Speakers communicate using language-specific information distributions

Anonymous CogSci submission

Abstract

What role does communicative efficiency play in how we organize our utterances? In this paper, we present a novel method of examining how much information speakers in a given language communicate in each word, surveying numerous diverse languages. We find that speakers produce frequent and informative words at regular parts of their utterances, depending on language they use. The information distribution for each language is derived in part from the features and genealogy of the language. This robust information distribution characterizes both spoken and written communication, and emerges in children's earliest utterances. However, in real-time communication, in-context word predictability allows listeners to process information at a constant, optimal rate, regardless of the information distribution in the language they understand.

Keywords: information theory; communication; language modeling; computational modeling

Introduction

We use language for a variety of purposes like greeting friends, making records, and signaling group identity. But, these purposes all share a common goal: Transmitting information that changes the mental state of the listerner or reader (Austin, 1975). For this reason, language can be thought of as a code, one that allows speakers to turn their intended meaning into a message that can be transmitted to a listener or reader, and subsequently converted by the listener back into an approximation of the intended meaning (Shannon, 1948). How should we expect this code to be structured?

If language has evolved to be a code for information transmission, its structure should reflect this process of optimization (Anderson & Milson, 1989). The optimal code would have to work with two competing pressures: (1) For listeners to easily and successfully decode messages sent by the speaker, and (2) For speakers to easily code their messages and transmit them with minimal effort and error. A fundamental constraint on both of these processes is the linear order of spoken language–sounds are produced one at a time and each is unavailable perceptually once it is no longer being produced.

Humans accomodate this linear order constraint through incremental processing. People process speech continuously as it arrives, predicting upcoming words and building expectations about the likely meaning of utterances in real-time rather than at their conclusion (Kutas & Federmeier, 2011; Pickering & Garrod, 2013; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). Since prediction errors can lead to severe processing costs and difficulty integrating new information on the part of listeners, speakers should seek to minimize prediction errors. However, the cost of producing more predictable utterances is using more words. Thus, the optimal strategy for speakers seeking to minimize their production costs is to produce utterances that are just at the prediction capacity of listeners without exceeding this capacity (Aylett & Turk, 2004; Genzel & Charniak, 2002). In other words, speakers should maintain a constant rate information of as close to the listener's fastest decoding rate as possible.

This Uniform Information Density hypothesis has found support at a variety of levels of language from the structure of individual words, to the syntactic structure of utterances (Jaeger & Levy, 2007; Piantadosi, Tily, & Gibson, 2011; see Gibson et al., 2019 for a review). Further, speakers make lexical choices that smooth out the information in their utterances (Jaeger & Levy, 2007; Mahowald, Fedorenko, Piantadosi, & Gibson, 2013). However, while speakers can control which of several near-synonyms they produce, or whether to produce an optional complementizer like "that," they cannot control the grammatical properties of their native language like canonical word order that impose top-down constraints on the structure of utterances. While speakers may produce utterances as uniform in information density as their languages will allow, these top-down constraints may impose significant variation.

How significant are these top-down constraints? One previous paper analyzed the information content in English sentences and found a surprising three-step shape where information first rises, then plateus, and then sharply rises again at the ends of sentences (Yu, Cong, Liang, & Liu, 2016). We build on these ideas, asking (1) Whether this shape depends on listener's predictive models, (2) Whether this shape varies across linguistic contexts, and (3) Whether this shape is broadly characteristic of a diverse set of languages or varies predictably from language to language. We find that languages are characterized by highly-reliable but cross-linguisticly variable structures that co-vary with typological features, but that predictive coding flattens these shapes across languages, in accord with predictions of the Uniform Information Density hypothesis.

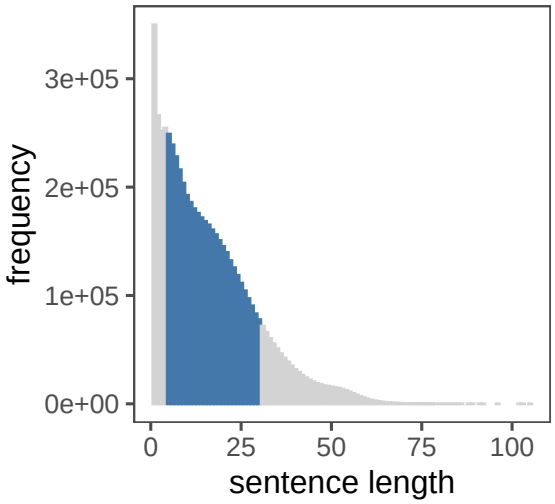

Figure 1: Sentence length distribution in the British National Corpus. Lengths included in analysis are dark.

