Developmental, Typological, and Diachronic Applications of Cognitively Plausible Language Models

Dissertation

zur Erlangung des akademischen Grades eines

Doktors der Philosophie

der Philosophischen Fakultät

der Universität des Saarlandes

vorgelegt von Julius Steuer

aus WADERN

Saarbrücken, 2025

Dekanin:

Univ.-Prof. Dr. Nine Miedema

Berichterstatter:

Univ.-Prof. Dr. Dietrich Klakow

Prof. Dr. Ethan Gotlieb Wilcox

Tag der letzten Prüfungsleistung: 14.04.2026

Abstract

The linguistic competence of today's neural-network-based language models has given rise to a new paradigm in cognitive modeling: Since human language processing is thought to be partially predictive in nature, the processing effort associated with a linguistic unit –a phoneme in a phonological word, a word in sentence, or a sentence in a document– can be approximated by the probabilistic predictions of a language model. The measure or correlate of processing effort is the negative logarithmic probability of the linguistic unit in context, or its surprisal. Surprisal mirrors how well the linguistic unit in question fits the language model's training data, i.e., how likely the unit is given the training data and the context units. While there is broad evidence for a correlation of processing effort and surprisal, the nature of today's transformer-based language models, who during training are exposed to several orders of magnitude more data than a human during language acquisition, has led to a disconnect between a model's linguistic competence and its applicability to cognitive modeling. In this thesis, I explore several strategies to obtain plausible surprisal estimates from language models by controlling the model's training data. I show that the disconnect a model's linguistic competence and cognitive plausibility is most severe in the case of reading times, even if the language model is trained on a developmentally plausible corpus. I then develop a framework in which a language model can serve as an approximation of an average reader of scientific literature in a diachronic setting. Finally, a study on cross-linguistic surprisal of personal pronouns provides an example of a research question that can be tackled with surprisal from a large multilingual model. Throughout the thesis, I highlight avenues for future research that involve tightly controlling inductive biases in the training data, tokenization, and model architecture, as well as methods to test if a language model actually acquired the linguistic structure under consideration.

Zusammenfassung

Die sprachliche Kompetenz der heutigen, auf neuronalen Netzwerken basierenden Sprachmodelle hat zu einem neuen Paradigma in der kognitiven Modellierung geführt: Da die Verarbeitung menschlicher Sprache als teilweise prädiktiv angesehen wird, kann der mit einer sprachlichen Einheit verbundene Verarbeitungsaufwand –ein Phonem in einem phonologischen Wort, ein Wort in einem Satz oder ein Satz in einem Dokument– durch die probabilistischen Vorhersagen eines Sprachmodells approximiert werden. Das Mass oder Korrelat des Verarbeitungsaufwands ist die negative logarithmische Wahrscheinlichkeit der sprachlichen Einheit im Kontext, oder ihr Surprisal. Surprisal spiegelt wider, wie wahrscheinlich die Einheit angesichts der Trainingsdaten und des Kontexts ist. Zwar gibt es zahlreiche Belege für eine Korrelation zwischen Verarbeitungsaufwand und Surprisal, doch hat die Natur der heutigen transformerbasierten Sprachmodelle zu einer Diskrepanz zwischen der sprachlichen Kompetenz eines Modells und seiner Anwendbarkeit auf die kognitive Modellierung geführt. In dieser Arbeit untersuche ich verschiedene Strategien, um plausible Schätzungen von Surprisal aus Sprachmodellen zu erhalten. Ich zeige, dass die Diskrepanz zwischen der sprachlichen Kompetenz und der kognitiven Plausibilität eines Modells im Fall der Lesezeiten am grössten ist, selbst wenn das Sprachmodell auf einem entwicklungsplausiblen Korpus trainiert wurde. Anschliessend entwickle ich einen Rahmen, in dem ein Sprachmodell als Annäherung an einen durchschnittlichen Leser wissenschaftlicher Literatur in einem diachronen Kontext dienen kann. Schliesslich liefert eine Studie zur sprachübergreifenden Überraschung von Personalpronomen ein Beispiel für eine Forschungsfrage, die mit Hilfe von Surprisal aus einem grossen mehrsprachigen Modell angegangen werden kann. Über die gesamte Arbeit zeige ich Wege für zukünftige Forschung auf, die eine Kontrolle des induktiven Biases in den Trainingsdaten, der Tokenisierung und der Modellarchitektur beinhalten.

