Cognition 163 (2017) 128–145

Contents lists available at ScienceDirect

Cognition

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

Words cluster phonetically beyond phonotactic regularities

Isabelle Dautriche [a,b,*,1], Kyle Mahowald [c,*,1], Edward Gibson [c], Anne Christophe [a], Steven T. Piantadosi [d]

[a] Laboratoire de Sciences Cognitives et Psycholinguistique (ENS, CNRS, EHESS), Ecole Normale Supérieure, PSL Research University, Paris, France
[b] School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, United Kingdom
[c] Department of Brain and Cognitive Science, MIT, United States
[d] Department of Brain and Cognitive Sciences, University of Rochester, United States

ARTICLE INFO

Article history:
Received 7 April 2015
Revised 16 January 2017
Accepted 1 February 2017
Available online 22 March 2017

Keywords:
Linguistics
Lexical design
Communication
Phonotactics

ABSTRACT

Recent evidence suggests that cognitive pressures associated with language acquisition and use could affect the organization of the lexicon. On one hand, consistent with noisy channel models of language (e.g., Levy, 2008), the phonological distance between wordforms should be maximized to avoid perceptual confusability (a pressure for *dispersion*). On the other hand, a lexicon with high phonological regularity would be simpler to learn, remember and produce (e.g., Monaghan et al., 2011) (a pressure for *clumpiness*). Here we investigate wordform similarity in the lexicon, using measures of word distance (e.g., phonological neighborhood density) to ask whether there is evidence for dispersion or clumpiness of wordforms in the lexicon. We develop a novel method to compare lexicons to phonotactically-controlled baselines that provide a null hypothesis for how clumpy or sparse wordforms would be as the result of only phonotactics. Results for four languages, Dutch, English, German and French, show that the space of monomorphemic wordforms is clumpier than what would be expected by the best chance model according to a wide variety of measures: minimal pairs, average Levenshtein distance and several network properties. This suggests a fundamental drive for regularity in the lexicon that conflicts with the pressure for words to be as phonologically distinct as possible.

1. Introduction

de Saussure (1916) famously posited that the links between wordforms and their meanings are arbitrary. As Hockett (1960) stated: "*The word 'salt' is not salty, 'dog' is not canine, 'whale' is a small word for a large object; 'microorganism' is the reverse.*" Despite evidence for non-arbitrary structure in the lexicon in terms of semantic and syntactic categories (Bloomfield, 1933; Monaghan, Shillcock, Christiansen, & Kirby, 2014), the fact remains that here is no systematic reason why we call a dog a 'dog' and a cat a 'cat' instead of the other way around, or instead of 'chien' and 'chat.' In fact, our ability to manipulate such arbitrary symbolic representations is one of the hallmarks of human language and makes language richly communicative, since it permits reference to arbitrary entities, not just those that have iconic representations (Hockett, 1960).

Because of this arbitrariness, languages have many degrees of freedom in what wordforms they choose and in how they carve up semantic space to assign these forms to meanings. Although the mapping between forms and meanings is arbitrary, the particular sets of form-meaning mappings chosen by any given language may be constrained by a number of competing pressures and biases associated with learnability and communicative efficiency. For example, imagine a language that uses the word 'feb' to refer to the concept HOT, and that the language now needs a word for the concept warm. If the language used the word 'fep' for WARM, it would be easy to confuse with 'feb' (HOT) since the two words differ only in the voicing of the final consonant and would often occur in similar contexts (i.e. when talking about temperature). However, the similarity of 'feb' and 'fep' could make it easier for a language learner to learn that those sound sequences are both associated with temperature, and the learner would not have to spend much time learning to articulate new sound sequences since 'feb' and 'fep' share most of their phonological structure. On the other hand, if the language used the word 'sooz' for the concept WARM, it is unlikely to be phonetically confused with 'feb' (HOT), but the learner might have to learn to articulate a new set of sounds and would need to remember two quite different sound sequences that refer to similar concepts.

