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 can use language for a number of diverse yet useful purposes, such as greeting friends, making records and signaling group identity. All of these tasks share a common unifying purpose: changing the mental state of the listener or reader through the information we transmit to them (Austin, 1975). Language can naturally 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).

Beyond its utility as a metaphor, this coding perspective on language is powerful as a framework for rational analysis. 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.

Listeners use a strategic solution which allows them to interpret words in rapid succession: *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 transmission of information, with the optimal rate of information transfer as close to the listener's fastest decoding rate as possible.

Previous research has shown evidence for optimal coding at the word (Piantadosi, Tily, & Gibson, 2011) and phrasal (Jaeger & Levy, 2007) levels, among others. Genzel & Charniak (2002) found a specific information trajectory at the level of sentences in paragraphs: a constantly increasing and nearly linear rate of information transmission. Speakers have cognitive control over the speed at which they say a word and whether or not they insert complementizers. But speakers do not have control over the broad properties of their language such as canonical word order that affect the organization of utterances top-down. Does this whole-utterance level follow the predictions of optimal coding? The utterance level may show strong effects of variation between languages, as specific languages have properties that constrain how speakers may form utterances in those languages, such as canonical word order. These properties vary widely from language to language.

Previous work (Yu, Cong, Liang, & Liu, 2016) studied this within-sentence level only in written English, finding an unexpected three-step distribution regardless of sentence length: little information in the first words of sentences and the most information in the final word. The the distribution found was robustly different from the linearly increasing trend in sentences from Genzel & Charniak (2002), and also did not resemble the uniform distribution of information that one might expect from a communicative efficiency account, in which each word has approximately equal information close to the channel capacity.

In this paper, we propose a novel method for modeling prototypical information structure in a language. We compute a unique information distributions for each language arising

from word frequency. We find that this information distribution characterizes all data within a given language, both spoken and written, from the time child speakers pronounce their first utterances. We link the information curves within language families to their structures such as morphology and syntax, such that the information curve can be partially predicted by the features of the language. When considering communication and predictive processing, listeners decode language at the same constant and optimal rate regardless of language. Previous uniform information density work has focused on English, but we analyze well over a hundred diverse languages from across the world and from a variety of language families.

Methods

We build a statistical model of a fluent listener: how much information do you tend to decode at each part of an utterance in a language? We begin by approximate language experience and linguistic knowledge by training on a corpus. We test on the same corpus to approximate the listener's real-time experience, avoiding overfitting by smoothing. We compute the prototypical information distribution over all sentences in the corpus.

Surprisal Shannon (1948) defined information as "the reduction in uncertainty about one variable given that the value of another variable is known". The *lexical surprisal* (Levy, 2008) metric applies Shannon's definition of information to words. This measure defines the information in a word as the predictability of the word based on previously heard or seen words in its context, as in the formula below. The surprisal of a word is inversely proportional to the predictability of that word, such that less common and less predictable words carry more information. Frequency is intimately tied rto information content in words, with much of the differences between word frequencies being explained by information content cross-linguistically (Piantadosi et al., 2011).

$$\text{surprisal}(\text{word}) = -\log P(\text{word})$$

Incorporating context However, when reading or listening, people don't just consider each word as an isolated linguistic signal. Instead, listeners use the words they have already heard to predict and decode the word they are currently hearing. Following this incremental processing paradigm, we can also condition the surprisal of a word in its context. In the formula below, w_i denotes the word currently being read or heard, while w_{i-1} denotes the first word before the current word, w_{i-2} denotes the second word before the current word, and so on.

$$\text{surprisal}(w_i|w_{i-1}w_{i-2}...) = -logP(w_i|w_{i-1}w_{i-2}...)$$

When we use a word or two of context in our surprisal calculations, then the set of reasonable final items in our ngrams is greatly restricted. For example, in the sentence "I take my coffee with cream and sugar", when hearing "cream and", a listener might automatically predict "sugar", but there are few possible continuations with even the two words "cream and". This is linked to the N400 component in neurological processing (Kutas & Federmeier, 2011).

