Information in sounds and letters in English, French and German speech

Josef Klafka and Dan Yurovsky

2019-08-13

Introduction

How do speakers tend to spread information among the sounds in English words? A salient feature of African American Vernacular English is the reduction of final consonants and consonant clusters. Why doesn't this reduction create communication problems, namely identifying and disambiguating words? One possibility is that the information important for identifying a word is not always uniformly distributed across all of its characters or sounds, but instead usually concentrated at the beginning or middle of the word. For these words, reducing final sounds might reduce the articulatory complexity of words without much cost for the listener in processing these words and inferring their intended meaning. By the time a listener hears only the first sounds of a word, they may already have identified which word they hear.

We tackle this problem using four exploratory analyses:

1. We compute the information of characters in English words from baserates of character production (unigram) and then conditioning on the prior characters in each word (bigram and trigram). This analysis roughly captures the difficulty that a listener might have in predicting each character.
2. We repeat this analysis at the level of phones rather than characters, arguably a better proxy for speakers' and listeners' representations of spoken English.
3. We compute hold-out entropy for individual phones in English words. This is roughly equivalent to asking "if you knew every other sound in a word, how much trouble would you have predicting what sounds appears in position X", where position X is held out.
4. We compare our English by-letter analysis to results for French and German, for which we expect different information profiles. We expect that French words will have less information on average at the end of words compared to English, as French speakers do not pronounce the finals sounds in many words in their language. We expect German words will have more information on average in their final letters compared to English, as German nouns are case-marked with affixes.

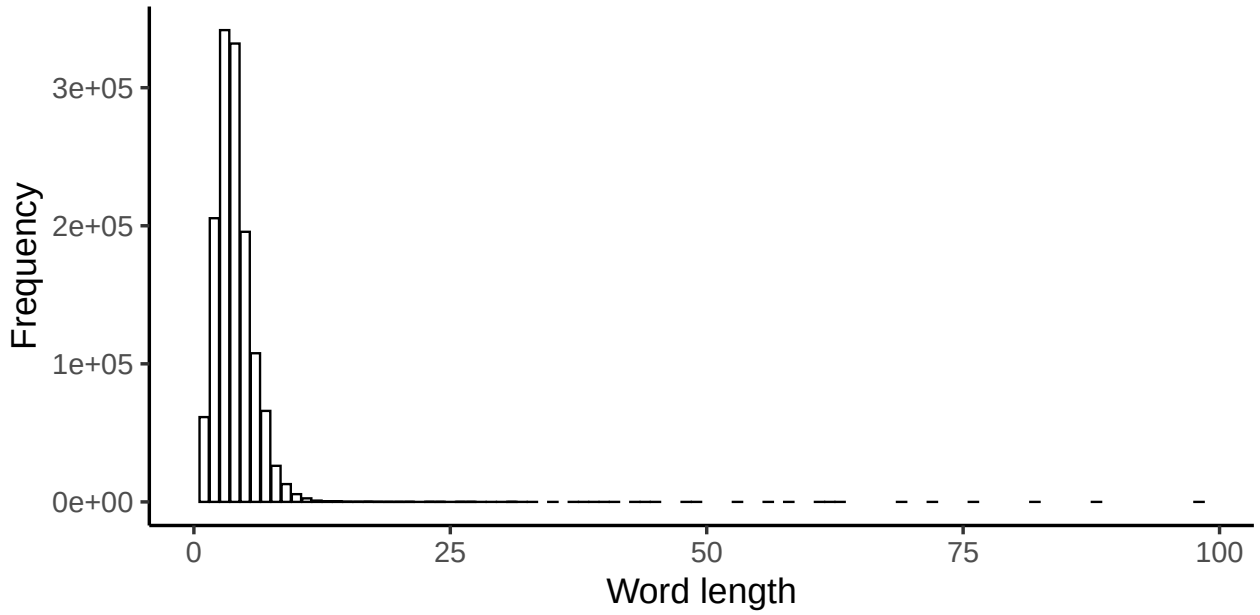
Across all of our analyses, we find that the letters and sounds at the end of words in English have less information than the first letters and sounds in a word, and that the middle and final letters in English words are easily predictable from their preceding letters. We find that letters at the end of words in French and German also have little information compared to letters at the beginning of words in those languages, like in English. This is unexpected for German, though we do find that the ends of words in French contain less information than the ends of words in English, even looking solely at character baserates.

We use three corpora from the CHILDES-Talkbank system (MacWhinney, 2000). See the Appendix for more technical information about methods.

English Letters

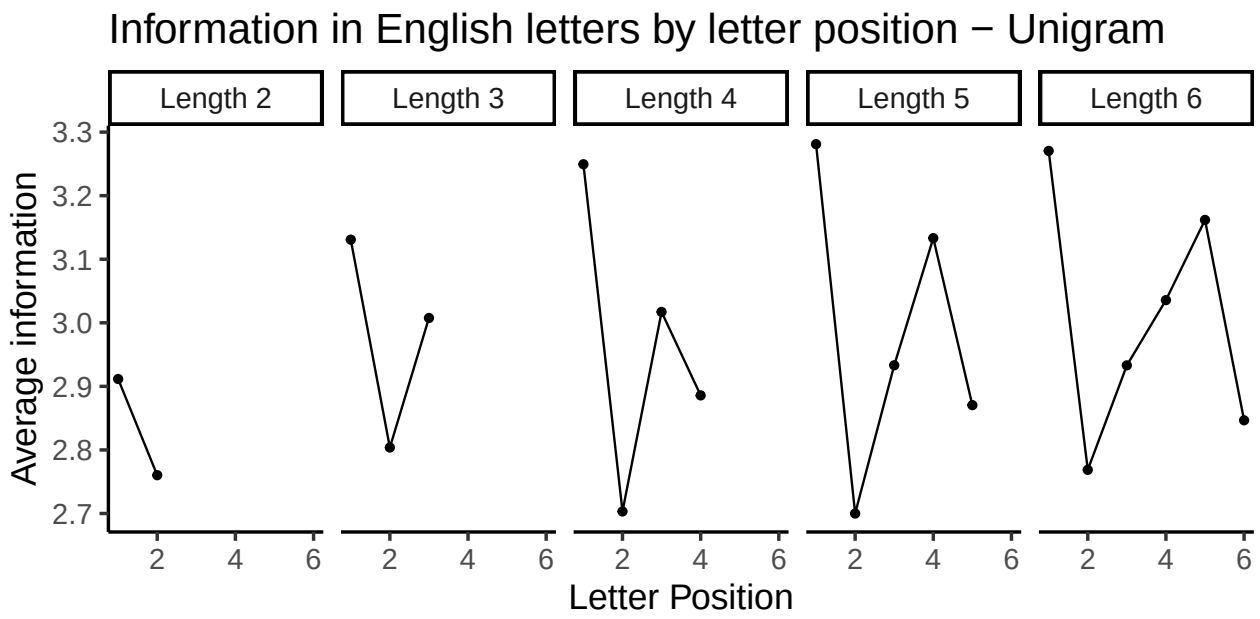
We used the Providence corpus from CHILDES as data for English (Evans & Demuth, 2012). In the plots below, each plot represents the information within letters at different positions in words, divided by words of different lengths. Word lengths are given by number of letters. Below is a histogram of the distribution of word lengths in the Providence corpus. We observed that most words (84%) have 5 letters or fewer, while 99% of words in the corpus have fewer than 10 letters. There are a tiny number of outliers which have 10 or more letters–by manual inspection we see that these are mainly onomatopoeia.

```
plot_hists()
```



