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 different tasks such as greeting friends, taking notes and signaling group identities. All of these tasks share a common unifying purpose: changing the mental state of the listener or reader through the information we transmit (**???**). 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 (**???**).

Beyond its utility as a metaphor, this coding perspective on language is powerful qs a framework for rational analysis. If language has evolved to be an code for information transmission, it's structure should reflect this process of optimization (**???**). 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 (**???**; **???**; **???**). This solution creates new guidance for speakers: 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 **NOT SURE WHAT NEW GUIDANCE MEANS**. 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 (**???**; **???**). 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. How does this pressure affect how speakers structure individual utterances? 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.

(**???**) studied this utterance level in written English sentences using a contextless entropy model based on word frequency. **TOO JARGONY** They found a distinctive three-step distribution regardless of sentence length, with little information in the first words of sentences and the most information in the final word. This was surprising, as the distribution they found was robustly different from the linearly increasing trend in sentences from (**???**), 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.

** I THINK YOU NEED TO INTRODUCE GENZEL BEFORE YU**

WHAT WE ARE GOING TO DO IN THIS PAPER

Methods

TOO VERBOSE HERE, I THINK. ALSO, I WOULDN'T INTRODUCE ALL OF THE NGRAMING AND DTWING HERE. I BET ACTUALLY SOME KIND OF DIAGRAM REPRESENTATION OF OUR PIPELINE WOULD BE HELPFUL.

We use the surprisal metric to quantify word-level information, ngram language modeling with smoothing as a cognitive model of a speaker and dynamic time warping to aggregate across utterance lengths. Theme: quantify information and sequential predictive processing in linguistic communication.

(**???**) defined information as "the reduction in uncertainty about one variable given that the value of another variable is known". The *lexical surprisal* (**???**) 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.

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

The surprisal of a word is also correlated with the processing cost of a word, shown by evidence from e.g. eye-tracking (**???**) and ERP (**???**) studies. Considered without context, the surprisal of an individual word is inversely proportional to the frequency of that word, so that simply the less often a person has seen a word, the more information that word holds. For example, "flower" has less information than "azalea" because "flower" is much more common than "azalea". Though the two words have the same length in number of letters, we might predict that it is more difficult to process "azalea" when reading it here than when reading "flower". Frequency is intimately tied information content in words, with much of the differences between words frequencies being explained by information content cross-linguistically (Piantadosi et al., 2011).

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.

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

$$= -log\frac{P(w_i, w_{i-1}w_{i-2}, ...)}{P(w_{i-1}w_{i-2}...)}$$

I DON'T THINK YOU NEED THE SECOND PART OF THAT EQUATION 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".

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 (**???**). Models which incorporate more than two words of context have issues with overfitting to the corpus and only predicting observed sequences, often generalizing poorly.

In our contextual models, we face another issue of overfitting: we only train our model on those utterances which occur in the corpus and test our model on the same utterances. This ignores possible other utterances which the speakers could have produced, e.g. the words "I", "saw" and "bears" are in the corpus vocabulary: while "I saw bears" might not be in the corpus, that's a possible utterance for the speaker to form. To combat this issue, we use modified Kneser-Ney smoothing as implemented in the KenLM toolkit (**???**). 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 (**???**).

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.

However, we then run into an issue: we have several dozen information distributions, one for each utterance length. How can we aggregate across these distributions to find the length-

agnostic distribution for a language? Recall that we are working with sequences of surprisal, with each sequence separated at regular time points. 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. The DTW barycenter algorithm (Petitjean, Ketterlin, & Gançarski, 2011) finds a prototypical time series given variable length series.

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 actually a step further than prior work, and indicates the tendencies speakers gravitate towards instead of examining individual stimuli in psycholinguistic experiments.

The frequency-based surprisal metric gives us an idea of when in their utterances speakers say frequent i.e. independently information-rich words. The context-based surprisal metric show us how speakers tend to distribute the information in utterances relative to real-time processing in communication. We expect a priori that our frequency-based surprisal curve will be flat. No one part of the sentence will on average have words that are more frequent than another across utterance lengths. Similarly, we expect that there will be a small smoothing effect for our contextual surprisal metric such that the word in each position of an utterance is more predictable than its frequency-based counterpart.

Experiment 1: English speech and writing

We first turn to working with written English in the British National Corpus (BNC; **???**). 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, (**???**) 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 (**???**) 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:bnc-unigrams).

What about context? So far we've only discussed the frequency-based metric, considering words on their own without any explicit incorporation of prior context. As previously discussed, listeners decode information and process what they hear incrementally, using prior heard words to ease the comprehension process. We now include two words of context (trigrams) for each word in our measurements. We observe a flattening effect of context across both modalities and all speaker populations. 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 are in Figure @ref(fig:bnc-trigrams). We also computed bigram curves with one word of context for each prediction: these bigram curves resemble the trigram curves.

TRIGRAM CURVES

Experiment 2: Child and child-directed speech

So far, we have only looked at the distribution of information in words in English, both with and without context. We have examined child speech and child-directed speech at a variety of ages, as well as writing samples selected to be representative of British English as a whole. But this only captures the picture for English.