Contents

1 Introduction

1.1 Motivation

The primary definition of the term language model language model (LM), as given by, e.g., H. Li (2022), is that of a probability distribution over sequences of (natural) language. This is with good reason: All modern LMs are trained to predict words from other words, and virtually all modern LMs based on the transformer architecture (Vaswani et al., 2017) predict structured sequences of words, i.e., phrases, sentences, paragraphs or documents. While this definition is correct in a mathematical sense, it abstracts away from the actual modeling decisions, such as the LM's architecture, training algorithm, training data and its tokenization, [1]. For most of the history of language modeling, none of these parts of a LM was thought or conceived of as representative of the human language faculty in a meaningful way. N-gram models, which dominated in the language modeling community for 30 years, are first and foremost large lookup tables of counts combinations of words, or n-grams, with some statistical machinery to handle out-of-distribution n-grams on top. The underlying assumptions, 1) the upper bound on a meaningful context size for a word's probability, and 2) the statistical independence of n-grams, were a consequence of limited computational resources. Even when estimated

1 Tokenization algorithms can be seen as LMs because they too maximize the probability of a sequence of linguistic units, e.g., via a unigram LM in the SentencePiece tokenizer (Kudo and Richardson, 2018)

from large amounts of texts, n-gram counts are notoriously unreliable, and much of the improvement in LM quality came from improvements in n-gram smoothing and the handling of out-of-vocabulary tokens. Whatever the nature of the human language faculty, the human brain clearly does not store n-gram counts, and humans are (mostly) able to relate the current word (e.g., while reading) to a long context, if only through a compressed representation of this context. Modeling this conditioning on a compressed representation of a long context became possible with the advent of Recurrent Neural Networks (Elman, 1990) and, more importantly, the rapid improvements in computational power and efficiency required to train them. These recurrent models were in term superseded by the transformer architecture (Vaswani et al., 2017), which in turn was superseded by large language models (LLMs). LLMs are no longer exclusively trained on a language modeling task, but also adapted to a wide range of tasks after the initial pre-training[2] to instruction following (instruction tuning), or to use intermediate steps when exposed to a difficult problem (chain-of-thought reasoning). Their previously unheard-of abilities in a wide range of language tasks (see, e.g., Annepaka and Pakray 2025) have made them interesting objects of study by cognitive scientists, and they certainly capture important aspects of human language processing (see Futrell and Mahowald (2025) for a recent overview). This has led some researchers to assume that transformer-based LMs are also good models of human cognition and language processing, e.g. Piantadosi (2023), a statement which has been strongly rejected by others (Chomsky, 2023; Katzir, 2023; Kodner et al., 2023). In the last few years, the field has partially moved away from relying on transformer-based LMs pre-trained on the entire internet to evaluate theories about human language processing (Timkey and Linzen, 2023; Mueller and Linzen, 2023), but transformer-based LMs are still widely used in research that makes use of surprisal theory (Levy, 2008; Hale, 2001a), as they provide a straightforward way to estimate word probabilites and in contrast to, e.g., neural parsing models, can be trained on raw text data. Except for the brief period around 1990, when progress in model architecture was coming from the cognitive sciences, cognitive modeling has always been somewhat of an afterthought of model

2 An LM trained only on a language modeling task is also called a *foundation model*.

architecture, which is mainly driven by the available hardware (it is no coincidence that transformers lend themselves to distributed training on tens of thousands of GPUs). While it is refreshing to see progress and independent thought on the architectural side, linguistic research that makes use of information-theoretic concepts like surprisal is still reliant on good estimates of in-context word probabilities. This thesis represents an attempt of finding cognitively plausible language modeling approaches that work for a number of linguistic research questions, spanning historical linguistics, linguistic typology and psycholinguistic modeling. I will put special emphasis on how modeling decisions affect the comparability and applicability of probability distributions induced by LMs, and which changes in the modeling decisions are required to answer specific linguistic research questions.

1.2 Research Questions

My thesis investigates language modeling decisions for 3 distinct linguistic research questions. In each of them, I set out my choices of model architecture, training data and tokenization in order to obtain surprisal values that are both cognitively plausible and comparable across different architectures, languages, or different registers and diachronic stages of the same language. I evaluate LM surprisal as a correlate of human language processing effort and the interplay of model size and pre-training regime with the explanatory power of surprisal. It does so in two fundamentally different ways:

Surprisal as a predictor of psycholinguistic measures or its psychometric predictive power (PPP). Here surprisal is used as one of many predictors of some psycholinguistic response measure in a statistical model, e.g., reading times. The quality of fit of surprisal to the psycholinguistic measure is expressed as the difference in log-likelihood between a baseline model without surprisal as a predictor and the full model. This aspect is treated in Chapter 3, answering

RQ1: Can neural LMs be considered good models of human linguistic competence when exposed to similar constraints as a human during language acquisition?

Surprisal as a measure of processing effort in the absence of experimental data. If psycholinguistic measures of processing effort are impossible or too costly to collect (e.g, for historical data or in a large multilingual corpus), surprisal can serve as a proxy measure. In a statistical model surprisal takes the place of the response variable, and linguistic features of the linguistic material are used as predictors. This aspect is treated in Chapters 4-6, answering the following research questions:

RQ2: Can surprisal from LMs recover cross-linguistic differences in the structuring of pronoun systems? (Chapter 4)

RQ3: Does scientific English become "less optimal" over time with respect to the memory-surprisal trade-off, when controlling for (a LM's) vocabulary, tokenization and training budget? (Chapters 5-6)

1.3 Contributions

My work so far was rooted in surprisal theory. Surprisal theory posits that human processing effort when encountering a word is proportional to the word's predictability in context, or surprisal. I have used surprisal as a correlative of online measures of human processing effort (Steuer et al., 2023), but the largest part of my work was on obtaining meaningful and cross-lingually comparable surprisal estimates from LMs. I have either trained these LMs myself, or chose them according to the criteria of the project, e.g., to make surprisal estimates comparable across languages (Talamo et al., 2025a) or across different time periods of the same language (Steuer et al., 2024a; Steuer et al., under review). In these studies, my aim was to explain synchronic and diachronic variation through adaptations to processing effort, and to reformulate facts we already know about language(s) with surprisal theory.

