Global Cocktail Parties and an Arms-Race in Language Evolution

Thomas T. Hills (t.t.hills@warwick.ac.uk)

James Adelman (j.s.adelman@warwick.ac.uk)
Department of Psychology, Gibbet Hill Road
Coventry, CV4 7AL UK

Abstract

Many species change their signals in response to crowding, consistent with information theoretic accounts of signalling in the presence of noise. In this article we explore the hypothesis that the statistical structure of language has evolved to enhance reception and processing in response to competition in the language market. In particular, we investigate how concreteness has changed over the last 200 years. We take a big data approach, combining large text corpora (including the Google Ngram corpus, the Corpus of Historical American English, and presidential speeches) with the recent collection of concreteness norms for over 40,000 English words (Brysbaert et al., 2013). Across corpora we find that concreteness has steadily increased since the 1800s. This takes place both within and across word classes, indicating that the rise in concreteness is systemic and not limited to changes in grammar. By comparing recent concreteness norms with older norms, we show that the observed changes in concreteness are not due to a bleaching effect caused by the loss of concreteness as words age. We further investigate how the statistical distribution of other properties of words change in a way that may indicate that language is becoming more distinctive. We discuss a number of potential explanations and implications of these changes, including changes in literacy, gender, the Flynn effect, and the influence of competition in the marketplace of ideas.

Keywords: Big data, language evolution, concreteness, Flynn effect, information theory

Introduction

Recent trends in communication indicate massive increases in the amount of information we must process on a daily basis. Borrowing from the idea of an attention economy, this leads to potential crowding in a language market. This is likely to lead to competition among message producers. Similar to a global cocktail party problem, crowding and the resultant attention economies of information markets mimic the inclusion of noise in information theoretic accounts of signal transduction. In such accounts, signals should adapt to noise in systematic ways to insure information transfer.

If language adapts to the needs of its users, language should change in response to information crowding. Indeed, numerous other species, for example birds, adapt their communication in response to crowding in ways that facilitate processing and distinctiveness. How should human languages evolve in response to competition in a language market? One approach is to evolve messages that are processed more rapidly, more easily comprehended, and better recalled. These properties of language have been well studied by psycholinguists and, indeed, one specific aspect of language is well marked by its ability to enhance these processing capacities. That is, concreteness.

Concreteness from a psycholinguistic perspective refers to the capacity for a concept to be perceived through direct experience. Among psycholinguistic variables, the relationship between concreteness and cognitive processing is easily among the most profound. Concrete words are more rapidly perceived in lexical decision tasks and more easily recalled in memory tasks. Much of this research follows on Paivio's dual-coding theory, which proposed that the combination of visual and verbal contributions to linguistic information enhance processing. Other theories have since been proposed to account for the cognitive utility of concreteness, all of which speak to different aspects of the processing advantage of concrete words. Beyond psycholinguistic tasks, concrete language has been shown to be both more interesting and easier to understand.

Other theories do not make this prediction. For example, Dunbar's theory of language as social grooming would suggest no change in relation to message quality. Knowledge-based theories of the Flynn effect might also be interpreted to suggest that language is becoming more abstract as humans increase their capacity for symbolic processing and abstract thinking associated with increased scores on intelligence tests.

We investigated these questions using the Google Ngram (1-gram) corpus of American English, which provides a collection of over 355 billion words since the 1800s as published in books. We also performed the analyses on a text corpus collected independently of the Google Ngram corpus, the Corpus of Historical American English, collected to represent a balanced and representative corpus of American English containing more than 400 million words of text from, for example, newspapers and magazines, from 1810 to 1990. We further supplemented these corpora with corpora representing the Inaugural Addresses given by U.S. presidents, starting with George Washington in 1789, and the State of the Union addresses, starting with Harry Truman in 1945.

Using the Brysbaert norms, we calculated a concreteness index, C_y, for each year, y, in each of our corpora as follows,

$$C_y = \sum_i^n c_i p_{i,y}$$

where c_i is the concreteness for word i as found in the Brysbaert et al. concreteness norms and p_i is the proportion of word tokens represented by word i in year y. This was computed for all n words in the concreteness norms and separately for each year in the corpora.

