 languages, even controlling for frequency, familiarity, imageability, and concreteness. While word lengths are systematically related to usage—both frequency and contextual predictability—our results reveal a systematic relationship with meaning as well. They point to a general regularity in the design of lexicons and suggest that pragmatic pressures may influence the structure of the lexicon.

1. Introduction

Human languages are systems for encoding information about the world. A defining feature of a symbolic coding system is that there is no inherent mapping between the form of the code and what the code denotes (Peirce, 1931)—the color red holds no natural relationship to the meaning 'stop,' the numeral *3* holds no natural relationship to three units, and in language, the word "horse" looks or sounds nothing like the four-legged mammal it denotes. This arbitrariness of the linguistic sign has long been observed as a fundamental and universal property of natural language (Hockett, 1960; Saussure, 1916, 1960). And, despite the growing number of cases suggesting instances of non-arbitrariness in the lexicon (see Dingemanse, Blasi, Lupyan, Christiansen, & Monaghan, 2015; Schmidtke, Conrad, & Jacobs, 2014, for reviews), there is clear evidence for at least some degree of arbitrariness in language based only on the observation that different languages use different words to denote the same meaning (e.g., the word for horse in English is "horse" but is "at" in Turkish).

However, the arbitrary character of language holds only from the perspective of the analyst observing a language system from the outside; from the perspective of an individual speaker, the goal of communication provides a strong constraint on arbitrariness. Perhaps this communicative constraint—roughly, that if my words were any different, I couldn't use them to talk to you—is why language doesn't *seem* arbitrary to us. Put another way, Saussure's (1916, 1960) insight was an insight because the form of language typically feels just right for the use to which we put it, namely talking to other people (Sutherland & Cimpian, 2015).

A rich body of theoretical work has explored communicative regularities in the use of particular forms to refer to particular types of meanings in context—the study of *pragmatics* (Clark, 1996; Grice, 1975; Horn, 1984). Broadly, this work argues that language users assume certain regularities in how speakers refer to meanings, and through these shared assumptions, the symmetry of the otherwise arbitrary character of language is broken. For example, consider a speaker who intends to refer to a particular apple on a table. Because language is *a priori* arbitrary, there are a range of ways the speaker could convey this meaning (e.g., "the apple," "the banana," "the green apple," "the green apple next to the plate," etc.), but the speaker is constrained by pragmatic pressures of the communicative context. If the listener also speaks English, the phrase "the banana" will be an unhelpful way to refer to the apple. Furthermore, if there is only one apple on the table, the phrase "the green apple" will be unnecessarily verbose given the referential context. These constraints might lead a speaker to

* Corresponding author at: Stanford University, Department of Psychology, Jordan Hall, 450 Serra Mall (Bldg. 420), Stanford, CA 94305, United States.
E-mail address: mll@stanford.edu (M.L. Lewis).

http://dx.doi.org/10.1016/j.cognition.2016.04.003
0010-0277/

select "the apple" as the referring expression, because it both allows the listener to correctly identify the intended referent while also minimizing effort on the part of the speaker.

In the present paper, we examine whether principles of communication influence the otherwise arbitrary mappings between words and meanings in the lexicon. This hypothesis is motivated by a regularity first observed by Horn (1984), who noted that pragmatic language users tend to consider the effort that speakers have exerted to convey a meaning. For example, consider the utterance "Lee got the car to stop," which seems to imply an unusual state of affairs. Had the speaker wished to convey that Lee simply applied the brakes, the shorter and less exceptional "Lee stopped the car" would be a better description. The use of a longer utterance licenses the inference that there was some problem in stopping—perhaps the brakes failed—and that the situation is more complex.

We ask whether speakers reason the same way about the meanings of words, breaking the symmetry between two unknown meanings by reference to length. Specifically, we test the following hypotheses:

Complexity Hypothesis 1: Speakers have a bias to believe that longer linguistic forms refer to conceptually more complex meanings.

Complexity Hypothesis 2: Languages encode conceptually more complex meanings with longer linguistic forms.

These two hypotheses are in principle independent from one another, and we test them separately. We see them as potentially emerging together from the same interactive forces, however, and we return to this relationship in Section 12.

An important construct for our hypothesis is the notion of conceptual complexity. One theoretical framework for understanding this construct is through conceptual primitives (e.g., Locke, 1847). Conceptual primitives can be thought of as the building blocks of meaning, similar to the notion of geons in the study of object recognition (Biederman, 1987). Within this framework, a more complex meaning would be one with more primitives in it. In a probabilistic framework, having more units would also be correlated with having a lower overall probability. We adopt this framework of conceptual primitives in our working definition of complexity.