The main content of this thesis is based on the following four publications:

Steuer, Julius, Marie-Pauline Krielke, Stefania Degaetano-Ortlieb, Elke Teich, and Dietrich Klakow (under review). "Modeling the Memory-Surpisal Trade-Off over

Time: Communicative Efficiency Decreases with Lexico-Grammatical Change in Scientific English". In: *Submitted to ACL Rolling Review*.

Talamo, Luigi, Julius Steuer, and Annemarie Verkerk (2025). "They saw it, onu, 它, coming: An information theoretic study of cross-linguistic variation in personal pronouns," in: *Linguistics*, to appear.

Steuer, Julius, Marie-Pauline Krielke, Stefan Fischer, Stefania Degaetano-Ortlieb, Marius Mosbach, and Dietrich Klakow (May 2024). "Modeling Diachronic Change in English Scientific Writing over 300+ Years with Transformer-based Language Model Surprisal". In: *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC-COLING 2024*. Ed. by Pierre Zweigenbaum, Reinhard Rapp, and Serge Sharoff. Torino, Italia: ELRA and ICCL, pp. 12–23. URL: https://aclanthology.org/2024.bucc-1.2.

Steuer, Julius, Marius Mosbach, and Dietrich Klakow (Dec. 2023). "Large GPT-like Models are Bad Babies: A Closer Look at the Relationship between Linguistic Competence and Psycholinguistic Measures". In: *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*. Ed. by Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. Singapore: Association for Computational Linguistics, pp. 142–157. DOI: 10.18653/v1/2023.conll-babylm.12. URL: https://aclanthology.org/2023.conll-babylm.12.

1.4 Outline

In Chapter 2 I will first provide an introduction to the transformer architecture, decoder-only LMs, their usual training procedure and the relationship between model size, model quality and cognitive plausibility. I will then introduce several surprisal-based measures of processing effort and their role in the evaluation of the linguistic competence of a

LM. Finally, I will relate these concepts to the main content of this thesis in Chapters 4-6. I will discuss future research directions at the end of Chapters 3, 4 and 6, and again in my conclusion.

2

Background

This chapter introduces the definitions, concepts, and techniques central to the research presented in this thesis. Concepts specific to only one of the four chapters will be discussed in the respective chapter, e.g., the mathematical definition of the memory-surprisal trade-off (Hahn et al., 2021) in Chapter 6 or the personal pronouns in Chapter 4.

2.1 Decoder-only language models

The original transformer architecture as introduced by Vaswani et al. (2017) consists of two basic building blocks: an ENCODER that converts a sequence of word embeddings $S_{ENC} = w_1, w_2, ..., w_{|S_{Enc}|}$ representing any length of sequentially ordered text into contextualized embeddings $S' = w'_1, w'_2, ..., w'_{|S|}$ through a self-attention mechanism, and a DECODER that uses cross-attention to simultaneously attend to S' as well as words generated up to timestep t, $S_{DEC} = u_1, u_2, ..., u_{t-1}$ when generating word u_t. At training time, the ground truth (word embeddings) of the decoded sequence is masked at indices $i > t - 1$, preventing the flow of information from future timesteps [1].

The original transformer architecture was developed for a machine translation setting in which the encoder reads in the source sentence and the decoder produces the translation. For pure language modeling analogous to n-gram and recurrent neural

1 This is the principal difference to masked LMs like BERT (Devlin et al., 2019)

models – with which this thesis is mainly concerned – the transformer architecture was simplified to only include the decoder with its self-attention mechanism. Examples for (open-source) decoder-only LMs are Pythia (Biderman et al., 2023), mGPT (Shliazhko et al., 2022) and Llama 3 (Grattafiori et al., 2024). These models have in common their pre-training task, next-word prediction, which will be explained in the next section.

When training a decoder-only transformer LM from scratch, we begin by initializing its parameters θ_{TF} randomly[2], that is, the embedding layer, the attention module and feedforward networks. A forward pass through the model of a sequence S entails the retrieval of the query, key and value vectors of each $u_i \in S_{DEC}$ and arranging them into matrices Q, K, V. In the first decoder layer, these vectors are retrieved from the word embedding table, in higher layers they are replaced by the last layer's contextualized embeddings. Attention scores are then computed by multiplying the Q and K matrices, scaling them by the square root of the key dimension $\sqrt{d_k}$, and multiplying the V matrix with the softmax of the resulting vector:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{2.1}$$

with softmax applied to the i^{th} element of a vector z defined as

$$\text{softmax}(z)_i = \frac{e^{z_i}}{\sum_{k=0}^{|z|} e^{z_k}} \tag{2.2}$$

In practice, this operation is performed multiple times per input: for each attention head h_i, there is a set of matrices W_i^Q, W_i^K, W_i^V that project the input embeddings to the dimensions d_q, d_k, d_v of the query, key and value vectors (in practice, $d_q = d_k = d_v$). This allows the model to attend to different features of the input sequence, but requires a merging operation of the heads: After each attention layer, the heads h_i are concatenated to a matrix H and then projected to the hidden dimension d_{model} of the decoder:

$$\text{MultiHead}(Q, K, V) = HW^O \tag{2.3}$$

2 There are multiple ways to arrive at a good initialization of a model's parameters, which is beyond the scope of this thesis.