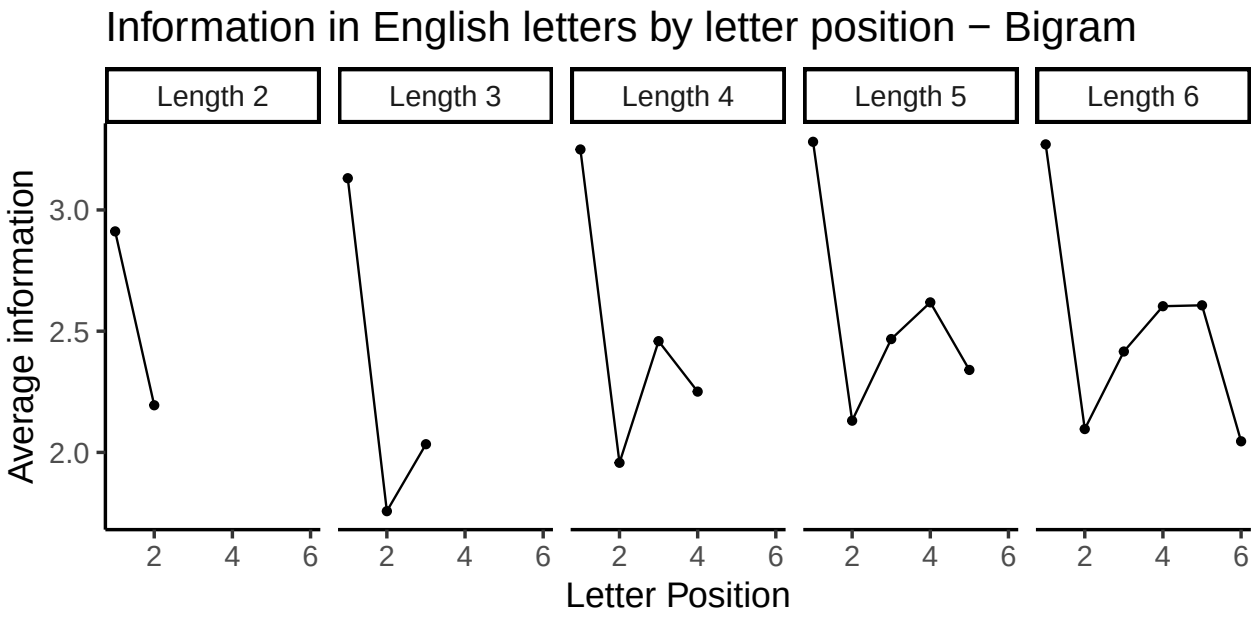
We observed a distinctive shape for information at the level of individual letters in English. The first letter has relatively high information, the second letter has less information, the third letter has high information, and the final letter has less information (if the word has more than three letters).

```
plot_curves()
```



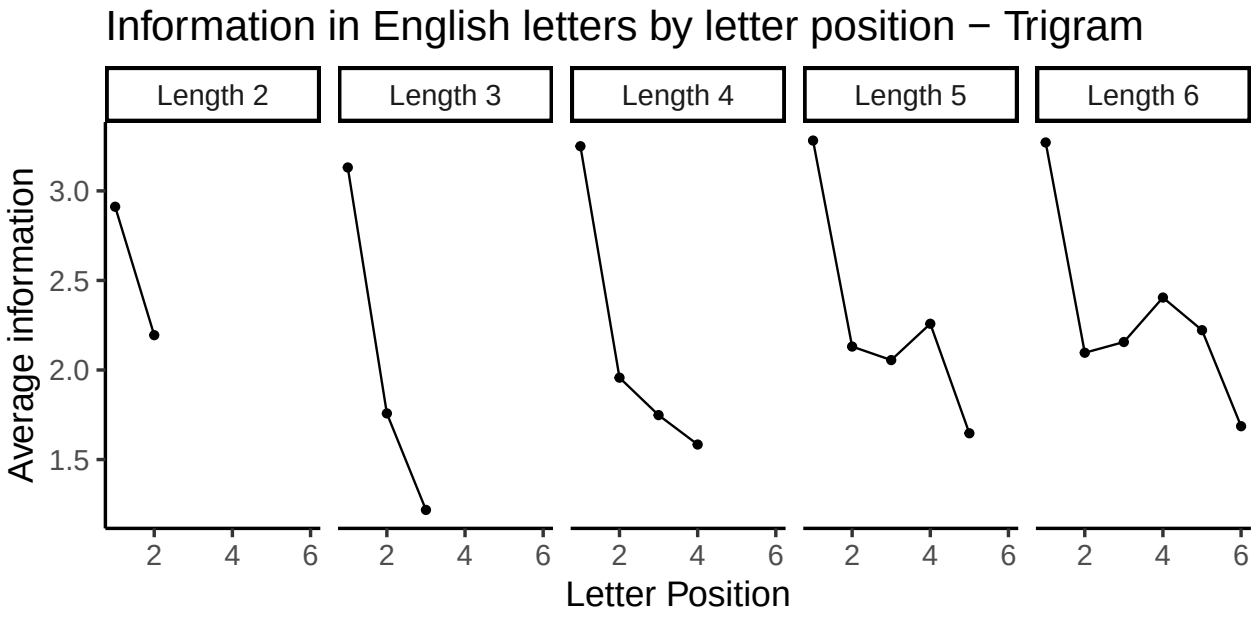
Now we ask: if you know the previous letter, how much information is in the next letter in English? There is relatively less information in each letter knowing the previous letter, but we see the same general information trajectory across letters as in the individual letter trajectory.

```
plot_curves(select_gram = "Bigram")
```



