A predictability-distinctiveness trade-off in the historical emergence of word forms

Aotao Xu (a26xu@uwaterloo.ca)
Computer Science Program, University of Waterloo

Christian Ramiro (chrisram@berkeley.edu)
Cognitive Science Program, University of California, Berkeley

Yang Xu (yangxu@cs.toronto.edu)
Department of Computer Science, Cognitive Science Program, University of Toronto

Abstract

It has been proposed that language evolves under the joint constraints of communicative expressivity and cognitive ease. We explore this idea in the historical emergence of word forms. We hypothesize that new word forms that enter the lexicon should reflect a trade-off between predictability and distinctiveness. An emergent word form can be highly predictable if it efficiently reuses elements from the existing word forms, resulting in low cognitive load. An emergent word form should also be sufficiently distinctive from the existing lexicon, facilitating communicative expressivity. We test our hypothesis by examining the properties of 34,478 emergent word forms over the past 200 years of Modern English. We show how word forms at future time $t+1$ are bounded statistically between n-gram generated word forms (highly predictable) and slang words that are outside the standard lexicon (highly distinctive) at time t. Our work supports the view of cognitive economy in lexical emergence.

Keywords: word form; lexicon; lexical emergence; language evolution; cognitive economy

Introduction

The lexicon is a central locus of human thought, but it undergoes constant change over time. In particular, new words may emerge due to changing sociocultural needs, resulting in growth of the lexicon. Taking the English lexicon as an example, it has grown by approximately tenfold over the past millennium, with more than 150,000 word forms having emerged from the period of Old English to the present day (Figure 1a). Here we ask what principles might underlie the historical emergence of word forms above and beyond the external sociocultural factors that could influence lexical emergence.

Our starting point is the idea that language evolves under the dual considerations of communicative function and cognitive effort (Labov, 2011; Jespersen, 1959; Otto, 1956; Kirby, Tamariz, Cornish, & Smith, 2015), a prominent proposal that has been framed similarly in linguistics as the principle of least effort (Zipf, 1949) and in cognitive psychology as the principle of cognitive economy (Rosch, 1978). This proposal also relates to a growing line of research that explores design principles of language through the lens of efficient communication (Piantadosi, Tily, & Gibson, 2012; Kemp & Regier, 2012; Kemp, Xu, & Regier, 2018). Most relevant to the current study is work by Labov who suggests that words may be selected under the joint constraints of least effort (cf. Zipf, 1949)—a drive for cognitive ease of production, and the competing force of communicative informativeness (Labov, 2011). There is evidence for each of these constraints in the design of word forms. For example, it has been shown that word forms that conform to well-formed phonotactic properties can facilitate production (Edwards, Beckman, & Munson, 2004), and words that sound similar to many existing words, or having dense lexical neighbourhoods, tend to reduce speech error (Stemberger, 2004). On the other hand, separate work has suggested that perceptual distinctiveness matters in the lexicon because it minimizes confusion and facilitates clarity in communication (Flemming, 2004; Meylan & Griffiths, 2017).

We extend previous work by exploring principles in the historical emergence of novel word forms. We believe that the same proposal of language evolution should apply to explaining how new word forms enter the lexicon over time. In particular, we hypothesize that the emergence of word forms should trade off *predictability* against *distinctiveness*. An emergent word form is highly predictable if it efficiently recombines elements from existing word forms, resulting in low cognitive effort in production and memory. Our notion of predictability is rooted in classic work by Shannon (1951) on the information analysis of English text. However, this criterion of predictability is likely in competition with distinctiveness: An emergent word form should be sufficiently distinctive from words in the existing lexicon, hence generating minimal confusion and facilitating communicative expressivity. Predictability and distinctiveness trade off against each other because a highly predictable word form is necessarily similar in form to existing words, so it is unlikely to be distinctive. Similarly, a highly distinctive word form is necessarily novel in its composition, so it is unlikely to be very predictable. Here we examine the possibility that the emergent word forms in history are shaped under these two joint forces, such that they appear sandwiched between (plausible) word forms that are highly predictive and those that are highly distinctive (see Figure 1b for illustration).

