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

**A R T I C L E    I N F O**

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

Keywords:
Semantic bleaching
Language evolution
Concreteness
Historical linguistics
Culturomics

**A B S T R A C T**

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.

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

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,
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,

$$ C_y = \sum_{i} n_i p_{i,y} $$

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 public policy. An additional example is the word going, which in the early nineteenth century lost its connection with motion and public policy.

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)
provide a 45-year window of comparison with the Brysbaert et al. norms. Normalizing the ratings for both shows that there are no systematic changes in word concreteness over the approximately 900 words used for comparison (Fig. 3). The median change is centered around zero and a paired t-test finds no significant difference in concreteness ($t(876) = -0.79$, $p = .45$). This suggests that a
systematic change of concreteness within words is unlikely to explain the apparent rise in concreteness we see across words in the data.

Changes in the word classes people use may also change concreteness over time and may be influenced by the same selective forces described above. For example, articles—such as the, a, and an—have low relative concreteness in comparison with other word classes. Names, on the other hand—like China and Butterfingers—have the highest concreteness. Articles and names belong to closed and open word classes, respectively, which refer to the finite or potentially unlimited size of word classes. Closed word classes are generally less concrete than open word classes (see Table 1). If preferences for different word classes changed over time—e.g., if people dropped more articles and used more proper names—then this would increase concreteness.

Consistent with our general hypothesis that American English has changed to increase its overall concreteness, closed class words have declined over the past two hundred years (see Fig. 4). These changes also explain some of the overall change in concreteness presented in Fig. 2 (predicting concreteness based on proportion of closed class words: Google Ngrams: $\beta = -1.39$, $R^2 = 0.74$, $p < .001$; Corpus of Historical American English: $\beta = -3.91$, $R^2 = 0.57$, $p < .001$; Inaugural Address: $\beta = -1.79$, $R^2 = 0.36$, $p < .001$). However, as shown in Fig. 5, within the three word classes representing the largest portion of the data (nouns, verbs, and prepositions)—representing more than 50% of the words in each corpus—concreteness rises (with some variation) over the period from 1800 to 2000, in all cases. Thus, concreteness also tends to increase within word classes, indicating that rising concreteness is systemic—operating at multiple levels within the structure of language.

A potential driving force on the evolution of learnability in American English is related to the influx of second language learners. Though correlation does not imply causation, it is nonetheless useful with data of this kind to attempt to rule out such potential explanations. As shown in Fig. 6, the rise in concreteness in Google Ngrams and the Corpus of Historical American English are tightly correlated with U.S. immigration patterns. The Pearson correlation with Google Ngrams concreteness and total U.S. population is slightly weaker, $r = .76 (r = .42, .91)$. This is consistent with theories of language change in response to the cognitive abilities of language learners (e.g., Lupyan & Dale, 2010).

Several additional forces can be ruled out as complete explanations for the rise in concreteness. According to data from the National Center for Education Statistics, literacy has been increasing decade over decade in the U.S. since its earliest recorded date in 1870, with measures above 95% since the 1930s (Snyder, 1993). Comparing this decadal literacy data from 1870 to 1980 with our concreteness data, literacy is not statistically correlated with concreteness in the Google Ngrams, $r = .45 (r = .16, .81)$, though it is with the Corpus of Historical American English, $r = .77 (r = .36, .93)$. Matching our analysis to the same time period, this is lower than the correlation between immigration and concreteness in the Google Ngrams, $r = .76 (r = .32, .92)$, but approximately the same as that found between immigration and concreteness in the Corpus of Historical American English, $r = .74 (r = .29, .92)$. Immigration remains the stronger predictor, even though literacy may be a contributing factor.

Gender may have also influenced the historical trajectory of language change. There has undoubtedly been an increase in the proportion of language media produced by females over the last two centuries. We applied our concreteness score to a corpus of blogs made available by Mukherjee and Liu (2010) and found that females’ language was more concrete than males’ language ($M = 2.50$ for females and $M = 2.48$ for males, $t(2941) = 5.54$, $p < .001$, Cohen’s $d = 0.20$). The approximate difference between males and females in concreteness—0.02 units on the concreteness scale—is small relative to the overall rise in concreteness shown in Fig. 2. Changes in the ratio of males to females in language media is

<table>
<thead>
<tr>
<th>Word class</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nouns</td>
<td>3.50</td>
<td>1.02</td>
</tr>
<tr>
<td>Numbers</td>
<td>3.49</td>
<td>0.38</td>
</tr>
<tr>
<td>Verbs</td>
<td>2.92</td>
<td>0.76</td>
</tr>
<tr>
<td>Pronouns</td>
<td>2.76</td>
<td>0.71</td>
</tr>
<tr>
<td>Adjectives</td>
<td>2.50</td>
<td>0.72</td>
</tr>
<tr>
<td>Prepositions</td>
<td>2.29</td>
<td>0.64</td>
</tr>
<tr>
<td>Determiners</td>
<td>2.11</td>
<td>0.55</td>
</tr>
<tr>
<td>Adverbs</td>
<td>2.06</td>
<td>0.53</td>
</tr>
<tr>
<td>Articles</td>
<td>1.68</td>
<td>0.54</td>
</tr>
<tr>
<td>Conjunctions</td>
<td>1.64</td>
<td>0.54</td>
</tr>
</tbody>
</table>