What about if you know the last two letters and are trying to predict the next letter? We see decreasing information for each letter within words as you move left-to-right. All letters in English words, aside from the first letter, are easily predicable from only knowing the preceding two letters for each. This is especially noticeable for shorter words with only three or four letters, but still holds true for words of length five or six letters.

```
plot_curves(select_gram = "Trigram")
```



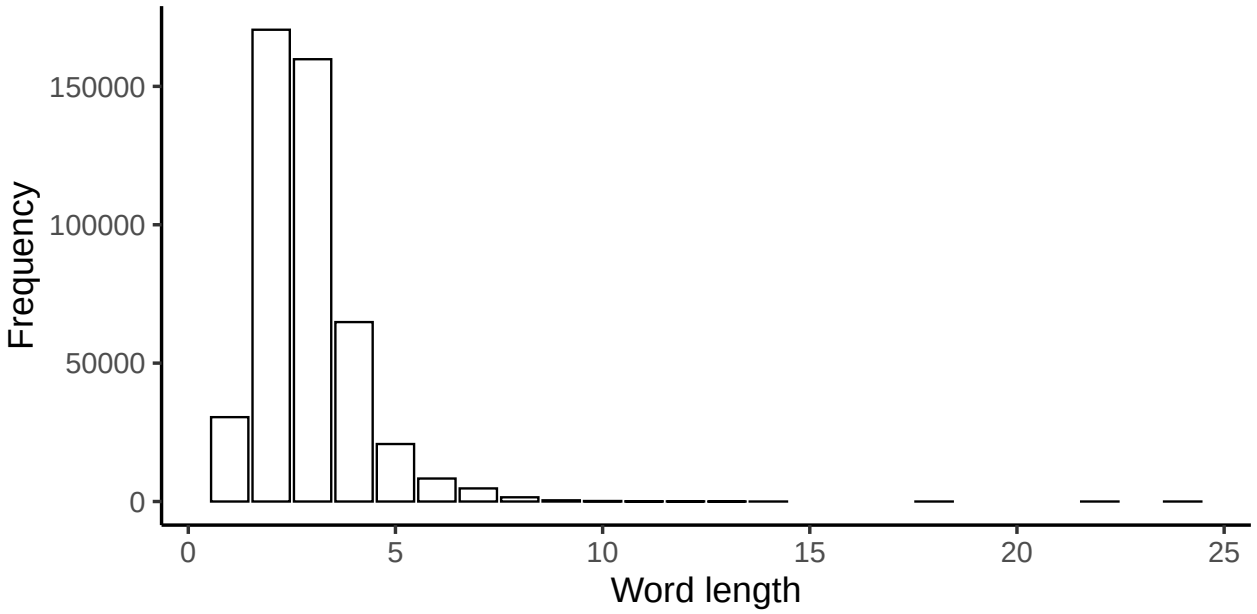
English Phones

In order to analyze the individual phones in English, we split each word in the Providence corpus into its constituent phones, the sounds that make up the word. We then divide up the words in the corpus by their length in number of phones, similar to how we divided words by number of letters. For example, the word

"food" has three phones: one for the "f", one for the long "oo", and one for the "d". We take all words like this that have three phones, and for each phone position we calculate the average information in that position.

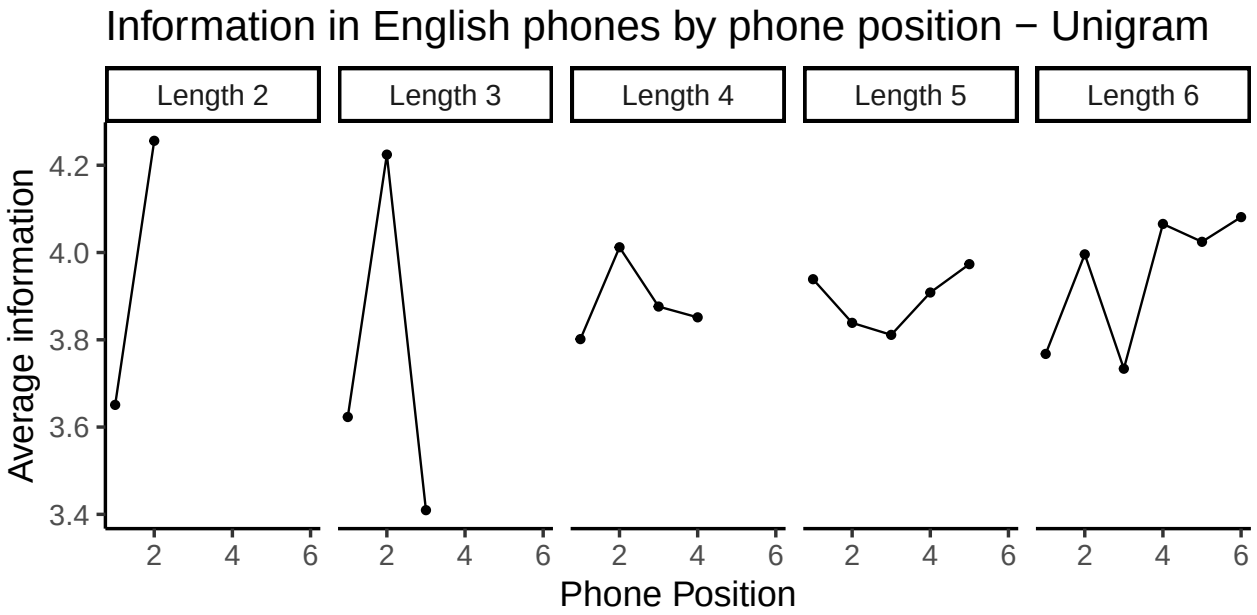
As seen in the histogram below, almost all words in the Providence corpus (92%) have four or fewer phones. 99.5% of words have seven or fewer phones.

```
plot_hists(select_rep = "phone")
```



Unlike in our letter analysis, there isn't as much of a consistent shape to the information distribution for phones in the individual (unigram) condition. There is some evidence that the first phone has relatively little information while the second phone has a lot of information, which is the opposite of what we observed consistently with the information in English letters.

