On␣the␣Proper␣Treatment␣of␣Tokenization␣in␣Psycholinguistics*

Mario Giulianelli mario.giulianelli@inf.ethz.ch
Luca Malagutti luca.malagutti@inf.ethz.ch
Juan Luis Gastaldi juan.luis.gastaldi@inf.ethz.ch
Brian DuSell brian.dusell@inf.ethz.ch
Tim Vieira tim.f.vieira@gmail.com
Ryan Cotterell rcotterell@inf.ethz.ch
ETH zürich

Abstract

Language models are widely used in computational psycholinguistics to test theories that relate the negative log probability (the surprisal) of a region of interest (a substring of characters) under a language model to its cognitive cost experienced by readers, as operationalized, for example, by gaze duration on the region. However, the application of modern language models to psycholinguistic studies is complicated by the practice of using tokenization as an intermediate step in training a model. Doing so results in a language model over *token* strings rather than one over character strings. Vexingly, regions of interest are generally misaligned with these token strings. The paper argues that token-level language models should be (approximately) marginalized into character-level language models before they are used in psycholinguistic studies to compute the surprisal of a region of interest; then, the marginalized character-level language model can be used to compute the surprisal of an arbitrary character substring, which we term a focal area, that the experimenter may wish to use as a predictor. Our proposal of marginalizing a token-level model into a character-level one solves this misalignment issue independently of the tokenization scheme. Empirically, we discover various focal areas whose surprisal is a better psychometric predictor than the surprisal of the region of interest itself.

https://github.com/rycolab/psycho-toke

1 Introduction

Language models (LMs) have become a popular tool for computational psycho- and neurolinguists, who use them to instantiate and test executable linguistic theories (Futrell et al., 2019; Schrimpf et al., 2021; Baroni, 2022). While there are various ways to operationalize theories of language processing using LMs (e.g., Caucheteux et al., 2023; Hoover et al., 2023; Giulianelli et al., 2024a,b; Frank, 2024), their most common use in this area is to produce a specific quantity of interest: the probability of a character string given a context, conventionally mapped to the string's negative log probability, also known as its surprisal. Surprisal has been posited to correlate with the difficulty incurred by a comprehender processing that word (Hale, 2001; Levy, 2008). And, estimates of surprisal obtained from neural LMs have proven to be significant predictors of a broad range of psycholinguistic measurements of processing difficulty (Goodkind and Bicknell, 2018; Shain et al., 2020; Wilcox et al., 2023; Michaelov et al., 2024, *inter alia*), providing ample empirical support for the role of surprisal in psycholinguistics theory, in addition to other predictors, e.g., word length and unigram surprisal.

Modern language models do not provide direct access to contextual probabilities at the character level. Instead, they provide a distribution over strings of *tokens*, supercharacter units that are induced during a pre-processing step, e.g., by the byte-pair encoding tokenizer (BPE; Sennrich et al., 2016). However, in computational psycholinguistics, it is often necessary to compute the surprisal of an arbitrary character substring of the stimulus. For example, the predictability of the first three characters of a word, which can be viewed parafoveally, is known to be an important predictor of whether the region is going to be skipped by the reader (Rayner et al., 1982; Blanchard et al., 1989; Rayner et al., 2011, *inter alia*). And, to properly model a region's skip rate under surprisal theory, we should use the surprisal of the first three characters, rather than the surprisal of the entire region, as a predictor. A complication arises when the first three characters of the region do not correspond to a token, and we have to marginalize over all token strings that start with those three characters.

Because performing such a marginalization over token strings is computationally expensive and therefore requires approximation (Cao and Rimell, 2021; Chirkova et al., 2023; Vieira et al., 2024), it has yet to be adopted among computational psy-

*The gray bars above each character of the title are proportional to its character-level surprisal under GPT-2.

cholinguists. Indeed, the literature lacks guiding principles for how computational psycholinguists should properly apply token-level language models to the field's inherently character-level problems. For instance, a number of recent studies have blurred the line between algorithmic and linguistic concerns (Oh et al., 2021; Nair and Resnik, 2023; Giulianelli et al., 2022; Beinborn and Pinter, 2023; Oh and Schuler, 2024; Pimentel and Meister, 2024; Yee et al., 2024, *inter alia*). In particular, two recent studies (Oh and Schuler, 2024; Pimentel and Meister, 2024) suggest it is important to include a word's trailing whitespace in the computation of the word's surprisal to account for the mismatch between tokens and words.[1] Both are motivated by a peculiarity[2] of the BPE tokenizer itself rather than a deeper appeal to linguistic theory: Examining standard practice in experimental eye-tracking research reveals that regions of interest are typically defined to include the *preceding* whitespace rather than the trailing one (Rayner, 1979; Pollatsek and Rayner, 1982; McConkie et al., 1988, *inter alia*). We attribute the focus on trailing whitespace to the obfuscated relationship between token-level and character-level surprisal.

Our paper clarifies the proper role of tokenization in surprisal theory: We take the stance that tokenization is irrelevant. First, we note psycholinguistic stimuli should be viewed as *character* strings rather than *token* strings. This follows from the observation that human linguists construct psycholinguistic stimuli without regard for any given LM's token alphabet. Then, as is common, the stimuli are divided into regions of interest (character substrings of the stimulus) for which the experimenter gathers a psycholinguistic measurement. Note that treating regions of interest as character strings does not prevent the experimenter from claiming they represent morphemes, words, or phrases, all of which are built from characters in text form. Finally, the experimenter collects the psycholinguistic measurements associated with each region. None of the above steps makes use of tokenization schemes, completing our argument.

A problem with respect to tokenization first arises when the experimenter *analyzes* the measurements they collected by means of a language and the regions of interest they decomposed their stimuli into do not neatly align with a string of tokens, or when the experimenter wishes to compute the surprisal of a sub- or super-string of the region. In this paper, we contend that the solution to this problem is to marginalize the token-level LM into a character one before using it to compute surprisal. Moreover, on this view, such marginalization constitutes an algorithmic problem, but not a theoretical one for psycholinguistics.

Because computational psycholinguists have yet to convert pretrained token-level LMs into character-level ones through marginalization, we contend that they have yet to explore many potentially effective surprisal-based predictors. In our experimental section, we test the degree to which the choice of various substrings of the stimulus that overlaps the regions of interest, which we term *focal areas*, affect recent empirical findings in surprisal theory. We perform such an exploration by means of Vieira et al.'s (2024) approximate marginalization scheme. Across four datasets of eye-tracked reading times, we consistently find that computing the surprisal of the entire region of interest (with leading or trailing whitespace) rarely leads to the most effective surprisal-based predictor. For instance, as hinted at above, we observe that on the CELER dataset (Berzak et al., 2022), the surprisal of the first three characters is a significantly better predictor of skip rate. On the Provo and MECO datasets (Luke and Christianson, 2018; Siegelman et al., 2022), the surprisal of the first characters of a region—either determined by a fixed-size or a dynamically sized focal area extending over typical rightward word identification spans—is on par with the surprisal of the full region. Finally, on the UCL dataset (Frank et al., 2013), including a look-ahead focal area that peeks at the subsequent region significantly improves reading time predictions.

2 Formalizing Psycholinguistic Stimuli

We now offer an abstract formalization of the stimuli present in a sentence-processing experiment.

2.1 Alphabets and Strings

We overview the building blocks of digitalized language: alphabets and strings. An **alphabet** is a finite, non-empty set. We use capital Greek letters to denote alphabets, e.g., we use Σ for an alphabet of **characters**, typically bytes. Let Σ^*

[1] See Marantz (2001), Haspelmath (2011), and Krauska and Lau (2023) for a discussion of various difficulties in defining the notion of a word in linguistic theory; thus, in the remainder of the paper, we use the term region of interest.

[2] The peculiarity in question is that standard implementations of BPE have the property that whitespace may occur at the beginning of a token but at no other place in the token.

be Σ's Kleene closure, i.e., the set of all strings formed from Σ, and let $\Sigma^+ \stackrel{\text{def}}{=} \Sigma^* \setminus \{\varepsilon\}$, where ε is the empty string. Given a string $\boldsymbol{\sigma} = \sigma_1 \cdots \sigma_N$ of N characters, we write $\boldsymbol{\sigma}_{(i,j)} = \sigma_{i+1} \cdots \sigma_{j-1}$, $\boldsymbol{\sigma}_{[i,j)} = \sigma_i \cdots \sigma_{j-1}$, $\boldsymbol{\sigma}_{(i,j]} = \sigma_{i+1} \cdots \sigma_j$ and $\boldsymbol{\sigma}_{[i,j]} = \sigma_i \cdots \sigma_j$ for $1 \leq i \leq j \leq N$. Furthermore, we write $\boldsymbol{\sigma} \preceq \boldsymbol{\sigma}''$ if $\boldsymbol{\sigma}$ is a prefix of $\boldsymbol{\sigma}''$ and $\boldsymbol{\sigma} \prec \boldsymbol{\sigma}''$ if $\boldsymbol{\sigma}$ is a proper prefix of $\boldsymbol{\sigma}''$. We denote string concatenation with juxtaposition: $\boldsymbol{\sigma}\,\boldsymbol{\sigma}'$.

