How to Make a Proceedings Paper Submission

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: Add your choice of indexing terms or keywords; kindly use a semi-colon; between each term.

Introduction

One of the defining features of human language is its power to transmit information. 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 (**???**). 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 because it allows a framework for rational analysis. If language has evolved to be an optimal code for information transmission, what would be the optimal structure for this code (**???**)? The optimal code would have to work with two competing pressures: (1) a pressure for listeners to easily and successfully decode messages sent by the speaker, and (2) a pressure 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. However, the cost of producing more predictable utterances is using more words. Thus, the optimal strategy is 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.

How do speakers structure individual utterances? This 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. 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.

In this paper, we expand on this body of prior work in a number of novel ways. We replicate the results from (**???**) with a metric tied to incremental word processing. We find their same distribution of information based on word frequency in English speech as well as in English writing. We extend our metric to include context, and show that the addition of context for each word smoothes out language-specific distributions. We expand the study of information density to the largest set of languages considered so far, and incorporate contextual and typological information into our analysis. Speakers will tend to distribute information in a language constrained but not determined by the morphology, syntax and phonology of that language. Using child speech corpora, we find that as soon as a child starts speaking, they tend to dis-

tribute information in their utterances according to the characteristic distribution in their language.

Methods

(**???**) defined information as "the reduction in uncertainty about one variable given that the value of another variable is known". We use a metric proposed for the study of information transmission more generally by Shannon and applied to words specifically by (**???**): lexical surprisal. This measure defines the information in a word based on the ratio of possible continuations of the sentence after to before the word is seen. Equivalently, we can compute surprisal with 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 a 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, 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 (**???**).

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.

$$\begin{aligned}\text{surprisal}(w_i|w_{i-1}w_{i-2}...) &= -logP(w_i|w_{i-1}w_{i-2}...) \\ &= -log\frac{P(w_i, w_{i-1}w_{i-2}, ...)}{P(w_{i-1}w_{i-2}...)}\end{aligned}$$

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. "Flower" may contain less information than "azalea" when we consider the words independently of their context, but with context this can be reversed. Flower appears in a variety of contexts, and so the information content of a word like "flower" in a particular context may be higher than "azalea". If you only have azaleas in your garden, then hearing someone say "in that garden, look at the flowers" may be higher surprisal for you: you expect them to say "azalea". This prediction does not require many words for context. 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". Hearing "I" restricts the next word to a verb, or possibly an adverb, and since the listener has heard the speaker refer to themselves in the first person singular, their set of possible completions is significantly restricted.

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, which a speaker may not have produced as the utterance "I saw bears" in the corpus but could have produced that utterance. 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 more complete explanation of modified Kneser-Ney smoothing, 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. By computing these averages for each word position in an utterance, we compute a low-dimensional approximation to the average distribution of information in the corpus. With the surprisal metric, we base the information contained in each word on how often the word is encountered in its context in the corpus. As long as the corpus is representative of the language or population we study, then the distribution of information is approximated for that language or population as a whole.

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: BNC & Switchboard

Experiment 2: Wikipedia

Conclusion

Acknowledgements

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

References