The Communicative Lexicon Hypothesis

Steven T. Piantadosi†, Harry J. Tily‡, Edward Gibson†
{ **piantado@mit.edu, hjt@stanford.edu, egibson@mit.edu** }
† MIT Department of Brain and Cognitive Sciences
43 Vassar Street, Building 46, Room 3037
Cambridge, MA 02139
‡ Stanford University Department of Linguistics
450 Serra Mall
Stanford, CA 94305

Abstract

Recent work suggests that variation in online language production reflects the fact that speech is information-theoretically efficient for communication. We apply this idea to studying the offline, structural properties of language, asking whether lexical properties may similarly reflect communicative pressures. We present evidence for the Communicative Lexicon Hypothesis (CLH): human lexical systems are efficient solutions to the problem of communication for the human language processor. While the relationship between sounds and meanings may be arbitrary, pressure for concise and error-correcting communication — within the constraints imposed by human articulatory, perceptual and cognitive abilities — has influenced which sets of phonological forms have emerged in the lexicons of human languages. We present two tests of the CLH: first, we show that word lengths are better predicted by a word's average predictability in context than its overall frequency. Second, we show that salient (lexically stressed) parts of words are more informative about a word's identity, in English, German, Dutch, Hawai'ian, and Spanish.

Keywords: Rational analysis; lexicon; word length; lexical stress; surprisal.

Introduction

Many human cognitive systems appear to implement good solutions to the problems they are required to solve (Oaksford & Chater, 1999). Decay patterns in human memory, for instance, can be interpreted as modeling the probability of needing to retrieve a given element from memory (Anderson & Milson, 1989), and patterns in phonetic perception can be derived by considering a Bayes-optimal phoneme recognizer (Feldman & Griffiths, 2007). Along these lines, Hockett (1960) identified thirteen features of human language that make it well-designed for communication. For instance, language can be conveyed in the vocal-auditory channel, leaving the rest of the body free to simultaneously perform other tasks, and language productively allows users to create never-before-uttered sentences that are immediately comprehensible to other speakers of the language. Another property desirable for communication is that more frequent words are shorter: Zipf (1935) argued that this means on average people can expend less effort and communicate more efficiently, since the most commonly uttered words take the least effort and time to articulate.

A good communicative system must balance several mutually incompatible goals. For instance, a language should communicate meanings as concisely as possible, so words should be short. But a language should communicate meanings unambiguously and with as little confusion as possible, so words should sound as different as possible. Unfortunately, the limits of human ability make it impossible to satisfy both of these objectives simultaneously: the number of short wordforms that can be differentiated accurately by the human articulation and perception systems is very limited. At one extreme, a language could have only one wordform which is used for every meaning. Such a language would be highly ambiguous but very concise. At the other extreme, one could have long and distinct wordforms, which would be different from each other and thus not confusable even with considerable noise, but utterances would then become very long. A better solution would be a language with some intermediate number of wordforms that sound distinct enough to be identified accurately in context, but are not so distinct that utterances become overly long.

Good communicative features can be found at the phonetic level within the speech channel, such as shortening and reducing less informative parts of words (e.g. Aylett & Turk 2004; Bell et al., 2003; Jurafsky et al., 2001; Pluymaekers, Ernestus & Baayen, 2005; van Son & Pols 2003). Here we ask whether—similar to known properties of speech production—characteristics of the *lexicon itself* may reflect communicative pressures. We present evidence for the *communicative lexicon hypothesis* (CLH): human lexical systems are efficient solutions to the problem of communication for the human language processor.

It is important to emphasize that the CLH makes predictions only relative to the limitations and capacities of the human language processor; for instance, it would likely be possible to design a better system for communication if humans had superior perceptual or cognitive abilities. The CLH holds that the lexicon is particularly well-structured for the specific communication mechanisms that humans use. Kuperman, Ernestus & Baayen (to appear) present a finding that exemplifies the kind of property we would expect from language under the CLH. They show that for the four languages they study—English, Dutch, German, and Italian—speech units with medium duration are used more frequently than those with particularly high or low duration. Kuperman and colleagues argue that particularly long sound speech units are inefficient because they are harder to produce and make utterances longer, but particularly short speech units are also inefficient since they may lead to more frequent mishearing.