2.2 Regions of Interest

In many psycholinguistic experiments, participants are presented with a character string $\boldsymbol{\sigma}$ as a **stimulus**, and various neural or behavioral responses to the stimulus are measured. To focus those measurements on parts of the string, the experimenter generally divides $\boldsymbol{\sigma}$ into regions of interest. The most common regions of interest considered in psycholinguistic experiments are white-spaced separate substrings of the stimulus, often referred to as words. However, one can just as easily experiment with regions of interest based on smaller (e.g., morphemes) or larger (e.g., constructions or sentences) units. We now give a formal definition of a region of interest.

Definition 1. *Let $\boldsymbol{\sigma} \in \Sigma^*$ be a non-empty stimulus. Let N be its length. A **region of interest** (ROI) $\boldsymbol{\rho} = [i, j)$ with $1 \leq i < j \leq N$ is a non-empty interval that correspond to a substring of $\boldsymbol{\sigma}$, which we denote as $\boldsymbol{\sigma}_{\boldsymbol{\rho}_k}$. We refer to $\boldsymbol{\sigma}_{\boldsymbol{\rho}_k}$ as the ROI's **yield** or simply the ROI when clear from context. We say that a sequence of regions of interest $\langle \boldsymbol{\rho}_k \rangle_{k=1}^K$ is **segmentative** if $\boldsymbol{\sigma}_{\boldsymbol{\rho}_1} \cdots \boldsymbol{\sigma}_{\boldsymbol{\rho}_K} = \boldsymbol{\sigma}$.*

2.2.1 Example 1: Self-paced Reading

In a self-paced reading experiment (Aaronson and Scarborough, 1976; Just et al., 1982; Rayner, 1998; Enochson and Culbertson, 2015), the experimenter decomposes each stimulus $\boldsymbol{\sigma} \in \Sigma^+$ into a sequence of ROIs $\langle \boldsymbol{\rho}_k \rangle_{k=1}^K$. The stimulus is then presented to the participant one ROI at a time, and the participant must click a button to progress to the next ROI. The measurement associated with each ROI is the time elapsed between the initial presentation of an ROI and the participant's pressing of a button.

As an example, consider the following stimulus taken from the UCL corpus (Frank et al., 2013):

(1) Anne lost control and laughed.

When viewed as a string of characters $\boldsymbol{\sigma} \in \Sigma^+$ (where Σ is the set of Unicode symbols), (1) is best thought of as the following Unicode string:

(2) `Anne␣lost␣control␣and␣laughed.`

where we have visualized the whitespace ␣ for clarity. While it may seem like a triviality at first blush, graphical markers of boundaries, e.g., whitespace, do play a role in reading behavior (Pollatsek and Rayner, 1982). Moreover, the whitespace ␣ becomes relevant when we extract a surprisal estimate from a language model, as discussed in §3. Consider the following natural first pass at a sequence of ROIs, and related substrings $\langle \boldsymbol{\sigma}_{\boldsymbol{\rho}_k} \rangle_{k=1}^K$, for (2):

(3) ⟨`Anne`, `lost`, `control`, `and`, `laughed`⟩

with whitespaces and the sentence-final period omitted to accommodate the self-paced paradigm. In our terminology, such a sequence of ROIs is called non-segmentative. However, a segmentative sequence of ROIs is generally desired, and the following two sequences are natural choices:

(4) ⟨`Anne`, `␣lost`, `␣control`, `␣and`, `␣laughed`⟩

(5) ⟨`Anne␣`, `lost␣`, `control␣`, `and␣`, `laughed`⟩

Indeed, Oh and Schuler (2024) take the stance that (5) is a better choice than (4). Specifically, they argue that, in a self-paced reading experiment where the regions of interest are generated by splitting the stimulus $\boldsymbol{\sigma}$ on the whitespace symbols ␣, the reader knows the character string displayed on the screen *must* be followed by the whitespace symbol ␣. We, however, contest this point. We agree, of course, that in such a setup, the participant knows that the symbols displayed must be followed by a whitespace ␣. However, the reader symmetrically knows that the character string was *preceded* by a whitespace ␣, but the surprisal of this preceding whitespace is attributed to the previous ROI. Thus, that the reader has knowledge of an ROI's surrounding whitespace, as is endemic to the self-paced reading paradigm, neither implies that the surprisal of those whitespace symbols should be lumped in with the ROI's surprisal nor gives us a reason to include the trailing whitespace and to omit the preceding whitespace if we are forced to choose.

Building on this, one alternative is to include the relevant whitespace in *all* regions of interest:

(6) ⟨`Anne␣`, `␣lost␣`, `␣control␣`, `␣and␣`, `␣laughed.`⟩

This choice, however, leads to a non-segmentative sequence of ROIs. Such a non-segmentative sequence is undesirable because it makes it impossible to cleanly divvy up the measurements among the ROIs due to the overlap among them. Nevertheless, choosing a surprisal-based predictor in a manner that takes into account both the preceding *and* trailing whitespace may be a good and useful idea. To accommodate this, we augment the notion of an ROI with that of a focal area, discussed in §3.3.

2.2.2 Example 2: Eye-tracked Reading

We now turn our attention to eye-tracked reading, another widely used psycholinguistic method for studying real-time sentence processing (Rayner, 1998; Rayner et al., 2006; Frank et al., 2013). In an eye-tracked reading experiment, participants naturally read a stimulus $\boldsymbol{\sigma}$ displayed on a screen while a camera tracks their eye movements. Unlike the self-paced reading task, where the experimenter's flexibility in defining different types of ROIs is limited by the task's ROI-by-ROI design, the eye-tracked reading paradigm does not provide an inherent notion of an ROI. Suppose a participant is presented with the stimulus string in Ex. (2). It would be a natural decision for the experimenter to design segmentative ROIs with yields:

(7) ⟨`Anne`, `␣lost␣control`, `␣and␣laughed.`⟩

if they were interested in studying verb phrases, and to then allot the measurements accordingly.

More frequently, however, the experimenter would define a segmentative sequence of ROIs split around whitespace boundaries. This might result in sequences of ROIs such as (4) and (5). This choice of ROIs might be meaningful, for instance, for a study on fixation duration under the hypothesis that whitespaces aid word identification processes and should thus have an effect on saccade latency (Fisher, 1975; Malt and Seamon, 1978). An alternative to segmentative ROIs such as (4) and (5) would be to exclude whitespaces from ROIs as in (3), with reading time measurements post-processed such that only fixations to the three whitespace-separated substrings composing the sentence are retained. This would be in line with evidence that readers use word boundary information in saccade planning to decide *where*—rather than *when*—to move their gaze (Rayner and Pollatsek, 1981; Pollatsek and Rayner, 1982), and that therefore whitespaces should not have an effect on saccade latency. However, excluding fixations on whitespace would discard data.

2.3 Much Ado About Trailing Whitespace

Two recent studies (Oh and Schuler, 2024; Pimentel and Meister, 2024) suggest that it is important to include the trailing whitespace and exclude the preceding whitespace in the definition of ROIs[3] in the context of surprisal theory; indeed, Pimentel and Meister (2024) state that the exclusion of a trailing whitespace is incorrect. With this backdrop, we offer an alternative view on this prescription. First and foremost, ROIs are typically determined by the experimenter who collected the dataset, and their choice of ROIs is already reflected in the psycholinguistic measurements reported; see §2.4. Thus, whether or not we wish to post-pend a trailing whitespace to an ROI's yield should primarily be informed by the data collection itself. Luckily, the role of whitespace in eye-tracking studies is already heavily investigated. In the traditional eye-tracking paradigm, ROIs are typically defined such that their yields include the preceding whitespace (Rayner, 1979; Pollatsek and Rayner, 1982; McConkie et al., 1988, *inter alia*). The justification for this choice stems from the fact that "people tend to direct their gaze to a point just left of the center of a word more frequently than to other locations" (McConkie et al., 1988). Moreover, as it is standard practice for ROIs to be segmentative (including whitespace in their yields) as the reader is taken to be fixating on one ROI at a time, the trailing whitespace must be excluded.

The situation is slightly different in the context of the digital corpora annotated with eye-tracked reading times used for larger-scale surprisal studies. For instance, Provo (Luke and Christianson, 2018), MECO (Siegelman et al., 2022), CELER (Berzak et al., 2022), and PoTeC (Jakobi et al., 2024) all divide the separating visual whitespaces equally between the preceding and the trailing ROI. When the yields of the ROIs are presented to a participant on a screen, as is the case with eye-tracking studies, it is possible to divide the whitespace displayed on the screen in half, associating the fixations in each half to the corresponding ROI. However, in the context of character strings one cannot perform such a splitting: The whitespace symbol ␣ is indivisible. Thus, in the case of these corpora, that is up to the modeler to determine how they wish to associate the indivisible whitespace symbol ␣ with the ROI; the data cannot inform the decision. In these cases, we view whether one includes a preceding or trail-

[3] Both studies refer to ROIs as words.

ing whitespace in the yields of the ROI as an empirical question and not a matter of correctness. No eye-tracking study to the authors' knowledge, however, associates a trailing whitespace with an ROI.

Finally, we note there is no inherent linguistic reason a trailing whitespace belongs to its preceding ROI, but contend that the ROIs used in surprisal studies should be the same ROIs the experimenter collected the data under. However, in §3.3, we discuss a more general abstraction for choosing what substring one should compute the surprisal of that is more detached from the specific choice of ROI in an attempt to resolve the tension between choosing the most useful surprisal-based predictors and respecting the data as they were collected.