Ideally, we would like to measure the predictability of each word in an utterance using all of the information available to that word. For example, in an utterance of twenty words, we would like to use the previous 19 words of context to predict the 20th word. However, we would need to train on a corpus of many trillion word tokens to predict with this amount of context. Regardless of computational constraints, we want to directly compare how predictable each word is regardless of its position in an utterance. We therefore use a simplifying *Markov assumption*: we condition our next predictions on a fixed-size context window instead of all preceding words.

$$\text{surprisal}(w_i|w_{i-1}w_{i-2}...) \approx \text{surprisal}(w_i|w_{i-1}w_{i-2})$$

We train two types of ngram language models independently on a corpus. One of our models is frequency-based: we do not incorporate context into our surprisal calculations. To incorporate context into our models, we train bigram and trigram language models, which incorporate one and two words of context for each processed word, respectively. Although these models may seem to use an inconsequential amount of context when predicting the next word, bigram and trigram models introduce a great deal of improvement over unigram models across tasks (Chen & Goodman, 1999). Models which incorporate more than two words of context have issues with overfitting to the corpus and only predicting observed sequences, often generalizing poorly.

Language model smoothing Once we have fitted our language model, we can compute the surprisal of a continuation by simply taking the negative log-probability of that word's ngram probability. To find the average information for a given position in a corpus, we take all utterances of a given length, and for each word position in utterances of that length, we compute the average of the surprisals for all of the non-unique words that occur in that position, conditioned or not conditioned on context.

To avoid overfitting our models to the corpus, we use modified Kneser-Ney smoothing as implemented in the KenLM toolkit (Heafield, Pouzyrevsky, Clark, & Koehn, 2013). Briefly, this smoothing technique discounts all ngram frequency counts, which reduces 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). For a longer explanation explanation of modified Kneser-Ney smoothing and comparison to other ngram smoothing methods, see Chen & Goodman (1999).

Aggregating time sequences To aggregate across utterance lengths, we use the DTW barycenter algorithm (Petitjean, Ketterlin, & Gançarski, 2011). The dynamic time warping (DTW) algorithm (Sakoe & Chiba, 1978) compares time sequential data, first used for speech recognition to unite stretched and shifted sound patterns. This algorithm finds a prototypical time series given variable length series. Something about length of barycenters.

The flexibility of the surprisal metric we employ in this paper allows us to calculate the anticipated information for an individual utterance, as most work with the metric has done in the past. Averaging together the surprisal values for a word position within utterances is a step further than prior work, and indicates the tendencies speakers gravitate towards instead of examining individual stimuli in psycholinguistic experiments.

Experiment 1: English speech and writing

We first turn to working with written English in the British National Corpus (BNC; Leech, 1992). The BNC is a collection of spoken and written records (90% written) from the turn of the century, intended to be a representative sample of British English. 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.

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 compare this result with spoken English conversations from the Switchboard telephone conversation corpus (**???**). The spoken and written English corpora have the same information trajectory.

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).

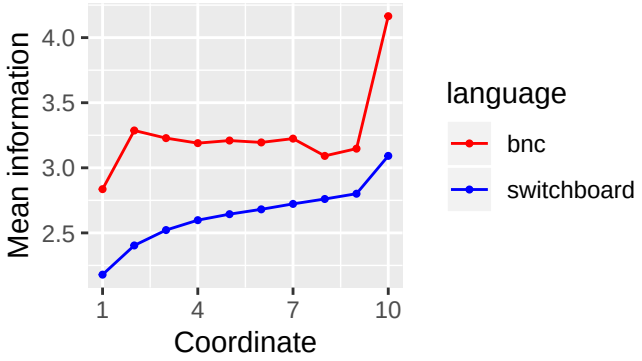

Figure 1: BNC and Switchboard frequency-based trigram curves

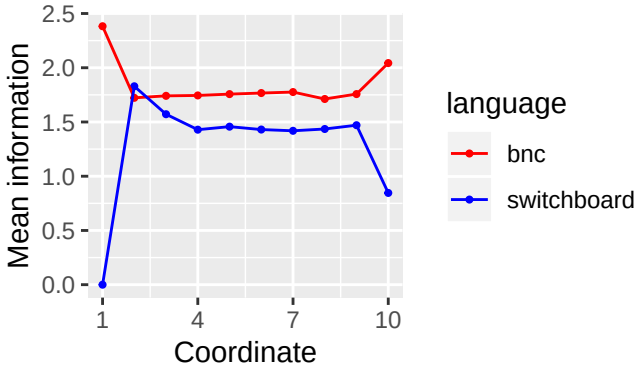

Figure 2: BNC and Switchboard context-based trigram curves

We also computed bigram curves with one word of context for each prediction: these bigram curves resemble the trigram curves.

