Disambiguatory Signals are Stronger in Word-initial Positions

Disambiguatory Signals are Stronger in Word-initial Positions

Tiago Pimentel[ð] **Ryan Cotterell**[ð,ʕ] **Brian Roark**[ə]
[ð]University of Cambridge [ʕ]ETH Zürich [ə]Google
tp472@cam.ac.uk, ryan.cotterell@inf.ethz.ch, roark@google.com

Abstract

Psycholinguistic studies of human word processing and lexical access provide ample evidence of the preferred nature of word-initial versus word-final segments, e.g., in terms of attention paid by listeners (greater) or the likelihood of reduction by speakers (lower). This has led to the conjecture—as in Wedel et al. (2019b), but common elsewhere—that languages have evolved to provide more information earlier in words than later. Information-theoretic methods to establish such tendencies in lexicons have suffered from several methodological shortcomings that leave open the question of whether this high word-initial informativeness is actually a property of the lexicon or simply an artefact of the incremental nature of recognition. In this paper, we point out the confounds in existing methods for comparing the informativeness of segments early in the word versus later in the word, and present several new measures that avoid these confounds. When controlling for these confounds, we still find evidence across hundreds of languages that indeed there is a cross-linguistic tendency to front-load information in words.[1]

1 Introduction

The psycholinguistic study of human lexical access is largely concerned with the incremental processing of words—whereby, as individual sub-lexical units (e.g., phones) are perceived, listeners update their expectations of the word being spoken. One common tenet of such studies is that the disambiguatory signal contributed by units early in the word is stronger than that contributed later—i.e. **disambiguatory signals are front-loaded in words**. This intuition is derived from ample indirect evidence that the beginnings of words are more important for humans during word processing—including, e.g., evidence of increased attention to word beginnings (Nooteboom, 1981, *inter alia*) or evidence of increased levels of phonological reduction in word endings (van Son and Pols, 2003b).

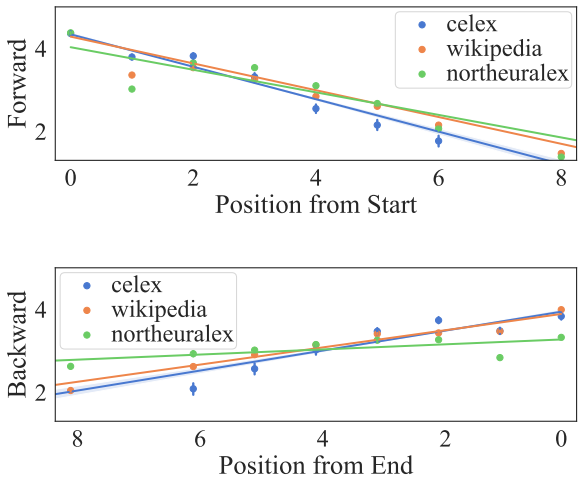

Figure 1: Forward and Backward Surprisals with LSTM model from Pimentel et al. (2020). The bottom plot has been flipped horizontally such that it visually corresponds to the normal string direction.

To analyse this front-loading effect, researchers have investigated the information provided by segments in words. van Son and Pols (2003a,b) showed that, in Dutch, a segment's position in a word is a very strong predictor of its conditional surprisal, with later segments being more predictable than earlier ones—a result which we show to arise directly from its definition in §3.3.1. Recently King and Wedel (2020) and Pimentel et al. (2020) confirmed the effect on many more languages.

Their analysis, however, presents an inherent confound between the amount of conditional information available to a model and the surprisal of the subsequent segment—see Fig. 1 for results illustrating this. Using the LSTM training recipes from Pimentel et al. (2020),[2] we calculated the conditional surprisal at each segment position within the words across all languages in three datasets.[3] The top-half of Fig. 1 shows that, indeed, positions

[1]Our code is available at https://github.com/tpimentelms/frontload-disambiguation.

[2]https://github.com/tpimentelms/phonotactic-complexity

[3]See §3 and §5 for specifics on training and data. Each segment corresponds to a single phone in CELEX and NorthEuraLex, and to a single grapheme in Wikipedia.

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 31–41
April 19 - 23, 2021.

earlier in the string have higher surprisal than positions later in the string, supporting the thesis of higher informativity earlier in words. The bottom-half shows that modelling the strings right-to-left instead of left-to-right reverses the resulting effect.

This decouples conditional surprisal from the disambiguatory strength. To expose this decoupling, consider an artificial language where every word contains a copy of its first half, e.g., *foofoo*, *barbar*, *foobarfoobar*, etc. The first and second halves of these words have identical disambiguatory strength; they are the same so one could disambiguate the word as easily from its second half as from the first. In contrast, conditional surprisal would be nearly zero for the second halves of words because the second half is perfectly predictable from the first half.

In natural languages, measuring conditional entropy in a left-to-right fashion inherently forces a reduction of conditional entropy in later segments because of a language's phonotactic constraints. However, the disambiguatory strength of later segments is not inherently less than that of earlier segments. For instance, in a language like Turkish, which has vowel harmony, knowledge of any of the vowels in a word will provide information about the word's other vowels in a similar way. As such, knowledge of vowels towards the front of a word is as disambiguating as of vowels towards its end.