2.4 Psycholinguistic Data

Saccade latency, discussed at the end of §2.2.2, is one example of the type of data gathered in a psycholinguistic experiment. More generally, during experimentation, we say a psycholinguist collects a **measurement**, abstractly denoted $\psi(\boldsymbol{\rho}_k) \in \mathbb{R}$ for each ROI $\boldsymbol{\rho}_k$. The measurement is typically a neural or behavioral response of the participant to consuming the stimulus segment $\boldsymbol{\rho}_k$, e.g., the time elapsed between keystrokes in a self-paced reading experiment (Just et al., 1982), the duration of a participant's first fixation on the ROI (Rayner, 1998), or the voltages produced by neural activity corresponding to that fixation (Donchin, 1979). Throughout the rest of the paper, notationally, we will write $\{\boldsymbol{\sigma}^{(m)}\}_{m=1}^{M}$ for a dataset of M stimuli. Additionally, we will write $\boldsymbol{\rho}_k^{(m)}$ for the k^{th} ROI of the m^{th} stimulus, and, correspondingly, $\psi(\boldsymbol{\rho}_k^{(m)})$ for the measurement of $\boldsymbol{\rho}_k^{(m)}$.

To explain or better understand the measurements for each ROI, the psycholinguist constructs a set of explanatory variables to predict the measurements. Regression analysis is applied to gain insight into the underlying aspects of human language processing that generated the measurements.

3 Language Models as Predictors

In psycholinguistics, LM-derived predictors are commonly used to predict measurements such as the participants' reading time for an ROI. We now give an overview of the necessary background.

3.1 Language Modeling

A **language model** p is a probability distribution over Σ^*. We define p's **prefix probability** $\overrightarrow{p}(\boldsymbol{\sigma})$ as the probability that a string drawn from p begins with a particular string $\boldsymbol{\sigma} \in \Sigma^*$:

$$\overrightarrow{p}(\boldsymbol{\sigma}) \stackrel{\text{def}}{=} \sum_{\boldsymbol{\sigma}' \in \Sigma^*} \mathbf{1}\{\boldsymbol{\sigma} \preceq \boldsymbol{\sigma}'\} p(\boldsymbol{\sigma}') \, . \tag{1}$$

Prefix probabilities are primarily used to compute the conditional probability of the continuation $\boldsymbol{\sigma}' \in \Sigma^*$ given a preceding context $\boldsymbol{\sigma}$:[4]

$$p(\boldsymbol{\sigma}' \mid \boldsymbol{\sigma}) = \frac{\overrightarrow{p}(\boldsymbol{\sigma} \cdot \boldsymbol{\sigma}')}{\overrightarrow{p}(\boldsymbol{\sigma})}. \tag{2}$$

We can factorize a language model p as

$$p(\boldsymbol{\sigma}) = p(\text{EOS} \mid \boldsymbol{\sigma}) \prod_{t=1}^{|\boldsymbol{\sigma}|} p(\sigma_t \mid \boldsymbol{\sigma}_{[1,t)}), \tag{3}$$

where each $p(\sigma_t \mid \boldsymbol{\sigma}_{[1,t)})$ is a conditional probability over the set $\Sigma \cup \{\text{EOS}\}$, $\text{EOS} \notin \Sigma$ is a distinguished end-of-string symbol, and

$$p(\text{EOS} \mid \boldsymbol{\sigma}) \stackrel{\text{def}}{=} \frac{p(\boldsymbol{\sigma})}{\overrightarrow{p}(\boldsymbol{\sigma})}. \tag{4}$$

The human language model. Much work in computational psycholinguistics builds on the assumption that humans process language probabilistically, i.e., that humans have an internal language model. We denote the hypothetical construct of a human language model as p_{H}. Because the true human language model is unknown, we must approximate it via another language model, which we will call p. To the extent that p is a good approximation of p_{H}, we would expect estimates derived from p to be a reliable proxy of the probabilities prescribed by the human language model.

3.2 Surprisal Theory

One popular information-theoretic framework for deriving computational predictors from language models is surprisal theory (Hale, 2001; Levy, 2008). Surprisal theory states that the predictability of an ROI's yield in the context of its preceding character string is a useful predictor for the ROI's psycholinguistic measurements. To define surprisal formally, we introduce additional notation. Let $\boldsymbol{\sigma} \in \Sigma^+$ be a stimulus divided into K ROIs $\langle \boldsymbol{\rho}_k \rangle_{k=1}^{K}$. Then, the **surprisal** of an ROI $\boldsymbol{\rho}_k = [i, j)$ in context $\boldsymbol{\sigma}_{[1,i)}$ is

$$\iota(\boldsymbol{\sigma}_{[i,j)} \mid \boldsymbol{\sigma}_{[1,i)}) \stackrel{\text{def}}{=} -\log p(\boldsymbol{\sigma}_{[i,j)} \mid \boldsymbol{\sigma}_{[1,i)}). \tag{5}$$

[4]Note that $\overrightarrow{p}(\varepsilon \mid \boldsymbol{\sigma}) = 1$ for all $\boldsymbol{\sigma} \in \Sigma^*$ and Eq. (2) is only well-defined when $\overrightarrow{p}(\boldsymbol{\sigma}) > 0$, a condition which will always be satisfied for softmax-normalized language models.

We remark again on a key latent assumption embedded in surprisal theory: It is assumed that the language model p used to compute Eq. (5) well-approximates the human language model p_{H}, as discussed above. Surprisal theory then posits that the surprisal of an ROI in context is a good predictor of many measurements ψ that seek to operationalize processing difficulty, e.g., reading time. Empirically, this result has been demonstrated in many studies (Smith and Levy, 2013; Goodkind and Bicknell, 2018; Shain et al., 2020; Merkx and Frank, 2021; Wilcox et al., 2023, *inter alia*).

3.3 Focal Areas

Computing the surprisal of the ROI's entire yield in context is often too coarse-grained. To allow for additional modeling freedom, we further associate every ROI with one more focal areas, i.e., the portion of the ROI's substring or the characters surrounding it to which the experimenter assigns a special status in terms of computing surprisal values. We can regard a focal area as a string-valued feature of a ROI which tells us which surprisal to compute. Indeed, we take the stance that there is no inherent reason why the substring of the stimulus one computes the surprisal of should be identical to the ROI's yield.

Definition 2. *Let $\boldsymbol{\sigma} \in \Sigma^*$ be a non-empty stimulus. Let N be its length, and let $\boldsymbol{\rho}_k = [i, j)$ be a region of interest. Then, a **focal area** of $\boldsymbol{\rho}_k$ is a non-empty interval $[\alpha_k, \beta_k)$ with $i \leq \beta_k$, $\alpha_k \leq j$ and $1 \leq \alpha_k < \beta_k \leq N$ corresponding to the substring $\boldsymbol{\sigma}_{[\alpha_k, \beta_k)}$ of the stimulus. We refer to $\boldsymbol{\sigma}_{[\alpha_k, \beta_k)}$ as the focal area's **yield** or simply the focal area when clear from context.*

Remark 1. *Def. 2 states that a focal area must overlap with its corresponding* ROI, *i.e., have a non-empty intersection. This constraints the focal area from being fully disassociated from the* ROI.

The surprisal of ROI $\boldsymbol{\rho}_k$'s focal area $[\alpha_k, \beta_k)$ is

$$\iota(\boldsymbol{\sigma}_{[\alpha,\beta)} \mid \boldsymbol{\sigma}_{[1,\alpha)}) \stackrel{\text{def}}{=} -\log p(\boldsymbol{\sigma}_{[\alpha,\beta)} \mid \boldsymbol{\sigma}_{[1,\alpha)}). \quad (6)$$

Focal areas allow the modeler to express that, in some circumstances, the psycholinguist hypothesizes, or assumes, that the non-focal areas of the region will not have an influence on the measurements collected for that ROI, or that characters outside of the ROI's yield will.[5]

[5]Indeed, the experimenter may devise multiple focal areas they believe have distinct influences on the measurements.

Example: Modeling skipped ROIs. To understand the utility of focal areas, consider an eye-tracked reading experiment where the measurement of interest for any given ROI is its skip rate, i.e., the proportion of experimental trials in which the participant did not fixate on that ROI (Rayner et al., 2011). The experimenter might design a stimulus such as (2) and the respective segmentative sequence of ROIs with trailing whitespaces, as in (5). A reader's decision to skip a target region, say $\boldsymbol{\sigma}_{\boldsymbol{\rho}_3} =$ `control␣`, cannot be made when fixating on the entire ROI—otherwise, we could not say that `control␣` had been skipped. Instead, the decision must be made when fixating on characters preceding the region. However, when fixating on the preceding characters, the reader has partial access to characters to the right (Rayner et al., 1982; Underwood and McConkie, 1985; McConkie and Zola, 1987), which belongs to the subsequent ROI. Thus, the decision to skip the target region should depend at most on the first few characters of the ROI's yield. To determine the exact number of characters in the focal area, the psycholinguist may build on prior empirical evidence (see §3.4). For example, Rayner et al. (1982) found that when the first three characters of the ROI's yield to the right of the fixation were available, and the remainder of the characters were replaced, the reading rate was not substantially affected. In line with such evidence, the psycholinguist may design a focal area on $\boldsymbol{\sigma}_{[11,14)} =$ `con` consisting of the first three characters of `control␣`, the yield of ROI $\boldsymbol{\rho}_3$. The skip rate for $\boldsymbol{\rho}_3$ would then be modeled using the surprisal of the focal area, $\iota($`con` $\mid$ `Anne␣lost␣`$)$, as a predictor.