With $W^O \in \mathbb{R}^{hd_v \times d_{model}}$. The output of the attention layer is then passed to a feedforward network, which projects each position first to a hidden dimension d_{FFN} and then to d_{model}, with a ReLU function in between:

$$\text{FFN}(x) = \textbf{max}(0, xW_1 + b_1)W_2 + b_2. \tag{2.4}$$

Layer-wise application of θ_{TF}, that is, $\mathrm{MultiHead}()$ and $\mathrm{FFN}()$, yields the contextualized embeddings $S'_{DEC} = u_1, u_2, ..., u_{|S|}$.

2.2 Language model training

In principle, every $u_i \in S'_{DEC}$ could be the input to what is called the language modeling head, that is, a fully connected feedforward network $\mathrm{LM}()$ with randomly initialized parameters θ_{LM} that projects u_i to a vector of the size of the vocabulary $|V|$ by multiplying it with the embedding matrix $W^E \in \mathbb{R}^{d_{model} \times |V|}$. Usually, the contextualized embedding of the last input token is passed on to $\mathrm{LM}()$.

$$\mathrm{LM}(u_i) = u_i W^E \tag{2.5}$$

Then, softmax is applied to transform the resulting vector into a probability distribution over the vocabulary:

$$\hat{y} = \mathrm{softmax}(z) \tag{2.6}$$

This probability vector is the basis for the calculation of the language modeling loss L by plugging it into the loss function $\mathcal{J}()$, commonly cross-entropy:

$$L = \mathcal{J}(\hat{y}) = -\log \hat{y}_{i+1} \tag{2.7}$$

Where $i + 1$ is the index in the vocabulary of the next word in the ground truth. This loss function is – apart from the base of the logarithm – identical to surprisal, which for surprisal theory has the additional advantage that a LM is trained on the same objective

as it is evaluated on. The loss L is then fed to a Backpropagation (Rumelhart et al., 1986)[3] algorithm, e.g., AdamW (Loshchilov and Hutter, 2019).

2.3 Information-theoretic measures of human processing effort

2.3.1 Surprisal as a metric of lexical and syntactic expectancy

Surprisal theory Levy (2008) and Hale (2001b) posits that the processing difficulty encountered by a human hearer or reader is proportional to the negative logarithmic probability of a word given the t preceding words:

$$\text{surp}(w_n|w_{n-t}, ..., w_{n-1}) = -\log_2 p(w_n|w_{n-t}, ..., w_{n-1}) \tag{2.8}$$

Surprisal as derived by Levy (2008) is a measure of syntactic reanalysis, quantifying the shift in probability mass of a distribution over possible parses based on the requirements of the next word. If word w_n is predictable from its context $w_{n-t}, ..., w_{n-1}$, the conditional probability in the above formula is expected to be large, corresponding to a small surprisal value. For example, in a sentence containing a fixed expression such as German *Die Bank musste Insolvenz anmelden* 'The bank had to file for bankruptcy', the surprisal of *anmelden* 'file for, register', is expected to be low since it is the only possible continuation. Correspondingly, the processing difficulty encountered by a human is predicted to be low. If the number of possible continuations given the preceding context of words is large, that is, if the context contains little information about upcoming words, the probability of the actual next word is expected to be low, corresponding to a high surprisal value. Sticking to our example, a sentence like *Die Bank musste Geld ...* 'The bank had to ... money' has multiple possible continuations (*auszahlen* 'pay', *überweisen* 'transfer', *zurückstellen* 'set aside', ...), each of which would receive a large share of the

3 Rumelhart independently developed Backpropagation for neural networks, but he was not the first to do so: (LeCun, 1988) credits Pontryagin with the first discovery of the algorithm.

probability mass. This formal definition of surprisal is based on the assumption that the underlying stochastic process operates on syntactic representations of the input, meaning that the words of the sentence and their lexical content are important only inasmuch they change the probability distribution over possible parses. While it is common in computational studies of (human) language processing to use surprisal estimates from (large) LMs as a predictor of psycholinguistic measures that are assumed to correlate with human syntactic language processing Oh and Schuler (2023a) and Wilcox et al. (2020), there is a growing body of evidence that LM surprisal primarily captures lexical expectancy effects Arehalli et al. (2022) and J. Huang et al. (2023), while there is a growing body of evidence for a logarithmic effect of a word's predictability given context on processing effort Meister et al. (2021), Shain et al. (2022), and Wilcox et al. (2023b).