[*] Corresponding authors at: School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, United Kingdom (I. Dautriche). Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02139, United States (K. Mahowald).

E-mail addresses: isabelle.dautriche@gmail.com (I. Dautriche), kylemaho@mit.edu (K. Mahowald).

[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.cognition.2017.02.001
0010-0277/© 2017 Elsevier B.V. All rights reserved.

Here, we investigate how communicative efficiency and learnability trade off in the large-scale structure of natural languages. We have developed a set of statistical tools to characterize the large-scale statistical properties of the lexicons. Our analysis focuses on testing and distinguishing two pressures in natural lexicons: a *pressure for dispersion* (improved discriminability) versus a *pressure for clumpiness* (re-use of sound sequences). Below, we discuss each in more detail.

1.1. A pressure for dispersion of wordforms

Under the noisy channel model of communication (Gibson, Bergen, & Piantadosi, 2013; Levy, 2008; Shannon, 1948), there is always some chance that the linguistic signal will be misperceived as a result of errors in production, errors in comprehension, inherent ambiguity, and other sources of uncertainty for the perceiver. A lexicon is maximally robust to noise when the expected phonetic distance among words is maximized (Flemming, 2004; Graff, 2012), an idea used in coding theory (Shannon, 1948). Such dispersion has been observed in phonological inventories (Flemming, 2002; Hockett & Voegelin, 1955; Liljencrants & Lindblom, 1972) in a way that is sensitive to phonetic context (Steriade, 2001; Steriade, 1997). The length and clarity of speakers' pronunciations are also sensitive to context predictability and frequency (e.g., Aylett & Turk, 2004; Bell et al., 2003; Cohen Priva, 2008; Pluymaekers, Ernestus, & Baayen, 2005; Raymond, Dautricourt, & Hume, 2006; Van Son & Van Santen, 2005), such that potentially confusable words have been claimed to be pronounced more slowly and more carefully. Applying this idea to the set of wordforms in a lexicon, one would expect wordforms to be maximally dissimilar from each other, within the bounds of conciseness and the constraints on what can be easily and efficiently produced by the articulatory system. Indeed, a large number of phonological neighbors (i.e., words that are one edit apart like 'cat' and 'bat') can impede spoken word recognition (Luce, 1986; Luce & Pisoni, 1998), and the presence of lexical competitors can affect reading times (Magnuson, Dixon, Tanenhaus, & Aslin, 2007). Phonological competition may also be a problem in early stages of word learning: Young toddlers fail to use a single-feature phonological distinction to assign a novel meaning to a wordform that sounds similar to a very familiar one (e.g., learning a novel word such as 'tog' when having 'dog' in their lexicon, Dautriche, Swingley, & Christophe, 2015; Swingley & Aslin, 2007).

1.2. A pressure for clumpiness of wordforms

Dispersion of wordforms in the lexicon may be functionally advantageous. Yet, it is easy to see that a language with a hard constraint for dispersion of wordforms will have many long, therefore complex, words (as words need to be distinctive). A well designed lexicon must also be composed of simple signals that are easily memorized, produced, processed and transmitted over generations of learners. In the extreme case, one could imagine a language with only one wordform. Learning the entire lexicon would be as simple as learning to remember and pronounce one word. While this example is absurd, there are several cognitive advantages for processing words that are similar to other words in the mental lexicon. Words that overlap phonologically with familiar words are considered to be easier to process because they receive support from stored phonological representations. There is evidence that words that have many similar sounding words in the lexicon are easier to remember than words that are more phonologically distinct (Vitevitch, Chan, & Roodenrys, 2012) and facilitate production as evidenced by lower speech error rates (Stemberger, 2004; Vitevitch & Sommers, 2003). They also may have shorter naming latencies (Vitevitch & Sommers, 2003) (but see Sadat, Martin, Costa, & Alario, 2014 for a review of the sometimes conflicting literature on the effect of neighborhood density on lexical production). Additionally, words with many phonological neighbors tend to be phonetically reduced (shortened in duration and produced with more centralized vowels) in conversational speech (Gahl, 2015; Gahl, Yao, & Johnson, 2012).This result is expected if faster lexical retrieval in production is associated with greater phonetic reduction in conversational speech as it is assumed for highly predictable words and highly frequent words (Aylett & Turk, 2006; Bell et al., 2003). In sum, while words that partially overlap with other words in the lexicon may be difficult to recognize (Luce, 1986; Luce & Pisoni, 1998), they seem to have an advantage for memory and lexical retrieval.

