Optimization of American English, Spanish, and Mandarin Chinese over time for efficient communication

John K Pate (johnpate@buffalo.edu)
Department of Linguistics, 609 Baldy Hall
Buffalo, NY 14260 USA

Abstract

Frequent words tend to be short, and many researchers have proposed that this relationship reflects a tendency towards efficient communication. Recent work has sought to formalize this observation in the context of information theory, which establishes a limit on communicative efficiency called the channel capacity. In this paper, I first show that the compositional structure of natural language prevents natural language communication from getting close to the channel capacity, but that a different limit, which incorporates probability in context, may be achievable. Next, I present two corpus studies in three typologically-diverse languages that provide evidence that languages change over time towards the achievable limit. These results suggest that natural language optimizes for efficiency over time, and does so in a way that is appopriate for compositional codes.

Keywords: Communicative efficiency; Uniform Information Density; Smooth Signal Hypothesis; Noisy channel

Introduction

Natural language researchers have long been interested in the prospect that natural language is organized for efficient communication: frequent words tend to be shorter than rare words, allowing talkers to produce shorter word-forms on average. More recent work (Plotkin & Nowak, 2000; Genzel & Charniak, 2002; Aylett & Turk, 2004; Levy & Jaeger, 2007; Jaeger, 2010; Piantadosi, Tily, & Gibson, 2011; Seyfarth, 2014) has sought to formalize this work in the context of information theory by proposing that natural language communicates information close to a limit called the *channel capacity*, although it has left open the question of how closely the limit is approached.

In this paper, I first show that it is not possible for a compositional code like natural language to transmit information close to the channel capacity. Specifically, average signal lengths must exceed the entropy of the distribution over messages by at least the Kullback-Leibler divergence of the true probability distribution over messages from a fully-factorized probability distribution in which every component of the message is statistically independent. Natural language communication must then underperform the channel capacity by at least this Kulback-Leibler divergence.

However, in light of recent work (Piantadosi et al., 2011) that showed a stronger relationship between a word's length and its average probability in context than its unigram probability, I investigate the possibility that language changes over time so that the optimal length of a word, as computed from its average probability in context, better matches its actual length. I present two corpus studies over 350 years in American English, Spanish, and Mandarin Chinese that compare actual word lengths to optimal word lengths in context and in isolation. The first 'backward-looking' study computes optimal word lengths using modern-day language data, and finds that the mismatch between optimal and actual word lengths is smaller for older words. Moreover, the mismatch drops more rapidly, as a function of word age, when the optimal word length is computed relative to a context-sensitive trigram model than when it is relative to a unigram model.

The second 'forward-looking' study divides the 350-year period into 25-year partitions, uses language data from each partition to compute optimal word lengths for each partition, and trains a regression model to predict whether the word-form appears in the next 25-year partition as a function of the mismatch between the word's actual length and its optimal length. This study finds that words with larger mismatches are less likely to 'survive' to the next partition, and moreover finds a stronger effect of mismatch relative to the context-sensitive trigram model than relative to the unigram model. Together, these two studies provide evidence that natural language lexicons change over time in a way that reflects communicative efficiency pressures on a compositional code.

I start by presenting previous work on information-theoretic approaches to language production, along with the minimal technical background necessary for this paper. I then show why natural language does not approach information-theoretic bounds, and use this result to suggest a new bound for compositional codes that may be achievable by considering probability in context. Finally, I present two corpus studies that find evidence that three typologically-diverse languages have changed to approach this new bound.

Background

Linguists have proposed that language is adapted for communication in a general sense for decades. Zipf (1949) proposed the 'Principle of Least Effort' to explain the observation that frequent words tend to be short: frequent words tend to be short so that talkers usually only have to say short words. Lindblom (1990) proposed Hyper- & Hypo-articulation theory to explain the observation that vowels in careful speech tend to be less centralized in formant space: talkers provide more distinct vowels when they believe errors are more likely.