The studies we present here test two more predictions of the CLH: (i) word length should be better predicted by a word's typical predictability in context than by its raw frequency, (ii) salient parts of a word should be more informative about the word's identity.

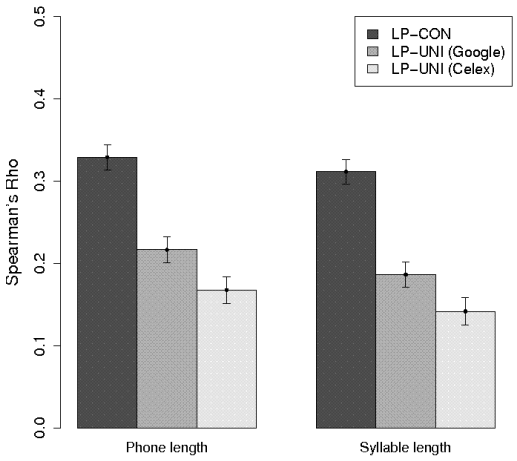

Figure 1: Nonparametric correlations between word length measures and both surprisal and frequency measures. All correlations are significant ($p < 0.001$) but *LP-CON* is a much better predictor of word length. Error bars show bootstrapped 99% confidence intervals.

Probability in context predicts word length better than overall frequency

Zipf (1935) noted a desirable feature of human lexical systems: more frequent words tend to be shorter. For instance, frequent words like "a" and "the" are only a few phonemes long, while less frequent words like "antipasto" and "reductionism" are longer. Zipf called this property the *Law of Abbreviation* and argued that it results from "an underlying law of economy": language can be expressed more concisely if the most commonly uttered words are short. It would be extremely inconvenient if a very frequent word like "of" took as long as "reductionism" to pronounce[1]. Zipf identified both truncations and substitutions as possible mechanisms for shortening frequent words. For instance "movies" is a truncated form of "moving pictures", which came into use as moving pictures became a more frequent conversation topic[2]. Thus, speakers minimize articulatory effort by making frequent words short.

An alternative hypothesis is that word lengths are determined by the average predictability of the word in context (cf. MacDonald & Shillcock, 2001; van Son & Pols, 2003). People can guess upcoming linguistic material with considerable accuracy (Shannon, 1951; Altmann & Kamide, 1999) and recent processing research has shown that comprehenders process predictable material more easily than unlikely material (Levy, 2008; Shillcock & MacDonald, 2003). Why would word lengths reflect this predictability measure instead of raw frequency? The CLH provides two closely related explanations. First, comprehenders need less information from the phonetic input to identify a likely word than an unlikely word. Given that human communication takes place across a "noisy channel" of imprecise articulation and perception, there is some probability that any given phoneme will be incorrectly transmitted. By assigning more phonemes only to words that will be hard to reconstruct if the phonological signal is mistransmitted because of their low frequency, a lexicon can keep the probability of any word failing to be understood small. By keeping words that tend to be easily predicted short, the language can simultaneously respect the pressure for conciseness. Second, the theory of *uniform information density* (UID) holds that speakers try to produce language at a relatively constant information rate[3] (e.g. Levy 2005, Levy & Jaeger, 2007). That is, when speakers are about to produce something very surprising—or unexpected—they slow down, allowing themselves or the listener more time to process the increased amount of information. When words are highly predicted, they convey little information and can be produced and processed more quickly. These effects can be seen at both the syntactic (e.g. Levy & Jaeger 2007) and phonetic levels (e.g. Aylett and Turk, 2004). If the lexicon were constructed to optimize information density, a word's length would depend on how expected it typically is.

Thus the CLH predicts that word length should depend on a word's average predictability in context, and not its overall frequency: such a system may represent a tradeoff between making a lexicon more concise, but still less prone to error, and also uniformly informative in normal communication[4]. To our knowledge, the only work to explicitly address the relationship between predictability and length is Manin (2006), who showed that subjects' ability to guess an upcoming word is strongly correlated with the word's length. We present results showing that predictability in context, as measured by forward trigram predictability estimated from written trigram frequencies, is a significantly better predictor of word length than overall frequency. We define a word's average predictability in context, *LP-CON*, as its average negative log probability in trigram contexts:

$$\textit{LP-CON}(Z) = -\frac{1}{cnt(Z)} \sum_{XY} cnt(XYZ) \log P(Z \mid XY) \quad (1)$$

[1]This principle is well understood in information theory: in efficient coding schemes—such as Huffman coding—the shortest average message length is achieved by assigning short code words to more frequent messages. Samuel Morse used a similar idea in designing his code, assigning more frequent letters shorter code words.