2.3.2 Estimating surprisal with causal language models

The conditional next-word probabilities are obtained by first passing a prefix of words (that is, the words $w_{n-t}, ..., w_{n-1}$) to the actual transformer model θ_{TF}, which yields a vector representation or embedding h_{n-1} of the prefix. This prefix is then passed to a feedforward neural network θ_{LM}, which projects h_{n-1} to a vector z_{n-1} of the size of the model's vocabulary. This vector is transformed into a probability distribution over the vocabulary by applying the softmax function, i.e., $\hat{y}_n = \text{softmax}(z_{n-1})$. The surprisal of w_n is then obtained by indexing the resulting probability vector $\hat{y_n}$ with the id i assigned to w_n by the models' tokenizer and taking the negative logarithm of the indexed probability:

$$\text{surp}(w_n | w_{n-t}, ..., w_{n-1}) = -\log_2 \hat{y}_n^i \tag{2.9}$$

We then calculate average surprisal $\hat{S}_t$ at context size t over all $w_n \in W$:

$$\hat{S}_t = \frac{1}{|W|} \sum_{n=0}^{|W|} \text{surp}(w_n | w_{n-t}, ..., w_{n-1}) \tag{2.10}$$

2.3.3 Psychometric predictive power of language model surprisal

Introduced by Frank and Bod (2011), psychometric predictive power (PPP) measures how much the likelihood of a statistical model – in most cases, and also in this thesis, a linear mixed model (LMM) – describing some cognitive response measure improves when computational estimates of that cognitive measure are added to it as predictor variables. In this thesis, and much of the recent psycholinguistic literature, the response measure is reading times either collected in a self-paced reading setting or in an eye-tracking study, and the added predictor is surprisal, or a derived measure like entropy (surprisal's expected value) or an UID measure as in Meister et al. (2021).

The first step in the process is the fitting of base model LMM_0 that predicts the reading time response measure RT from a set of control variables, e.g. word length (in characters) $\mathrm{length}(w)$, word unigram surprisal (as estimated from a corpus) $\log(\mathrm{freq}(w))$, and word position in a stimulus $\mathrm{pos}(w, s)$. On top of these control variables, aLMM allows us to incorporate random intercepts that control for the effect of the individual stimulus (the sentence in question), word, and participant. Usually, these will be nested, i.e., each participant sees multiple stimuli that each have multiple words [4] Using the lmerTest R package (Kuznetsova et al., 2017), LMM_0 can be defined as:

```
LMM_0 <- lmer(
    log(RT) ~
        log_2(freq(w)) + length(w) + pos(w, s)
        + (1|word)
        + (1|participant/stimulus),
    data = .)
```

The base model is then compared to an augmented model LMM_1, which differs from LMM_0 in having at least one more predictor variable, e.g., the surprisal of the word in context $\mathrm{surp}(w|c)$:

4 Of course, words can occur multiple times across stimuli.

```
LMM_1 <- lmer(
    log(RT) ~
        surp(w|c) + log(freq(w)) + length(w) + pos(w, s)
        + (1|word)
        + (1|participant/stimulus),
    data = .)
```

After fitting bothLMMs to the data, they are assigned a likelihood, more precisely the sum of log-likelihoods of the individual data points given the model. If LMM_1 is assigned a *lower* log-likelihood than LMM_1, contextual surprisal increases the likelihood of the data given the model, i.e., it has greater explanatory power than LMM_1. The relative difference in log-likelihood is calculated as

$$\Delta_{\text{LL}} = \text{LL}(\text{LMM\_1}) - \text{LL}(\text{LMM\_0}) \tag{2.11}$$

Since the two LMMs differ only in one parameter (contextual surprisal), we can compare their log-likelihoods directly without recourse to a an adjusted measure like the Akaike information criterion. If Δ_{LL} is positive, contextual surprisal (or any other variable used in its stead) improves the fit of the model to the data and we conclude that LMM_1 is to be preferred over LMM_0.

The relationship of surprisal and reading times has been corroborated by a long list of studies, starting with Goodkind and Bicknell (2018) over Meister et al. (2021), Wilcox et al. (2020), and Shain et al. (2022) to Wilcox et al. (2023a). Outside of English, the same effect was found in a multilingual dataset by Wilcox et al. (2023b), and several studies Meister et al. (2021), Shain et al. (2022), and Wilcox et al. (2023b) provided evidence both for the importance of the next-word prediction task and the logarithmic nature of the effect. Arehalli et al. (2022) and Oh et al. (2022) showed that incorporating a parsing training objective or generally a syntactic inductive bias that goes beyond the surface form of sentences improves the fit of surprisal to reading times.