Study 1: Information in Written English

Describe the Genzel Idea and how to Generalize it to within-sentence Yu did a thing but it doesn't generalize, we do it with surprisal Using their word entropy metric without context, Yu et al. (2016) found a distinctive three-step distribution for information in written English sentences in the corpus. The first word tended to contain little information. While the middle words of sentences each had more information than the first word, they found a flat and non-increasing rate of information transmission across the middle of sentences. The final word contained the most, though not most, of the information out of any in the sentence, with a noticeable spike in information. They found the same distribution across sentence lengths, from sentences with 15 words to sentences with 45 words.

Corpus

Following Yu et al. (2016), we selected the British National Corpus (BNC) for analysis (British National Corpus Consortium, 2007). The BNC is a collection of spoken (10%) and written (90%) from the turn of the century, intended to be a representative sample of British English.

Pre-processing

We began the XML version of the corpus, and used the `justTheWords.xsl` script provided along with the corpus to produce a text file with one sentence of the corpus on each line. Compound words (like "can't") were combined, and all words were converted to lowercase before analysis. This produced a corpus of just over six million utterance of varying lenghts. From these, we excluded utterances that were too short to allow for reasonable estimation of information shape (fewer than 5 words), and utterances that were unusually long (more than 30 words). This exclusion left us with 67.64% of the utterances (Fig @ref{fig:bnc-lengths}.

Estimating information

To estimate how information is distributed across utterances, we computed the lexical surprisal of each word under two different models (Levy, 2008; Shannon, 1948). Intuitively, the surprisal of a word is a measure of how unexpected it would be to read that word, and thus how much information it contains. First, following Yu et al. (2016), we estimated a unigram model which considers each word independently, asking how unexpected that word would be in the absence of any context: surprisal(word) $= -\log P(\text{word})$. This unigram surprisal measure is a direct transformation of the word's frequency and thus less frequent words are more surprising.

Second, we estimated a trigram model in which the surprisal of a given word (w_i) encodes how unexpected it is to read it after reading the prior two words (w_{i-1} and w_{i-2}): surprisal$(w_i) = -logP(w_i|w_{i-1},w_{i-2})$. This metric encodes the idea that words that are low frequency in isolation (e.g. "meatballs") may become much less surprising in certain contexts (e.g. "spaghetti and meatballs") but more surprising in others (e.g. "coffee with meatballs"). In principle, we would like to encode the surprisal of a word given all of the prior sentential context (surprisal$(w_i) = -P(w_i|w_{i-1}w_{i-2}...w_1)$. However, the difficulty of correctly estimating these probabilities from a corpus grow combinatorically with the number of prior words, and in practice trigram models perform well as an approximation (see e.g. Chen & Goodman, 1999; Smith & Levy, 2013).

Model details We estimated the surprisal for each word type in the British National Corpus using the KenLM toolkit (Heafield, Pouzyrevsky, Clark, & Koehn, 2013). Each utterance was padded with a special start-of-sentence word "$\langle s \rangle$" and end of sentence word "$\langle /s \rangle$". Trigram estimates did not cross sentence boundaries, so for example the surprisal of the second word in an utterances was estimated as (surprisal$(w_2) = -P(w_2|w_i, \langle s \rangle)$.

Naïve trigram models will underestimate the surprisal of words in low-frequency trigrams (e.g. if the word "meatballs" appears only once in the corpus following exactly the words "spaghetti and", it is perfectly predictable from its prior two words). To avoid this underestimation, we used modified Kneser-Ney smoothing which discounts all ngram frequency counts–reducing the impact of rare ngrams on probability calculations–and interpolates lower-order ngrams into the calcuations. These lower-order ngrams are weighted according to the number of distinct contexts they occur as a continuation (e.g. "Francisco" may be a common word in a corpus, but likely only occurs after "San" as in "San Francisco", so it receives a lower weighting; see Chen & Goodman, 1999).