Plotkin and Nowak (2000) proposed an explicit model of word formation over the course of language change in an information-theoretic framework, and showed, analytically and via simulation, that it approached information-theoretic bounds as the vocabulary size increases. However, their model considered words in isolation, but natural language utterances consist of sequences of words. This paper will show that codes with the compositional structure that characterizes natural language cannot approach these information-theoretic bounds, and focus on optimizing sentence lengths.

Subsequent work has mainly consisted of corpus studies that show that synchronic samples of natural language exhibit the correlations on would expect, under an information-theoretic account, between different measures of word length or distinctiveness and word probability, both overall and in context (Aylett & Turk, 2004; Frank & Jaeger, 2008; Bell, Brenier, Gregory, Girand, & Jurafsky, 2009). Piantadosi et al. (2011) revisited Zipfian distributions, and compared both log word probability and average log word probability in context, operationalized as a trigram model, with word lengths, operationalized as the length of the word's spelling in letters. They found that, of the two probability measures, average log probability in context exhibited a stronger correlation with word lengths. At first glance, this result would seem to contradict an information-theoretic approach: the optimal length of a word in isolation simply is its negative log probability, with an appropriate choice of base for the logarithm. However, I soon show that average probability in context is a more appropriate measure for optimizing *sentence* lengths.

In this paper, I return to Plotkin and Nowak's (2000) proposal that language adapts towards these bounds over time, using a corpus-based methodology and an emphasis on the relationship between words and their sentential contexts.

Information theory

In the information-theoretic framing of language, a talker has a message m that is a sequence of message characters m_i from some alphabet of message characters $\mathcal{M}$. For example, $\mathcal{M}$ may be the set of lexical entries or, as in CCG, a set of (syntactic category, semantic category) pairs. The message cannot be transmitted directly, so the talker encodes it into a signal s that is a sequence of signal characters s_i from some alphabet of signal characters $\mathcal{S}$.[1] For example, $\mathcal{S}$ may be the inventory of syllables or phonemes of the language.

An efficient code has two properties. First, it is *short*: the number of signal characters per message character, on average, is low. Second, it is *robust*: the probability that the listener fails to identify the correct message is low. The length of the shortest possible code depends only on the probability distribution over messages $P(\boldsymbol{m})$, and is given by the entropy of that distribution:

$$H(\boldsymbol{m}) = \sum_{m \in \boldsymbol{m}} P(m) \log_b \left(\frac{1}{P(m)} \right) \tag{1}$$

where $\boldsymbol{m} = \mathcal{M}^+$ is the set of all messages (i.e. all sequences of message characters). The log term is called the Shannon information of m, and the entropy is just the expected Shannon information under the probability distribution $P(m)$. If we set b to be the size of the signal alphabet $|\mathcal{S}|$, then the Shannon information of m is the optimal signal length in signal characters for message m. Adjusting signal lengths to match the Shannon entropy, called *source coding* eliminates redundancy in the code, and achieves property 1: short codes.

Listeners may encounter noise in real-world situations, due to slips of the tongue on the part of the talker, distraction or cognitive overload on the part of the listener, dialect differences, environmental noise, or other sources of noise. Noise can be countered by adding redundancy to the signal. For example, a word may differ from all other words by several phonemes, allowing the listener to recover the intended word even if some phonemes are mis-perceived or masked by environmental noise. While the resulting code is more robust, it is also longer, and we might worry that signals will have to become arbitrarily long to drive the error rate toward zero.

The Noisy Channel theorem shows that an arbitrarily low error rate can be achieved with signals that are not arbitrarily long, as long as they do not exceed the channel capacity (Shannon, 1948). The channel capacity depends on both $H(\boldsymbol{m})$ and the uncertainty about the intended signal, given the received signal. Adding redundancy, such as pronouncing words more slowly, that anticipates likely noise is called *channel coding*, and makes signals robust but still short.