The role of EOS in focal areas. Because EOS $\notin \Sigma$, by the definition of a stimulus as a character string and ROIs as intervals corresponding to substrings of the stimulus, EOS cannot be included in an ROI's yield. However, when analyzing wrap-up effects (Meister et al., 2022), it may be prudent to abuse the definition and include EOS anyway.

3.4 Selecting Focal Areas

We now explain our construction of various focal areas based on insights from human language processing and the psychology of reading.

Dynamically sized focal areas. The perceptual span during reading, which is the range of visual information available around the fixation point, is relatively limited for readers of alphabetical orthographies such as English. It typically extends

	Leading Whitespace	Trailing Whitespace
Full ROI	⟨␣lost, ␣control, ␣and, ␣laughed.⟩	⟨lost␣, control␣, and␣, laughed.⟩
Fixed	⟨␣lo, ␣co, ␣an, ␣la⟩	⟨los, con, and, lau⟩
Dynamic (7)	⟨␣lost, ␣cont, ␣an, ␣laug⟩	⟨lost␣, contr, and, laugh⟩
Dynamic (8)	⟨␣lost, ␣contr, ␣and, ␣laugh⟩	⟨lost␣, contro, and␣, laughe⟩
Look-ahead (3)	⟨␣lost␣co, ␣control␣an, ␣and␣la, ␣laughed.⟩	⟨lost␣con, control␣and, and␣lau, laughed.⟩
Look-ahead (4)	⟨␣lost␣con, ␣control␣and, ␣and␣lau, ␣laughed.⟩	⟨lost␣cont, control␣and␣, and␣laug, laughed.⟩
Look-ahead (5)	⟨␣lost␣cont, ␣control␣and␣, ␣and␣laug, ␣laughed.⟩	⟨lost␣contr, control␣and␣l, and␣laugh, laughed.⟩
Look-ahead (6)	⟨␣lost␣contr, ␣control␣and␣l, ␣and␣laugh, ␣laughed.⟩	⟨lost␣contro, control␣and␣la, and␣laughe, laughed.⟩
Look-ahead (7)	⟨␣lost␣contro, ␣control␣and␣la, ␣and␣laughe, ␣laughed.⟩	⟨lost␣control, control␣and␣lau, and␣laughed, laughed.⟩
Look-ahead (Full)	⟨␣lost␣control, ␣control␣and, ␣and␣laughed., ␣laughed.⟩	⟨lost␣control␣, control␣and␣, and␣laughed., laughed.⟩

Table 1: Yields of the focal areas for the stimulus $\boldsymbol{\sigma}$ = Anne␣lost␣control␣and␣laughed. using two segmentative sequences of ROIs (see §2.2) and ten focal areas (see §3.4). The first ROI is skipped to ensure all focal areas are well-defined.

from about 3–4 characters to the left of the fixation point to approximately 14–15 characters to the right (McConkie and Rayner, 1975, 1976; Rayner and Bertera, 1979; Rayner et al., 1981; den Buurman et al., 1981; Underwood and McConkie, 1985) for English. However, the span within which words can actually be identified, known as the **word identification span**, is narrower, generally extending no more than 7–8 characters to the right of the fixation (Rayner et al., 1982; McConkie and Zola, 1987; Underwood and McConkie, 1985). Furthermore, readers' preferred viewing location, i.e., the location where they typically land after a saccade, tends to be a character between the beginning and the middle of the ROI (O'Regan, 1980; Rayner, 1979), approximately at position $\lceil |\boldsymbol{\sigma}_{\rho_k}|/2 \rceil - 1$ for the k^{th} ROI (Rayner and Pollatsek, 1981). Thus, the size of the focal area on the initial characters of the upcoming region $\boldsymbol{\rho}_{k+1}$ should vary depending on the length of $\boldsymbol{\rho}_k$. With a **preferred viewing location** on the character at position $v = \lceil |\boldsymbol{\sigma}_{\rho_k}|/2 \rceil - 1$, and a **rightward word identification span** of $s \in \{7, 8\}$ characters, the focal area for region $\boldsymbol{\rho}_{k+1}$ should include the first $\min(|\boldsymbol{\sigma}_{\boldsymbol{\rho}_{k+1}}|, \max(0, v + s - |\boldsymbol{\sigma}_{\boldsymbol{\rho}_k}|))$ characters of the region. See the row labeled Fixed in Tab. 1 for an example.

Fixed-size focal areas. Alternatively, the design of a focal area could be fixed in size. Research indicates that the initial characters of parafoveal ROIs are crucial not merely due to their proximity to the fixation point but because they aid in initiating lexical access and integrating information across fixations (Inhoff, 1989, 1990; Inhoff and Tousman, 1990). Multiple studies show that previewing exactly the first three characters of a word, even with the remaining characters replaced by visually similar ones, enhances reading speed (Rayner et al., 1982; Lima and Inhoff, 1985; Lima, 1987), and that parafoveal previews also allow readers to skip words up to three characters long (Blanchard et al., 1989). Consequently, a fixed-size focal area, consistently covering the first $\min(|\boldsymbol{\rho}_k|, 3)$ characters of a region, might be an effective predictor for that ROI's collected measurements. See the rows labeled Dynamic in Tab. 1 for an example.

Look-ahead focal areas. Finally, we design a focal area that looks ahead, i.e., one that includes characters to the right of the ROI's yield and into the next ROI's yield. The argument for designing a look-ahead focal area stems from the fact we may wish to model the structural integration cost that could arise if the ROI corresponds to the end of a constituent—or, symmetrically, the additional processing cost that could arise when creating a new constituent (Gibson, 2001; Futrell et al., 2020). However, without look-ahead into the next ROI's yield, it can be difficult to judge whether it is necessary to integrate a new constituent. Finally, we remark that focal areas that admit look-ahead resolve the problem of how to associate whitespace with ROIs as they detether defining a sequence of ROIs from the surprisal computation (see §2.3): A sequence of ROIs that incorporates preceding whitespace, to respect how the psycholinguistics measurements were collected, can still be associated with a surprisal value that includes the surprisal of that ROI's trailing whitespace. Of course, whether including this trailing whitespace helps remains an empirical question. In our experiments, we use look-aheads of 3 to 7 characters as well as a look-ahead peeking into the entire upcoming ROI. See the rows labeled Look-ahead in Tab. 1 for an example.

Focal areas in past studies. In most past studies, experiments predicting measurements of reading behavior typically compute the surprisal of the entire ROI without accounting for specific focal areas (Goodkind and Bicknell, 2018; Wilcox et al., 2020, 2023; Shain et al., 2024, *inter alia*). Most of these studies use LMs that rely on the BPE tokenizer and, thus, the ROIs' yields are defined to include the *preceding* whitespace due to an oddity of BPE (see Footnote 2), but recall §2.3 for two exceptions.

4 Marginalizing Out Token Strings

The previous two sections (§2 and §3) have formalized psycholinguistic stimuli, their regions of interest, and focal areas at the character level. Indeed, in this discussion, the character-level model used to compute surprisal is agnostic as to whether the model underlyingly uses tokenization or not.

However, tokenization has evolved into a standard practice in constructing language models. Rather than constituting a distribution over Σ^*, the set of all character strings, most modern language models are distributions p_Δ over Δ^* where Δ is an alphabet of **tokens**. To **encode** a character string as a token string, we apply a function of type $\tau\colon \Sigma^* \to \Delta^*$. To **decode** a token string to a character string, we apply a function of type $\kappa\colon \Delta^* \to \Sigma^*$ (cf. Gastaldi et al., 2024). For the purposes of this paper, we assume this pair of functions satisfy:

- Exactness: $\forall \boldsymbol{\sigma} \in \Sigma^*\colon \kappa(\tau(\boldsymbol{\sigma})) = \boldsymbol{\sigma}$.[6]
- Multiplicativity: $\kappa(\varepsilon) = \varepsilon$, and $\forall \delta_1 \cdots \delta_N \in \Delta^*\colon \kappa(\delta_1 \cdots \delta_N) = \kappa(\delta_1) \cdots \kappa(\delta_N)$.

BPE satisfies both of these properties.

The probability of a character string $\boldsymbol{\sigma}$ can be computed from a language model over tokens p_Δ using the following marginalization:

$$p_\Sigma(\boldsymbol{\sigma}) = \sum_{\boldsymbol{\delta} \in \Delta^*} \mathbf{1}\{\boldsymbol{\sigma} = \kappa(\boldsymbol{\delta})\}\, p_\Delta(\boldsymbol{\delta}). \quad (7)$$

Similarly, the prefix probability is given by

$$\overrightarrow{p_\Sigma}(\boldsymbol{\sigma}) = \sum_{\boldsymbol{\delta} \in \Delta^*} \mathbf{1}\{\boldsymbol{\sigma} \preceq \kappa(\boldsymbol{\delta})\}\, p_\Delta(\boldsymbol{\delta}). \quad (8)$$

Vieira et al. (2024) show that $\overrightarrow{p_\Sigma}(\boldsymbol{\sigma})$ can be computed with a finite summation:

$$\overrightarrow{p_\Sigma}(\boldsymbol{\sigma}) = \sum_{\boldsymbol{\delta} \in \mathcal{C}(\boldsymbol{\sigma})} \overrightarrow{p_\Delta}(\boldsymbol{\delta}), \quad (9)$$

where $\overrightarrow{p_\Delta}$ is the prefix probability of p_Δ, calculated as in Eq. (1), and the **prefix-cover** $\mathcal{C}$ is defined as

$$\mathcal{C}(\boldsymbol{\sigma}) \stackrel{\text{def}}{=} \textbf{if } \boldsymbol{\sigma} = \varepsilon\colon \{\varepsilon\} \quad (10)$$
$$\textbf{else}\colon \{\boldsymbol{\delta}' \cdot \delta \in \Delta^+ \mid \kappa(\boldsymbol{\delta}') \prec \boldsymbol{\sigma} \preceq \kappa(\boldsymbol{\delta}' \cdot \delta)\}.$$

[6]But, not necessarily, $\forall \boldsymbol{\delta} \in \Delta^*\colon \tau(\kappa(\boldsymbol{\delta})) = \boldsymbol{\delta}$. Thus, we do not require (τ, κ) to form a bijection over (Σ^*, Δ^*).