Table 1
Concreteness for a selection of word classes in the Brysbaert et al. norms. Closed word classes are shown in italics.
therefore likely be a contributor to the rise in language concreteness, even if it only contributes to a fraction of the effect.

3. Conclusions

Though numerous accounts of language evolution have suggested that language should evolve to be learnable (Christiansen & Chater, 2008; Deacon, 1997; Smith & Kirby, 2008), the present work is the first to demonstrate the diachronic evolution of a property associated with learnability in a natural language. In particular, the evidence demonstrates a systematic rise over two centuries in one of the most well understood psycholinguistic variables associated with learnability—concreteness. We further demonstrate that this effect appears to permeate language in a variety of forms (e.g., newspapers, books, and public speeches) and across word classes (e.g., nouns, verbs, and prepositions). We also show that this change is unlikely to be explained by a prominent observation from historical linguistics, semantic bleaching, because a comparison between concreteness norms collected 45 years apart shows no evidence for a systematic loss of concreteness among nouns, while during the same time period we observe a systematic rise in concrete nouns (see Fig. 5). Finally, we describe some suggestive

![Fig. 5. Concreteness within word classes (i.e., nouns, verbs, or prepositions) for the Google Ngram corpus (GN), the Corpus of Historical American English (COHA), and presidential inaugural addresses (IA). Because of the variation in the IA data, we present the smoothed data using a locally-weighted polynomial ("lowess" in R, scaling factor \( f = 2/3 \)). Pearson correlations with time (95% CI): nouns, GN, \( r = .39 \) (.27, .50); COHA, \( r = .96 \) (.88, .98); IA, \( r = .42 \) (.17, .61); verbs, GN, \( r = .79 \) (.74, .84); COHA, \( r = .86 \) (.65, .95); IA, \( r = .68 \) (.51, .80); prepositions, GN, \( r = .94 \) (.92, .95); COHA, \( r = .99 \) (.97, 1.00); IA, \( r = .17 \) (.10, .41).

![Fig. 6. Comparison of immigration patterns in the U.S. with concreteness computed from Google Ngrams (GN) and the Corpus of Historical American English (COHA). Pearson correlations between immigration and concreteness (95% CI): Google Ngrams, \( r = .96 \) (.62, 1.00); COHA, \( r = .86 \) (.61, .96).]
evidence that the observed changes in concreteness are unlikely to be fully explained by factors such as literacy and gender, but are nonetheless highly correlated with changes in population size and immigration. These latter results are consistent with recent hypotheses that language adapts to the cognitive demands of its users (e.g., Lupyan & Dale, 2010).

Culture is a marketplace for information. This leads to the prediction that any broad medium of communication should evolve over time in response to changes in the communication environment. Moreover, aspects of these signals should be enhanced in the presence of competition or among new learners. In particular, the distribution of information that is processed more quickly, better remembered, and more easily learned should increase as information markets become more crowded and involve more learners. In the present work we take a culturomics approach to address these issues—which has the strength of big data and the weakness of limited information about individual data points. This approach has substantial potential for addressing existing hypotheses in cognition as well as producing new lines of research. In particular, it is quite likely that concreteness is but one change among many in modern American English; changes in rhetorical style or topics of conversation that increase memorability, processing, and retrieval would all be equally consistent with reducing processing difficulty, which is the broader concept we hypothesize to be causative.

These represent exciting research prospects and, in combination with laboratory studies on language memorability, also offer new inroads to understanding how to enhance the effectiveness of communication for key messages in competitive language environments such as in education or emergency situations. In addition, language efficiency is likely to have pronounced effects on the cognitive capacities of its users. For example, the documented rise in symbolic intelligence associated with the Flynn effect (Flynn, 2012; Fox & Mitchum, 2013) could be both cause and consequence of increased concreteness in language. That is, an increase in concrete language may have enhanced our individual capacity to process complex information (Sadoski, 2001) at the same time as it allows us to communicate clear but increasingly complicated ideas.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cognition.2015.06.009.

References


Brylbaert, M., Warriner, A. B., & Kuperman, V. (2013). Concreteness ratings for 40 thousand generally known English word lemmas. Behavior Research Methods, 46, 904–911.


Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social users (e.g., Lupyan & Dale, 2010).

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social users (e.g., Lupyan & Dale, 2010).