All the corpora we examined showed an increase in concreteness over time. The Google Ngrams show this rise over the last 200 years, with a temporary drop following the second world war (Figure 1). The Corpus of Historical American

English shows a similar rise with a slightly less dramatic fall in the 1950s (relative to the overall increase). On the same figure we also show the Google Ngram data restricted either to the words in the lexicon at year 1800 or using only those words in the concreteness norms that were known by more than 95% of participants. In both cases, the rise in concreteness remains indicating that the rise is not due to new words entering the lexicon or unfamiliar words leaving the lexicon. We find a similar rise in concreteness for Presidential Inaugural Addresses (e.g., Figure 2) and State of the Union Addresses (not shown). Further, we observe that changes in concreteness are due not only to changes in the statistical distribution of word classes (e.g., dropping articles), but also occur within word classes (e.g., nouns, verbs, and prepositions).

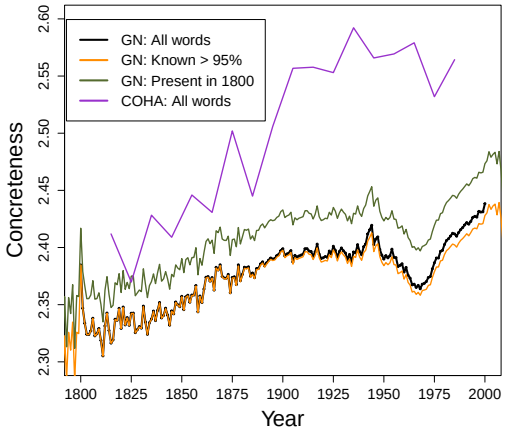

Figure 1: Concreteness in the Google Ngrams and COHA.

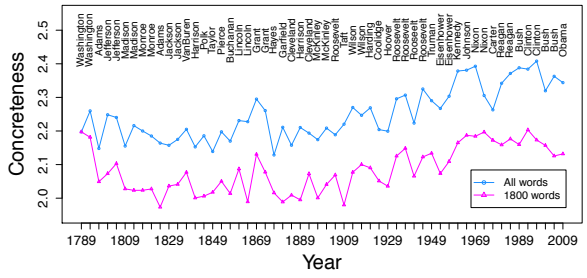

Figure 2: Concreteness in Presidential Inaugural Addresses.

One explanation for the observed rise in concreteness is that some words may tend to lose concreteness over time. This has been referred to as semantic bleaching, similar to grammaticalization. An example is the word *disaster*, which originally referred to dire and acute events, but can now usefully refer to everything from one's hair to public policy. If bleaching were systematic, it would explain the observed rise in concreteness. The words that were more often used in the past may have been more concrete in the past, but due to bleaching now appear more abstract. Figure 3 shows a comparison of the Brysbaert norms with the 925 word Paivio concreteness norms, which found no systematic change in concreteness over the last 45 years.

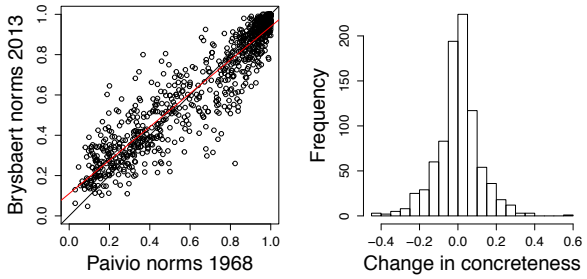

Figure 3: Comparison of Paivio and Brysbeart norms.

We further investigated the statistical structure of word collocations and the relative distributions of word usage (Zipf distributions) over successive years. Our results support the notion that language is evolving–in our lifetimes–to compensate for increased crowding in the information marketplace.

References

Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, *17*(9), 814–823.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2013). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 1–8.

Hills, T. (2012). The company that words keep: comparing the statistical structure of child-versus adult-directed language. *Journal of Child Language*, 1–19.

Hills, T., Mata, R., Wilke, A., & Samanez-Larkin, G. (2013). Mechanisms of age-related decline in memory search across the adult life span. *Developmental psychology*.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological review*, *119*(2), 431.

Hills, T. T., Maouene, J., Riordan, B., & Smith, L. B. (2010). The associative structure of language: Contextual diversity in early word learning. *Journal of memory and language*, *63*(3), 259–273.

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009). Longitudinal analysis of early semantic networks preferential attachment or preferential acquisition? *Psychological Science*, *20*(6), 729–739.

Segev, E., & Hills, T. (2014). When news and memory come apart a cross-national comparison of countries mentions. *International Communication Gazette*, *76*(1), 67–85.