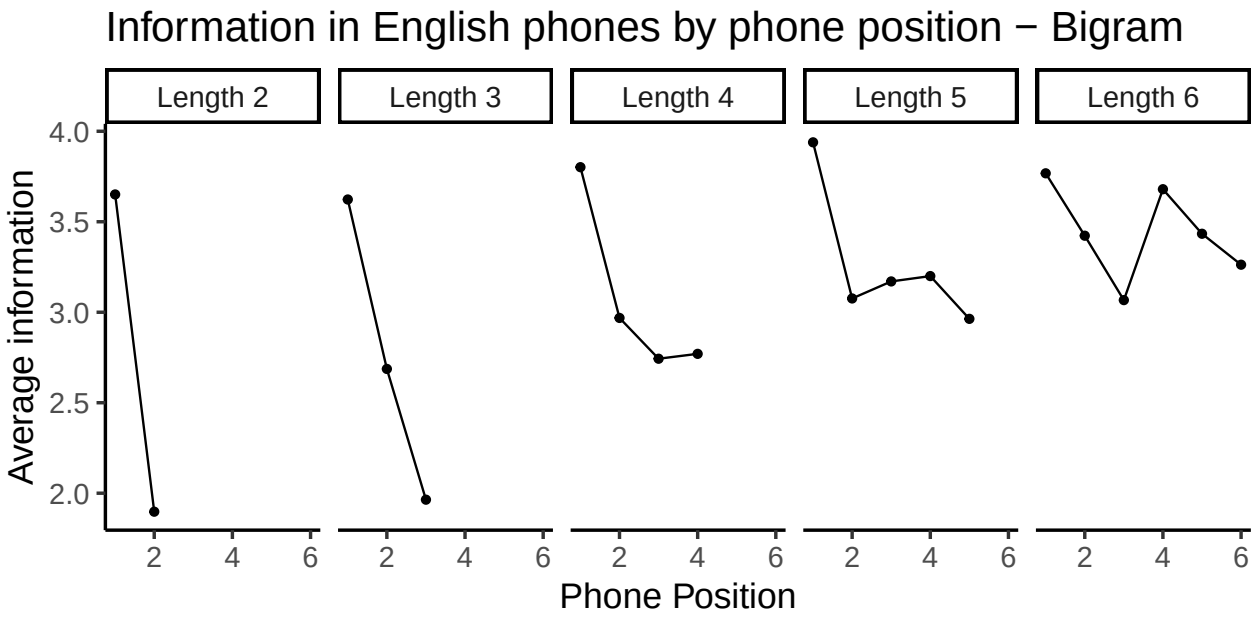
```
plot_curves(select_rep = "phone")
```



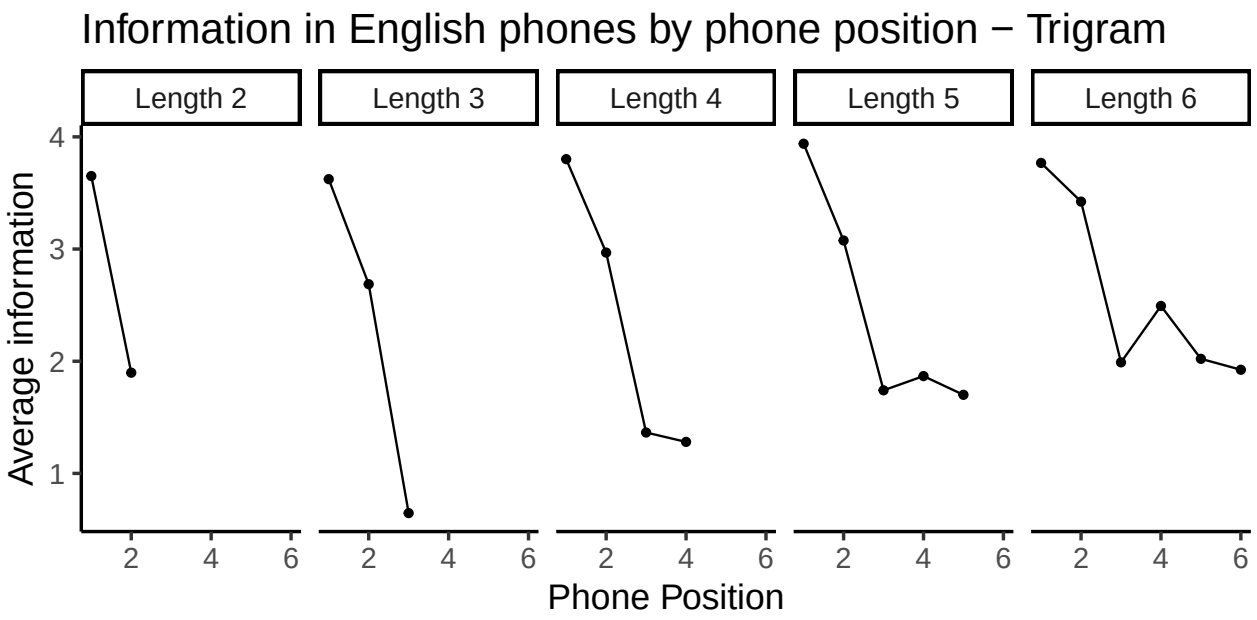
With bigrams and trigrams for English phones, the next phone is very predictable if you know the phone or two that came before it. Especially if you know the preceding two phones, then each phone past the first two

is predictable and carries relatively little information in English. This is the same as our conclusion from the letter analysis in English. The middles and especially the ends of words in English contain relatively little information, because those part of words in speech or reading are easily predictable from what came before them.

```
plot_curves(select_rep = "phone", select_gram = "Bigram")
```




```
plot_curves(select_rep = "phone", select_gram = "Trigram")
```



Hold-out entropy

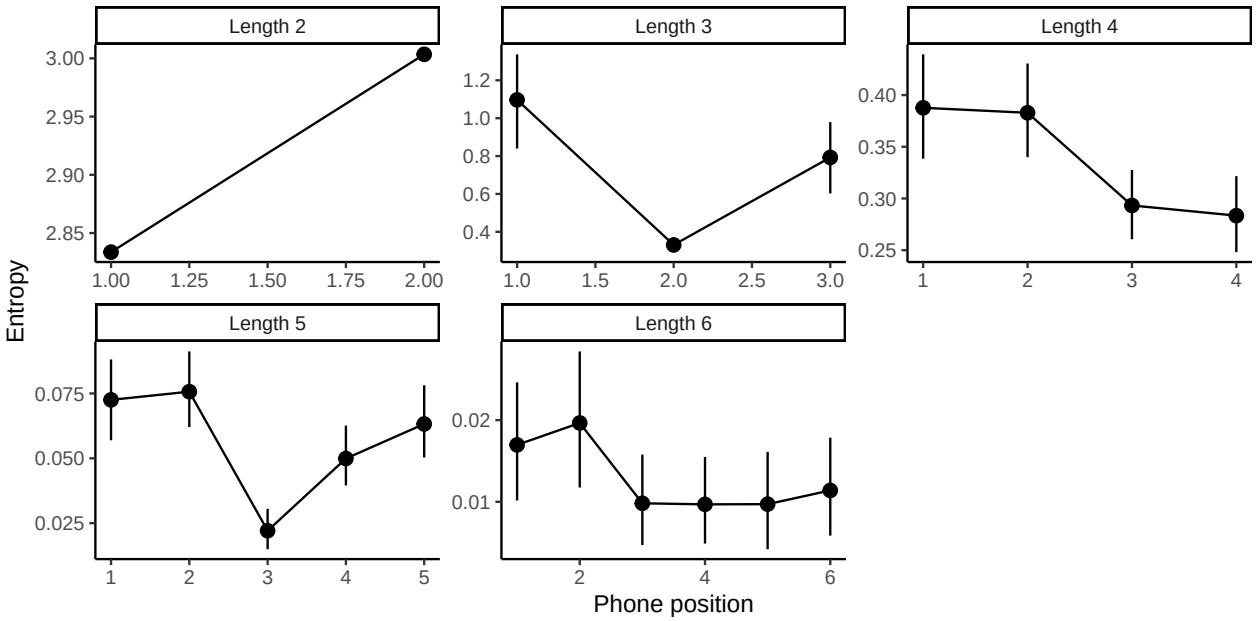
For this next analysis, we took all words in the Providence corpus with between two and six phones. For each word length, we held out the phones in one phone position. For example, in all words with four phones, we held out the first phone. For each phone we held out, we computed the entropy (a measure of variation) in the observed words that had the remaining phones. Holding out the /f/ in "food" gives the continuations