Averaging curves To aggregate across utterance lengths, we used dynamic time warping barycenter averaging (Petitjean, Ketterlin, & Gançarski, 2011). The dynamic time warping (DTW) algorithm (Sakoe & Chiba, 1978) compares time sequential data, first used for speech recognition to unite

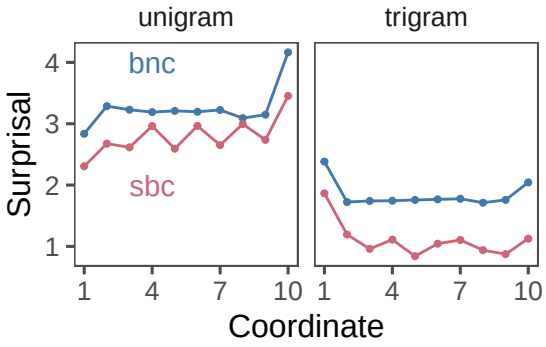

Figure 2: BNC and Switchboard frequency-based trigram curves

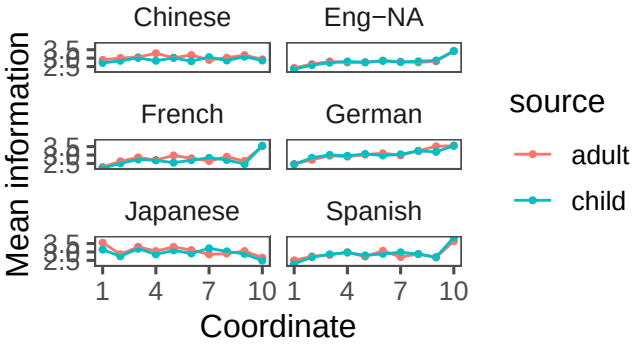

Figure 3: CHILDES frequency-based trigram curves

stretched and shifted sound patterns. This algorithm finds a prototypical time series given variable length series. Something about length of barycenters. **Talk about choosing size based on something about utterance length distribution Maybe we even want to show this distribution?**

Results

We replicate the Yu et al. (2016) result using the surprisal metric in place of the entropy metric. We use the frequency-based or "contextless" surprisal metric, which determines the average distribution of information based on word frequencies in a corpus. A priori we expect that the frequency-based metric will produce a flat distribution of information across word positions in the BNC. We find the same frequency-based information trajectory as Yu et al. with little information in the first words of utterances and the most information in the final word, see Figure @ref(fig:bncunigrams).

We now include two words of context (trigrams) for each word in our measurements. We observe a flattening effect of context for both spoken and written English. After the first word or two, where the listener does not have access to prior context, then they decode information at a flat and more or less uniform rate. The contextual information curves for the BNC and Switchboard are in Figure @ref(fig:bnctrigrams). We also computed bigram curves with one word of context for each prediction: these bigram curves resemble the trigram curves.

Experiment 2: English in Other contexts

Spoken adult-adult Spoken adult-child Spoken child

We now turn to developmental data, and show that frequency-based information curves characterize child speech from the time a child first begins speaking as well as adult speech, regardless of utterance length.

We compare this result to the one for the Santa Barbara corpus of spoken American English (Du Bois, Chafe, Meyer, Thompson, & Martey, 2000).

We're going to examine child and child-directed speech from CHILDES (MacWhinney, 2000) to capture the developmental picture. We use corpora for Spanish, German, French, Mandarin Chinese and Japanese as well as English. Mandarin and Japanese are not natively written using the Latin alphabet, and moreover words are not segmented in their native scripts. Instead of the native scripts, we use transliterations from the corpus for each of the Mandarin and Japanese utterances into pinyin for Mandarin and romanji for Japanese. In these transliterations, words are previously segmented.

We observe a distinct and characteristic frequency-based information trajectory for each language, robust across each utterance length. We see the same distribution of information for both parents and children. The parent often has more information on average at each word position in their utterances. This is an effect of the surprisal metric: parents speak more utterances than their children in most of the corpora, which inflates the number of tokens they use and increases the surprisal of hearing a rare word. We include the frequency-based information curve from the North American English CHILDES collection for comparison. See Figure @ref(fig:childesunigrams)

For the trigram information curves, we see the same contextual smoothing effect as in English. While the frequency-based information curves may depend based on the language, the contextual information curves show the same trajectory cross-linguistically. Using more than two words of context is difficult for parent-child speech corpora because the utterances are so short on average (less than 10 words). Based on our results from the CHILDES collections, we hypothesize that the frequency-based information curves may vary based on the genealogy and typology of the languages in question. However, this does not extend to the information curves with two words of context in particular, where all languages we have seen so far are characterized by the same information distribution. See Figure @ref(fig:childestrigrams).

To make a claim about how languages on a larger scale, we need to use larger corpora and a much larger number of languages.