We test our hypothesis by examining new word forms that entered the Modern English lexicon over the past 200 years. At each future decade $t+1$, we compare the actual emergent words against a control set of computer-generated words and slang words that did not enter the standard lexicon up to the previous decade t. We show how the actual word forms are interleaved between the highly predictable and distinctive control words in terms of their statistical properties.

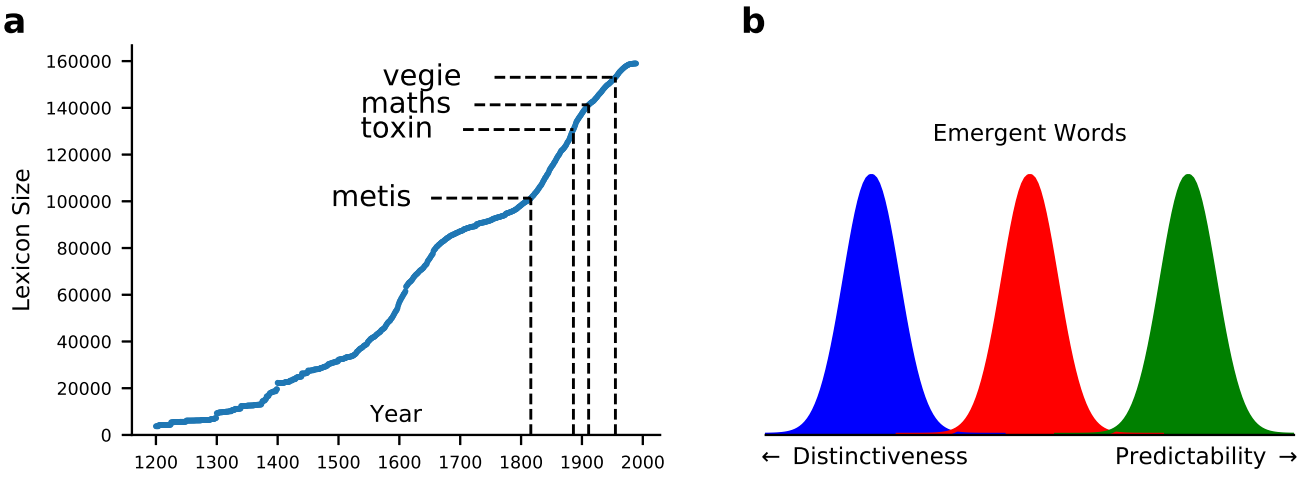

Figure 1: Illustration of the phenomenon of lexical emergence and our hypothesis. a) Growth of the English lexicon over the past 800 years based on data from the Historical Thesauraus of English (HTE). Lexicon size represents the number of unique word forms that exist during this period, and examples of emergent words are shown at respective years of emergence. b) The hypothesis that emergent words (middle) at future time $t+1$ should reflect a trade-off between predictability and distinctiveness among the space of plausible word forms (on the two sides) given the current lexicon at time t, or effectively sandwiched.

Materials and Methods

We analyzed 34,478 word forms during each decade from 1800 to 1980 as recorded in the Historical Thesauraus of English (HTE) (Kay, Roberts, Samuels, & Wotherspoon, 2017) (`https://ht.ac.uk/`) which is based on the Oxford English Dictionary. We considered single-word lexemes that are composed of 26 English letters (a-z) and took the first year recorded in HTE for a given word form as its emerging date. We focused our analysis on the Modern English period due to both orthographic and phonetic changes in words during the remote periods of Old English and Middle English (Baugh & Cable, 1993, p279), and the lack of control words for the same periods that is critical for our analyses. We grouped word forms according to the lengths of their orthographic forms, and we considered lengths ranging from 4 to 9 because lengthier words are more likely formed due to rule-based compositional strategies (Krott, 1996). The grouping by word length is necessary because longer words are by chance more distinctive in form than shorter words, so a principled analysis of our trade-off hypothesis should be independent of length. We focus on reporting results based on orthographic forms, although we observed similar results with phonological forms that we do not include here due to space limit.