For our purposes, the crucial observation is that it is not possible to get arbitrarily close to the channel capacity if it is not possible to obtain a source code that is arbitrarily close to the entropy of the distribution over messages. The next section shows that, for compositional codes like natural language, optimal source coding is not possible.[2]

Compositional codes and optimality

This section shows that optimal source coding is not possible for a compositional code like natural language. If a code is optimal and compositional, then it follows that the components of every message are statistically independent. However, this is not true for natural language, since, e.g., transitive verbs tend to appear with at least two noun phrases.

By 'compositional,' I mean only that natural language messages consist of components that are realized the same way across different messages, and that the length of the signal for a message is the total length of the signal for each component of that message. Setting l_m to be the length of the signal for message m and l_{m_i} to be the length of the signal for component m_i, compositionality provides:

$$l_m = \sum_{i=1}^{|m|} l_{m_i} \tag{2}$$

For example, if a message is a sequence of lexical entries, and a signal is a sequence of phones, then Equation 2 says that the length of a sentence in phones is the sum of the lengths

[1] All the results follow straightforwardly for structured messages as long as there is a deterministic linearization, such as reverse Polish notation for tree-structured messages.

[2] I here consider only discrete signals and messages. The continuous case requires either a limit on the power of the signal or for source and channel coding to be considered simultaneously.

of the phonological forms of the lexical entries in that sentence.[3] Equation 2 is not trivial. Arithmetic codes, for example, encode each message as a number between 0 and 1 that is determined by the conditional probability of each message character given the previous message characters; an individual message character is not directly expressed in any part of the signal, and l_{m_i} is not even defined.

Now assume that the length of the sentence l_m for each message m is optimal. Because the optimal signal length for a message m is $-\log_b(P(m))$, the probability distribution over messages $P_{\boldsymbol{m}}$ can be recovered from l_m by exponentiating:

$$P_{\boldsymbol{m}}(m) = b^{-l_m} \tag{3}$$

Since only sentence lengths are assumed to be optimal, component signal lengths l_{m_i} may not be negative log probabilities. They do, however, assume an *implicit* distribution for which they are least sub-optimal (MacKay, 2003, Ch. 5):

$$Q_m(m_i) = \frac{b^{-l_{m_i}}}{z} \;\; ; \;\; z = \sum_{m_i \in \mathcal{M}} b^{-l_{m_i}} \tag{4}$$

Equations 2, 3, and 4 imply that each message component is statistically independent:

$$P_{\boldsymbol{m}}(m) = b^{-l_m} = b^{\sum_{i=1}^{|m|} -l_{m_i}} = b^{\sum_{i=1}^{|m|} \log_b(zQ_m(m_i))} \tag{5}$$

$$= z^{|m|} \prod_{i=1}^{|m|} Q_m(m_i) \overset{def}{=} Q_{\boldsymbol{m}}(m) \tag{6}$$

There does not seem to be any notion of message in natural language that allows for statistically independent message components. For example, messages may be high-level event representations, but such messages that include transfer tend to include at least three entities (a giver, a receiver, and a thing being transferred), and such messages that include edible entities tend to include entities that can eat. Alternatively, messages may be syntactic analyses, but such messages with a determiner tend to have at least one noun, and such messages with a complementizer tend to have at least two main verbs. While other framings are possible, they do not appear to satisfy the independence assumption above. Thus, natural language is not information-theoretically optimal.

More specifically, the average signal length of the best compositional source code must exceed the entropy of the true distribution over messages $P_{\boldsymbol{m}}$ by at least the Kullback-Leibler divergence of the fully factorized distribution $Q_{\boldsymbol{m}}$ from the true distribution:

$$H(P_{\boldsymbol{m}}) + \mathrm{KL}(P_{\boldsymbol{m}} || Q_{\boldsymbol{m}}) \tag{7}$$

Intuitively, language uses at least an extra $\mathrm{KL}(P_{\boldsymbol{m}}||Q_{\boldsymbol{m}})$ signal characters per message character because it incorrectly assumes the message characters are statistically independent.