The contributions of this paper are threefold. First, we document and demonstrate the shortcomings of existing methods for measuring the informativeness of individual segments in context, including the confound with the amount of conditional information discussed above. Second, we introduce three surprisal-based measures that control for this confound and enable comparison of word-initial versus -final positions in this respect: unigram, position-specific and cloze surprisal (see §3). Finally, we find robust evidence across many languages of stronger disambiguatory signals in word initial than word-final positions. Out of a total of 151 languages analysed across three separate collections, 82 of them present a higher cloze surprisal in word beginnings than in endings—with similar patterns arising with the other two measures.

2 Background and Related Work

Psycholinguistic evidence. Lexical access has long been a topic of interest for psycholinguists, leading to many distinct models being proposed for this process (Morton, 1969; Marcus, 1981; Marslen-Wilson, 1987). Far earlier, though, Bagley (1900) had already demonstrated that earlier segments in words were more important for word recognition than later segments; specifically, they found that, when exposed to words with word-initial or word-final consonant deletions, listeners found the word-initial deletions more disruptive. Fay and Cutler (1977) showed mispronunciations are more likely in word endings, while Bruner and O'Dowd (1958) showed that recognizing written words with flipped initial characters was harder than with word final ones—demonstrating that the initial part of the word was more "useful" for readers. More recently, Wedel et al. (2019a) found evidence in support of Houlihan (1975), showing neutralizing rules tend to target word endings more significantly than beginnings in both suffixing and prefixing languages.

Nooteboom (1981) investigated the ease of recovering lexical items from either word beginnings or endings, finding that people had an easier time recovering words from their beginnings. For this, he examined words for which the first and second halves each completely identified them in a large Dutch dictionary—controlling for both segments' length and uniqueness. Later on, though, Nooteboom and van der Vlugt (1988) showed this difference vanishes when priming people with the length of the word—proposing the difference comes not from how informative segments were, but from the difficulty in time aligning later segments in mental lexicons. Connine et al. (1993) also found no difference in priming effects with non-words that differed from real words in either word initial or medial positions, suggesting initial positions have no special status in word recognition.

Psycholinguistic evidence is key to understanding how lexical access works in human language processing, and can help us understand why lexicons may evolve to provide more disambiguatory signals earlier in words.[4] Given the incremental nature of human lexical processing, however, such evidence cannot provide direct evidence of the nature of the lexicon uninfluenced by incrementality.

Computational evidence. To the best of our knowledge, van Son and Pols (2003b,a) were the first to use computational methods coupled with an

[4]Note that there are many possible reasons why the effects we demonstrate in this paper may arise, from the demands of lexical access to constraints on articulation. We provide no evidence for any of the possible explanations, evolutionary or otherwise, just methods for measuring the effect.

information theoretic definition of informativeness to investigate this question. They showed that segments in the beginning of words carry most of a word's information, as measured by their contextual surprisal using a plug-in tree structured probabilistic estimator. Although assessing a less-biased sample of words than Nooteboom (1981),[5] this study is also limited to a single language (Dutch), hence cannot assess whether this is a general phenomenon or specific to that language.

Further, van Son and Pols (2003a,b) use absolute word positions in their analysis. Word length correlates strongly with frequency, hence while early positions are present in all words, later positions only exist for a much smaller sample of typically lower frequency words. Thus this comparison amounts to asking if later positions in longer and infrequent words have lower surprisal than earlier positions in all (frequent or infrequent) words. We analyse this confounding factor in §6.

Wedel et al. (2019b) and King and Wedel (2020) applied a methodology similar to that of van Son and Pols (2003a) to show, for many diverse languages, that more frequent words contain less informative segments in word initial positions, while less frequent types carry more informative ones. They further showed that segments in later word positions were less informative (given the previous ones) than average in rarer words. While controlling for length, King and Wedel (2020) also compared words' forward and backward uniqueness points—nodes in a trie from which only one leaf node can be reached, i.e., where the word is uniquely identified—showing they happened earlier in forward strings.

While these studies provide evidence from more diverse sets of languages, they follow van Son and Pols (2003a) in studying closed lexicons.[6] As we show in §3.3.1, the use of probabilistic trie models on a closed lexicon yields a trivial effect of higher informativity at word initial positions. Furthermore, such studies cannot account for out-of-vocabulary words (e.g., nonce, proper name or otherwise unknown words) or derivational morphology, which are key parts of lexical recognition. Lexical access is also somewhat robust to segmental misordering (Toscano et al., 2013) and sounds later in a word help determine the perception of earlier ones (Gwilliams et al., 2018). In contrast, a trie over a closed lexicon is deterministic. Beyond this, Luce (1986) showed in a corpus study that the probability of a word type being uniquely identifiable before its last segment was only 41%—and 19% of types were identified only by the end of word, being proper prefixes of other words, such as *cat* and *cats*. They conclude that uniqueness point statistics may only be useful for long word analysis.

In Pimentel et al. (2020), we analysed several languages' phonotactic distributions, focusing on presenting a trade-off between phonotactic entropy and word length across languages. As a control experiment we analysed the correlation between a segment's surprisal and its word position across 106 languages. We did not control for word length and did not run per-language experiments, though—so we could have just been capturing the effect that later positions will mostly be present in languages with longer words (which, as we find, have lower information on average).[7]

While this last work avoids many of the issues raised earlier in this section, it fails to control the key confound mentioned earlier: it relies on left-to-right conditional probabilities to calculate surprisal. Thus segments early in the word have less conditional information and hence are generally of lower probability—a trivial effect that does not indicate a segment's disambiguatory signal strength.

