Real Words, Possible Words, and New Words

Janet B. Pierrehumbert (jbp@northwestern.edu)
Linguistics Department
Northwestern University
2016 Sheridan Rd.
Evanston IL 60208

Keywords: lexicon; phonotactics; morphology; nonwords; word frequency; word formation

Introduction

Phonologists and psycholinguists draw a three-way distinction amongst real words, possible words, and impossible words.The distinction between real words and possible words provides the foundation for lexical decision experiments. The distinction between possible words and impossible words reveals implicit cognitive generalizations about words in a language, and thereby contributes to the understanding of language acquisition and processing. Left to the side in this vast body of theory and experimental results is a real understanding of new words. Is a new word just a new random selection from the possible words? No. First of all, some possible words are more possible than others. Second, there's an important distinction between the creation of a new word, and its adoption by the linguistic community. The creation of a new word is a manifestation of an individual person's cognitive system. But to be widely adopted, it must successfully compete with other words to be used in discourse.

In this paper, I review a series of results on how and why some possible words are more possible than others. Then, I will introduce work in progress that looks at the interaction of social and cognitive factors in processing new words.

Phonotactics

The phonology of a language is a grammar for its sound structure. The simplest type of grammar is a diphone grammar. Many studies have revealed gradient effects of diphone statistics in predicting the inventory of word types and the extent to which nonwords are judged to be well-formed. These include Frisch, Large, and Pisoni (2010); Hay, Pierrehumbert, and Beckman (2004). For English, diphone statistics alone can provide a powerful method for bootstrapping the lexicon from continuous speech (Daland & Pierrehumbert, 2011). In a widely used algorithm for generating nonwords, diphone statistics are the only treatment of the phonological grammar per se (Rastle, Harrington, & Coltheart, 2002).

However, constraints at larger time-scales are also found in phonology. These, too, make gradient and cumulative contributions to the well-formedness of nonwords. To capture effects of syllable structure, it is necessary to use triphone statistics and/or an explicit hierarchical structure (Coleman & Pierrehumbert, 1997; Hay et al., 2004; Pierrehumbert, 1994). Stress modulates the likelihood of different phones at larger time-scales (Coleman & Pierrehumbert, 1997). Cross-linguistically, a common constraint mitigates against sequences of consonants with the same place of articulation, regardless of the intervening vowels. In Arabic, this constraint displays a cumulative interaction between the similarity of the consonants and their distance (Frisch, Pierrehumbert, & Broe, 2004).

In general, local constraints can make detailed reference to segmental features, whereas constraints involving long spans of phonemes tend to refer to more general classes. This generalization follows from learnability considerations. Forming a statistical generalization requires a big enough sample of word types to distinguish a significant pattern from a simpler null hypothesis about the grammar (Pierrehumbert, 2001). The means by which the cognitive system combines precise local statistical constraints with broad non-local statistical constraints is not yet well understood.

Morphology

Morphology is the theory of how words are made from meaningful parts. Several studies just cited involve morphological structure as well as phonological structure. In Hay et al. (2004), diphone statistics of bisyllabic nonwords predicted well-formedness judgments, but only given the best morphological parse of the nonword. In Frisch et al. (2010), the Arabic statistical patterns pertain to verbal roots, which are a morphological abstraction from the surface forms. The surface forms include obvious violations of the constraints, due to the operation of the non-concatenative Arabic word formation system.

New words are judged to be much better if they have a valid morphological analysis. In fact, productive morphology is the dominant source of new words. In languages such as Turkish and Finnish, the morphology is so productive that the lexicon cannot be construed as a stable, shared, inventory of words (Creutz & Lagus, 2007), and morpheme-based systems perform better than word-based systems in speech engineering (Hirsimäki, Pylkkönen, & Kurimo, 2009). Learning morphology involves learning statistics about relations of words to each other (Pierrehumbert, 2003, 2006). The best known predictor of the productivity of a morphological pattern is the number of word types that exhibit the pattern, and the transparency with which they exhibit it, including both semantic and phonological transparency (Racz, Pierrehumbert, Hay, & Papp, in press). Exploiting the fact that meaningful units are found in more different combinations than arbitrary units, unsupervised learning algorithms that lack any overt semantics perform remarkably well (Creutz & Lagus, 2007).

Heterogeneity

Because they emerge from high-order comparisons amongst words, the morphological systems of individual people should be highly sensitive to their individual vocabularies. It can be difficult to draw the line between rare words and novel words. In Table 1, words occurring at frequencies of $1/1000$ are known to everyone. But words with frequencies of $1/10,000,000$ include some words which seem reasonably familiar on the basis of their parts, and others like *trangia* that are known to some people but not to others.