2.3.4 Bigger is not always better

Until the advent of GPT2 (Radford et al., 2019), PPP correlated positively with LM quality (in terms of perplexity, the exponential of average surprisal) and therefore, usually model size (Wilcox et al., 2020; Wilcox et al., 2023a). It became clear at an early point that the very large context windows (compared to human attention spans) of decoder-only transformer LMs represent a challenge to applying surprisal from these models to cognitive modeling (e.g., Kuribayashi et al. 2022). At the same time, it was widely recognized that the absence of an explicit syntactic bias in the training data lead to a substantial underprediction the effect on reading times of rare syntactically ambiguous constructions like garden-path sentences (Arehalli et al., 2022; K.-J. Huang et al., 2023). Byung-Doh Oh and William Schuler established a negative correlation of the PPP of surprisal for human reading times in LMs larger than GPT2-small (Oh and Schuler, 2023b; Oh and Schuler, 2023a). It has since become clear that this effect is to a large part explained by the fact that larger models tend to overpredict low-frequency tokens, and this effect scales with model size (Oh et al., 2024)[5]. Similarly, Aoyama and Wilcox (2025) found that the effect of deteriorating PPP is driven by the emergence of attention heads specialized on syntactic structure during training, but did not identify the exact reason for this deterioration. However, previous research by Chen et al. (2024) showed that these specialized heads enable the acquisition of grammatical structure as measured by BLiMP (Warstadt et al., 2020a), highlighting the interplay of PPP and a model's linguistic competence.

5 It is conceivable that this effect sets in earlier than the purported "ideal" size of GPT2-small as it is connected to the general training regime of transformer LMs. I am however not aware of any research into this direction.

2.4 Evaluating the linguistic competence of a model

2.4.1 Why should we care?

Piantadosi (2023) argued for taking LMs as good theories of human language processing, because they definitely can use language to solve a wide range of tasks. Of course this was not well received by Chomsky (2023) and, e.g., Katzir (2023), who argue that there is a principal problem with the way in which LMs learn language. This criticism is mostly directed towards (very) large LMs, the likes of GPT4 or Llama 3; smaller models of a few hundred million parameters trained on developmentally plausible data, as introduced in the BabyLM challenge (Warstadt et al., 2023a), have largely been welcomed and are now used to answer a variety of research questions (Futrell and Mahowald, 2025). Still, the debate on the significance of very large models for human language processing is ongoing, mostly revolving around the way in which the linguistic capabilities of a model are to be evaluated. To name one example, Dentella et al. (2023) showed that LLMs are significantly outperformed by humans if asked to evaluate the grammaticality of a sentence by *generating* an answer instead of evaluating surprisal, as is done, e.g., in the now ubiquitous BLiMP Warstadt et al. (2020a). In their reply, J. Hu et al. (2024) criticized their approach on the grounds that this evaluation strategy notoriously understimates both LLMs' and humans' linguistic capacities. This claim was in turn rejected by Leivada et al. (2024), who showed that relying on sequence surprisal alone outside of the typical minimal pair setting drastically decreases LLM performance on grammaticality judgment tasks. Whatever the outcome of this debate, it is clear that there is still some resistance in the linguistic community to accept LLMs as a means to the end of understanding human language processing, based on their very nature of statistical models of text. At the same time, the reliance on neural LMs of some kind for linguistic modeling is accepted, provided that they plausibly represent aspects of the human language faculty.

2.4.2 Evaluating the linguistic competence of a language model

While in principle *all* linguistic tasks that can be solved by a LM can be considered evaluations of its linguistic competence (question answering, summarization, spellchecking, etc.), in the psycholinguistic literature this term has a somewhat narrower meaning. This thesis, in Chapter B, adopts the definition of linguistic competence by Mahowald et al. (2023) and joins to it the dimension of *processing effort*. Mahowald et al. (2023) recognize two dimensions of linguistic competence:

Formal linguistic competence is defined as the "capacity required to produce and comprehend a given language, i.e., the ability to distinguish grammatically correct from incorrect formations, based either on "knowledge of and flexible use of linguistic rules" or "non-rule-like statistical regularities" (Mahowald et al., 2023). An example for the former mechanism would be the regular formation of past tense verbs in English (*look*:*looked*), and for the latter the formation of irregular or ablauting past tense verbs (*go*:*went*,*tread*:*trod*).

Functional linguistic competence is defined as "non-language-specific cognitive functions that are required when we use language in real-world circumstances" (Mahowald et al., 2023) , i.e., the ability to perform cognitive tasks *with* language. GLUE is an example for a benchmark that test this dimension of linguistic competence, with some if its tasks (CoLA (Warstadt et al., 2019)) also testing for aspects of formal linguistic competence.

The dichotomy between formal and functional linguistic competence can be understood in terms of Wittgenstein's definition of the meaning of a word as its use in a language (Wittgenstein, 1953). The debate on whether statistical learners (i.e. LMs) can learn the meaning of a linguistic unit (word, phrase, text, etc.) in Wittgenstein's sense is still ongoing, with much division between positions that strongly deny that LMs can have such a property (Bender and Koller, 2020) and positions that advocate that they might have it, e.g., under the condition that the LM's predictions are grounded in extralinguistic reality (Bisk et al., 2020).