We used two standard measures to quantify the statistical properties of word forms along the predictability–distinctiveness dimension: letter n-gram probability and lexical neighbourhood density. We quantify the two measures for word forms at a future decade $t+1$ based on statistical properties of the existing lexicon at a decade earlier at t. Formally, we define the probability of an emergent word form w of length $|w|$ by using the n-gram probabilities of its constituent letters (or phonemes), extending the classic work by Shannon on information analyses of English words (Shannon, 1951):

$$p^{t+1}(w) = \prod_{i=1}^{|w|} p^t(l_i|l_{<i}) \quad (1)$$

$$= p^t(l_1|\cdot) \times p^t(l_2|l_1) \times p^t(l_3|l_2,l_1) \times \quad (2)$$
$$\ldots \quad \times p^t(l_{|w|}|l_{|w|-1},\ldots,l_1)$$

Equations 1-2 effectively estimate how probable a novel word form w would be at decade $t+1$ given the n-gram statistics at the current decade t. We considered n-gram of up to order 5 because statistics of higher orders are sparse and prohibitively expensive to compute. Under this measure, a highly predictable word form at $t+1$ for a given length should be one that maximizes the n-gram probability based on the lexical statistics at t. On the contrary, a highly distinctive word form should have low predictability that minimizes the same probability measure.

To ensure the robustness of our approach, we considered lexical neighbourhood density as an alternative measure. We define the neighbourhood density of an emergent word form w based on how similar it is to existing word forms v in the lexicon at time t (L^t), grounded in the psycholinguistic study of English word forms by Bailey and Hahn (2001):

$$ND^{t+1}(w) = \sum_{v \in L^t} e^{-d(w,v)} \quad (3)$$

Equation 3 effectively estimates how crowded a novel word form w would be at decade $t+1$ given existing word forms at the current decade t. We used the standard Levenshtein

edit distance for calculating $d(\cdot,\cdot)$ that considers if two word forms are similar or distant based on the number of edits required to match the forms via insertion, deletion, or substitution (Yarkoni, Balota, & Yap, 2008; Bailey & Hahn, 2001). For example, the edit distance between "cat" and "maths" is 3 since the edit involves one substitution and two insertions. Similar to the case of the *n*-gram measure, a highly predictable word form at a given length should be one that maximizes neighbourhood density at $t+1$. On the contrary, a highly distinctive word form should not be crowded and hence minimizes its lexical neighbourhood density.

To evaluate the hypothesis that emergent word forms trade off predictability against distinctiveness, we considered a set of control word forms that are representative of the extremities of this dimension yet did not formally enter the English lexicon. Our goal is to assess whether the trade-off hypothesis might explain why certain word forms have entered the lexicon over time, whereas other plausible forms have not appeared. Because the set of all possible word forms is enormous (e.g., there are over 10 million possible word forms of length 5 that did not appear in English up to 1980), we chose control words by focusing on word forms that are either likely to be very predictable or distinctive.

We first obtained the *predictable control set* by generating word forms according to the *n*-gram probability measure in Equations 1-2. At each yet-to-emerge decade, we sampled these word forms from the *n*-gram statistics obtained from the previous decade in a sample size that matches the number of the emergent words. The sample does not intersect with the lexicon, but it can intersect with the set of actual emerging words. We then partitioned these control words by length and calculated their *n*-gram probabilities and neighbourhood densities according to Equations 1-3. This control word set approximates the extremity of predictability because the candidates are directly generated from the distribution of the existing lexicon, so they should be statistically equivalent to the existing word forms in the lexicon. Because *n*-gram probability correlates with neighbourhood density (Sanders & Chin, 2009), we also expect this word set to have high (but not necessarily the maximal) neighbourhood density. If the trade-off hypothesis is correct, the emergent word forms should generally have lower but not near-identical *n*-gram probability and neighbourhood density to this control set.