"food" and "mood" for example.

The higher the entropy when holding out phones in a given position, the more information is contained in that phone on average. We observed that word-initial phones tend to have more information in them than word-medial or word-final phones, so less is lost by dropping the final sounds in a word than in dropping the initial sounds.

```
split_lengths <- adult_phonemes %>%
  group_by(word_length) %>%
  nest() %>%
  filter(word_length > 1) %>%
  filter(word_length %in% 1:6 ) %>%
  mutate(entropy = map(data, hold_out_entropy))

split_lengths %>%
  select(-data) %>%
  unnest() %>%
  ggplot(aes(x = phone_order, y = empirical_stat)) +
  facet_wrap(~ word_length,
             labeller = as_labeller(function(x) paste0("Length ", x)),
             scales = "free") +
  geom_pointrange(aes(ymin = ci_lower, ymax = ci_upper)) +
  theme_classic() +
  geom_line() +
  xlab("Phone position") +
  ylab("Entropy")
```



French and German

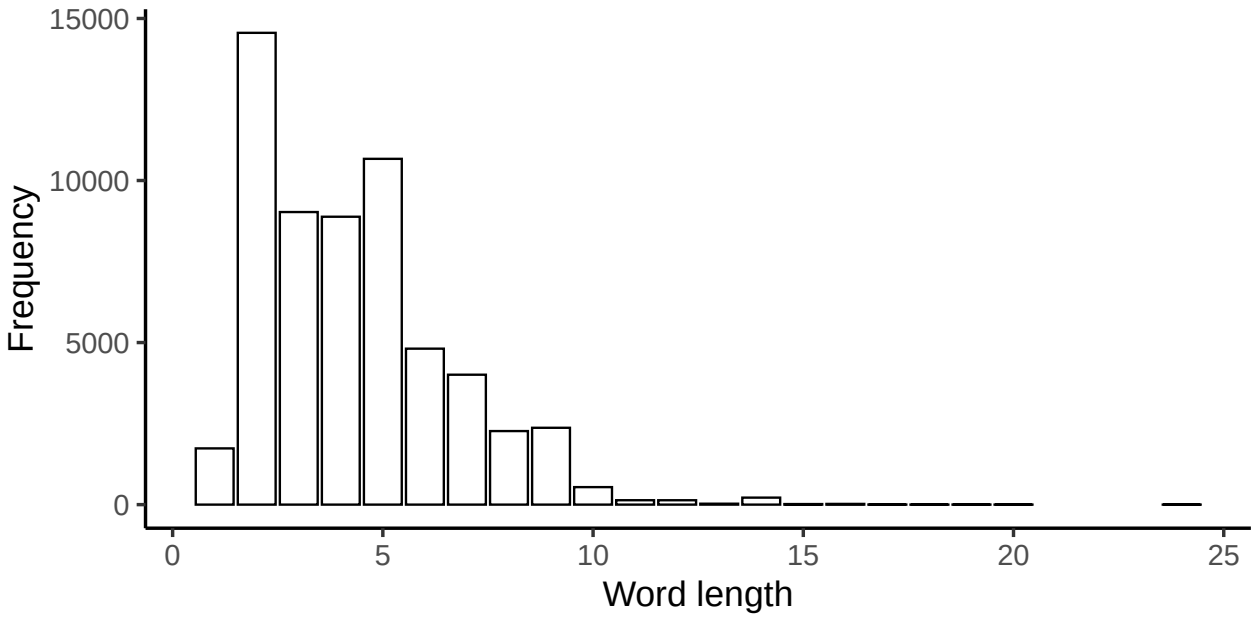
For our final analysis, we examined how the information within words in French and German compare to English. We used the French Palasis corpus (Palasis, 2009) and German Wagner corpus (Wagner, 1985) from CHILDES for our data. For many French words, the final letters are not pronounced although those letters are still written in the word. For example, "ils sauraient" meaning "they would know" is pronounced without the final "ent" in "sauraient". We predicted that the ends of French words will be less important

than in English. German has a much richer nominal case system than English or French, and case markings in German are suffixes at the ends of nouns. We predicted that in German, the ends of words may be more important than in French or English.