Optimizing towards the new bound

While the bound in Equation 7 shows that natural language does not approach the channel capacity, natural languages may still adapt over time for communicative efficiency towards the less efficient bound. In fact, the findings of Piantadosi et al. (2011) suggest that languages adapt to minimize $\mathrm{KL}(P_{\boldsymbol{m}}||Q_{\boldsymbol{m}})$. Piantadosi et al. examined how a word's length (in orthographic letters) relates to its unigram probability and its probability in context (operationalized as a smoothed trigram model). Across all eleven languages they examined, word lengths had a stronger relationship with average probability in context than unigram probability.

I propose the following interpretation of their result. Message characters are lexical entries, signal characters are orthographic letters, and probability in context is $P_{\boldsymbol{m}}$. While $Q_{\boldsymbol{m}}$ is determined by word lengths, $P_{\boldsymbol{m}}$ is determined by a stochastic grammar and lexicon together with typical real-world situations. Their results suggest that, as a speech community gains experience with the use of lexical entries in real-world situations, the grammar, including the lexicon, adapts so that $P_{\boldsymbol{m}}$ is better approximated by $Q_{\boldsymbol{m}}$. This adaptation could be achieved by adjusting the grammar, narrowing or broadening word meanings, or deleting lexical entries whose length often differs substantially from their optimal in-context length.

The next two sections present corpus studies that look at adaptation of this sort over centuries in three languages.

Corpus studies

I now present two corpus studies that find evidence of optimization relative to probability in context over time for English, Spanish, and Mandarin Chinese. The first 'backwards-looking' study relates a word's mismatch with its optimal length to its age. If the lexicon evolves over time for efficient communication, the lengths of oldest words should most closely match their optimal lengths. Moreover, to the extent that efficiency pressures respect sentence length, there should be a stronger relationship between a word's age and its mismatch with optimal lengths under a trigram model than between a word's age and its mismatch under a unigram model.

The 'forwards-looking' study uses a sequence of language models, estimated in 25-year partitions, to predict whether a word appears in the next partition based on how well its length matches its optimal length under each language model. If language change reflects efficiency pressures, words with many extra characters should be less likely to remain in use; and if efficiency pressures respect sentence lengths, the effect of mismatches under trigram models should be stronger.

Corpus study 1 – Looking backwards

In this study, I used a large dataset containing texts from about 1990 to about 2010 for each of the three languages to compute synchronic unigram and trigram language models for each language. The language models are used to compute optimal lengths for each word in and out of context by subtracting the optimal lengths from the actual lengths to quantify extra characters. I used Google books, a dataset of scanned books, to

[3] Composition operations that involve copying, such as *Suffixaufnahme* in Old Georgian (Michaelis & Kracht, 1996), present an interesting wrinkle. If they can be handled by introducing an integer coefficient for each l_{m_i}, the ultimate independence result of this section still holds. In any case, they make the signal longer, so they should not present a more efficient bound than Equation 7.

Table 1: Study 1 dataset sizes

	Language Model	Regression	
Dataset	Tokens	Unigrams	Trigrams
English	71,531,906	81,742	24,965,851
Spanish	279,744,284	545,708	74,144,973
Chinese	26,800,660	3,182	13,642,166