3 Measures of Disambiguatory Strength

3.1 A Lexicon Generating Distribution

In this work, instead of the lexicon itself, we investigate the probability distribution from which it is sampled. The distribution is unobserved, but we can get glimpses of it via the sampled lexicon:

$$\left\{\mathbf{w}^{(n)}\right\}_{n=1}^{N} \sim p(\mathbf{w}) = \prod_{t=1}^{|\mathbf{w}|} p(w_t \mid \mathbf{w}_{<t}) \qquad (1)$$

The distribution $p(\mathbf{w})$ is defined over the entire space of possible phonological wordforms $\mathbf{w} \in \Sigma^*$, where Σ is a language-specific alphabet and the operator * indicates its Kleene closure.[8] This dis-

[5] Nooteboom (1981) looked at words completely identifiable by both their first and second halves in a large Dutch dictionary—this resulted in a study with only 14 words.

[6] The closed lexicon assumption is incorporated implicitly in the probabilistic trie models used by van Son and Pols (2003a,b) and King and Wedel (2020)—i.e. they assign zero probability to any form not in their training sets—and in the uniqueness point analysis of King and Wedel (2020).

[7] We note this issue only applies to the control experiment, and has no bearing on the key findings of that paper.

[8] We pad all strings with the end-of-word (EOW) symbol. For simplicity, we assume the alphabet includes EOW throughout the rest of the paper.

tribution should assign high probability to likely wordforms (attested or not) and low probability to unlikely ones. Using Chomsky and Halle's (1965) classic example from English, *brick* (attested) and *blick* (unattested) would have high probability, whereas **bnick* (unattested) would have a low probability.

3.2 Entropy and Conditional Entropy

Shannon's entropy is a measure of how much information a random variable contains. Consider a segment w_t at word position t, which is a value of the random variable W_t. The average information (surprisal) relayed per segment is:

$$\mathrm{H}(W_t) \equiv \sum_{w_t \in \Sigma} p(w_t) \log \frac{1}{p(w_t)} \tag{2}$$

A random variable is maximally entropic if it is a uniform distribution, in which case $\mathrm{H}(W_t) = \log(|\Sigma|)$. Conditional entropy measures how much information the knowledge of a variable conveys, given some previous knowledge. The average information transmitted per segment, given the previous ones in a word, is

$$\mathrm{H}(W_t \,|W_{<t}) \equiv \sum_{\mathbf{w}_{\leq t} \in \Sigma^*} p(\mathbf{w}_{\leq t}) \log \frac{1}{p(w_t \mid \mathbf{w}_{<t})} \tag{3}$$

where $\mathbf{w}_{\leq t} = \mathbf{w}_{<t} \circ w_t$. We note the conditional entropy is always smaller or equal to the entropy, i.e. $\mathrm{H}(W_t \mid W_{<t}) \leq \mathrm{H}(W_t)$.

3.3 Plug-in Estimators, Context Size, and Disambiguatory Strength

Our criticism of previous work investigating the disambiguatory strength of word-initial vs. word-final segments can be mainly divided in two parts: (i) the use of maximum likelihood plug-in estimators of the conditional entropy, by e.g. van Son and Pols (2003b); (ii) the use of left-to-right conditional entropy in itself, by all previous information-theoretic work in this vein.

3.3.1 A Critique of van Son and Pols (2003b)

We present a *reductio ad absurdum* which shows that van Son and Pols's (2003b) method will lead to the conclusion that word-initial segments are more informative even if all segments were equally entropic and sampled independently—a nonsensical finding. Accordingly, assume the probability distribution $p(w_t \mid \mathbf{w}_{<t})$, from which each segment in a word is sampled, was independent, e.g. define

$$\hat{p}(\mathbf{w}) = \prod_{t=1}^{|\mathbf{w}|} \hat{p}(w_t \mid \mathbf{w}_{<t}) = \prod_{t=1}^{|\mathbf{w}|} \hat{p}(w_t) \tag{4}$$

Assume now that a large, but finite, lexicon is sampled from it $\left\{\hat{\mathbf{w}}^{(n)}\right\}_{n=1}^{N} \sim \hat{p}(\mathbf{w})$. Further consider modelling this sampled lexicon with a probabilistic trie structure, similarly to what was done by van Son and Pols (2003a,b),[9] i.e.

$$q_{\text{trie}}(w_t \mid \mathbf{w}_{<t}) = \frac{\text{count}(w_t, w_{t-1}, \ldots, w_0)}{\text{count}(w_{t-1}, \ldots, w_0)} \tag{5}$$

where w_0 is the beginning-of-word symbol. Such a model uses all N words to approximate the distribution of the first segment—i.e. $\text{count}(w_0) = N$. Yet after $t-1$ segments, an exponentially smaller sample is used to capture the distribution—i.e. $\mathbb{E}[\text{count}(w_{t-1}, \ldots, w_0)] = N/|\Sigma|^{t-1}$. Using this model as a plug-in estimator of the entropy will lead to negatively biased estimates, where the error is approximately (Basharin, 1959):

$$\begin{aligned} \mathrm{H}(W_t \mid W_{t-1}) - \mathbb{E}\left[\hat{\mathrm{H}}\right] &\approx \frac{(|\Sigma| - 1)\, \log e}{\text{count}(w_{t-1}, \ldots, w_0)} \\ &\approx \frac{|\Sigma|^{t-1}(|\Sigma| - 1)\, \log e}{N} \end{aligned} \tag{6}$$

where $\hat{\mathrm{H}}$ is a plug-in estimate of the entropy. The error grows exponentially in t due to the $|\Sigma|^{t-1}$ factor. However, by assumption, $\mathrm{H}(W_t \mid W_{t-1})$ is constant—we have equally entropic and independent segments. Thus, the only way for this difference to increase is for the second term to decrease as a function of t. It follows that the estimated cross-entropies decrease as a function of t due to a methodological technicality. Indeed, in the extreme case, every position after a word's uniqueness point would be estimated to have zero entropy. Thus, van Son and Pols's (2003a) method only reveals a trivial effect.