Experiment 3: Large-scale data and linguistic features

We pulled corpora for 159 diverse languages from Wikipedia, each of which had at least 10,000 articles. To compare languages more rigorously, we used two databases of language

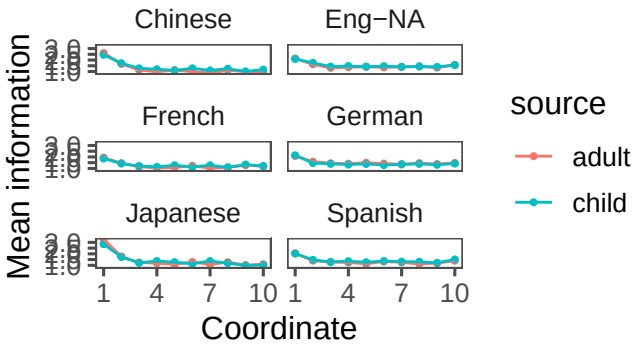

Figure 4: CHILDES context-based trigram curves

similarity features. To target lexical differences between languages, we used the 40-item Swadesh list (Swadesh, 1955), retrieved from the ASJP database (Wichmann et al., 2016). The Swadesh list is a well-known method for comparing lexical similarity between languages, by quantifying the similarity between the words on the list for pairs of languages, and is often used to compare genetic relationships between languages. We computed the average normalized Levenshtein distance, a string edit distance measure (LDN; Holman et al., 2008) between each pair of our Wikipedia languages.

As our surprisal metric is a lexical measure, we expect the Levenshtein distance to be high. To describe more structural relationships, we used the World Atlas of Language Structures (WALS; Dryer & Haspelmath, 2013) to describe the morphology, syntax, phonology, etymology and semantics–in short the structures in each language. As WALS is a compiled database from dozens of papers from different authors, most of the features and languages are fairly sparse. We used a iterative imputation algorithm for categorical data Multiple Imputation Multiple Correspondence Analysis (MIMCA; Audigier, Husson, & Josse, 2017) to fill in the missing features.

Conclusion

In this paper we did model and it showed unique distributions for unigrams and same distribution for trigrams. Developmental angle.

Follow-up. Possible questions one might have.

Acknowledgements

Place acknowledgments (including funding information) in a section at the end of the paper.

References

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, *96*(4), 703.

Audigier, V., Husson, F., & Josse, J. (2017). MIMCA: Multiple imputation for categorical variables with multiple correspondence analysis. *Statistics and Computing*, *27*(2), 501–518.

Austin, J. L. (1975). *How to do things with words*. Oxford university press.

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.

British National Corpus Consortium. (2007). *British national corpus version 3 (BNC XML edition)*. Oxford: Oxford University Computing Services. Retrieved from URL: http://www.natcorp.ox.ac.uk/

Chen, S. F., & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, *13*(4), 359–394.

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *WALS online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from `https://wals.info/`

Du Bois, J. W., Chafe, W. L., Meyer, C., Thompson, S. A., & Martey, N. (2000). Santa barbara corpus of spoken american english. *CD-ROM. Philadelphia: Linguistic Data Consortium*.

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 199–206).

Gibson, E., Futrell, R., Piandadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*.

Heafield, K., Pouzyrevsky, I., Clark, J. H., & Koehn, P. (2013). Scalable modified kneser-ney language model estimation. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 690–696).

Holman, E. W., Wichmann, S., Brown, C. H., Velupillai, V., Müller, A., Bakker, D., & others. (2008). Advances in automated language classification. *Quantitative Investigations in Theoretical Linguistics*, 40–43.

Jaeger, T. F., & Levy, R. P. (2007). Speakers optimize information density through syntactic reduction. In *Advances in neural information processing systems* (pp. 849–856).

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the n400 component of the event-related brain potential (erp). *Annual Review of Psychology*, *62*, 621–647.

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

MacWhinney, B. (2000). *The childes project: The database* (Vol. 2). Psychology Press.

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313–318.

Petitjean, F., Ketterlin, A., & Gançarski, P. (2011). A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, *44*(3), 678–693.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.

Pickering, M. J., & Garrod, S. (2013). An integrated the-

ory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(4), 329–347.

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *26*(1), 43–49.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, *21*(2), 121–137.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Wichmann, S., Müller, A., Wett, A., Velupillai, V., Bischoffberger, J., Brown, C. H., ... others. (2016). The asjp database. *Max Planck Institute for the Science of Human History, Jena*.

Yu, S., Cong, J., Liang, J., & Liu, H. (2016). The distribution of information content in english sentences. *arXiv Preprint arXiv:1609.07681*.