One source of wordform regularity in the lexicon comes from a correspondence between phonology and semantics and/or syntactic factors. Words of the same syntactic category tend to share phonological features, such that nouns sound like nouns, verbs like verbs, and so on (Kelly, 1992). Similarly, phonologically similar words tend to be more semantically similar within a language, across a wide variety of languages (Dautriche, Mahowald, Gibson, & Piantadosi, 2016; Monaghan et al., 2014). The presence of these natural clusters in semantic and syntactic space therefore results in the presence of clusters in phonological space. Imagine, for instance, that all words having to do with sight or seeing had to rhyme with 'look'. A cluster of '-ook' words would develop, and they would all be neighbors and share semantic meaning. One byproduct of these semantic and syntactic clusters would be an apparent lack of sparsity among wordforms in the large-scale structure of the lexicon. There is evidence that children and adults have a bias towards learning words for which the relationship between their semantics and phonology is not arbitrary (Imai & Kita, 2014; Imai, Kita, Nagumo, & Okada, 2008; Monaghan, Christiansen, & Fitneva, 2011, 2014; Nielsen & Rendall, 2012; Nygaard, Cook, & Namy, 2009). However such correspondences between phonology and semantic may affect some aspects of the production system: speech production errors that are semantically and phonologically close to the target (e.g., substituting 'cat' by 'rat') are much more likely to occur than errors than are purely semantic (e.g., substituting 'cat' by 'dog') or purely phonological (e.g., substituting 'cat' by 'mat') in spontaneous speech (the *mixed error effect*, e.g., Dell & Reich, 1981; Goldrick & Rapp, 2002; Schwartz, Dell, Martin, Gahl, & Sobel, 2006).

Another important source of phonological regularity in the lexicon is *phonotactics*, the complex set of constraints that govern the set of sounds and sound combinations allowed in a language (Hayes & Wilson, 2008; Vitevitch & Luce, 1998). For instance, the word 'blick' is not a word in English but plausibly could be, whereas the word 'bnick' is much less likely due to its implausible onset *bn-* (Chomsky & Halle, 1965).[2] These constraints interact with the human articulatory system: easy-to-pronounce strings like 'ma' and 'ba' are words in many human languages, whereas some strings, such as the last name of Superman's nemesis *Mister Mxyzptlk*, seem unpronounceable in any language.[3] Nevertheless, the phonotactic constraints of a language are often highly language-specific. While English does not allow words to begin with *mb*, Swahili and Fijian do. Phonotactic constraints provide an important source of regularity that aids production, lexical access, memory and learning. For

[2] There are many existing models that attempt to capture these language-specific rules. A simple model is an n-gram model over phones, whereby each sound in a word is conditioned on the previous n-1 sounds in that word. Such models can be extended to capture longer distance dependencies that arise within words (Gafos, 2014) as well as feature-based constraints such as a preference for sonorant consonants to come after less sonorant consonants (Albright, 2009; Goldsmith & Riggle, 2012; Hayes, 2012; Hayes & Wilson, 2008).

[3] Though as a anonymous reviewer pointed out, some have succeeded in doing so (https://en.wikipedia.org/wiki/Mister_Mxyzptlk#Pronunciation).