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.

Experiment 2: Child and child-directed speech

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

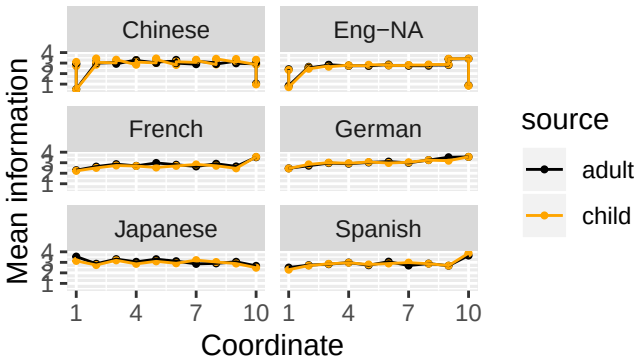

Figure 3: CHILDES frequency-based trigram curves

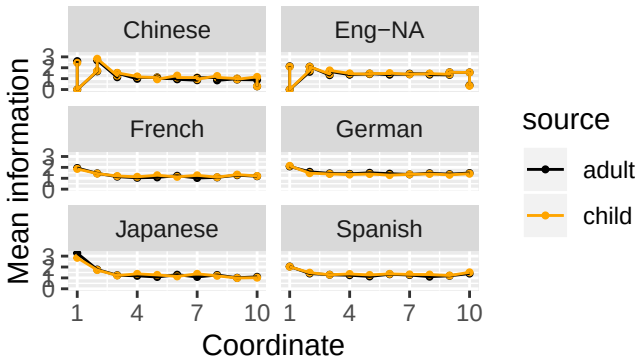

Figure 4: CHILDES context-based trigram curves

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)

English, Spanish, French and German feature similar information curve shapes, with slight variations. The German information curve features lower information for longer towards the beginnings of utterances, possibly due to the grammatical restriction that the second word in German utterances must be a verb (V2). Spanish features a larger spike in the amount of information in the final word of utterances. For Japanese and Mandarin, we observe completely different frequency-based information curve trajectories. The Japanese frequency-based information curve trajectory begins high and finishes low, the mirror image of the German and Romance language information curves. The Mandarin curve begins low and finishes low, but features high information in the middle of utterances. We hypothesize this may be due to Japanese and Mandarin speakers typically ending their utterances with particles, which are common and thus contain little information on their own.

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 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.

WALS plot goes here

Lexical plot goes here.

What did we learn from this? Unigrams vary kind of with features across languages. Trigrams are flat all-around.

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.

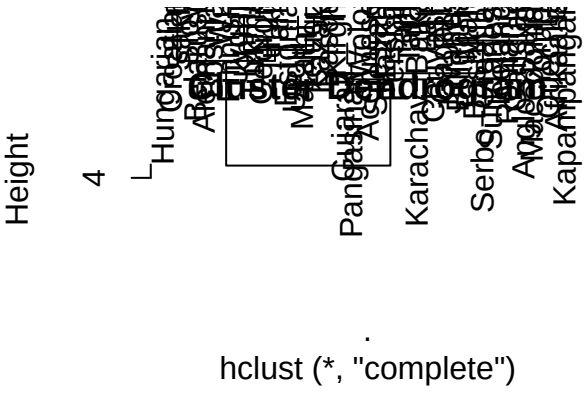

Figure 5: CHILDES context-based trigram curves

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.

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/`

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).

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.

Leech, G. N. (1992). 100 million words of english: The british national corpus (bnc).

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

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

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 theory 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.

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*.