It is common to evaluate on combinations of the three tasks covering both types of linguistic competence, but not on all three at the same time: Meister et al. (2021) looked at the correlation of various formalizations of the uniform information density (UID) hypothesis, reading times and human acceptability judgments, but leaves out an evaluation of formal linguistic competence. This, however, is understandable given their problem setting: it would have been difficult to align, e.g., the BLiMP stimuli to their graded measures of language processing effort and acceptability, as the main focus of their work was *not* LM quality. Similarly, the first iteration of BabyLM Challenge focused on the two forms of linguistic competence introduced by Mahowald et al. (2023), with my work (Steuer et al., 2023) pointing out that it is necessary to also evaluate a model's PPP.

3

Large GPT-like Models are Bad Babies: A Closer Look at the Relationship between Linguistic Competence and Psycholinguistic Measures

As part of the BabyLM challenge (Warstadt et al., 2023b; M. Y. Hu et al., 2024), in this work we looked into modeling constraints of LMs that are supposed to reflect the linguistic competence of a *single* human by pre-training a series of decoder-only LMs on a developmentally plausible corpus, varying model size and the number of training steps. Research on the cognitive plausibility of language models (LMs) has so far mostly concentrated on modeling psycholinguistic response variables such as reading times, gaze durations and N400/P600 EEG signals, while mostly leaving out the dimension of what Mahowald et al. (2023) described as formal and functional linguistic competence, and developmental plausibility. We address this gap by training a series of GPT-like LMs of different sizes on the strict version of the BabyLM pre-trained corpus, evaluating on the challenge tasks (BLiMP, GLUE, MSGS) and an additional reading time prediction task.

The content in this chapter is based on:
Julius Steuer et al. (Dec. 2023). "Large GPT-like Models are Bad Babies: A Closer Look at the Relationship between Linguistic Competence and Psycholinguistic Measures". In: *Proceedings of the BabyLM Challenge at the 27th Conference on Computational*

Natural Language Learning. Ed. by Alex Warstadt et al. Singapore: Association for Computational Linguistics, pp. 142–157. DOI: 10.18653/v1/2023.conll-babylm.12. URL: https://aclanthology.org/2023.conll-babylm.12

Julius Steuer conceptualized the research, conducted experiments and led the paper writing. Marius Mosbach and Dietrich Klakow developed the initial research questions, and Marius Mosbach helped with the paper writing.

3.1 Introduction

In recent years several approaches have been taken to test LMs for cognitive plausibility. This is usually done by using output probabilities of the LM as a predictor for a model's preference towards certain linguistic structures (Roark et al., 2009; Wilcox et al., 2020). Another strain of research uses the output probabilities as a correlate of psycholinguistic measures, e.g., N400 and P600 EEG signals (Heilbron et al., 2019 and recently J. Li and Futrell, 2023) and (self-paced) reading times (Fernandez Monsalve et al., 2012). A natural question that arises is whether cognitive plausibility should be attributed to the model architecture itself, or to the training regime in combination with the training dataset. Little research has been done on the actual neurological plausibility of large language models (LLMs), but Schrimpf et al. (2021) showed that the architecture of BERT-like models is already plausible for the next word prediction task before training: model predictions with only the language modeling head trained are already predictive of human brain activity during reading *and* correlate well with the predictions of the fully trained model. In contrast, no correlation between brain activity and model predictions was found for models trained on GLUE (A. Wang et al., 2019), a natural language understanding (NLU) benchmark. This finding may mirror an underlying difference in language processing between *formal* and *functional linguistic competence* as introduced by Mahowald et al. (2023) and discussed in Chapter 2.4.2. Our study, however, does not attempt to find arguments in favour of either position, but to study the implications of this dichotomy for the paradigm of cognitive modeling.