[2]Zipf also identified two kinds of substitutions: durable and temporary. Durable substitutions are changes to shorter words such as "car" for "automobile", while temporary substitutions consist of using shorter words temporarily in discourse — for instance, pronouns such as "it" obviate the need to repeatedly use a longer word.

[3]Levy (2005) shows conditions under which UID is optimal for communication.

[4]It is worth mentioning that because predictability in context is a better approximation of the predictive distribution of words than raw frequency, it can be used to construct a code with shorter average lengths. Making use of context in this way would require *context-dependent* codes where words have different forms depending on the context. These can be seen, for instance, in the shortening of highly predictable words, such as changing "give them" to "givem." However, learnability and processing concerns may rule out more widespread use of such alternations.

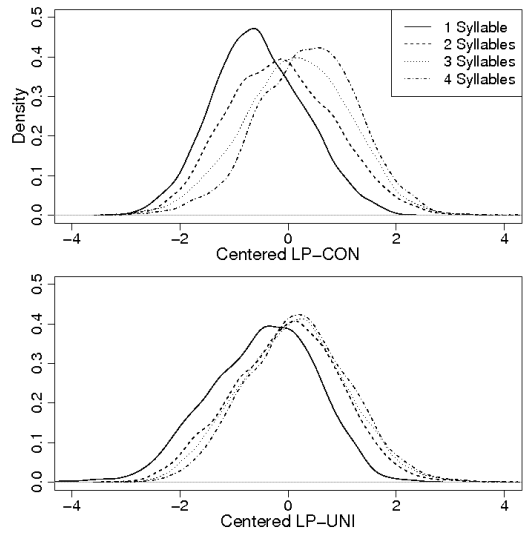

Figure 2: Distributions of centered versions of *LP-CON* and *LP-UNI* (computed using Google) for words of each syllable length. Similar results were found using *LP-UNI* from CELEX. The *LP-CON* curves are more spread out, indicating greater discriminability.

where $cnt(Z)$ and $cnt(XYZ)$ are the corpus counts of Z and XYZ respectively. This measure can be interpreted as the log probability of a word Z averaged over all trigram contexts XY it appears in, weighted by how often XY precedes Z. This measure differs from a word's unigram frequency, or overall negative log probability:

$$\textit{LP-UNI}(Z) = \log \frac{cnt(Z)}{\sum_X cnt(X)}. \quad (2)$$

We computed *LP-CON* and *LP-UNI* using the Google Web-1T corpus (Brants & Franz, 2006), a large database of English n-grams computed from approximately 1 trillion words of English text taken from web pages. We also calculate *LP-UNI* using the CELEX database (Baayen, Piepenbrock & Gulikers, 1995). In practice, frequency and contextual probability are moderately correlated ($R^2 = 0.36$ for the Google unigram counts). Our first study aims to determine which of these measures better predicts word length. If word forms were structured to minimize articulatory effort, one would expect *LP-UNI* to better predict word length; conversely, if words are well-designed for communication across a noisy channel with a processor that can predict upcoming material, word length should be determined mainly by *LP-CON*.

Results of several correlation analyses are shown in Figure 1. Because the relationship between predictability and predictability and word length appears to be nonlinear, we performed a Spearman rank correlation between *LP-CON* and word length, as well as *LP-UNI* and word length. The results of this correlation analysis can be interpreted as describing how well any arbitrary monotonic function could predict word length from each measure. As Figure 1 shows, *LP-CON* is a better predictor of word length as measured either by number of phones or number of syllables. These results indicate that predictability as measured by surprisal in local linguistic context better predicts word length than raw frequency.