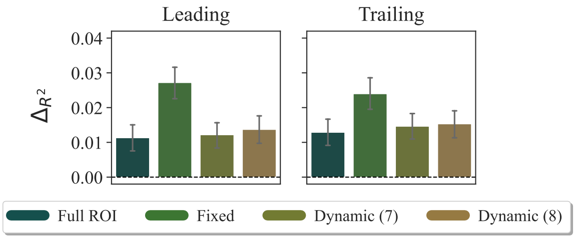

Figure 1: **Skip Rate**. Predictive power Δ_{R^2} of an ROI's character-level surprisal, calculated with varying focal areas and two ROI types: leading (left) or trailing (right) whitespace. All Δ_{R^2} scores are significantly above zero ($p < 0.001$). Error bars represent 95% confidence intervals. The black dotted line corresponds to the baseline regressor, including ROI length and frequency. The target regressor includes the length, frequency, and full ROI surprisal of the previous two ROIs.

Unfortunately, $|\mathcal{C}(\boldsymbol{\sigma})|$ can be exponential in $|\boldsymbol{\sigma}|$; thus, we use the beam summing algorithm proposed by Vieira et al. (2024) as a practical approximation algorithm.[7] Lastly, to compute the character-level conditional distribution, we use Eq. (2), albeit with our approximation to $\overrightarrow{p_\Sigma}(\boldsymbol{\sigma})$.

5 Predictive Power of Focal Areas

Our experimental design is discussed in App. B. We consider ROIs with leading and trailing whitespaces and experiment with the focal areas described in §3.4. Additional results may be found in App. C.

Skip rate. The results for skip rates in the CELER dataset (Berzak et al., 2022) are shown in Fig. 1 with predictive power expressed as the difference in R^2 between target and baseline regressors. The trends observed are consistent across different ROI types. Among the predictors examined, the surprisal of the fixed-size focal area, which corresponds to the first three characters of an ROI, emerged as the strongest predictor of skipping behavior, with a Δ_{R^2} significantly higher than all other predictors ($p < 0.001$). The surprisal of the dynamically sized focal area with a word identification span of 8 characters is the second-best predictor, followed by that of the dynamically sized focal area with a word identification span of 7 characters. The surprisal of the full ROI is the weakest predictor, with a Δ_{R^2} approximately two times lower than that of the fixed-size focal area. These results are consistent with findings that English readers process information to the right of the currently fixated ROI collected with human subjects. In particular, they provide new evidence that upcoming ROIs are skipped because they are

[7]See App. A for more details.

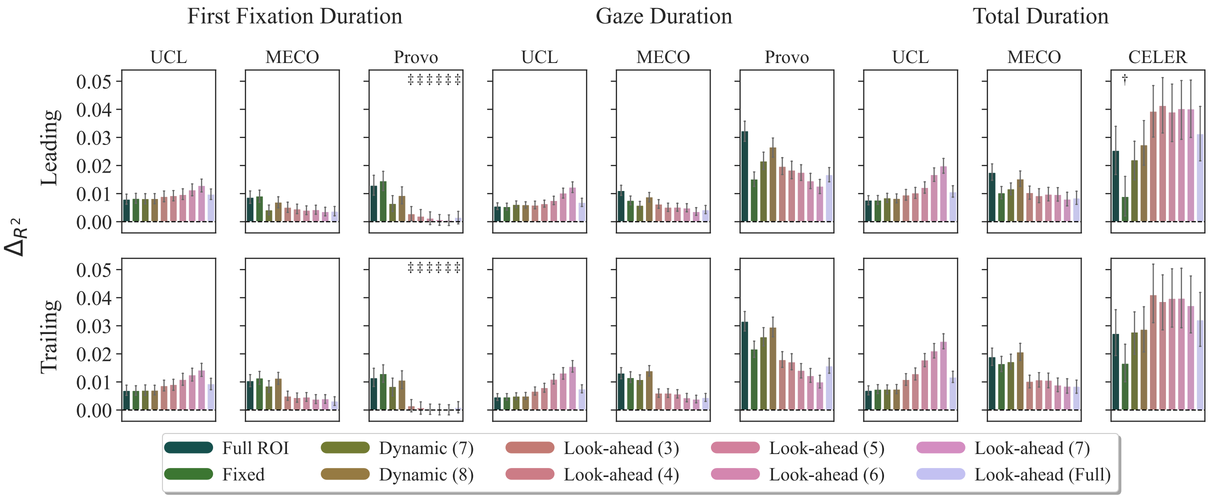

Figure 2: **First Fixation, Gaze, and Total Duration**. Predictive power Δ_{R^2} of an ROI's character-level surprisal, calculated with varying focal areas and according to two ROI types: with leading (top) or trailing (bottom) whitespace. All Δ_{R^2} scores are significantly above zero with $p < 0.001$, unless marked with † ($p < 0.01$) or ‡ ($p \geq 0.01$). Error bars represent 95% confidence intervals. The black dotted line represents the baseline regressor, including ROI length and frequency. The target regressor includes the length, frequency, and full ROI surprisal of the previous two ROIs to account for spillover effects.

partially read rather than filled in from contextual cues (McConkie and Rayner, 1975; Rayner, 1975; Rayner et al., 1982), and that the first three characters of the upcoming ROI have a special status (Lima and Inhoff, 1985; Lima, 1987). Moreover, the fact that the predictive power of the surprisal of fixed-size focal areas is highest for ROIs with leading whitespace (13% higher than fixed-size focal areas with trailing whitespace; $p < 0.01$) suggests a dual model of skipping decisions. On the one hand, parafoveal preview provides cues for lexical identification (Inhoff, 1989, 1990; Inhoff and Tousman, 1990) and on the other, word boundary information to the right of the currently fixated ROI that is used for saccade planning (Rayner and Pollatsek, 1981; Pollatsek and Rayner, 1982).

First fixation, gaze, and total duration. The results for the UCL, Provo, MECO, and CELER datasets are shown in Fig. 2, with predictive power expressed as the difference in R^2 between target and baseline regressors. Look-ahead focal areas significantly improve reading time predictions on the UCL dataset (first fixation duration, gaze duration, and total duration) and the CELER dataset (total duration), compared to using focal areas over the entire ROI. In the Provo and MECO datasets, the predictive power of surprisal from the fixed-size and dynamically sized focal areas is comparable to that of the full ROI's surprisal. For first fixation duration, the fixed-size focal area surprisal emerges as the best predictor, alongside the full ROI's surprisal. For gaze and total duration, the surprisal from the dynamically sized focal area, with a rightward word identification span of 8 characters, is on par with the full ROI's surprisal; their relative ranking varies across the two types of ROI. These findings align with psycholinguistic evidence that the perceptual span of English readers extends to the right of the current fixation (McConkie and Rayner, 1975; den Buurman et al., 1981; Underwood and McConkie, 1985). The variability in results across datasets reinforces the view that ROI and focal area definitions are an empirical matter and points to a nuanced perspective on how character-level information influences reading behavior. The strong predictive power of look-ahead focal areas, which span both the current and upcoming ROI, indicates that parafoveal information affects saccade latency for the *currently* fixated ROI—an effect which could be connected to the assessment of longer-horizon prediction errors (Giulianelli et al., 2024b) or to the front-loading of integration costs that may occur when previewing the upcoming region (Gibson, 2001; Futrell et al., 2020). However, the predictive power of fixed-size and dynamically sized focal areas, which span only part of the ROI and model saccade latency based on the *preceding* fixation, suggests that parafoveal information is also used to preprocess upcoming ROIs.

6 Conclusion

We treat the role of tokenization in psycholinguistics. We recommend predictors be derived from character-level surprisal, allowing the modeler to explore a wider range of useful predictors.

Limitations

Our analyses are conducted exclusively on English stimuli and measurements collected from L1 English readers. Additionally, we focus solely on eye-tracking data, as it is more natural to conceptualize focal areas in this context. We do not analyze self-paced reading, where the challenges are likely even more complex due to the variability in how the method is applied. For example, some studies use a moving-window paradigm in which words are masked by dashes (Just et al., 1982), preserving whitespace information, whereas others rely on centered presentation (Aaronson and Scarborough, 1976), which omits whitespace information by design. Further complications may arise from differences between word-by-word and chunked presentation (see, e.g., Tremblay et al., 2011), where both single-word and multi-word ROIs may be considered, as well as from paradigms presenting multiple alternative ROIs at a time (Forster et al., 2009; Boyce et al., 2023). How these variations interact with surprisal predictors remains poorly understood, and future work is necessary to model self-paced reading data more comprehensively using focal area predictors.