Although identifying a general set of conceptual primitives might rank among the deepest challenges for cognitive science, some work has attempted this task. A body of research has sought to understand the innate conceptual primitives in young children ("core knowledge"; Kinzler & Spelke, 2007). The proposed set of concepts in this work, however, is restricted to those present only in early development (e.g., "agent"), and is therefore not suitable for the broad scope of our current project. Wierzbicka (1996) has also sought to identify conceptual primitives, but with a more general focus. This work compares lexical systems across languages to identify common primitives. The hypothesis is that there exists universal and innate semantic primitives which are the building blocks of meaning in human language. Under this view, all meanings can be derived from a set of numerable semantic primitives and a syntax for combining them. Our work here does not directly address the character of the underlying primitives, nor whether they are universal or innate. Rather, it assumes only that such units exist for a speaker and that lexical meanings can vary in the number of their compositional primitives.

In the remainder of the Introduction, we first review prior work suggesting that communicative principles are reflected in the structure of the lexicon. We then review work related to accounts of our particular linguistic feature of interest—variability in the length of forms. Then, in the body of the paper we test the complexity hypotheses above in nine experiments and a corpus analysis.

1.1. Pragmatic equilibria in the lexicon

The present hypotheses are motivated by the possibility that language dynamics take place over different timescales, and these different dynamics may be causally related to each other (Blythe, 2015; Christiansen & Chater, 2015; McMurray, Horst, & Samuelson, 2012). Our two hypotheses correspond to two distinct timescales. Hypothesis 1 corresponds to the timescale of minutes in a single communicative interaction—*the pragmatic timescale*. Hypothesis 2 corresponds to the timescale of language change, which takes place over many years—*the language evolution timescale*. We consider the possibility that communicative pressures at the pragmatic timescale may, over time, influence the structure of the lexicon at the language evolution timescale. Although a complexity bias at the language evolution timescale has not been previously explored, there are a number of other cases in which pragmatic equilibria are reflected in the structure of the lexicon. Here, we describe three such cases: semantic organization, ambiguity, and one-to-one structure.

Several broad theories of pragmatics include a version of two distinct pressures on communication: the desire to minimize effort in speaking (*speaker pressure*) and the desire to be informative (*hearer pressure*; Horn, 1984; Zipf, 1936). Importantly, these two pressures trade off with each other: The optimal solution to the speaker's pressure is a single utterance that can refer to all meanings, while the optimal solution to the hearer's pressure is a longer utterance that presents no ambiguity. The utterance that emerges is argued to be an equilibrium between these two tradeoffs.[1]

At the timescale of language evolution, there are a number of cases in which these pragmatic equilibria are reflected in the lexicon. The most well-studied of these cases is the size of the semantic space denoted by a particular word. Horn (1984) argues that the hearer has a pressure to narrow semantic space. This reflects the idea that the hearer's optimal language is one in which every possible meaning receives its own word. To understand this, consider the word "rectangle," which refers to a quadrilateral with four right angles. A special case of a "rectangle" is a case where the four sides are equal in length, which has its own special name, "square." Consequently, the term "rectangle" has been narrowed to mean a quadrilateral with four right angles, where the four sides are *not* equal. From the speaker's perspective, there is a pressure for semantic broadening. This is because the speaker's ideal language is one in which a single word can refer to a wide range of meanings. This phenomenon is exemplified by the broadening of brand names to refer to a kind of product. For example, "kleenex" is a product name for facial tissues, but has taken on the meaning of facial tissues more generally.

The opposition of these two semantic forces predicts an equilibrium in the organization of semantic space that satisfies the pressures of both speaker and hearer. A growing body of empirical work tests this prediction by examining the organization of particular semantic domains cross-linguistically (see Regier, Kemp, & Kay, 2015, for review). This work finds that languages show a large degree of similarity in how they partition semantic space for a particular domain, but also a large degree of variability. Such analyses demonstrate that the attested systems all approximate an equilibrium point between hearer and speaker pressures.

In one example of this kind of analysis, Kemp and Regier (2012) demonstrate this systematicity in the semantic domain of kinship. For each language, they developed a metric of the degree to which Horn's speaker and hearer pressures are satisfied. A language that better satisfies the hearer's pressure is one that is more complex, as

[1] Note that this analysis only reflects interlocutors' *non*-aligned utilities in a communication task. Of course, both speaker and hearer also have aligned utility derived from successful communication.