3.3.2 Conditional Entropy and Context Size

As previously mentioned, the conditional entropy measures how much information the knowledge of a variable conveys, given some previous information, and it is always smaller or equal to the entropy. For this reason, relying on left-to-right conditional

[9] This is in fact a simplification of van Son and Pols's (2003a) model, which in practice uses Katz smoothing.

entropies to estimate the strength of disambiguatory signals yields straightforward results; the availability of larger conditioning contexts in a word's final segments will naturally reduce its conditional entropy. This will negatively skew the estimated informativeness of the later parts of a word.

$$\mathrm{H}(W_t) \geq \mathrm{H}(W_t \mid W_{t-1}) \geq \mathrm{H}(W_t \mid W_{<t}) \quad (7)$$

This effect can also be easily demonstrated by the symmetrical nature of mutual information (MI), where the MI is defined as:

$$\begin{aligned}\mathrm{MI}(W_t; W_{t-1}) &= \mathrm{H}(W_t) - \mathrm{H}(W_t \mid W_{t-1}) \\ &= \mathrm{H}(W_{t-1}) - \mathrm{H}(W_{t-1} \mid W_t) \\ &= \mathrm{MI}(W_{t-1}; W_t) \end{aligned} \quad (8)$$

If we assume both segments had the same unconditional entropy, i.e. $\mathrm{H}(W_t) = \mathrm{H}(W_{t-1})$, then using left-to-right conditional entropies would suggest the later segment was less informative, while right-to-left conditioning would imply the opposite. Nonetheless, both their contextual and uncontextual disambiguatory strength would in fact be the same, if we estimated it with equal-sized contexts:

$$\begin{aligned}\mathrm{H}(W_t) = \mathrm{H}(W_{t-1}) \implies \\ \mathrm{H}(W_t \mid W_{t-1}) = \mathrm{H}(W_{t-1} \mid W_t)\end{aligned} \quad (9)$$

3.4 Cross-Entropy and Entropy

As mentioned above, the distribution $p(\mathbf{w})$ is not directly observable. We can, however, approximate it using character-level language models $p_\theta(\mathbf{w})$. We are interested in the entropy of variable W_t, as a proxy we measure its cross-entropy

$$\mathrm{H}_\theta(W_t) \equiv \sum_{w_t \in \Sigma} p(w_t) \underbrace{\log \frac{1}{p_\theta(w_t)}}_{\text{surprisal}} \quad (10)$$

where the surprisal is the information provided by a single segment instance w_t. The cross-entropy is an upper bound on the entropy, i.e. $\mathrm{H}(W_t) \leq \mathrm{H}_\theta(W_t)$, with their difference being the Kullback–Leibler (KL) divergence between both distributions. Since the KL-divergence is always positive, this upper-bound holds. Furthermore, the closer p_θ is to the true distribution p, the smaller the divergence is, and the tighter this bound. As such, the better our model is at estimating the true distribution, the better our estimates of the entropy will be.

Calculating eq. (10) still requires knowledge of the true p. We overcome this limitation by empirically estimating it on a held out part of the lexicon

$$\mathrm{H}_\theta(W_t) \approx \frac{1}{N} \sum_{n=1}^{N} \log \frac{1}{p_\theta(w_t^{(n)})} \quad (11)$$

3.5 Earlier vs. Later Word Entropy

For the remainder of this work, we will discuss information in terms of surprisal, since the entropy is its expected value. We analyse the distribution of disambiguatory information across word positions via three distinct measures—all of which control for the amount of conditioning per position:

- **Unigram Surprisal** $\mathrm{H}_\theta(W_t)$**:** the surprisal of individual segments.
- **Cloze Surprisal** $\mathrm{H}_\theta(W_t \mid W_{\neq t})$**:** surprisal of a segment given all others in the same word.
- **Position-Specific Surprisal** $\mathrm{H}_\theta(W_t \mid T = t, |W|)$: the surprisal of individual segments given their position in the wordform and the word's length.

The unigram surprisal captures the information provided by each segment when considering no context; while the cloze surprisal represents the information provided by a segment when one already knows the rest of the word. The position-specific surprisal represents a mid way between both, conditioning each segment only on its position and the word's length—being inspired by Nooteboom and van der Vlugt's (1988) experiments. These three measures of information control for the context size considered at each position, being thus better for an investigation of disambiguatory strength.

We used an unigram model (see §4) to estimate the unigram surprisal, and transformers (Vaswani et al., 2017) for cloze and position-specific surprisals. We also use the LSTM (Long-Short Term Memory, Hochreiter and Schmidhuber, 1997) model from Pimentel et al. (2020) for two other entropy measures which do not control for the amount of conditional information:

- **Forward Surprisal** $\mathrm{H}_\theta(W_t \mid W_{<t})$**:** the surprisal of a segment given the previous ones.
- **Backward Surprisal** $\mathrm{H}_\theta(W_t \mid W_{>t})$**:** the surprisal of a segment given the future ones.