French letters

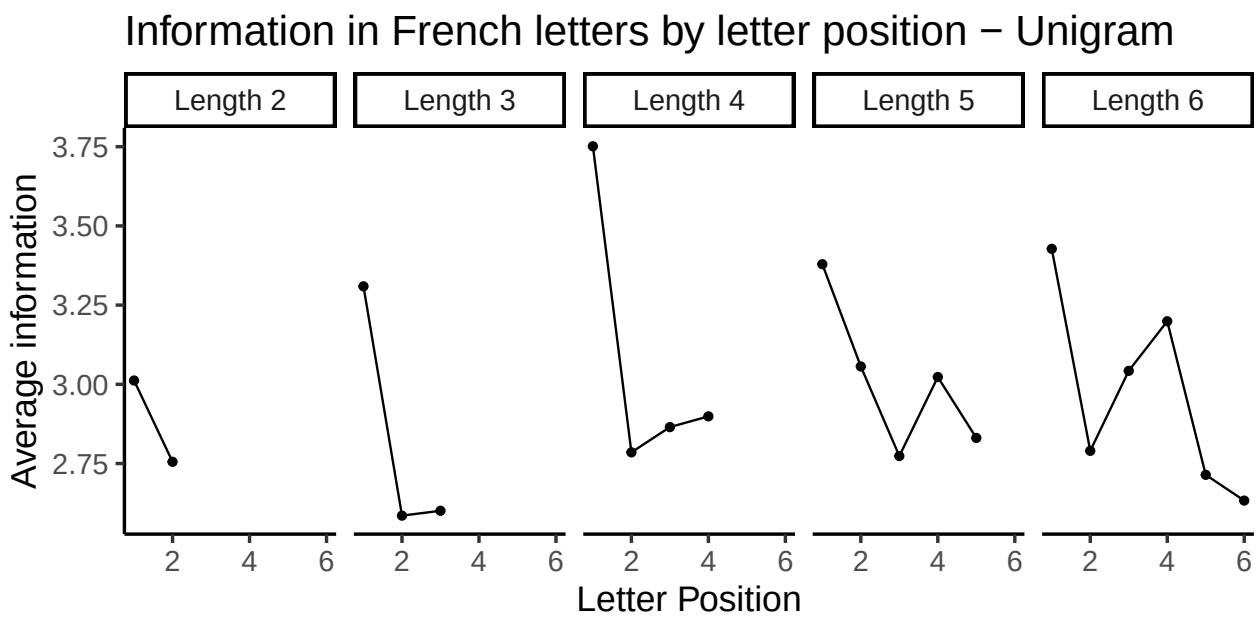
First, we examined the word length frequencies by number of letters in the French Palasis corpus. In the Palasis corpus, 99% of words have 10 letters or fewer. This French corpus has a higher share of words with six to ten letters than the English Providence corpus: 24% versus 16%. However, almost all words in both corpora have 10 or fewer letters. While there are relatively more medium length words in the French corpus compared to the English corpus, there are few long words in either corpus.

```
plot_hists(select_language = "French")
```

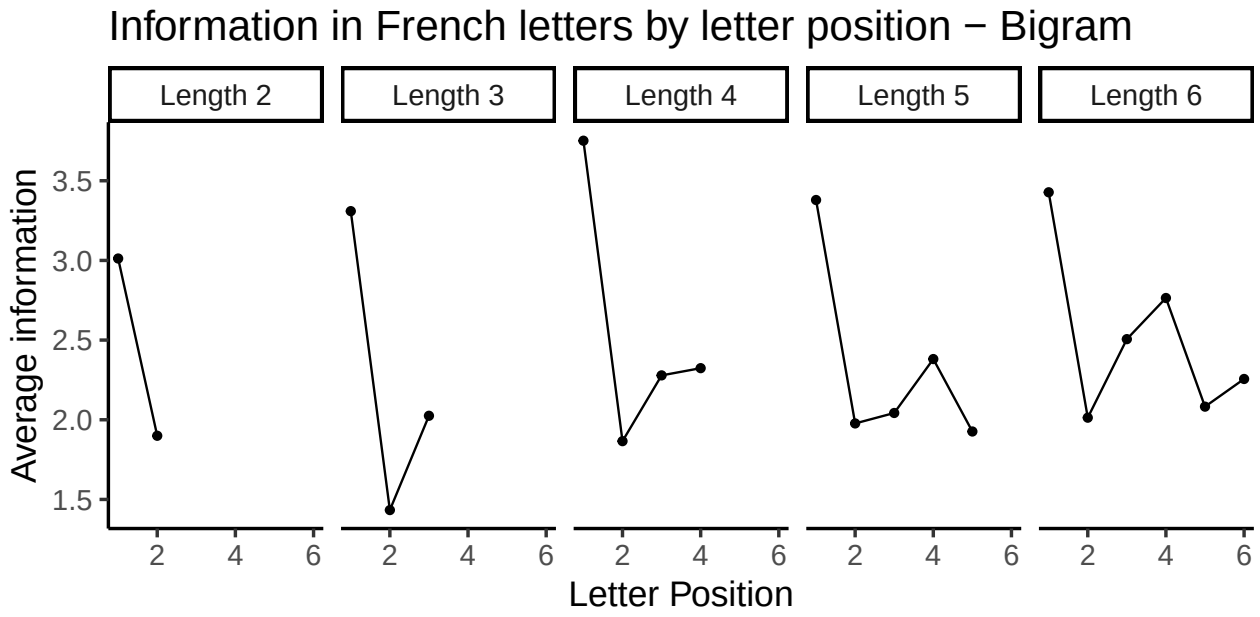


We observed that the final letters in French words, even considered without context, have little information. The gap between the information-rich first letter and final letters becomes more apparent with context, and we observe that the middle letters in French have relatively little information when considered with context, similar to English letters.

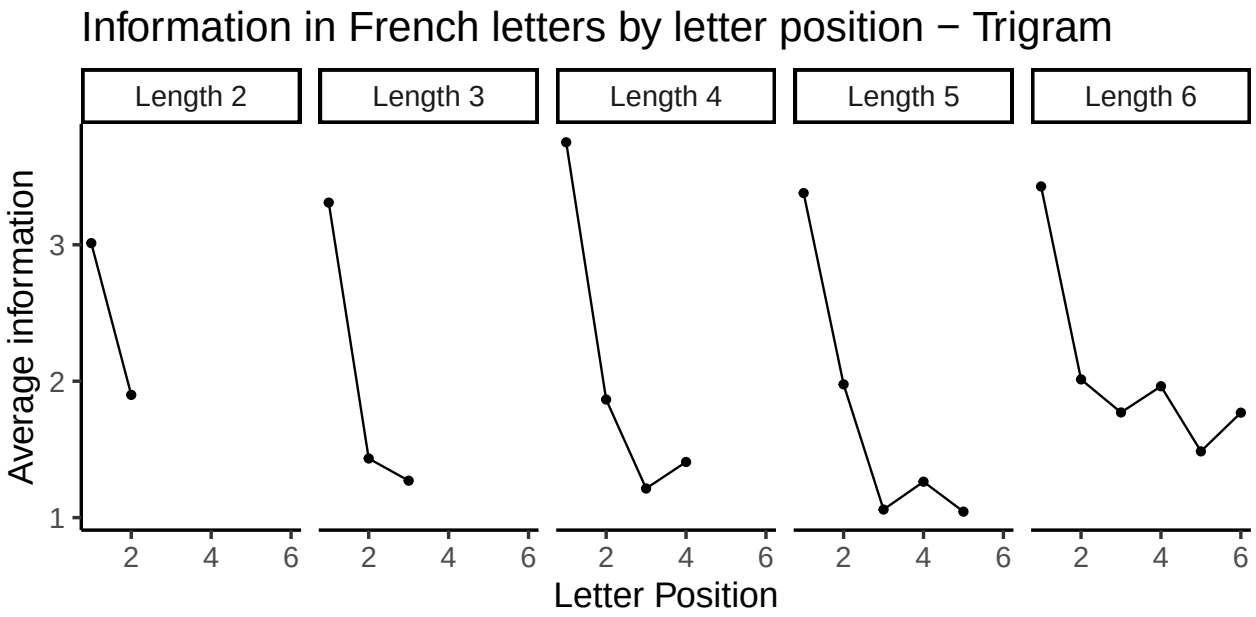
```
plot_curves(select_language = "French")
```




```
plot_curves(select_language = "French", select_gram = "Bigram")
```