We next obtained the *distinctive control set* by sampling word forms from slang that did not enter the standard English lexicon. Slang is likely to represent the extremity of distinctiveness because slang words are known to differ from the standard lexicon (Mattiello, 2008, 2013), and 2) they serve as a more conservative measure for plausible word forms (plausible because a subset of slang can eventually become actual words (Baugh & Cable, 1993, p293)) than random samples of non-existent word forms that can be distinctive but not permissible, e.g., "jxyzh" is very distinctive from existing words in English but it is not permissible based on the knowledge of English. We drew data from a large online resource, the Urban Dictionary (`https://www.urbandictionary.com/`), for this control set. We used word forms containing only the letters a-z conforming to the same selection standard with the emergent words. During each decade of interest, we excluded homographs of word forms or words that have overlapping lemma in the lexicon via the lemmatizer from the Natural Language Toolkit (NLTK) (Bird, Klein, & Loper, 2009). We then sampled from the rest of the 317,403 unique word forms in matching size to the emergent lexicon per length, and calculated the *n*-gram and neighbourhood statistics for these word forms. If the trade-off hypothesis is correct, the emergent word forms should generally have higher but not near-identical *n*-gram probability and neighbourhood density to this control set.

Results

We evaluated our hypothesis by first examining whether newly emerging word forms tend to fall between predictable control words and distinctive (slang) control words in terms of *n*-gram probability and lexical neighbourhood density. At each decade, we compared the actual emergent word forms to the two sets of control words of the same length under the two measures separately. We took the average values of the two measures for each word group and every length that we examined.

Figure 2 summarizes the results for these comparisons for every decade from 1800 to 1980 and word forms of lengths 4 to 9. In most cases, we observed that the emergent word group is situated in the middle between the predictable and distinctive control word sets, and the rank order of these three groups based on *n*-gram probability and neighbourhood density conforms to our prediction. Specifically, the predictable control words exhibit the highest mean predictability, manifested in the highest overall *n*-gram probability (or equivalently, the lowest overall negative logarithmic *n*-gram probability) and lexical neighbourhood density among the three groups. In comparison, the slang/distinctive control words exhibit the highest mean distinctiveness, manifested in the lowest overall *n*-gram probability and neighbourhood density. The emergent word group tends to fall in between the two control groups.

To evaluate the significance of these trade-offs, we tested a null hypothesis for each comparison between the emergent group and each of the control groups. The null hypothesis is that the mean estimate of the emergent word set does not differ in *n*-gram probability or lexical neighbourhood density from each of the control sets. We tested this by performing a two-tailed *t*-test for every comparison. Across different word lengths and time periods, we observed consistent evidence for rejecting the null hypothesis (see Figure 3; the variations in the magnitude of p values correlate with time and changing sample sizes, as the number of actual emergent words are different in every decade). These results show that the emergent words are significantly different from the control words.