We include the beginning- and end-of-word symbols in the forward and backward surprisal analysis, respectively, following previous work (Wedel

et al., 2019b; Pimentel et al., 2020; King and Wedel, 2020). However, we ignore them in the unigram, position-specific and cloze surprisal analyses. Position-specific and cloze surprisal are given information about word length, hence these symbols are unambiguously predictable. We analyse the impact of these symbols in §6.

4 Character-Level Language Models

In this paper, we make use of character-level language models to model the probability distributions p_θ and approximate the relevant cross-entropies.

Unigram. This might be the simplest language model still in use in Natural Language Processing. We use its Laplace-smoothed variant

$$p_\theta(w_t) = \frac{\text{count}(w_t) + 1}{\sum_{c' \in \Sigma} \text{count}(c') + |\Sigma|} \quad (12)$$

LSTM. This architecture is the state-of-the-art for character-level language modelling (Melis et al., 2020). Given a sequence of segments $\mathbf{w} \in \Sigma^*$, we use one hot lookup embeddings to transform each of them into a vector $\mathbf{z}_t \in \mathbb{R}^d$. We then feed these vectors into a k-layer LSTM

$$\mathbf{h}_t = \text{LSTM}(\mathbf{z}_{t-1}, \mathbf{h}_{t-1}) \quad (13)$$

where $\mathbf{h} \in \mathbb{R}^d$, $\mathbf{h}_0$ is a vector with all zeros and w_0 is the beginning-of-word symbol. We then linearly transform these vectors before feeding them into a softmax non-linearity to obtain the distribution

$$p_\theta(w_t \mid \mathbf{w}_{<t}) = \text{softmax}(W\mathbf{h}_t + b) \quad (14)$$

in this equation, $W \in \mathbb{R}^{|\Sigma| \times d}$ is a weight matrix and $b \in \mathbb{R}^{|\Sigma|}$ a bias vector.

Backward LSTM. To get the backward surprisals we use models with the same architecture, but reverse all strings before feeding them to the models. As such, we get the similar equations

$$\mathbf{h}_t = \text{LSTM}(\mathbf{z}_{t+1}, \mathbf{h}_{t+1}) \quad (15)$$

$$p_\theta(w_t \mid \mathbf{w}_{>t}) = \text{softmax}(W\mathbf{h}_t + b) \quad (16)$$

Transformer. Transformers allow a segment to be conditioned on both future and previous symbols. Our implementation starts similar to the LSTM one, getting embedding vectors $\mathbf{z}_t$ for each segment in the string $\mathbf{w} \in \Sigma^*$, except that we replace segment w_t with a **MASK** symbol. We then feed these vectors through k multi-headed self-attention layers, as defined by Vaswani et al. (2017). Finally, the representations from the last layer are linearly transformed and fed into a softmax

$$p_\theta(w_t \mid \mathbf{w}_{\neq t}) = \text{softmax}(W\mathbf{h}_t + b) \quad (17)$$

Position-Specific Transformer. To get position-specific surprisal values, we again use a transformer architecture, but instead of replacing a single segment with a **MASK** symbol, we replace all of them. This is equivalent to conditioning each segment's distribution on its position and the word length—i.e., estimating $p_\theta(w_t \mid t, |\mathbf{w}|)$.

5 Data

In order to estimate redundancy and informativeness of segments we use three different datasets, each with its own pros and cons. We focus on types instead of tokens—i.e., the datasets consist of lexicons—for a few different reasons. First, it is easier to get reliable samples of types than tokens for a language, specially low-resource ones. Second, it is a well known result that token frequency correlates with both word length (Zipf, 1949) and phonotactic probability (Mahowald et al., 2018; Meylan and Griffiths, 2017), so that would be a strong confound in the results. Third, morphology is more easily modeled at the type level than at token level (Goldwater et al., 2011).[10]

CELEX (Baayen et al., 2015) allows us to experiment exclusively on monomorphemic words, but covers only three closely related languages. It contains both morphological and phonetic annotations for a large number of words in English, Dutch and German. We follow Dautriche et al. (2017) in using only words labeled as monomorphemic in our study, leaving us with 4,810 words in German, 6,206 words in English and 7,045 words in Dutch.

NorthEuraLex (Dellert et al., 2019) spans 107 languages from 21 language families in a unified IPA format. This database is composed of concept aligned word lists for these languages, containing 1016 concepts, each of them translated in most languages. However, most of these languages are from Eurasia, hence the collection lacks the typological diversity we would ideally like.

Wikipedia allows us to investigate a broader and more diverse set of languages, but has no phonetic information (only graphemes) and lexicons

[10] For each of the analysed datasets, we use 80% of the word types for training, with the rest being equally split between development and test sets; only test set surprisal and cross-entropies are used in our analysis.

extracted from it may be "contaminated" with foreign words. We fetch the Wikipedia for a set of 41 diverse languages,[11] and tokenise their text using language-specific tokenisers from spaCy (Honnibal and Montani, 2017). When a language-specific tokeniser was not available, we used a multilingual one. We then filtered all non-word tokens—by removing the ones with any symbol not in the language's scripts—and kept only the 10,000 most frequent types in each language.