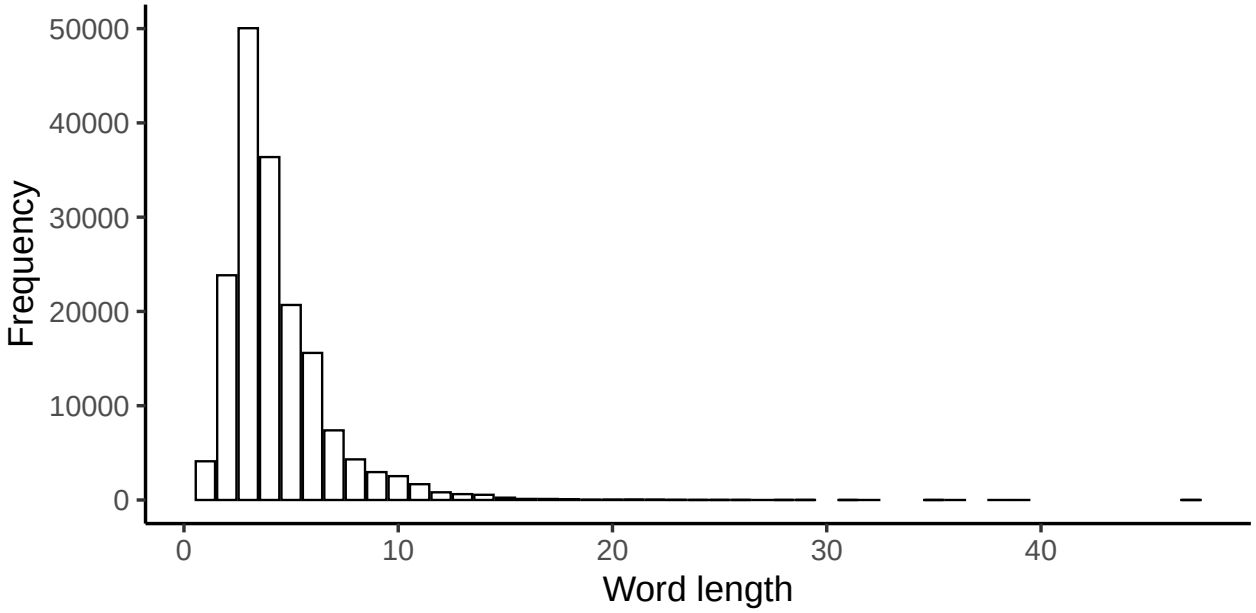

```
plot_curves(select_language = "French", select_gram = "Trigram")
```



German letters

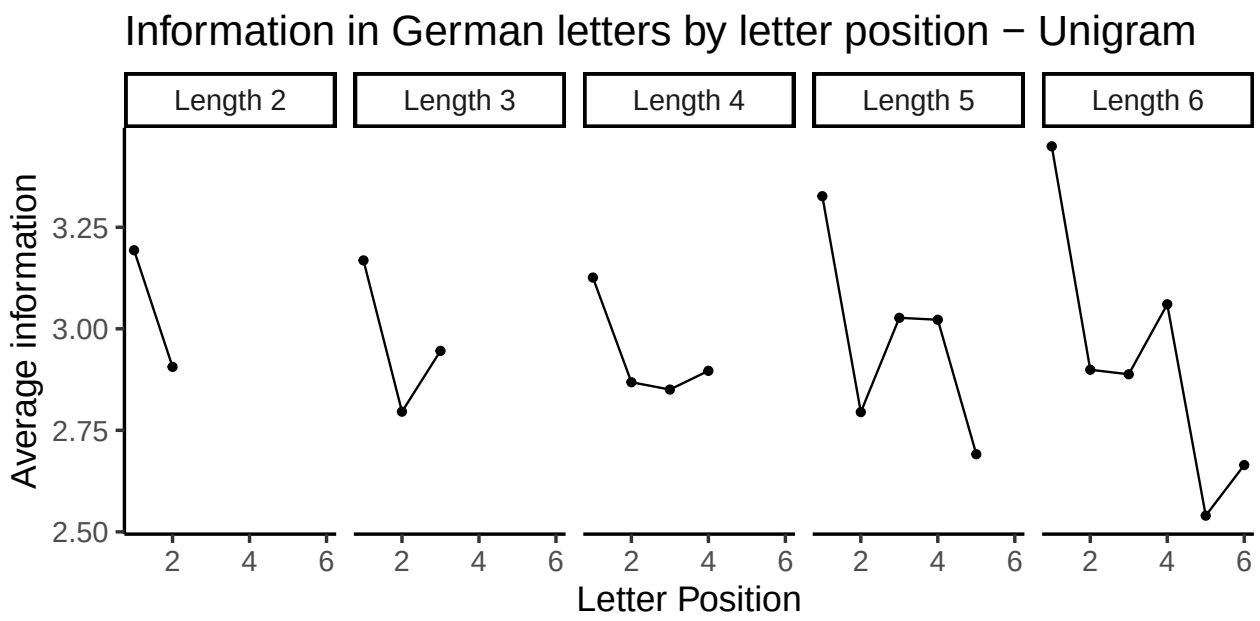
In the German Wagner corpus, the distribution of word lengths has a thicker tail than in English. 99% of words have 13 or fewer letters, compared to the Palasis and Providence corpora where over 99% of words had 10 or fewer letters. However, 97.4% of German words have 10 or fewer letters, so the corpus is similarly comprised of short and medium-length words.

```
plot_hists(select_language = "German")
```