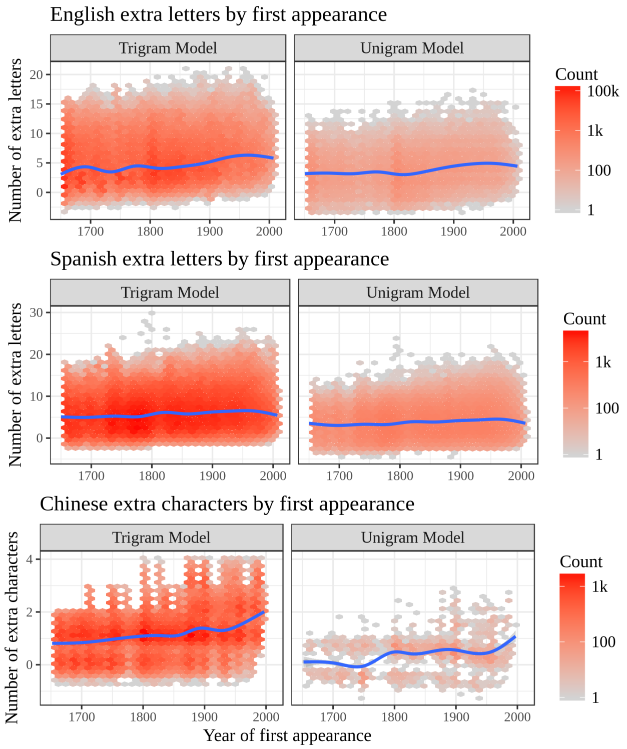

Figure 1: Heatmaps of actual length minus optimal word length for trigram (left) and unigram (right) models, as a function of the word's earliest appearance in Google books. The blue line is a GAM fit.

estimate when each word of the synchronic time slice first appeared, and perform a regression of the extra character measure against year of first appearance, probability model type, and their interaction, to identify how word length inefficiency varies as a function of word recency and probability model type. A positive coefficient for year of appearance will indicate that more recent words are longer than they should be, and a negative coefficient for the interaction will indicate a stronger relationship between year of appearance and inefficiency relative to a trigram model than between year of appearance and inefficiency in isolation.

Data I approximated a word's year of first appearance as the first year that it appeared in the Google Books unigram records in each language (Michel et al., 2010).

To estimate the American English language models, I used the spoken portion of the Corpus of Contemporary American English (CoCA) (Davies, 2008), which contains news broadcasts from 1990 to 2012. To estimate the Spanish and Chinese language models, I used 'story' documents from the 3rd edition of the Spanish Gigaword dataset of newswire text from 1993 to 2010 (Ângelo Mendonça, Jaquette, Graff, & DiPersio, 2011) and the Tagged Chinese Gigaword version 2.0 dataset of newswire text from 1991 to 2004 (Ren Huang, 2009), respectively. While written Chinese does not separate words with whitespace, this dataset is segmented into words.

For each language, I discarded punctuation and words that contained a symbol that was not part of the usual character set for that language, estimated unsmoothed trigram and unigram probabilities. The datasets for regression were obtained by discarding words that did not appear in Google Books after 1650, producing datasets with sizes as reported in Table 1.

These particular languages were chosen because they appeared in Google Books, allowing us to obtain an estimate of word age, and because they use words in very different ways. Spanish has relatively rich derivational and inflectional morphology, with agreement for person and number for verbs and number and gender for adjectives. While English also has relatively rich derivational morphology, it has little inflectional morphology with few agreement constraints. Mandarin Chinese occupies a morphological extreme, with no inflectional morphology or agreement.

Method For each word token in CoCA and Gigaword datasets, I computed the optimal length of the word under its unigram probability and probability in context, operationalized as its trigram probability. The numbers of 'extra' letters relative to each model e_{uni} and e_{tri} are then the actual length minus the optimal length:

$$e_{\text{uni}}(w) = l(w) - (-\log_b(P(w)))$$
$$e_{\text{tri}}(w_i|w_{i-2},w_{i-1}) = l(w_i) - (-\log_b(P(w_i|w_{i-2},w_{i-1})))$$

where b is the size of the signal alphabet. English and Spanish both have a mostly alphabetic orthography, with roughly one letter per sound, so I simply set b to the number of distinct letters in these datasets. For English $b = 27$ (a-z plus hyphen), and for Spanish, $b = 33$ (with some additional accented letters). Chinese orthography has one character per syllable, and so similarly provides a good indication of word length, but the alphabet size is more complicated. The strict phonotactics of spoken Chinese lead to a syllabic inventory of about $1,500$ syllables, but our Chinese dataset contained $6,780$ distinct characters (many characters are homophonous). I set $b = 1,518$, the number of distinct syllables in the CEDict pronouncing dictionary, to reflect the size of the 'syllable alphabet' for spoken Chinese (*CC-CEDICT*, 2016).[4]

I performed linear regressions of extra letters against the word's year of first appearance, probability model type, and an interaction between the word's first appearance and probability model type. To make the regressions easier to interpret, I subtracted 1650 from the year of first appearance, so that the oldest words had a year of first appearance of zero.