6 Experiments and Results

Forward Surprisal. We first replicate the results from van Son and Pols (2003a,b), Wedel et al. (2019b), and Pimentel et al. (2020), which show that surprisal decreases with as the words position advances. On average, forward surprisal, i.e. $\mathrm{H}_\theta(W_t \mid W_{<t})$, could decrease for two reasons: (i) words indeed front-load disambiguatory signals; or (ii) the trivial fact that conditioning reduces entropy. For each word, we first get the forward surprisal for each segment in it. We then group surprisal values in two groups: word initial (when they are in the first half of its word) and final (when in the second half), ignoring mid positions in words with uneven lengths; we average these initial and final surprisals per word, getting a single value of each per word. This way we compare earlier vs. later word positions while ignoring any length effect—words with all lengths will possess segments in both groups. For each analysed language, we then use permutation tests (permuting word initial and final surprisals) to evaluate if one group is statistically larger than the other—using 100,000 permutations. All but one language in three analysed datasets had significantly larger surprisal in word initial positions[12]—the exception being Abkhaz in NorthEuraLex. These results can be seen in Tab. 1 and in Fig. 2 (left).

Backward Surprisal. If the result for forward surprisal is largely due to the amount of conditional information, then reversing the strings should lead to a roughly opposite effect. With this in mind, for each language, we again bin surprisals in word initial vs. final position, but now we evaluate languages using backward surprisal, i.e., $\mathrm{H}_\theta(W_t \mid W_{>t})$.[13] When using backward surprisal, many of the analysed languages have significantly higher surprisals in word final positions (see Tab. 1 and the right graph in Fig. 2). However, 11 languages in the NorthEuraLex dataset still have higher word initial surprisals, suggesting that initial positions in these languages are indeed largely more informative than final ones.[14] There does seem to be a large effect of the amount of conditional information and also some lexical effect of front-loading disambiguatory signals, however it is difficult to determine if there are cross-linguistic tendencies with these measures.

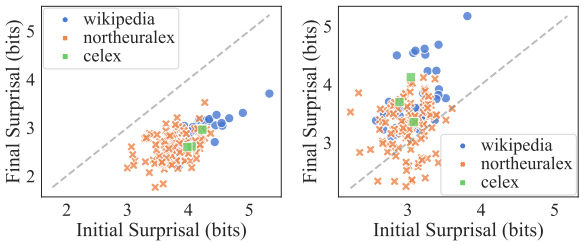

Figure 2: Word initial vs. final surprisals with: (left) Forward; (right) Backward.

Unigram Surprisal. To control for the conditioning aspect of the question: *do words front-load their disambiguatory signals?*, we can look at unigram surprisal $\mathrm{H}_\theta(W_t)$. This value tells us how uncommon the segments that appear in a certain position are, when analysed in isolation from the rest of the word—uncommon segments are more informative and provide stronger signal for disambiguation. In NorthEuraLex, 71 of the languages have significantly higher informativity in word beginnings than in endings—nonetheless, one language (Kildin Saami) has higher surprisals in word endings. In CELEX, Dutch and German have higher surprisals in initial positions, but English does not. And in Wikipedia, all languages but Hebrew and Bengali have higher surprisal in initial positions—with Bengali having higher surprisal in word endings. This experiment suggests that indeed most languages are biased towards providing stronger disambiguatory signals in word beginnings, even when we control for the

[11] These languages were: af, ak, ar, bg, bn, chr, de, el, en, es, et, eu, fa, fi, ga, gn, haw, he, hi, hu, id, is, it, kn, lt, mr, no, nv, pl, pt, ru, sn, sw, ta, te, th, tl, tr, tt, ur, zu.

[12] All statistical significance results in this work have been corrected for multiple tests with Benjamini and Hochberg (1995) corrections and use a confidence value of $p < 0.01$.

[13] We note that King and Wedel (2020) also used backward surprisal, although with a different objective in mind. In one of their experiments, they presented aggregate results of a comparison between the forward and backward surprisal.

[14] We also ran the same experiments with a probabilistic trie model like the ones used in van Son and Pols (2003b) and Wedel et al. (2019b), which showed an even stronger result reversal when using backward surprisal.

Dataset	# Languages	Surprisal				
		Forward	Backward	Unigram	Position-Specific	Cloze
CELEX	3	3 | 0	0 | 3	2 | 0	2 | 1	2 | 1
NorthEuraLex	107	106 | 0	11 | 31	71 | 1	24 | 4	45 | 1
Wikipedia	41	41 | 0	0 | 39	39 | 1	31 | 1	35 | 2

Table 1: Number of languages in the analysed datasets with significantly larger surprisals in initial | final positions.

amount of conditional information. Nonetheless, this is not a universal characteristic which all languages share and two analysed languages even had a statistically significant inverse effect.

Position-Specific Surprisal. While cloze surprisal makes explicit the non-redundant informativity a segment conveys, unigram surprisal analyses the same segments in isolation. Position-specific surprisal provides a midway analysis, incorporating the position as some previously-specified knowledge, but not conditioning on the other segments in the word. The position-specific surprisal is inspired by Nooteboom and van der Vlugt (1988) experiments, which prime individuals on word length and position. As can be seen in Tab. 1, position-specific surprisal again seems to favour initial positions over final, but only slightly. Interestingly, most languages present no significant difference and some the inverse effect (i.e. higher surprisal in final positions).

Position-specific Unigram models. To better understand the differences between the unigram and position-specific surprisal results, we trained position-specific unigram models—which count each segment's frequency per position—and then calculated their Kullback–Leibler (KL) divergence per position with the traditional unigram

$$\mathrm{KL}(p(w_t \mid t) \,||\, p(w_t)) \tag{18}$$
$$= \sum_{w_t \in \Sigma} p(w_t \mid t) \log \frac{p(w_t \mid t)}{p(w_t)}$$

We compare these KL divergences and find that, for all but four languages, the KL is largest in either the first or second segment positions.[15] This suggests that one of the reasons for higher unigram surprisal in initial positions is that the first two segments usually differ from the rest of the positions, potentially serving as markers for word segmentation.