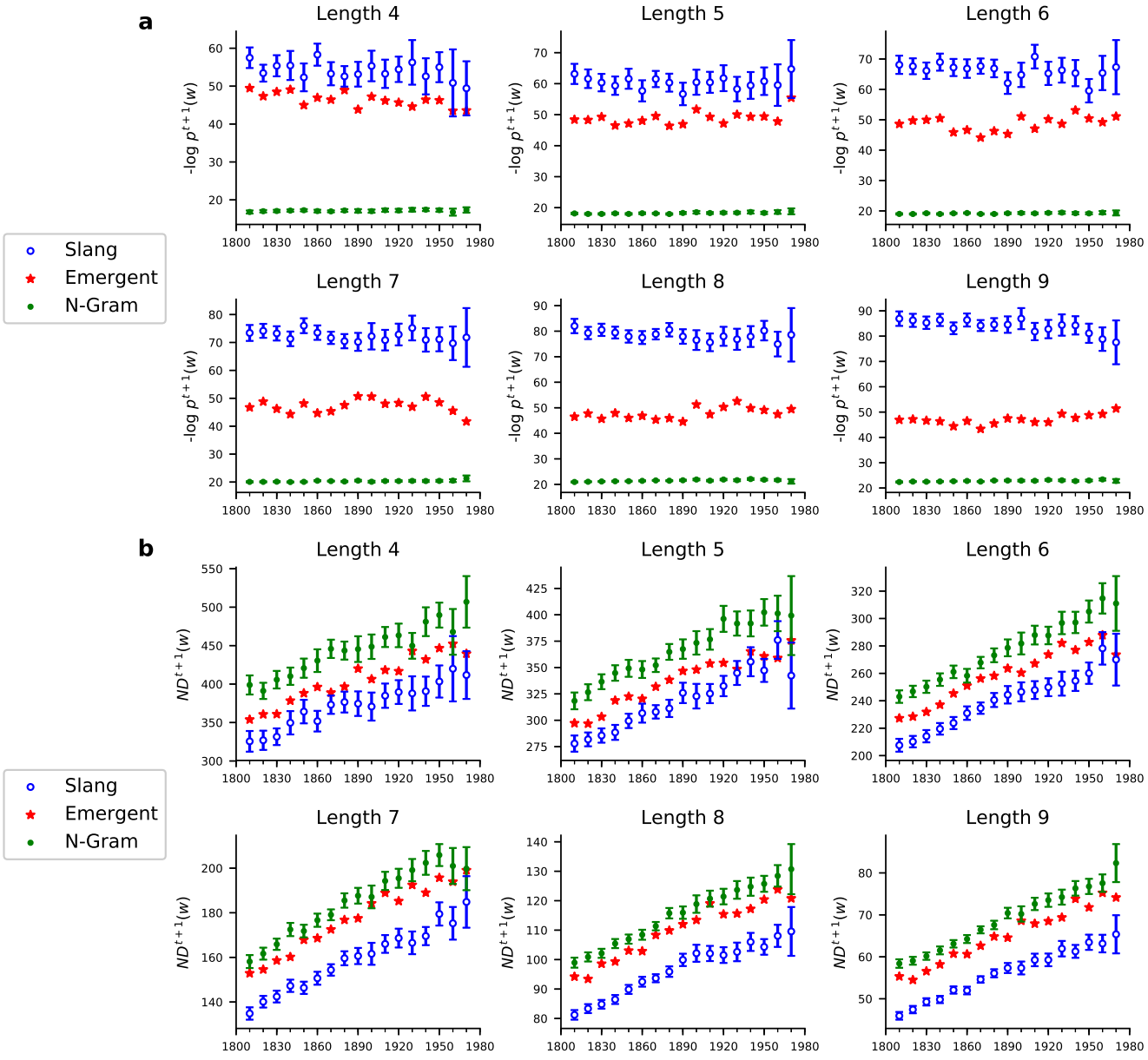

Figure 2: Summary of main results on the trade-off hypothesis of lexical emergence. Each of the panels summarizes the results computed under one of the two measures: a) *n*-gram probability (negative logarithm) and b) neighbourhood density. The vertical axes represent magnitudes under these two measures, and the horizontal axis represents the temporal dimension where each tick corresponds to one decade over the period between 1800 and 1980. Each subplot corresponds to the results for word forms of a different length as specified. Dots (green), stars (red), and circles (blue) correspond to the *n*-gram (predictable) control words, the actual emergent words, and the slang (distinctive) control words, respectively. Each error bar indicates a 95% confidence interval (constructed from the *t*-distribution) for the estimated mean value of the control group. This confidence level is uncorrected for multiple comparisons, and we expect 5% of all intervals to exclude emergent word groups by chance.