Next, we studied the ways in which *LP-CON* is a better predictor. To do this, we scaled the *LP-CON* and *LP-UNI* measurements for the whole lexicon, to give them equal variance across all words. We then looked at the distribution of these measurements within each word length, as measured by number of syllables. These plots are shown in Figure 2. The fact that the curves for *LP-UNI* are highly overlapping for words with more than one syllable indicates that *LP-UNI* cannot distinguish these word lengths well. The curves for *LP-CON* are considerably more spread out for longer words, indicating that *LP-CON* can better predict the words' lengths.

Together these results indicate that the predictability-based account provides a better theory of word length than only frequency. This does not imply that word frequency plays no role; indeed, it is difficult to uncover the independent effects of *LP-CON* and *LP-UNI* since they are correlated, their relationship is heteroskedastic, and each relate nonlinearly to length[5]. However, the results do indicate that between the simple predictability-based and frequency-based accounts of word length, the former is better supported.

Stressed syllables are more informative than unstressed syllables

In our first study, we provided evidence that human lexicons aim for both UID and conciseness while ensuring that the probability of miscommunication remains low. In our second study, we explore a second prediction of CLH. If external factors were to make some syllables have a lower probability of being misspoken or misheard than others, then a well-designed lexicon would make more use of those syllables to convey information. Analogously, if one is sending two boxes—perhaps one through UPS, which has a low chance of getting lost, and one through normal mail, which has a higher chance—it makes sense to put more in the box that is not likely to get lost, rather than the other way around.

The factor we investigated as potentially affecting the probability of error is stress. Stress is a per-syllable property that affects phonetic realization in terms of at least duration, energy, and spectral tilt (Bolinger, 1965; Lehiste & Peterson, 1959). All else equal, stressed syllables are less likely than unstressed syllables to be misproduced or misheard, due to the extra articulatory effort involved in producing a stressed syllable, the extra duration to allow the articulatory organs to reach their targets, and the salience of the stressed syllable to the hearer. The CLH therefore predicts that stressed syllables will be more informative about a word's identity than

[5] After transforming *LP-CON* and word length to approximately correct for nonlinearity and heteroskedasticity, a multiple regression revealed that transformed *LP-CON* is a significant predictor of length, over and above *LP-UNI*. However, the transformation makes it difficult to directly compare the sizes of their respective effects.

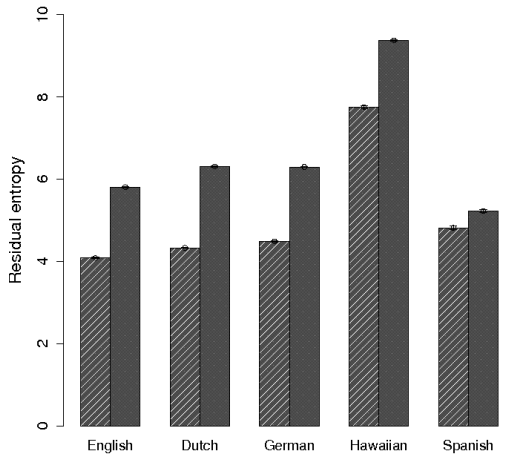

Figure 3: Residual entropy for stressed (striped bars) versus unstressed (solid bars) in five languages. Error bars show 99% confidence intervals.

unstressed syllables. Accordingly, Huttenlocher (1984) observed that deleting unstressed syllables led to less ambiguity in the lexicon than deleting stressed syllables.

Typologically, languages exhibit many different types of stress systems. In most languages, stress placement is determined by hierarchical principles: as well as one syllable within a word being picked out to bear word-level stress, one word in an intonational phrase will be selected to bear phrasal stress, and so on (see Hayes, 1995 for discussion). For simplicity, we consider only primary stress in cases where words have multiple stresses, and only lexical stress. This is fixed for a given lexical item, though languages differ in to what extent words differ idiosyncratically in stress placement versus being predictable from language-wide rules or semi-productive generalizations. In all different stress systems, the CLH predicts that a good lexicon would use stressed syllables to convey more information about the intended word, because they are more acoustically salient.

Here we consider five languages, chosen primarily due to the availability of data: English, Dutch, German, Spanish and Hawai'ian. However, these do differ somewhat in their stress patterns. English, German and Dutch stress is largely predictable from syllable weight, and few cases where lexical stress distinguishes between lexical items. Hawai'ian stress is also relatively predictable, though its phonology differs greatly in not having closed syllables, and therefore no notion of syllable weight beyond vowel length. Spanish has a greater number of lexically specific cases where stress distinguishes between otherwise homophonous words.