We now turn to a small number of typologically diverse languages, and conduct the same analysis, using monolingual adult-child speech corpora from CHILDES (**???**) to compare the results from these languages directly to our results from English. We use corpora for Spanish, German, French, Mandarin Chinese and Japanese. Similar to our English child speech collection, all of the language collections consist mainly of shorter utterances: most utterances in the corpora are under 10 words long. 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:childes-unigrams)

PLOT HERE FOR UNIGRAMS

English, Spanish, French and German feature similar information curve shapes, with slight variations. The German information curve features lower information for longer to-

wards 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 rigram 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:childes-trigrams).

PLOT HERE FOR TRIGRAMS

Experiment 3: Large scale data and linguistic features

To make a claim about how languages on a larger scale, we need to use larger corpora and a much larger number of languages. We pulled corpora for 159 diverse languages from Wikipedia, each of which had at least $10,000$ articles on the knowledge base. We split each article into sentences; the variance in sentence lengths for Wikipedia was significantly larger than for the CHILDES corpora we used in the previous section. Most sentences in Wikipedia contained between 10 and 30 words, unlike the CHILDES corpora which mainly contained utterances with under 10 words. We excluded the small fraction of utterances with more than 50 words since they were small in number and, from manual inspection, uncharacteristic of typical written sentences.

To more rigorously described the typological differences between languages, we used data from the World Atlas of Language Structures (WALS; **???**). The WALS database has data for 144 typological features in 2569 languages from across the world. These features describe aspects of morphology, syntax, phonology, etymology and semantics–in short the features describe 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. Even limiting ourselves to the 159 language corpora we pulled from Wikipedia and 122 features from WALS, there are nearly 20000 individual possible data values, fewer than half of which were already computed for those languages in the WALS database.

To fill in the missing data for the features we selected using statistical imputation, we used Multiple Imputation Multiple Correspondence Analysis (MIMCA; **???**). MIMCA begins with mean imputation, converts the categorical WALS features into a numerical contingency table with dummy coding, then repeatedly performs principle components analysis and reconstructs the contingency table. Our final result from the MIMCA algorithm was a fully imputed table with 122 feature values for each language.

However, the WALS features describe specific structural differences between languages, while our surprisal metric is word-based. To target lexical differences between languages, we computed the average normalized Levenshtein distance (LDN; **???**) over the 40 item Swadesh list (**???**), retrieved from the ASJP database (**???**). The Swadesh list is designed to include near-universal words that target basic cognitive concepts, and are useful in determining the genealogical similarities and differences between languages. The results of classifying languages using the Swadesh list and LDN are correlated with those using WALS features, but the Swadesh list and LDN do not suffer from the same sparsity problem as WALS (**???**).

We ran a hierarchical clustering algorithm on the frequency-based information curves using the hclust package from the R stats core library (**???**). We used the complete linkage algorithm for hierarchical clustering, with distances between information curves between languages computed using cosine distance between their embeddings in the slope space. The complete linkage algorithm at every step pairs each language or cluster of languages with its closest neighboring language or cluster. A sample from the dendrogram is shown in Figure @ref(fig:dendro). From a quick glance, the unigram information curves appear to reproduce some of the genealogical relationships between languages, although the dendrogram does not exactly replicate language genealogy for all 159 languages. This suggests using a first-pass quantitative method that the information curves do correspond in some measure to language families, but language families do not explain all of the variation and relationships between frequency-based information curves.

DENDRO

For our first quantitative analysis, we examined the effects of individual typological features on the shapes of the unigram information curves. We ran logistic regressions using the lme4 package in R (**???**), checking whether the cosine distance between two languages' embeddings in the slope space played a role in determining if those two languages had the same value for a given WALS feature. Individual WALS features do not necessarily have ordinal values. Some, such as the "Number of Cases" feature, are easy to quantify and order. Others are more difficult. For example, how does one order

"relative clauses appear after the nouns they modify", "relative clauses appear before the nouns they modify" and "free order of relative clauses and nouns"? We chose the identify relation to avoid deciding on the basis of individual features. We found that 100 out of the 122 features from WALS we examined were statistically significant ($p < .001$) in determining whether two languages had the same frequency-based information curve shape. The results for some important features are in Figure @ref(fig:linear-models).

INDIVIDUAL WALS FEATURES

We next compared how the cosine distance between two languages related to how many WALS features they had in common. r^2 value is .005734, which suggests that in aggregate there is not a correlation between how many WALS features languages have in common and the similarity of their frequency-based information curves. Figure @ref(fig:cosine-wals) displays the results. This result is surprising based on the significance of many WALS features in predicting the shapes of the frequency-based information curves, and we return to this result in the general discussion.

COSINE

For lexical features, we see a stronger correlation between the similarity of two languages in terms of their average LDN and the cosine distance between their information curves. Figure @ref(fig:cosine-ldn). We see a higher r^2 value here of .026, indicating that there is more correspondence between a language's lexical similarity to another language and their similarity in information curves. From these typological and lexical investigations, we conclude that the shape of a language's frequency-based information curve covaries with its typological and lexical similarity to other languages. However, most of the variation in frequency-based information curve shapes is not explained by typological properties in language.

LDN

Conclusion

Acknowledgements

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

References

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

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.

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.