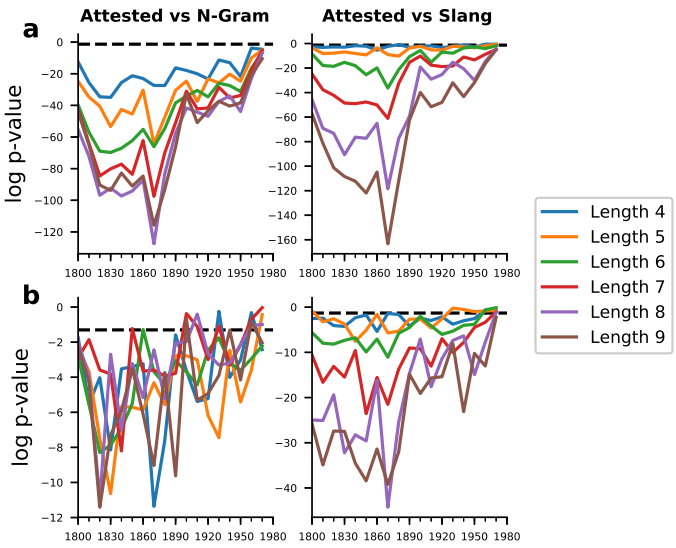

Figure 3: Time courses of p-values. Panels on the left summarize the comparison of the emergent word sets against the corresponding predictable control sets. Panels on the right summarize similar comparisons against the distinctive control sets. The comparisons were based on the measures from a) n-gram probability and b) neighbourhood density. The vertical axis indicates p-values from the t-tests in logarithmic scale, and the horizontal axis represents the time dimension in decades. The black dashed line represents the significance level $p = 0.05$. For each measure, we made 216 simultaneous uncorrected tests, so we expect 11 rejections by chance.

To assess the robustness of these findings, we performed similar analyses based on 1) word forms defined in phonological space as opposed to orthographical space; 2) alternative lexicons obtained by excluding morphologically derived words from the HTE data; 3) an alternative control set based on slang words from a historical resource as opposed to a modern resource. We found that the effects are robust to this variation in design choices, and we omit the details of these analyses due to space constraints. In sum, this set of results provide empirical evidence for our proposal that emerging word forms reflect a trade-off between predictability and distinctiveness and suggest why certain words have entered the lexicon over time, but others have not.

As a follow-up analysis, we assessed whether we can reliably predict emergent word forms from possible words that did not formally enter the lexicon. In particular, we performed a simple logistic regression analysis to predict the identity of each word form from the three groups: emergent words, predictable control words, and distinctive control words. We applied a logistic classifier with $L2$ penalty and the multinomial loss function using the `scikit-learn` package (Pedregosa et al., 2011). For each future decade, we trained the classifier using data from the previous decade t and used the same classifier to make predictions for data from $t+1$. We used three feature sets for classification: 1) n-gram probabilities of words from the three groups; 2) lexical neighbourhood densities of the same words; 3) a combination of their n-gram probabilities and neighbourhood densities.

In general, we observed that predictive accuracies of the three word groups are above chance (33.3% for a three-way classification) under all three feature choices for each decade and length that we considered (predictive accuracy when using neighbourhood density, mean = 43.0%, and standard deviation across word length groups and time periods, SD = 4.2%; using n-gram probability, mean= 61.0%, SD = 3.6%; using the combined features, mean= 61.0%, SD = 3.6%). We noted that the above-chance predictive accuracies are sustained over time, suggesting the trade-off holds generally and not just for certain periods in the history of Modern English. We also noted that the n-gram model performed generally better than the neighbourhood density model, partly because one of the control word groups was directly simulated using n-gram statistics.

Overall, these findings suggest that there are predictable differences in the compositional structure of emergent word forms and that of n-gram generated and slang word forms from the control groups.

Figure 4 further demonstrates the trade-off idea with three example word forms chosen from the three word groups in the 1930s, along with their nearest-neighbour word forms measured by edit distance from the same period. The emergent word form "macro" reflects a trade-off in neighbourhood density: It has fewer 1-edit lexical neighbours (6) than the highly predictable n-gram generatd word "codet" (9 neighbours), but it has more neighbours than the highly distinctive slang word "porph" that has the fewest neighbours (3).

Figure 4: Demonstration of the predictability-distinctiveness trade-off. Panels a), b), and c) show an example word form and its lexical neighbours from the lexicon in the 1930s under the n-gram control set, emergent word set, and slang control set, respectively. The examples are placed in the center, surrounded by their neighbours. Each example word is exactly one edit distance away from each of its neighbours.