Table 1: Some English words with different British National Corpus frequencies.

$1/1,000$	$1/100,000$	$1/10,000,000$
should	delicious	swampland
than	weird	thunk
only	understanding	escapologist
people	light	zirconium
also	duck	sitka
me	propaganda	trangia

It is well known that the use of some words is highly dependent on the choice of discourse topic. In a large-scale study of language in USENET discussion groups, Altmann, Pierrehumbert, and Motter (2011) found that most words with frequencies of $1/1000$ or less are at least somewhat concentrated by topic. Further, most are at least somewhat concentrated by speaker. The correlation between these two types of heterogeneity is only moderate; different people use different words to discuss the same topic. Given that the rank-frequency spectrum for words is very heavy-tailed, as observed by Zipf, most word types are rare, and we often encounter unfamiliar words in everyday lexical processing as we meet new people and discuss new topics. Since the real words in psycholinguistic experiments are words that all the subjects can be expected to know, there is a lot we don't understand about how most word types are processed.

In the Wordovators project, my colleagues and I are conducting large-scale experiments in the form of computer games to better understand how novel word types are created, remembered, and adopted (http://www.wordovators.org/). These experiments include experiments on the interaction of cognitive factors with social-indexical factors. Initial results already show significant differences depending on gender (Racz, Beckner, Hay, & Pierrehumbert, 2014) and on the social relevance of variability. The presentation at CogSci2014 will include breaking news for this project.

Acknowledgments

This project was made possible through a grant from the John Templeton Foundation. The opinions expressed in this publication are those of the author and do not necessarily reflect the views of the John Templeton Foundation.

References

Altmann, E. G., Pierrehumbert, J. B., & Motter, A. E. (2011). Niche as a determinant of word fate in online groups. *PLoS One doi:10.1371/journal.pone.0019009*, *6*(5).

Coleman, J., & Pierrehumbert, J. B. (1997). Stochastic phonological grammars and acceptability. In *3rd meeting of the ACL special interest group in computational phonology: Proceedings of the workshop* (p. 49-56). Somerset NJ: Association for Computational Linguistics.

Creutz, M., & Lagus, K. (2007). Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing (TSLP)*, *4*(1), 3.

Daland, R., & Pierrehumbert, J. B. (2011). Learning diphone-based segmentation. *Cognitive Sci*, *35*(1), 119–155.

Frisch, S. A., Large, N. R., & Pisoni, D. B. (2010). Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *J Mem Lang*, *42*(4), 481496.

Frisch, S. A., Pierrehumbert, J. B., & Broe, M. (2004). Similarity avoidance and the OCP. *Nat Lang Ling Th*, *22*(1), 179-228.

Hay, J. B., Pierrehumbert, J. B., & Beckman, M. E. (2004). Speech perception, well-formedness, and the statistics of the lexicon. In *Papers in Laboratory Phonology VI* (p. 58-74). Cambridge UK: Cambridge University Press.

Hirsimäki, T., Pylkkönen, J., & Kurimo, M. (2009). Importance of higher-order n-gram models in morph-based speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(4), 724–732.

Pierrehumbert, J. B. (1994). Syllable Structure and Word Structure: a Study of Triconsonantal clusters in English. In *Papers in Laboratory Phonology III* (p. 168-188). Cambridge, UK: Cambridge Univ. Press.

Pierrehumbert, J. B. (2001). Why phonological constraints are so coarse-grained. *Lang Cognitive Proc*, *16*, 691-698.

Pierrehumbert, J. B. (2003). Probabilistic phonology: Discrimination and robustness. In J. B. Hay, S. Jannedy, & R. Bod (Eds.), *Probabilistic linguistics.* MIT Press.

Pierrehumbert, J. B. (2006). The statistical basis of an unnatural alternation. In *Laboratory Phonology VIII* (p. 81-107). Berlin: Mouton de Gruyter.

Racz, P., Beckner, C., Hay, J. B., & Pierrehumbert, J. B. (2014). Rules, analogy, and social factors co-determine past tense formation in english. In *Joint 1-day workshop between SIGMORPHON and SIGFSM.* Somerset NJ: Association for Computational Linguistics.

Racz, P., Pierrehumbert, J. B., Hay, J. B., & Papp, V. (in press). Morphological emergence. In B. MacWhinney & W. O'Grady (Eds.), *The handbook of language emergence.* New York: Wiley.

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The arc nonword database. *Q J Exp Psychol*, *55A*, 1339-1362.