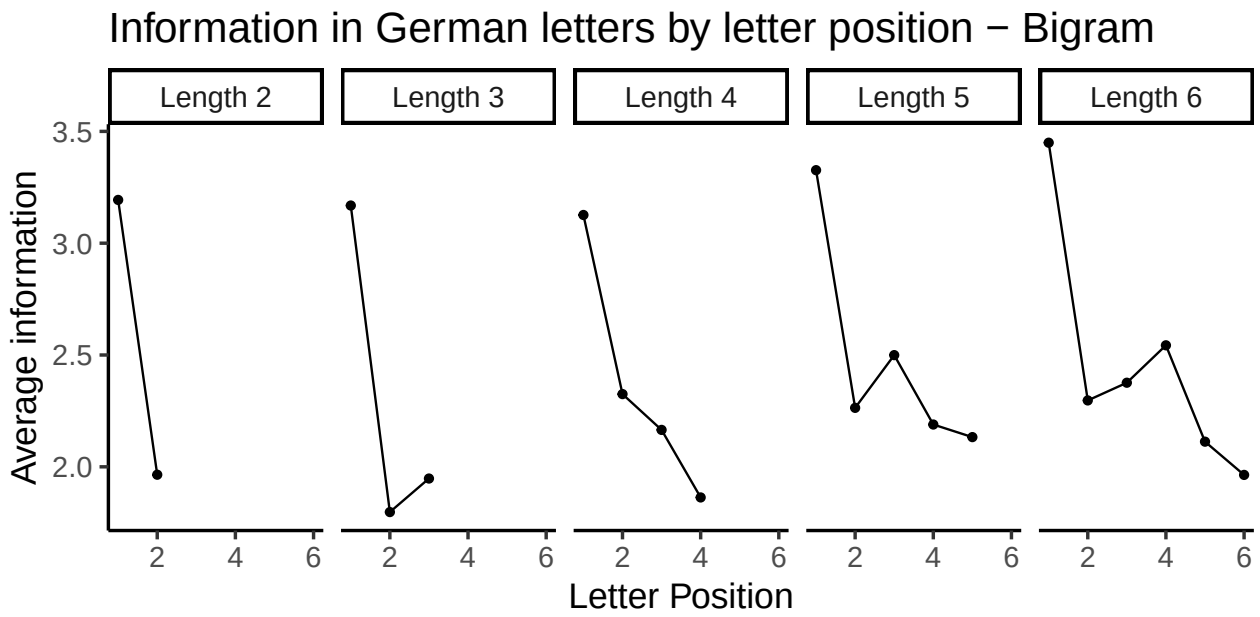


We did not observe that German had more information in final letters than English or French, even when examining the information in independent letters. The final letters of German words hold little information compared to the first letter of German words, which is even more noticeable with two letters of predictive context for each letter.

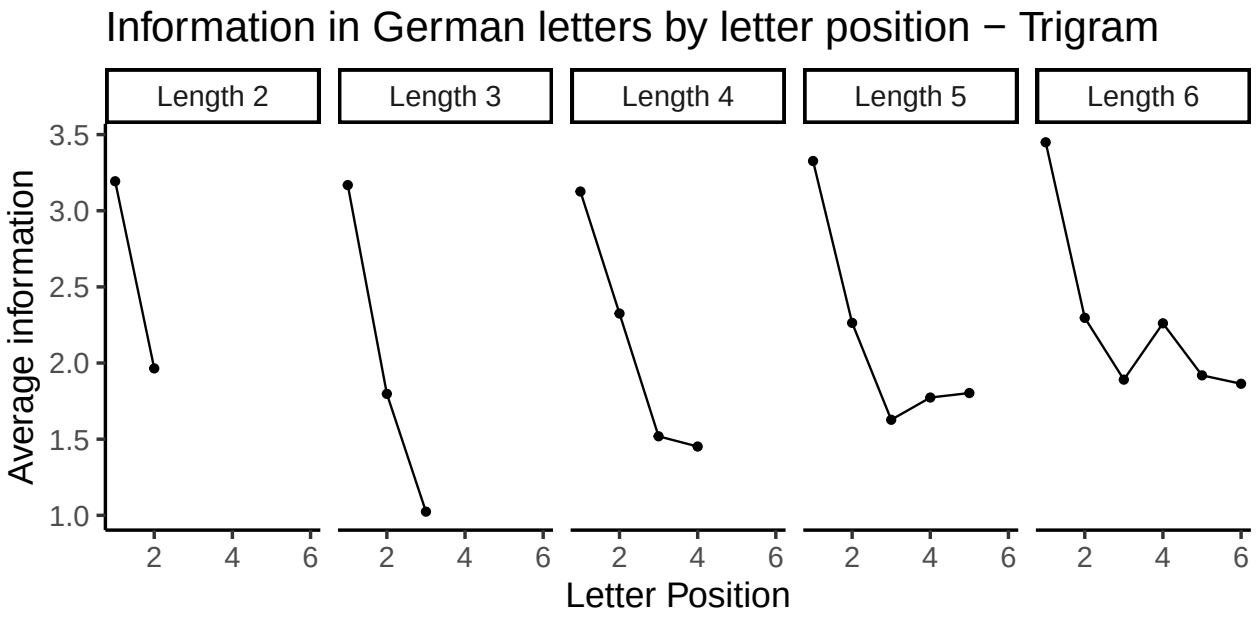
```
plot_curves(select_language = "German")
```




```
plot_curves(select_language = "German", select_gram = "Bigram")
```




```
plot_curves(select_language = "German", select_gram = "Trigram")
```



General Discussion

Across all three languages we examined and both letters and phones in English, we found that the first two or three letters and sounds in a word contain relatively more information than any other letters or sounds in the word. The final letters and sounds tended to contain little information compared to the initial letters or sounds. The middle and final letters in words in all three languages were easily predictable from the two letters that came directly before them. Even for a language like German which marks case at the end of nouns, listeners can identify and disambiguate the words using only the first few letters of those words.

Appendices

A: Methods

The metric we use for information is based on the lexical surprisal metric from Levy (2008), with formula given below. The surprisal s of the ith word w_i in an utterance is given by

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

where w_{i-1} denotes the word one previous to the target word. For our work, the surprisal s of the ith letter or sound ℓ_i in a word is given by

$$s(\ell_i) = -\log P(\ell_i|\ell_{i-1}\ell_{i-2}...)$$

where ℓ_{i-1} denotes the letter or sound one previous to the target letter or sound. For our unigram work, we did not use context, and so the conditional probability is instead a simple probability.

References

Evans, K. E., & Demuth, K. (2012). Individual differences in pronoun reversal: Evidence from two longitudinal case studies. *Journal of child language*, 39(1), 162-191.

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

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

Palasis, K. (2009). *Syntaxe générative et acquisition: le sujet dans le développement du système linguistique du jeune enfant* (Doctoral dissertation, Nice).

Wagner, K. R. (1985). How much do children say in a day? *Journal of Child Language*, 12, 475–487.