In general, one might expect that some syllables of a word are more informative about the word's identity than others. For example, in the word *accordion*, each syllable *a*, *kor*, *di*, *on* conveys some amount of information that the word is *accordion*. However, *a* is relatively less informative than, say, *di*, because fewer and less frequent words contain *di* than *a*. Therefore, just hearing *a* leaves the listener more unsure about the intended word than hearing *di*.

We operationalize the information conveyed by each syllable in the following way. Before hearing anything, we assume the comprehender's uncertainty about an upcoming word is the entropy of the lexicon: all words are likely in proportion to their frequency and the listener's uncertainty about the next word is given by

$$H = -\sum_{w \in L} P(w) \log P(w) \tag{3}$$

where $P(w)$ is the probability of a word w, and L is the set of words in the lexicon.

After a single syllable, only words beginning with that syllable are now possible, so uncertainty is reduced to just the entropy over those words. The syllable's information is defined as the size of this reduction in uncertainty. If each syllable were transmitted noiselessly, most words could be fully identified or nearly identified after just one or two syllables. However later syllables may still be informative, especially in the presence of noise. To avoid relying on any specific assumptions about articulation-perception error in testing our hypothesis, we calculated the information in each syllable of each the word individually, as though each was the only syllable heard.[6] We refer to the resulting measure as *residual entropy:*

$$H(w \mid s \in w) = -\sum_{w \text{ s.t. } s \in w} P(w \mid s \in w) \log P(w \mid s \in w) \tag{4}$$

where $P(w \mid s \in w)$ is the renormalized probability of w among all words containing the syllable s. Thus, when the residual entropy is large, then the syllable s is relatively uninformative about what word it occurs in. That is, there is considerable uncertainty–entropy–about the intended meaning. Conversely, when residual entropy is small, there is little uncertainty about the intended word.

We test whether stressed syllables are more informative than unstressed syllables by computing the average of $H(w \mid s \in w)$ for every stressed and unstressed syllable *token* in a language's dictionary[7]. Figure 3 shows residual entropy for stressed and unstressed syllables from the five languages. All differences between stressed and unstressed syllables are significant to $p < 0.001$ using a Mann-Whitney test. As this figure makes clear, stressed syllables tend to be more informative than unstressed syllables, and in English, Dutch, German, and Hawai'ian, this difference is a considerable portion of the overall uncertainty over the word's identity. Interestingly, the effect is smallest in the language that makes the most contrastive use of stress, Spanish. This might reflect the

[6] An additional motivation for not conditioning on position within the word is that doing so would lead to significant data sparsity.

[7] Because word frequency data was not available for Hawai'ian, we computed these measures assuming each word was equally probable. Assuming equiprobable words yielded similar results to Equation 4 for the other languages.

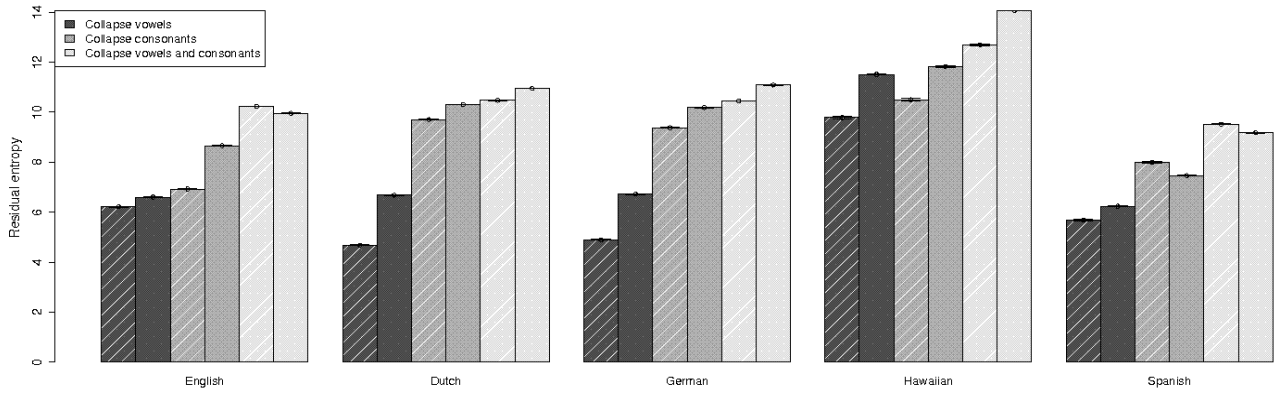

Figure 4: Residual entropy for stressed (striped bars) versus unstressed (solid bars) in five languages. Error bars show 99% confidence intervals.