Other limitations of our approach lie in the modeling assumptions made about the relationship between surprisal and reading behaviors. While we employ linear modeling based on established evidence that the relationship between surprisal and reading time is linear (Smith and Levy, 2008, 2013; Wilcox et al., 2023; Shain et al., 2024), this relationship has not yet been determined for skip rates. To our knowledge, no studies have examined skip rates using our focal area predictors, and the functional relationship between surprisal and skip rates remains to be determined. Future research should investigate skip rates with modeling approaches capable of capturing non-linear relationships, such as generalized additive models (GAMs; Wood, 2004, 2017). Finally, we do not account for individual differences between participants in our analysis, which could be more accurately modeled using mixed-effects models (Gelman et al., 2004).

Acknowledgements

We thank Shayne Sloggett for a lengthy discussion about the state of the self-paced reading paradigm as well as his sage insights into the eye-tracking literature. We also thank Patrick Haller, Ethan Gotlieb Wilcox, Yahya Emara, Ekaterina Vylomova, and Eleanor Chodroff for useful discussions about the psycholinguistic content of the paper, and Clemente Pasti, Robin Shing Moon Chan, Zeerak Talat, Anej Svete, and Vésteinn Snæbjarnarson for help with copy editing.

References

Doris Aaronson and Hollis Shapiro Scarborough. 1976. Performance theories for sentence coding: Some quantitative evidence. *Journal of Experimental Psychology: Human Perception and Performance*, 2(1).

Marco Baroni. 2022. On the proper role of linguistically oriented deep net analysis in linguistic theorising. In *Algebraic Structures in Natural Language*. CRC Press.

Lisa Beinborn and Yuval Pinter. 2023. Analyzing cognitive plausibility of subword tokenization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Yevgeni Berzak, Chie Nakamura, Amelia Smith, Emily Weng, Boris Katz, Suzanne Flynn, and Roger Levy. 2022. CELER: A 365-Participant Corpus of Eye Movements in L1 and L2 English Reading. *Open Mind*.

Harry E. Blanchard, Alexander Pollatsek, and Keith Rayner. 1989. The acquisition of parafoveal word information in reading. *Perception & Psychophysics*, 46(1).

Veronica Boyce, Roger Levy, Veronica Boyce, and Roger P. Levy. 2023. A-maze of Natural Stories: Comprehension and surprisal in the Maze task. *Glossa Psycholinguistics*, 2(1).

Marc Brysbaert, Boris New, and Emmanuel Keuleers. 2012. Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior research methods*, 44.

Kris Cao and Laura Rimell. 2021. You should evaluate your language model on marginal likelihood over tokenisations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. 2023. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, 7(3).

Nadezhda Chirkova, Germán Kruszewski, Jos Rozen, and Marc Dymetman. 2023. Should you marginalize over possible tokenizations? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Andrea Gregor de Varda, Marco Marelli, and Simona Amenta. 2023. Cloze probability, predictability ratings, and computational estimates for 205 English sentences, aligned with existing EEG and reading time data. *Behavior Research Methods*.

Rudy den Buurman, Theo Roersema, and Jack F. Gerrissen. 1981. Eye movements and the perceptual span in reading. *Reading Research Quarterly*, 16(2).

Emanuel Donchin. 1979. Event-related brain potentials: A tool in the study of human information processing. In *Evoked Brain Potentials and Behavior*. Springer.

Kelly Enochson and Jennifer Culbertson. 2015. Collecting psycholinguistic response time data using Amazon Mechanical Turk. *PloS one*, 10(3).

Dennis F. Fisher. 1975. Reading and visual search. *Memory & Cognition*, 3(2).

Kenneth I. Forster, Christine Guerrera, and Lisa Elliot. 2009. The maze task: Measuring forced incremental sentence processing time. *Behavior Research Methods*, 41.

Stefan L. Frank. 2024. Neural language model gradients predict event-related brain potentials. In *Proceedings of the Society for Computation in Linguistics 2024*.

Stefan L. Frank, Irene Fernandez Monsalve, Robin L. Thompson, and Gabriella Vigliocco. 2013. Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, 45.

Richard Futrell, Edward Gibson, and Roger P. Levy. 2020. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, 44(3).

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Juan Luis Gastaldi, John Terilla, Luca Malagutti, Brian DuSell, Tim Vieira, and Ryan Cotterell. 2024. The foundations of tokenization: Statistical and computational concerns. *Preprint*, arXiv:2407.11606.

Renato Lui Geh, Honghua Zhang, Kareem Ahmed, Benjie Wang, and Guy Van den Broeck. 2024. Where is the signal in tokenization space? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2004. *Bayesian Data Analysis*. Chapman and Hall/CRC.

Edward Gibson. 2001. The Dependency Locality Theory: A Distance-Based Theory of Linguistic Complexity. In *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*. The MIT Press.

Mario Giulianelli, Andreas Opedal, and Ryan Cotterell. 2024a. Generalized measures of anticipation and responsivity in online language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Mario Giulianelli, Arabella Sinclair, and Raquel Fernández. 2022. Construction repetition reduces information rate in dialogue. In *Proceedings of the Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*.

Mario Giulianelli, Sarenne Wallbridge, Ryan Cotterell, and Raquel Fernández. 2024b. Incremental alternative sampling as a lens into the temporal and representational resolution of linguistic prediction. *Preprint*, PsyArXiv:10.31234.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Meeting of the North American Chapter of the Association for Computational Linguistics*.

Martin Haspelmath. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica*, 45(1).

Jacob Louis Hoover, Morgan Sonderegger, Steven T Piantadosi, and Timothy J O'Donnell. 2023. The plausibility of sampling as an algorithmic theory of sentence processing. *Open Mind*, 7.

J. Hyönä and R. K. Olson. 1995. Eye fixation patterns among dyslexic and normal readers: Effects of word length and word frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(6).

Albert W. Inhoff and S. Tousman. 1990. Lexical integration across saccades in reading. *Psychological research*, 52(4).

Albrecht Werner Inhoff. 1989. Lexical access during eye fixations in reading: Are word access codes used to integrate lexical information across interword fixations? *Journal of Memory and Language*, 28(4).

Albrecht Werner Inhoff. 1990. Integrating information across eye fixations in reading: The role of letter and word units. *Acta Psychologica*, 73(3).

Albrecht Werner Inhoff and Keith Rayner. 1986. Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & Psychophysics*, 40(6).

Deborah N. Jakobi, Thomas Kern, David R. Reich, Patrick Haller, and Lena A. Jäger. 2024. PoTeC: A German naturalistic eye-tracking-while-reading corpus. *CoRR*, abs/2403.00506.

Marcel A. Just, Patricia A. Carpenter, and Jacqueline D. Woolley. 1982. Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, 111(2).

Alexandra Krauska and Ellen Lau. 2023. Moving away from lexicalism in psycho-and neuro-linguistics. *Frontiers in Language Sciences*, 2.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3).

Susan D. Lima. 1987. Morphological analysis in sentence reading. *Journal of Memory and Language*, 26(1).

Susan D. Lima and Albrecht W. Inhoff. 1985. Lexical access during eye fixations in reading: effects of word-initial letter sequence. *Journal of Experimental Psychology: Human Perception and Performance*, 11(3).

Steven G. Luke and Kiel Christianson. 2018. The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50(2).

Barbara C. Malt and John G. Seamon. 1978. Peripheral and cognitive components of eye guidance in filled-space reading. *Perception & Psychophysics*, 23(5).

Alec Marantz. 2001. Words. *WCCFL XX Handout, USC*.

G. W. McConkie, P. W. Kerr, M. D. Reddix, and D. Zola. 1988. Eye movement control during reading: I. the location of initial eye fixations on words. *Vision Research*, 28(10).

George W. McConkie and Keith Rayner. 1975. The span of the effective stimulus during a fixation in reading. *Perception & Psychophysics*, 17(6).

George W. McConkie and Keith Rayner. 1976. Asymmetry of the perceptual span in reading. *Bulletin of the Psychonomic Society*, 8(5).

George W. McConkie and David Zola. 1987. *Visual attention during eye fixations while reading*. Routledge.

Clara Meister, Tiago Pimentel, Thomas Clark, Ryan Cotterell, and Roger Levy. 2022. Analyzing wrap-up effects through an information-theoretic lens. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the Uniform Information Density hypothesis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Danny Merkx and Stefan L. Frank. 2021. Human sentence processing: Recurrence or attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*.

James A. Michaelov, Megan D. Bardolph, Cyma K. Van Petten, Benjamin K. Bergen, and Seana Coulson. 2024. Strong Prediction: Language Model Surprisal Explains Multiple N400 Effects. *Neurobiology of Language*, 5(1).

Sathvik Nair and Philip Resnik. 2023. Words, subwords, and morphemes: What really matters in the surprisal-reading time relationship? In *Findings of the Association for Computational Linguistics: EMNLP*.

Byung-Doh Oh, Christian Clark, and William Schuler. 2021. Surprisal estimators for human reading times need character models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*.

Byung-Doh Oh and William Schuler. 2023. Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times? *Transactions of the Association for Computational Linguistics*, 11.

Byung-Doh Oh and William Schuler. 2024. Leading whitespaces of language models' subword vocabulary poses a confound for calculating word probabilities. *Preprint*, arXiv:2406.10851.

J. Kevin O'Regan. 1980. The control of saccade size and fixation duration in reading: The limits of linguistic control. *Perception & Psychophysics*, 28.

Tiago Pimentel and Clara Meister. 2024. How to compute the probability of a word. *Preprint*, arXiv:2406.14561.