[4] I obtained similar results when using $b = 6,780$, the number of distinct characters.

Table 2: Coefficients of one linear regression each for American English, Mandarin, and Spanish, of extra letters (English, Spanish) or extra characters (Mandarin) against first appearance, a main effect of probability model type (with unigram coded as 1), and an interaction between first appearance and probability model. All coefficients are significant ($p < 0.01$).

	Am. English	Spanish	Mandarin
Intercept	3.289	4.408	1.067
Year of first appearance (since 1650)	0.006481	0.007399	0.00214
Which language model	-0.491	-2.506	-1.288
Years of first appearance (since 1650) × Which language model	-0.001465	-0.001438	-0.000349

Results Figure 1 presents hexagram-binned heatmaps with a Generalized Additive Model fit for each language of extra letters against year of first appearance, separated by language model. All cases show a broad trend where older words have fewer extra letters. The trend is roughly linear except for the latest decades; information-theoretic pressures may be different for recently-coined words.

Table 2 presents coefficients from the three regressions of extra characters against a word's year of first appearance, the probability model used, and their interaction. Each intercept expresses the number of predicted extra letters or characters under the trigram model for words that first appeared in 1650. Adding the 'Which Language Model' coefficient to the intercept obtains the predicted extra letters or characters under the unigram model for words that first appeared in 1650.

The 'Year of first appearance' coefficient expresses how many extra characters we expect a word to have for each year that it is younger than the oldest words. For all three languages, this coefficient is positive, indicating that younger words tend to be longer, than older words, relative to their ideal length under the trigram model. Dividing this coefficient into 1 obtains how old we expect a word to be before an additional letter or character has been 'optimized' away. American English optimizes one letter every 154 years, Spanish optimizes one letter every 135 years, and Mandarin Chinese optimizes one character every 467 years.[5]

Finally, the interaction between year of first appearance and model type expresses the effect of a word's year of appearance under the unigram model minus the effect of a word's year of appearance under the trigram model. The coefficients are negative but smaller in absolute magnitude than 'Year of first appearance,' which indicates that first appearance still has a lengthening effect relative to the unigram model, but a *weaker* one. American English optimizes an extra letter relative to the unigram model only every 199 years, Spanish optimizes an extra letter only every 168 years, and Mandarin optimizes an extra character only every 558 years.

These results show that words that first appeared in books recently tend to be further from their information-theoretically optimal lengths than words that first appeared in books several decades ago, and so provide evidence of optimization of the lexicon towards efficiency bounds.

Moreover, the extra characters relative to the trigram model decreased faster than the extra characters relative to the unigram model. This is a remarkable finding, since it is much harder to optimize for the trigram model – there are many trigram contexts but only one unigram 'context,' and, under this operationalization of 'word,' a word has only one length. However, as previously discussed, there are good reasons to optimize towards a context-sensitive probability model. Communicative efficiency ultimately depends on sentence lengths, not word lengths directly, so considering context can make sentences shorter even if it does not minimize the typical length of individual words.

Corpus study 2 – Looking forwards

This corpus study looks for evidence that a word is less likely to remain in use if it has more extra characters. For each language, I divided the 350 years of Google Books data described above into 14 partitions of 25 years each, and estimated a unigram and a trigram language model for each of the first 13 partitions to compute extra characters for each word and trigram in each partition under each probability model. To guard against OCR errors in Google Books, I computed extra characters only for words that also appeared in the language model datasets from Study 1. I then performed a logistic regression that predicted whether each word that appeared in partition n also appeared in partition $n+1$ using the extra characters measure, probability model type, and an interaction between extra characters and the probability model type.