fact that if stress has been "harnessed" for a discriminative purpose, there is less freedom for the language to adapt to placing stress on more informative syllables.

We also analyzed the informativeness of syllables within part of speech categories. Analysis within nouns, verbs, adjectives, and function words showed that stressed syllables are significantly more informative than unstressed syllables. This result was consistent across English, Dutch, and German, the languages for which part of speech information was available from CELEX.

It is well known that more phonemic contrasts tend to be licensed in stressed positions: English vowels tend to be reduced to a schwa in unstressed positions, for example, and heavy syllables (which contain long vowels or phones in the coda) are more likely to be stressed (Hayes, 1995). The CLH gives a formal explanation for these findings: more contrasts means that a syllable is more informative, but also more likely to be misspoken or misheard. Therefore, only syllables that are carefully articulated should carry important contrasts.

Altmann & Carter (1989) investigated whether stressed syllables still contain more information after removing contrasts caused by the greater variety of vowels in stressed position. Their results suggest that in their dataset, any greater informativeness of stressed syllables may be due to the vowel alone. To test whether vowel reduction is driving the effects presented above by computing the residual entropy, collapsing across vowels, consonants, and both. For example, these respectively count the syllable *kor* as *kVr*, *CoC* and *CVC*, where *C* is a symbol for consonants, and *V* is a symbol for vowels. These averages are shown in Figure 4. All stressed-unstressed comparisons within languages are significantly different at $p < 0.01$ using a Mann-Whitney test. In English, the difference in residual entropy between stressed and unstressed syllables is much smaller when collapsing across vowels than in Figure 3. However, the effect is still present and statistically significant, indicating that most, but not all, of this effect can be explained by vowel neutralization in English. In Dutch, German, Hawai'ian, and Spanish, however, collapsing across consonants impacts the difference in residual entropy between stressed and unstressed syllables *more* than collapsing across vowels.

Discussion & Conclusion

We have presented two studies that show ways in which lexical systems are good solutions to the problem of human communication. We applied the idea common in speech production literature of a "noisy channel" to the lexicon itself and confirmed two information-theoretic predictions of the CLH.

Our first study showed that word length is better predicted by a word's average predictability in context than by its raw frequency. This has the effect that words that are less predictable are typically represented by more syllables, and so have a lower probability of being mistransmitted due to noise. In addition, this may help achieve UID for the language processor. The fact that predictability is a better correlate of length than frequency is good evidence that there is pressure for an error-correcting code as well as just a concise one, and that characteristics of the processor may influence lexical forms. A plausible mechanism for this change can be found in studies of phonetic reduction: several studies have found that specific pronunciations of a word tend to be more reduced in terms of duration and deletion of segments when in a more predictable environment (e.g. Aylett & Turk 2004; Bell et al., 2003; Jurafsky et al., 2001; Pluymaekers, Ernestus & Baayen, 2005; van Son & Pols 2003). Similarly, it may be that speakers modify the effort they put into articulation depending on the local linguistic context (Lindblom, 1990). These might plausibly represent purely speaker-internal processing factors; if these kinds of effects were "fossilized" in the lexicon by learners, they could give rise to effects similar to those we have found. Along these lines, Bybee (2007) reviews evidence that more frequent words exhibit faster lexical change, and argues that the more a word is used in contexts for reduction, the more likely it is to change to a lexically reduced form.

