Cognition 143 (2015) 87–92

Contents lists available at ScienceDirect

Cognition

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

Short Communication

Recent evolution of learnability in American English from 1800 to 2000

Thomas T. Hills *, James S. Adelman

Department of Psychology, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK

ARTICLE INFO

Article history:
Received 23 February 2015
Revised 11 June 2015
Accepted 19 June 2015
Available online 24 June 2015

Keywords:
Semantic bleaching
Language evolution
Concreteness
Historical linguistics
Culturomics

ABSTRACT

Concreteness—the psycholinguistic property of referring to a perceptible entity—enhances processing speed, comprehension, and memory. These represent selective filters for cognition likely to influence language evolution in competitive language environments. Taking a culturomics approach, we use multiple language corpora representing more than 350 billion words combined with concreteness norms for over 40,000 English words and demonstrate a systematic rise in concrete language in American English over the last 200 years, both within and across word classes (nouns, verbs, and prepositions). Comparisons between new and old concreteness norms indicate this is not explained by semantic bleaching, but we find some evidence that the rise is related to changes in population demographics and may be associated with increasing numbers of second language learners or attention economics in response to crowding in the language market. We also examine the influence of gender and literacy. In sum, we demonstrate evolution in the psycholinguistic structure of American English, with a well-established impact on cognitive processing, which is likely to permeate modern language use.

For more than half a century, the central question of language acquisition has been how is the mind shaped to learn language (Chomsky, 1966; Pinker & Bloom, 1990). Recently this question has been inverted to ask how does language evolve to be learnable (Christiansen & Chater, 2008; Deacon, 1997; Smith & Kirby, 2008). Languages clearly do evolve over time (Labov, 2011; Lieberman, Michel, Jackson, Tang, & Nowak, 2007) and laboratory as well as computational studies have indicated that this evolution should be in the direction of learnability (Griffiths & Kalish, 2007; Hurford, 1989; Kirby, 2001; Smith & Kirby, 2008). To date, however, there is limited evidence for diachronic changes in learnability in a naturally occurring language. The challenge to documenting such changes rests in identifying a clear linguistic marker of learnability and tracking this marker in relation to diachronic corpora over a sufficient time scale and in a language environment likely to be under selective forces sufficient to produce linguistic change.

Two potential forces on the evolution of communication include the role of language learners and the influence of crowding. Language learners, because of the rate at which they learn linguistic properties, favor linguistic simplicity over complexity, and existing languages reflect this at multiple levels (Lupyan & Dale, 2010; Monaghan, 2014). Crowding may play a similar role. Recent trends in information production are likely to be associated with analogous and highly competitive environments for communicators in English (Eppler & Mengis, 2004; Varian & Lyman, 2000). Information creates markets for attention (Davenport & Beck, 2001; Hansen & Haas, 2001) and in competitive situations these should drive evolutionary forces in communication. Similar to a global cocktail party problem—where communicators compete for the attention of listeners—crowding and the resultant attention economies should drive language evolution in the direction of learnability (Shannon & Weaver, 1949). This has been demonstrated in birds by an increase in the acoustic distinctiveness of bird song in response to crowding (Grant & Grant, 2010; Luther, 2009). If language responds to the number of competing language producers as well as the demands of language learners, then evolution in language should occur in communities where information crowding is most likely to be taking place.

It is a basic principle of evolution that replicators that have selective advantages in survival and reproduction should be more likely to be found in future populations. Analogously, language that comes to mind faster, is more easily understood, and is more readily remembered is more likely to be produced by future speakers. Each of these features is well known by psycholinguists to be associated with concreteness. Concreteness refers to the capacity for a concept to be vividly imagined and identified with a specific referent. Words like *dog* and *computer* are more vividly imagined than words like *truth* and *feeling*, and people easily report this difference. In studies of language processing, concrete words are more rapidly recognized as valid words in lexical decision tasks (James,

* Corresponding author.
E-mail address: t.t.hills@warwick.ac.uk (T.T. Hills).

http://dx.doi.org/10.1016/j.cognition.2015.06.009
0010-0277/

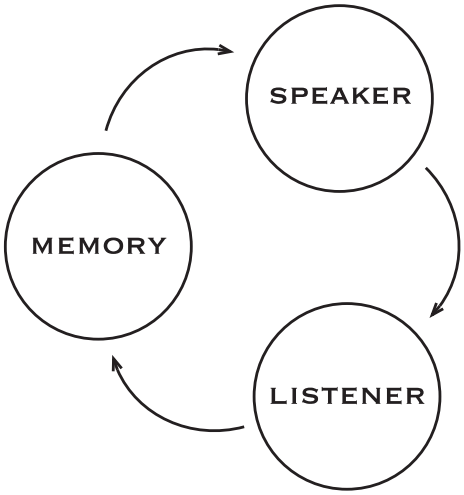

Fig. 1. Stages where forces favoring learnability can select for concreteness in the cognitive life cycle of language. Because of cognitive capacity limitations, each arrow in the diagram represents a stage in the life cycle of language where selection may favor more learnable concepts: Speakers recall concepts from memory and then produce them for listeners, who must attend to the message and then store the information for later production. Selection will occur anytime there is more information to attend to, encode, and produce than language users can cognitively process.