Results Table 3 presents strikingly consistent regression results across the three languages. The large intercepts indicate that most words carry over from one partition to the next. As the unigram model is again coded as 1, the negative main effect of extra characters indicates that words with more extra letters relative to a given partition's trigram model are less likely to persist in the next 25-year partition. Moreover, the positive coefficient of the interaction indicates that the effect of extra letters relative to the unigram model is weaker: the coefficients suggest the effect of unigram mismatch is about half the effect of trigram mismatch in English, two-thirds in Spanish, and about one-third in Mandarin.

Conclusion

This paper has answered an important question about natural language communication, whether talkers approach information-theoretic limits on efficiency, in the negative. Be-

[5] As the CEDict pronouncing dictionary has an average length of about 3.1 non-tone pinyin letters, or 2.8 phonemes, per character type, the optimization rate of Mandarin is similar to the others.

Table 3: Coefficients of a logistic regression each for American English, Mandarin, and Spanish, of appearance in the next 25-year partition against extra letters or characters, a main effect of probability model type (unigram coded as 1), and an interaction between extra characters and probability model. All coefficients are significant ($p < 0.01$).

	Am. English	Spanish	Mandarin
Intercept	11.772	7.477	10.415
Extra letters or characters	-0.316	-0.320	-2.360
Which language model	-1.015	-0.852	-2.415
Extra characters × Which language model	0.165	0.101	1.713

cause language is compositional and natural language messages are highly interdependent, natural language cannot approach information-theoretic limits on efficiency. I have used this result to propose a new bound that appreciates probability in context, and interpreted a previous result as evidence that languages optimize for this more appropriate bound.

I then performed two corpus studies that examined how the mismatch between a word's actual length and its optimal length relates to its preservation over the course of language change. The first 'backwards-looking' study found, using optimal lengths computed using fairly homogenous modern-day corpus data, that present-day words more closely match their optimal lengths if the word has been in use for a long time. Moreover, this first study found that the mismatch according to probability in context decreased more rapidly as words age. The second 'forwards-looking' study found that if a word's length more closely matches its optimal length under a language model computed in one 25-year partition, it is more likely to be retained in the next 25-year partition. Moreover, extra letters relative to probability in context was a stronger predictor than extra letters relative to a unigram model. Together, these results indicate that natural language lexicons develop over time towards an information-theoretic efficiency bound that is appropriate for compositional codes.

Acknowledgements

I thank Rui Chaves for feedback on a very early draft, and anonymous reviewers for helpful comments.

References

Ângelo Mendonça, Jaquette, D., Graff, D., & DiPersio, D. (2011). Spanish Gigaword Third Edition LDC2011T12 [Computer software manual]. Web download. Philadelphia, PA.

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.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*, 92–111.

CC-CEDICT. (2016). `http://www.mdbg.net/chindict/`. (Accessed: 2016 - 30 - 05)

Davies, M. (2008). *The Corpus of Contemporary American English: 520 million words, 1990 – present.* (Available online at `http://corpus.byu.edu/coca/`)

Frank, A., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of CogSci* (pp. 933–938).

Genzel, D., & Charniak, E. (2002). Variation of entropy and parse trees as a function of sentence number. In *Proceedings of the Association for Computational Linguistics.*

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*, 23–62.

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In *Proceedings of NIPS.*

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

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms.* Cambridge University Press. (`http://www.inference.phy.cam.ac.uk/mackay/itila/`)

Michaelis, J., & Kracht, M. (1996). Semilinearity as a syntactic invariant. In *Proceedings of LACL.*

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Brockman, W., … Aiden, E. L. (2010). Quantitative analysis of culture using millions of digitized books. *Science*, *331*.

Piantadosi, S., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526.

Plotkin, J. B., & Nowak, M. A. (2000). Language evolution and information theory. *Journal of Theoretical Biology*, *205*, 147–159.

Ren Huang, C. (2009). Tagged Chinese Gigaword Version 2.0 LDC2009T14 [Computer software manual]. Web download. Philadelphia, PA.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, *133*(1), 140–155.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423.

Zipf, G. K. (1949). *Human behavior and the principle of least effort.* Addison-Wesley.