Conclusion

We have shown that the historical emergence of English word forms follows a trade-off between predictability and distinctiveness. This trade-off is manifested in the properties of emergent words that straddle between 1) highly predictable computer-generated word forms that conform to statistical properties of the existing lexicon, and 2) highly distinctive word forms originated from slang that had not yet enter into the standard lexicon. We have suggested that such a trade-off may reflect the general principles of language evolution discussed in prior work, under the joint functional pressures for communicative expressivity and cognitive ease (Labov, 2011; Jespersen, 1959; Otto, 1956). Future research should explore whether the same set of principles holds in the emergence of word forms in languages other than English and how word forms interact with meaning (cf. Ramiro, Srinivasan, Malt, & Xu, 2018) in lexical evolution.

Acknowledgments

We thank the University of Glasgow for licensing of the HTE data. We also thank Barbara C. Malt, Peter Turney, Suzanne Stevenson, and Barend Beekhuizen for their constructive comments on this work. This research is supported by an NSERC DG grant and a Connaught New Researcher Award to YX.

References

Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, *44*(4), 568–591.

Baugh, A. C., & Cable, T. (1993). *A history of the english language*. London, UK: Routledge.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: analyzing text with the natural language toolkit*. Sebastopol, CA: O'Reilly Media, Inc.

Edwards, J., Beckman, M. E., & Munson, B. (2004). The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *Journal of speech, language, and hearing research*, *47*(2), 421–436.

Flemming, E. (2004). Contrast and perceptual distinctiveness. In B. Hayes, R. Kirchner, & D. Steriade (Eds.), *The phonetic bases of markedness.* (pp. 232–276). Cambridge, UK: Cambridge University Press.

Jespersen, O. (1959). *Language: Its nature, development and origin*. London: Allen & Unwin.

Kay, C., Roberts, J., Samuels, M., & Wotherspoon, I. (Eds.). (2017). *The historical thesaurus of english, version 4.21*. Glasgow, UK: University of Glasgow. Retrieved from `http://historicalthesaurus.arts.gla.ac.uk/`

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*, 1049–1054.

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*, 109–128.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.

Krott, A. (1996). Some remarks on the relation between word length and morpheme length. *Journal of Quantitative Linguistics*, *3*(1), 29–37.

Labov, W. (2011). *Principles of linguistic change, volume 3: Cognitive and cultural factors* (Vol. 36). Hoboken, NJ: John Wiley & Sons.

Mattiello, E. (2008). *An introduction to english slang: A description of its morphology, semantics and sociology* (Vol. 2). Monza, Italy: Polimetrica-International Scientific Publisher.

Mattiello, E. (2013). *Extra-grammatical morphology in English: Abbreviations, blends, reduplicatives, and related phenomena* (Vol. 82). Berlin, Germany: Walter de Gruyter.

Meylan, S. C., & Griffiths, T. L. (2017). Word forms–not just their lengths–are optimized for efficient communication. *arXiv preprint arXiv:1703.01694*.

Otto, J. (1956). *Language: Its nature development and origin.* Crows Nest, Australia: George Allen & Unwin Limited.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*, 280–291.

Ramiro, C., Srinivasan, M., Malt, B. C., & Xu, Y. (2018). Algorithms in the historical emergence of word senses. *Proceedings of the National Academy of Sciences*, *115*(10), 2323–2328.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognitionandcategorization* (pp. 27–48). Hillsdale: Lawrence Erlbaum.

Sanders, N. C., & Chin, S. B. (2009). Phonological distance measures. *Journal of Quantitative Linguistics*, *16*(1), 96–114.

Shannon, C. E. (1951). Prediction and entropy of printed english. *Bell System Technical Journal*, *30*(1), 50–64.

Stemberger, J. P. (2004). Neighbourhood effects on error rates in speech production. *Brain and Language*, *90*(1-3), 413–422.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond coltheart's n: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971–979.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Cambridge: Addison-Wesley.