Alexander Pollatsek and Keith Rayner. 1982. Eye movement control in reading: The role of word boundaries. *Journal of Experimental Psychology: Human Perception and Performance*, 8(6).

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8).

Keith Rayner. 1975. The perceptual span and peripheral cues in reading. *Cognitive Psychology*, 7(1).

Keith Rayner. 1979. Eye guidance in reading: Fixation locations in words. *Perception*, 8.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3).

Keith Rayner and James H. Bertera. 1979. Reading without a fovea. *Science*, 206(4417).

Keith Rayner, Kathryn H. Chace, Timothy J. Slattery, and Jane Ashby. 2006. Eye movements as reflections of comprehension processes in reading. *Scientific studies of reading*, 10(3).

Keith Rayner and Susan A. Duffy. 1986. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, 14(3).

Keith Rayner, Albrecht Werner Inhoff, Robert E. Morrison, Maria L. Slowiaczek, and James H. Bertera. 1981. Masking of foveal and parafoveal vision during eye fixations in reading. *Journal of Experimental Psychology: Human Perception and Performance*, 7(1).

Keith Rayner and Alexander Pollatsek. 1981. Eye movement control during reading: Evidence for direct control. *The Quarterly Journal of Experimental Psychology Section A*, 33(4).

Keith Rayner, Timothy J Slattery, Denis Drieghe, and Simon P Liversedge. 2011. Eye movements and word skipping during reading: Effects of word length and predictability. *Journal of Experimental Psychology: Human Perception and Performance*, 37(2).

Keith Rayner, Arnold D. Well, Alexander Pollatsek, and James H Bertera. 1982. The availability of useful information to the right of fixation in reading. *Perception & Psychophysics*, 31(6).

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45).

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Cory Shain, Idan Asher Blank, Marten van Schijndel, William Schuler, and Evelina Fedorenko. 2020. fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, 138.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10).

Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, et al. 2022. Expanding horizons of cross-linguistic research on reading: The multilingual eye-movement corpus (MECO). *Behavior research methods*, 54(6).

Nathaniel J. Smith and Roger Levy. 2008. Optimal processing times in reading: A formal model and empirical investigation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 30.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3).

Antoine Tremblay, Bruce Derwing, Gary Libben, and Chris Westbury. 2011. Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning*, 61(2):569–613.

N. R. Underwood and G. W. McConkie. 1985. Perceptual span for letter distinctions during reading. *Reading Research Quarterly*, 20(2).

Tim Vieira, Luca Malagutti, Juan Luis Gastaldi, Brian DuSell, Mario Giulianelli, John Terilla, Timothy J. O'Donnell, and Ryan Cotterell. 2024. From language models over tokens to language models over characters.

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the Predictions of Surprisal Theory in 11 Languages. *Transactions of the Association for Computational Linguistics*, 11.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. Cognitive Science Society.

Simon N. Wood. 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467):673–686.

Simon N. Wood. 2017. *Generalized Additive Models: An Introduction with R*, 2nd edition. Chapman and Hall/CRC.

Jun Sen Yee, Mario Giulianelli, and Arabella J. Sinclair. 2024. Efficiency and effectiveness in task-oriented dialogue: On construction repetition, information rate, and task success. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation*.

A A Note on Spurious Ambiguity

The potentially exponential size of $|\mathcal{C}(\boldsymbol{\sigma})|$ stems from **spurious ambiguity**, i.e., when character strings in Σ^* correspond to more than one token string in Δ^* (cf. Gastaldi et al., 2024). For example, due to its BPE-based tokenizer, GPT-2 can generate the character string `footprint` using its canonical tokenization `foot print`, but it can also generate hundreds of non-canonical tokenizations of the same word, e.g., `foot pr in t` and `f o o t p r i n t`, all of which have to be accounted for when marginalizing a token-level language model to a character-level language model. In the case that there are many such tokenizations, it is #P-hard to compute character-level surprisal exactly (Geh et al., 2024), justifying our reliance on an approximate beam summing algorithm.

B Experimental Design

Our experiments investigate the predictive power of the surprisal of different focal areas for reading time measurements across four eye-tracking datasets, presented in App. B.1. We consider ROIs including either a leading or a trailing whitespace, estimate the surprisal for the focal areas of ROIs using a parameterized language model (App. B.2), and then fit a statistical model to the average measurements[8] of human reading behavior using surprisal estimates as predictors, as outlined in App. B.3. We describe each component of our experimental setup in the following sections.

B.1 Data

We analyze four datasets annotated with reading time measurements collected in eye-tracking experiments with human particpants: UCL (Frank et al., 2013), Provo (Luke and Christianson, 2018), MECO (Siegelman et al., 2022), and CELER (Berzak et al., 2022). For MECO and CELER, both multilingual datasets, we include only data from English stimuli and participants with English as their first language (L1). These eye-tracking datasets provide several ROI-based measurements of reading times. Our study focuses on four specific measurements (Rayner, 1998): first fixation duration, gaze duration, total duration, and skip rate.

UCL (Frank et al., 2013). The UCL Corpus of eye-tracked reading times contains 205 stimuli extracted from three English novels. This datasets attempts to serve as a gold standard for evaluating computational psycholinguistic models of English sentence comprehension. It addresses limitations of previous datasets by using independent sentences that can be understood without extensive context or extra-linguistic knowledge. The corpus includes data from 43 subjects who were recruited from the University College London subject pool. Eye movements were recorded using a head-mounted EyeLink II eyetracker with a 500 Hz sampling rate. Stimuli range from 5 to 15 words, for a total of 1726 ROIs; measurements for the first ROI in a stimulus are omitted. We analyze first fixation duration, gaze duration, and total (right-bounded fixation) duration, and exclude go-past time, a measurement that includes the time spent by the reader in regressions to previous words.

Provo (Luke and Christianson, 2018). This corpus consists of 136 sentences of English text from a variety of genres, including online news articles, popular science, and public-domain works of fiction. These sentences were presented as part of 55 short passages, with an average length of 50 words and 2.5 sentences per passage. Eye movement data was collected from 84 native English speakers using an SR Research EyeLink 1000 Plus eye-tracker. Participants read the texts for comprehension while their eye movements were recorded. The Provo corpus was designed to facilitate the investigation of predictability effects in reading and offers a more naturalistic distribution of word predictability compared to traditional sentence completion norms. In this work, we analyze first fixation duration, gaze duration, and skip rate.

MECO (Siegelman et al., 2022). The Multilingual Eye Movement Corpus (MECO) contains eye-tracking data from L1 speakers (between 29 and 54 per language) for 12 simplified Wikipedia-style articles in 13 languages. In our analysis, we only include English stimuli and responses from 46 L1

[8] We average an ROI's measurements across participants, following a standard procedure used extensively in prior work (Smith and Levy, 2013; Wilcox et al., 2020; Meister et al., 2021; de Varda et al., 2023; Wilcox et al., 2023, *inter alia*). See Smith and Levy (2013) for experiments verifying that this leads to the same linear surprisal effects in eye-tracked reading time datasets.

speakers of English. We analyze first fixation duration, gaze duration, and total duration for comparability with previous work (Wilcox et al., 2023).

CELER (Berzak et al., 2022). The Corpus of Eye Movements in L1 and L2 English Reading (CELER) is a large-scale eye-tracking dataset focused on English reading, consisting of data from 365 participants, including 69 native English speakers (L1) and 296 non-native English speakers (L2). The dataset contains over 320,000 words, with each participant reading 156 newswire sentences from the Wall Street Journal. CELER includes reading time and eye movement data (collected using Eyelink 1000 and Eyelink 1000 Plus eye-trackers) for each sentence. Participants are asked comprehension questions to assess their understanding of the read text. In this paper, we consider only L1 English speakers reading sentences shared across all participants, and discarding sentences unique to a single reader. We analyze first fixation duration, gaze duration, total duration, and skip rate.

B.1.1 Data Preprocessing

We only apply a simple data filtering step: We skip the first region of every stimulus. This makes our analysis consistent across datasets (as measurements corresponding to the first region are not always available) and comparable with prior work, which adopts the same procedure (Frank et al., 2013; Goodkind and Bicknell, 2018; de Varda et al., 2023).

B.2 Language Models

All experiments are conducted using GPT-2 (Radford et al., 2019) in its `small` variant. Despite its size, GPT-2 `small` has been shown to have greater predictive power for reading time data than larger models (Oh and Schuler, 2023; Shain et al., 2024). As explained in §4, we use the beam summing algorithm proposed by Vieira et al. (2024) to compute the surprisal of focal areas, setting the beam size to 5.

B.3 Linear Modeling

Given a dataset of ROIs extracted from a corpus of psycholinguistic stimuli, our goal is to design a statistical model that explains the measurements associated with each stimulus in terms of surprisal predictors. We use linear modeling for our regression analyses as previous work has shown the relationship between surprisal and reading time measurements is largely linear (Smith and Levy, 2008, 2013; Wilcox et al., 2023; Shain et al., 2024). To investigate the predictive power of the surprisal of different focal areas, we employ a 2-by-10 design, with two ways of constructing regions of interest (see §2.2) and ten ways of defining a region's focal area (see §3.3). For regions of interest, we include segmentative sequences with a leading whitespace and segmentative sequences with a trailing whitespace. As focal areas, we consider the entire region, a fixed-size focal area covering the first three characters of the region, dynamically sized focal areas with a rightward word identification span of 7 and 8 characters, and look-ahead focal areas peeking into the next 3–7 characters as well as spanning over the entire next ROI. Tab. 1 shows the twenty sequences of focal areas that result from this experimental design for an example stimulus. The next sections describe the analysis procedure in detail, first presenting our metric of predictive power (App. B.3.1), then the specific predictors used in our linear models (App. B.3.2 and App. B.3.3).