[15] We use Laplacian smoothing in the position-specific unigrams and constrain the analysis to positions which appear in at least 75% of the analysed words in that language.

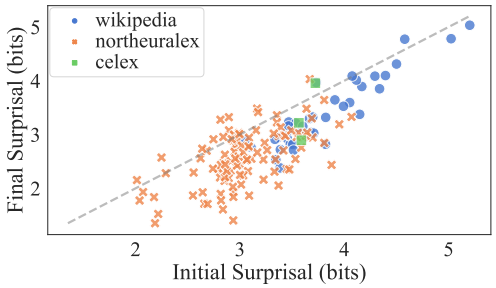

Figure 3: Word initial vs. final cloze surprisals.

Cloze Surprisal. When we condition a segment on all others in the same word, we measure how much uncertainty is left about that individual segment when considering everything else, or, in other words, how much information is passed only by that segment non-redundantly. Word initial surprisal is higher in most analysed languages (see Tab. 1). Nonetheless, two languages in Wikipedia, Thai and Bengali, have significantly higher surprisal in their final segments—while English in CELEX and Hungarian in NorthEuraLex also present this same inverse effect. Front-loading disambiguatory information, thus, is not established to be the linguistic universal it is believed to be, with only roughly half the analysed languages showing this property when we control for morphology (CELEX and NorthEuraLex). Fig. 3 plots the results for all languages analysed.

When we compare these results, we find an interesting pattern. Morphology seems to reduce non-redundant (cloze) information later in the words—while only half of the languages had significant surprisals in CELEX (which consists of monomorphemic words) and NorthEuraLex (base forms), most languages were significant in Wikipedia. Furthermore, English and Hungarian had significantly higher surprisals in word endings in CELEX and NorthEuraLex, while the opposite trend in Wikipedia—this is consistent with the fact that suffix morphemes are present in more types than word roots are, so morphology would make word endings less surprising.

	EOW	Non-EOW
Forward	1.14	3.55
Backward	0.89	3.61
Unigram	2.75	4.90
Position-specific	0.00	4.36
Cloze	0.00	3.23

Table 2: Average surprisal (in bits) of EOW vs. non-EOW segments averaged over all datasets.

Length as a Confounding Effect. We evaluate the impact of length as a confounding effect on previous methodologies. As mentioned in §2, by directly analysing surprisal–position pairs (as opposed to binning word initial vs. final positions), previous work confounds position and word length—i.e., only long words will have later word positions. In this study, we analyse forward surprisal–length pairs; instead of pairing a segment's surprisal with its position, we pair it with its word length. We then get the slope formed by a linear regression between these pairs of values and test for its significance per language by using a permutation test, in which we shuffle surprisal–length values. On the three datasets, all languages have statistically significant negative slopes, meaning long words have smaller surprisals on average than shorter ones.[16] A caveat, though, is that now we are confounding position into our length analysis. Constraining our analysis only to the first two segments in each word, we still find the same effect—though now one language (Hebrew) in Wikipedia and seven in NorthEuraLex are not significant. We can thus conclude that longer words have smaller surprisal values than shorter ones, even when controlling for the same word positions. This implies that directly using surprisal–position pairs for such an analysis is not ideal.

The Effect of End of Word in Surprisal. The end-of-word (EOW) symbol is a special "segment" which symbolises the end of a string. It is necessary when modelling the probability distribution over strings $\mathbf{w} \in \Sigma^*$, to guarantee that the overall distribution sums to 1. Nonetheless, it is expected to behave in a different way from other segments. If a speaker wants to reduce their production effort, although changing from one phone to another may help, the most efficient way is usually just ending the string earlier. Furthermore, since all realisable strings must eventually end, it will be present in all words, making it a very frequent symbol—in fact, Tab. 2 shows its average surprisal is much lower than that of other segments. As such, it is only natural it should be analysed on its own, separately from other segments. Through the same logic, other segments should also be analysed separately from EOW—or else, lower word final surprisals may be due to this symbol alone. As such, we analyse the surprisal of LSTM "language models" without the EOW symbol here.[17]

	EOW			No EOW		
	Initial	Final	Diff (%)	Initial	Final	Diff (%)
Forward	3.85	2.65	31.1 %	3.83	3.00	21.6 %
Backward	3.02	3.40	-11.3 %	3.63	3.39	6.7 %
Unigram	-	-	-	4.85	4.40	9.3 %
Position	-	-	-	4.36	4.17	4.3 %
Cloze	-	-	-	3.26	2.81	13.9 %

Table 3: Average surprisal per segment in word initial and final positions with and without EOW symbols.

Unsurprisingly, Tab. 3 shows the difference between word initial and final positions is considerably reduced when we remove the EOW symbol from the forward surprisal analysis. Surprisingly, we see that when we remove the beginning-of-word from the backward surprisal analysis, instead of a larger word final surprisal, we get a larger word initial value—even though we are still conditioning the models right-to-left. This result further supports the hypothesis that the disambiguatory signals are on average stronger in word initial positions.