Our second study showed that more acoustically salient

and well-articulated parts of a word are more informative about a word's identity, according to a formal, information-theoretic definition of informativeness. Again, this is predicted by our noisy channel model of communication: human language maximizes the rate at which information can be transmitted while keeping error probability low, so if some section of a word has a lower probability of being mistransmitted then more information can be packed into that section without affecting the overall chance of miscommunication. This effect was robust across the five languages examined, and is generally true even when collapsing across vowels or consonants, but not both. Thus, the CLH provides a rational explanation for the cross-linguistic generalization that stress licenses phonetic contrast.

In summary, we have shown that the CLH refines the rational "least effort" theory of Zipf (1935): lexicons are indeed organized for concise communication, but also show features of error-minimization and UID. Furthermore, some apparently universal properties of language—such as the distribution of word lengths and the licensing of more phonetic contrasts in stressed positions—can be explained by analyzing the lexicon as one component of a rational communicative system.

Acknowledgments

We'd like to thank Mike Frank, Dan Jurafsky, Adam Albright, Celeste Kidd, and members of Tedlab for helpful suggestions. This work was supported by an NSF Graduate Research Fellowship to the first author.

References

Altmann, G., & Carter, D. (1989). Lexical stress and lexical discriminability: Stressed syllables are more informative, but why? *Computer Speech and Language*, *3*, 265-275.

Altmann, G., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*, 247-264.

Anderson, J., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, *96*, 703-719.

Aylett, M., & Turk, A. (2004). The Smooth Signal Redundancy Hypothesis: A Functional Explanation for Relationships between Redundancy, Prosodic Prominence, and Duration in Spontaneous Speech. *Language and Speech*, *47*(1), 31–56.

Bolinger, D. (1965). *Forms of English: Accent, Morpheme, Order*. Cambridge: Harvard University Press.

Brants, T., & Franz, A. (2006). Web 1t 5-gram corpus version 1. *Technical report, Google Inc.*

Bybee, J. (2007). *Frequency of Use and the Organization of Language*. Oxford University Press, USA.

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences*, *3*, 57-65.

Feldman, N., & Griffiths, T. (2007). A rational account of the perceptual magnet effect. In *Proceedings of the cognitive science society.*

Freij, G., Fallside, F., Hquist, C., & Nolan, F. (1990). Lexical stress estimation and phonological knowledge. *Computer Speech & Language*, *4*, 1-15.

Hayes, B. (1995). *Metrical stress theory: Principles and case studies*. Chicago: University of Chicago Press.

Hockett, C. (1960). The origin of speech. *Scientific American*, *203*, 88-96.

Huttenlocher, D. (1984). *Acoustic-phonetic and lexical constraints in word recognition: lexical access using partial information*. MS. thesis, MIT.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. *Typological Studies in Language*, *45*, 229–254.

Kuperman, V., Ernestus, M., & Baayen, R. (in press). Frequency distributions of uniphones, diphones and triphones in spontaneous speech. *The Journal of the Acoustical Society of America*.

Lehiste, I., & Peterson, G. (1959). Linguistic considerations in the study of speech intelligibility. *J. Acoust. Soc. Am.*, *31(4)*, 428-435.

Levy, R. (2005). *Probabilistic models of word order and syntactic discontinuity*. Unpublished doctoral dissertation, Stanford University.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106(3)*, 1126-1177.

Levy, R., & Jaeger, F. (2007). Speakers optimize information density through syntactic reduction. In *Proceedings of the twentieth annual conference on neural information processing systems.*

Lindblom, B. E. F. (1990). Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling* (pp. 403–439). Kluwer.

Manin, D. (2006). Experiments on predictability of word in context and information rate in natural language. *J. Information Processes*, *6(3)*, 229-236.

McDonald, S., & Shillcock, R. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, *44*, 295–323.

McDonald, S., & Shillcock, R. (2003). Eye movements reveal the on-line computation of lexical probabilities. *Psychological Science*, *14*, 648-652.

Pluymaekers, M., Ernestus, M., & Baayen, R. (2005). Lexical frequency and acoustic reduction in spoken Dutch. *The Journal of the Acoustical Society of America*, *118*, 2561.

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

Van Son, R., & Pols, L. (2003). How efficient is speech? *Proceedings of the Institute of Phonetic Sciences*, *25*, 171–184.

Zipf, G. (1935). *The psycho-biology of language: an introduction to dynamic philology*. New York: Houghton Mifflin.