B.3.1 Predictive Power

For each combination of ROI and focal area type, we compare a **baseline regressor** including well-established predictors of reading behavior (frequency and length) to a **target regressor** which, on top of the baseline predictors, includes the surprisal of the focal area. To isolate the true predictive power contributed by a target predictor of interest (i.e., the surprisal of an ROI's focal area) from that of baseline predictors (i.e., an ROI's length and frequency), we inspect the difference in R^2 assigned to a held-out set by the baseline regressor and the target regressor, which we denote as Δ_{R^2}. We estimate Δ_{R^2} via 10-fold cross-validation, iterating over 10 random seeds. We fit the regressor on 9 data folds at a time by finding the coefficients that minimize the residual sum of squares, and then measure the regressor's R^2 on the 10^{th} fold to evaluate its fit. As our final measure of **predictive power**, we report the average Δ_{R^2} across folds and random seeds, with 95% confidence intervals. To assess the statistical significance of a target predictor's Δ_{R^2}, we run paired permutation tests with the cross-validation results;[9] see Giulianelli et al. (2024a) for a detailed description of these significance tests. Finally, for comparison with prior work (Goodkind and Bicknell, 2018; Wilcox et al., 2020, 2023, *inter alia*), we also calculate the average per-ROI difference in log-likelihood $\Delta_{\mathcal{L}}$ of the test set between the target and baseline regressor, and report $\Delta_{\mathcal{L}}$ as an additional metric of predictive power (see App. C).

B.3.2 Baseline Predictors

We consider two baseline predictors: the length of an ROI measured in characters and the logarithm of the ROI's frequency, obtained using the `wordfreq` software.[10] Both length and frequency are well-established predictors of reading times (Rayner, 1998). The impact of length is fairly intuitive, and the effectiveness of frequency as a predictor has been demonstrated in numerous studies (Inhoff and Rayner, 1986; Rayner and Duffy, 1986; Hyönä and Olson, 1995, *inter alia*).

B.3.3 Target Predictors

As the target predictor for a given ROI's measurements, we use the surprisal of the ROI's focal area, calculated as described in App. B.2. Finally, to account for spillover effects, we also include in the target regressor the length, log-frequency, and full ROI surprisal of the previous two ROIs (Just et al., 1982; Frank et al., 2013). The surprisal of the previous two ROIs is calculated according to the ROI definition (with leading or trailing whitespace) of the main target predictor. This makes predictive power scores comparable across ROIs and focal areas.

C Additional Experimental Results

Our experiments investigate the predictive power of focal areas for eye-tracked measurements of reading behavior. In §5 of the main paper, we report results on skip rate (CELER), first fixation duration (UCL, MECO, and Provo), gaze duration (UCL, MECO, and Provo), total duration (UCL, MECO, and CELER), using Δ_{R^2} as our metric of predictive power. On CELER, we find no significant predictors of first fixation duration and gaze duration. Here, we report further results on skip rate in CELER, using $\Delta_{\mathcal{L}}$ as a metric of predictive power (Fig. 3) and in Provo (Fig. 4, both with Δ_{R^2} and $\Delta_{\mathcal{L}}$). The latter follow the same trends as skip rate in CELER albeit with lower predictive power. Finally, Fig. 5 shows results on first fixation duration, gaze duration, and total duration on the same datasets as in §5 but using $\Delta_{\mathcal{L}}$ as a metric of predictive power.

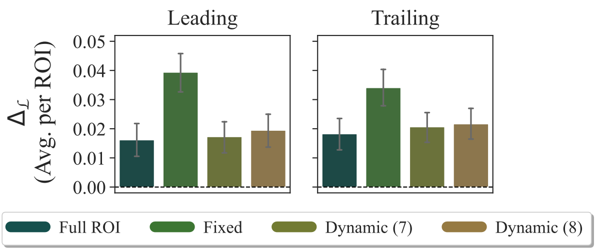

Figure 3: **Skip Rate (CELER)**. Predictive power $\Delta_{\mathcal{L}}$ of an ROI's character-level surprisal, calculated with varying focal areas and two ROI types: leading (left) or trailing (right) whitespace. All $\Delta_{\mathcal{L}}$ scores are significantly above zero ($p < 0.001$). Error bars represent 95% confidence intervals. The black dotted line corresponds to the baseline regressor, including ROI length and frequency. The target regressor includes the length, frequency, and full ROI surprisal of the previous two ROIs.

[9] We use the implementation provided by the `SciPy` library under `scipy.stats.permutation_test`.

[10] We use the Zipf frequency (Brysbaert et al., 2012), i.e., the base-10 logarithm of the number of times an ROI (with whitespaces removed) appears per billion words.

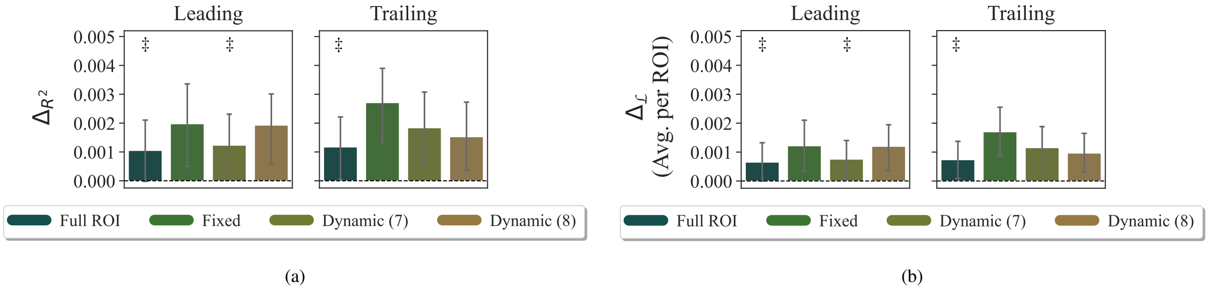

Figure 4: **Skip Rate (Provo)**. Predictive power (subfigure (a): Δ_{R^2}; subfigure (b): $\Delta_{\mathcal{L}}$) of an ROI's character-level surprisal, calculated with varying focal areas and two ROI types: leading or trailing whitespace (respectively left and right within each subfigure). All Δ_{R^2} and $\Delta_{\mathcal{L}}$ scores are significantly above zero with $p < 0.01$, unless marked with ‡ ($p \geq 0.01$). Error bars represent 95% confidence intervals. The black dotted line corresponds to the baseline regressor, including ROI length and frequency. The target regressor includes the length, frequency, and full ROI surprisal of the previous two ROIs.

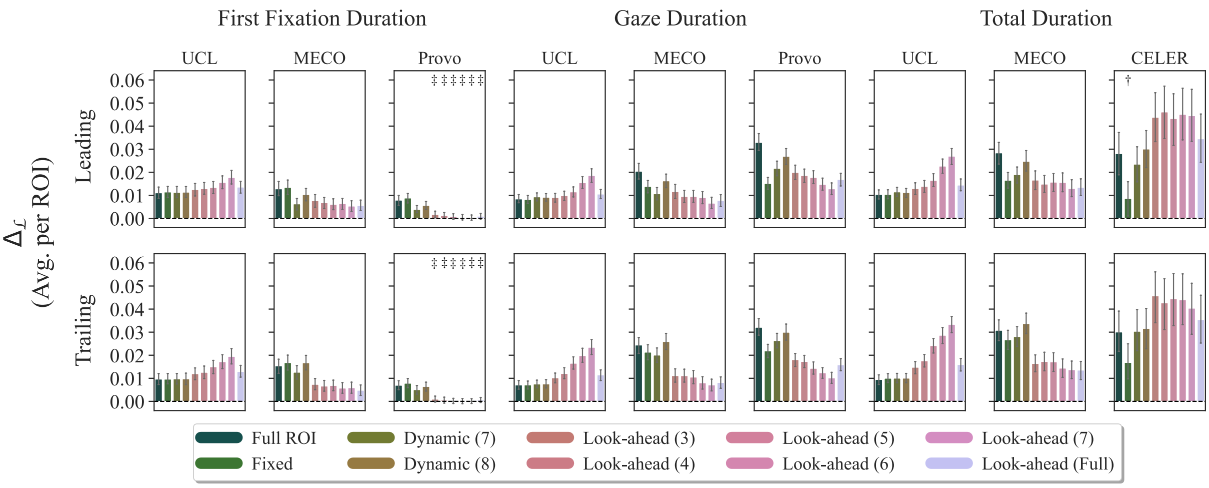

Figure 5: **First Fixation, Gaze, and Total Duration**. Predictive power $\Delta_{\mathcal{L}}$ of an ROI's character-level surprisal, calculated with varying focal areas and according to two ROI types: with leading (top) or trailing (bottom) whitespace. All $\Delta_{\mathcal{L}}$ scores are significantly above zero with $p < 0.001$, unless marked with † ($p < 0.01$) or ‡ ($p \geq 0.01$). Error bars represent 95% confidence intervals. The black dotted line represents the baseline regressor, including ROI length and frequency. The target regressor includes the length, frequency, and full ROI surprisal of the previous two ROIs to account for spillover effects.