1975), more easily recalled in memory tasks (Miller & Roodenrys, 2009; Romani, McAlpine, & Martin, 2008), and language composed of more concrete language is both more interesting and easier to understand (Sadoski, 2001). Concrete words are also more readily learned by both second and first language learners (e.g., De Groot & Keijzer, 2000). Together, these provide a cognitive basis for predicting selection for more concrete language during the stages of iterated-learning necessary for language evolution (e.g., Kirby, Cornish, & Smith, 2008; Smith & Kirby, 2008). In particular, these represent (1) what listeners attend to and comprehend, (2) what listeners encode, and (3) what speakers recall and, in turn, produce (see Fig. 1).

In the present work, we investigate historical changes in the distribution of concrete language in American English by using large collections of text combined with psycholinguistic measures of concreteness. Our approach follows on the "culturomics" approach of using "big data" to measuring behavioral and cultural change over time (e.g., Dodds, Harris, Kloumann, Bliss, & Danforth, 2011; Greenfield, 2013; Michel et al., 2011; Petersen, Tenenbaum, Havlin, Stanley, & Perc, 2012). In particular, we extend this approach to the study of a cognitive capacity-induced change in language by combining historical changes in text with recent large-scale collections of psycholinguistic norms (e.g., Brysbaert, Warriner, & Kuperman, 2013). This approach offers psychologists the opportunity to develop historical psychological accounts of behavioral change over cultural time and, as we demonstrate here, offers new questions and insights into how large-scale social change may influence communication in relation to cognitive processing.

1. Methods

To investigate diachronic changes in concreteness, we computed a concreteness rating over time for corpora using a recent collection of large-scale concreteness norms consisting of more than 40,000 words, each ranked for concreteness on a 5-point scale by multiple online participants (Brysbaert et al., 2013). These norms cover more than 50% of each of the corpora in any year. Concreteness ratings were combined with diachronic language corpora, consisting of book collections archived in the Google Ngrams corpus of American English (355 billion words, see Michel et al., 2011), newspaper and magazine language collected in the Corpus of Historical American English (400 million words, see Davies, 2009), and inaugural addresses produced by American presidents. Presidential speeches represent much smaller language samples ranging from 487 words for Roosevelt in 1945 to 7189 words for Harrison in 1841. To investigate the influence of gender, we used a corpus of approximately one million words from 2943 blogs coded for gender of the author (1393 female, 1550 male; Mukherjee & Liu, 2010). We compared language data with historic records of U.S. immigration (Migration Policy Institute., 2014). In the supplemental online material, we provide similar results to those presented below for English Fiction and state of the union addresses.

The concreteness scores, C_y, for each year, y, in the diachronic corpora were computed as a weighted average,

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

where c_i is the concreteness for word i taken from the concreteness norms (Brysbaert et al., 2013) and $p_{i,y}$ is the probability of word i in year y. This was computed for all n words in the concreteness norms. We also computed concreteness on a per document basis, as opposed to per word as we do above; the same patterns of results were found as we report here.

2. Results

All the corpora we examined show an increase in concreteness over time (Fig. 2; see also Figs. S2 and S4). The Google Ngrams show this rise over the last 200 years, with a temporary drop following the Second World War. The Corpus of Historical American English and presidential inaugural addresses show a similar rise. Fig. 2 also shows the Google Ngram data with concreteness computed only for words present in the lexicon at year 1800 or using only those words that were known by more than 95% of participants in the concreteness norms (Brysbaert et al., 2013). 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.

One possible explanation for the general rise in concreteness is that some words may tend to lose concreteness over time. This is referred to as semantic bleaching, desemanticization, or grammaticalization (Aitchison, 2003; Hopper & Traugott, 2003). An example is the word *disaster*, which originally referred to dire and acute events, but can now usefully refer to topics as diverse as hairstyles and public policy. An additional example is the word *going*, which in the early nineteenth century lost its connection with motion and became a grammatical marker associated with events in the near future (Millar, 2015). Not all diachronic linguistic patterns are associated with semantic bleaching. Metonymy is a common figure of speech where some specific aspect of an abstract concept is used in place of a more abstract reference, as in *The White House*. Nonetheless, if bleaching were sufficiently strong, it might explain the observed rise in concreteness: if people used words in the past because they were more concrete at the time, then if these words became more abstract through use, and were therefore replaced with other more concrete words in more contemporary speech, the net result would be that older language appears more abstract, even though it may have been equally concrete. Note that if bleaching took place equally across all words, then this would create the opposite effect of reduced concreteness over time.

Concreteness norms were not collected as far back as the 1800s. However, the Paivio norms (Paivio, Yuille, & Madigan, 1968)