7 Conclusions

In this work, we analysed the distribution of disambiguatory information in word positions. We present an in-depth critique of previous work, showing several confounding effects in their analysis. We then proposed the use of three new methods which corrected for these biases—namely unigram, position-specific and cloze surprisal. These models controlled for the amount of conditional information across word positions, allowing for an unbiased analysis of the lexicon. Using these models we show that the lexicons of most languages indeed front-load their disambiguatory signals. This effect, though, is not universal and the difference in disambiguatory information between word initial and final positions is much lower than previously estimated—ranging from 4% to 14%, depending on the used metric, instead of 31%.

[16] King and Wedel (2020) indeed present a similar correlation in their Figure 2.

[17] To be more precise, we actually ignore the beginning-of-word symbol when estimating backward surprisal.

References

R. H. Baayen, R. Piepenbrock, and L. Gulikers. 2015. CELEX2 LDC96L14.

William Chandler Bagley. 1900. The apperception of the spoken sentence: A study in the psychology of language. *The American Journal of Psychology*, 12(1):80–130.

Georgij P. Basharin. 1959. On a statistical estimate for the entropy of a sequence of independent random variables. *Theory of Probability & Its Applications*, 4(3):333–336.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Jerome S. Bruner and Donald O'Dowd. 1958. A note on the informativeness of parts of words. *Language and Speech*, 1(2):98–101.

Noam Chomsky and Morris Halle. 1965. Some controversial questions in phonological theory. *Journal of Linguistics*, 1(2):97–138.

Cynthia M. Connine, Dawn G. Blasko, and Debra Titone. 1993. Do the beginnings of spoken words have a special status in auditory word recognition? *Journal of Memory and Language*, 32(2):193–210.

Isabelle Dautriche, Kyle Mahowald, Edward Gibson, Anne Christophe, and Steven T. Piantadosi. 2017. Words cluster phonetically beyond phonotactic regularities. *Cognition*, 163:128–145.

Johannes Dellert, Thora Daneyko, Alla Münch, Alina Ladygina, Armin Buch, Natalie Clarius, Ilja Grigorjew, Mohamed Balabel, Hizniye Isabella Boga, Zalina Baysarova, et al. 2019. NorthEuraLex: A wide-coverage lexical database of Northern Eurasia. *Language Resources and Evaluation*, pages 1–29.

David Fay and Anne Cutler. 1977. Malapropisms and the structure of the mental lexicon. *Linguistic Inquiry*, 8(3):505–520.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2011. Producing power-law distributions and damping word frequencies with two-stage language models. *Journal of Machine Learning Research*, 12(Jul):2335–2382.

Laura Gwilliams, Tal Linzen, David Poeppel, and Alec Marantz. 2018. In spoken word recognition, the future predicts the past. *Journal of Neuroscience*, 38(35):7585–7599.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Kathleen Houlihan. 1975. *The Role of Word Boundary in Phonological Processes*. Ph.D. thesis, University of Texas at Austin.

Adam King and Andrew Wedel. 2020. Greater early disambiguating information for less-probable words: The lexicon is shaped by incremental processing. *Open Mind*, pages 1–12.

Paul A. Luce. 1986. A computational analysis of uniqueness points in auditory word recognition. *Perception & Psychophysics*, 39(3):155–158.

Kyle Mahowald, Isabelle Dautriche, Edward Gibson, and Steven T. Piantadosi. 2018. Word forms are structured for efficient use. *Cognitive Science*, 42(8):3116–3134.

Stephen Michael Marcus. 1981. ERIS-context sensitive coding in speech perception. *Journal of Phonetics*, 9(2):197–220.

William D. Marslen-Wilson. 1987. Functional parallelism in spoken word-recognition. *Cognition*, 25(1-2):71–102.

Gábor Melis, Tomáš Kočiský, and Phil Blunsom. 2020. Mogrifier LSTM. In *International Conference on Learning Representations*.

Stephan C. Meylan and Thomas L. Griffiths. 2017. Word forms—not just their lengths—are optimized for efficient communication. *arXiv preprint arXiv:1703.01694*.

John Morton. 1969. Interaction of information in word recognition. *Psychological Review*, 76(2):165.

S. G. Nooteboom and M. J. van der Vlugt. 1988. A search for a word-beginning superiority effect. *The Journal of the Acoustical Society of America*, 84(6):2018–2032.

Sieb G. Nooteboom. 1981. Lexical retrieval from fragments of spoken words: Beginnings vs endings. *Journal of Phonetics*, 9(4):407–424.

Tiago Pimentel, Brian Roark, and Ryan Cotterell. 2020. Phonotactic complexity and its trade-offs. *Transactions of the Association for Computational Linguistics*, 8:1–18.

Rob J. J. H. van Son and Louis C. W. Pols. 2003a. Information structure and efficiency in speech production. In *Eighth European Conference on Speech Communication and Technology*.

Rob J. J. H. van Son and Louis C.W. Pols. 2003b. How efficient is speech? In *Proceedings of the Institute of Phonetic Sciences*, volume 25, pages 171–184.

Joseph C. Toscano, Nathaniel D. Anderson, and Bob McMurray. 2013. Reconsidering the role of temporal order in spoken word recognition. *Psychonomic Bulletin & Review*, 20(5):981–987.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Andrew Wedel, Adam Ussishkin, and Adam King. 2019a. Crosslinguistic evidence for a strong statistical universal: Phonological neutralization targets word-ends over beginnings. *Language*, 95(4):e428–e446.

Andrew Wedel, Adam Ussishkin, and Adam King. 2019b. Incremental word processing influences the evolution of phonotactic patterns. *Folia Linguistica*, 40(1):231–248.

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press.