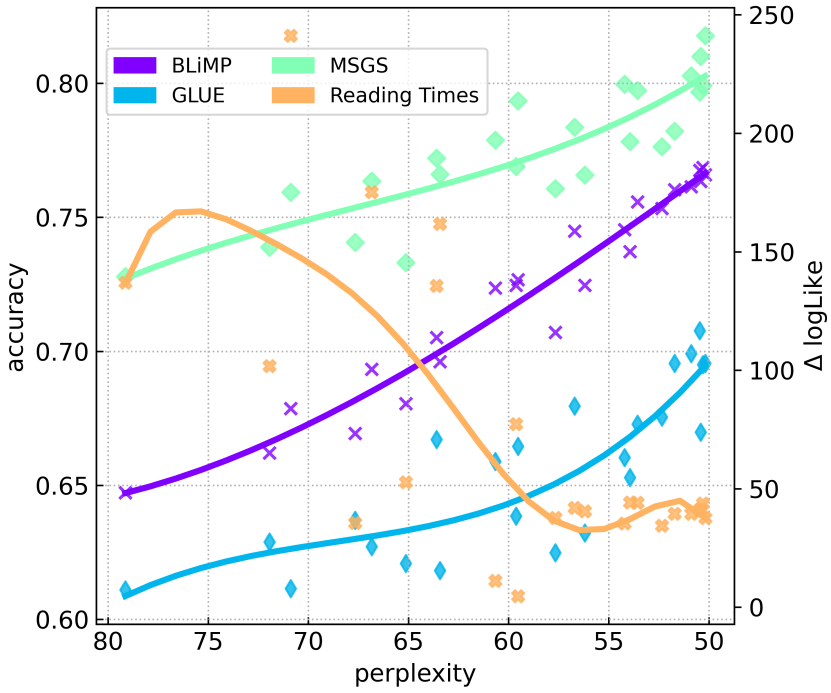


Figure 3.1: Our results show that LM performance on the BabyLM challenge tasks is negatively correlated with perplexity on the development set of the BabyLM corpus (lower perplexity leads to higher performance). In contrast, a *positive* correlation (Spearman's $\rho = 0.4784$, $p < 0.05$) was found between LM perplexity and the fit of LM surprisal to self-paced reading times from the Natural Stories corpus (Futrell et al., 2021) in terms of the difference in log-likelihood between a basline linear mixed-effects model and a model using LM surprisal as a predictor. Lines were fitted with 3 (challenge tasks) or 6 (reading times) degrees of freedom to the LMs' average performance on the task. See Section 3.6 for detailed results.

As stated earlier, the output probabilities of LMs lie at the basis of the application of LMs to cognitive language modeling, usually in the form of a probability distribution over a vocabulary of word forms given either surrounding words (masked language modeling) or preceding words (causal language modeling). Evidence for the use of surprisal (a word's negative logarithmic probability in context) instead of the actual probability comes from logarithmic effects of contextual probabilities on processing difficulty (Shain et al., 2022). Another approach is to evaluate the output probabilities of a LM over a number of classes that may or may not apply to the input sequence,

usually after fine-tuning the LM. The reliance of research in this direction on the output probabilities of LMs has already been criticized from multiple sides. There is a growing body of evidence that the performance of a LM in the typical language modellng task, next word prediction, and measures of formal linguistic competence are not correlated. J. Hu et al. (2020) found no correlation between LM perplexity and measures of formal linguistic competence, while K.-J. Huang et al. (2023) argue that LM surprisal should not be assumed to be a good predictor of psycholinguistic measures of processing difficulty that require more than just lexical information. This lack of correlation with psycholinguistic measures becomes more prominent with the increasing size of LMs (Oh and Schuler, 2022), and especially so in extreme cases of human processing difficulty: Arehalli et al. (2022) showed that surprisal from LSTM-based LMs underestimates garden-path effects on reading times, while successfully predicting reading times for most non-garden-path sentences. This finding has been corroborated for transformer-based LMs such as GPT-2 (Jurayj et al., 2022) and BERT (Irwin et al., 2023).

3.2 BabyLM

The BabyLM challenge (Warstadt et al., 2023a) introduces a novel constraint to cognitively plausible language modeling by limiting the token budget for LM pre-trained to 100 million (100M) tokens, roughly the same amount of tokens a 13-year old child has seen during language acquisition (Gilkerson et al., 2017). While the focus of the challenge is on the pre-trained procedure, the evaluation pipeline consists of the BLiMP (Warstadt et al., 2020a), MSGS (Warstadt et al., 2020b) and GLUE benchmarks, each of which aims to test for a specific dimension of linguistic competence.

BLiMP tests for *formal linguistic competence* by comparing model predictions at a critical word in pairs of grammatically acceptable and unacceptable sentences, with the sentence pair only differing with respect to a single feature, e.g., whether a determiner

agrees with its antecedent in gender or not. A model succeeds at the task if it assigns a higher probability to the critical word in the acceptable sentence.

GLUE is a benchmark that requires fine-tuning[1] of the LM. It tests for a wide range of NLU problems, e.g., question answering, natural language inference and linguistic acceptability judgements, and hence can be regarded as a proxy for the *functional linguistic competence* of a LM.

MSGS is a benchmark of binary classification tasks that tests whether a LM prefers *surface generalizations* over *syntactic generalization* by first fine-tuning on data consistent with both types of generalization. At inference time, items are consistent with only one type, potentially revealing a bias towards either generalization type.

Previous studies mainly provided insights into the relationship of pre-trained token budget and measures of formal and functional linguistic competence. Y. Zhang et al. (2021) showed that encoder-only LMs already perform well on formal tasks such as BLiMP at a budget of 10-100M tokens, while requiring substantially larger token budgets to perform well on functional tasks such as GLUE. While this research established correlations for pre-trained token budgets, similar relationships for *model size* at a fixed token budget have not yet been investigated. This study is dedicated to finding a relationship between model size and performance on these tasks, while simultaneously addressing the dimension of *processing effort*, which is not covered by the challenge tasks. This is done using the **strict** version of the BabyLM corpus, mainly because there is evidence that the fit with psycholinguistic measures profits from token budgets far larger than the 100M tokens in the corpus (Oh and Schuler, 2023a).However, we also implicity evaluate on models that are trained on token budgets of 10M tokens, corresponding rather to the **strict-small** track in Section 3.7.

1 During fine-tuning, we train all parameters of the pre-traineded LM as well as a randomly initialized classifier on top of the LM.

3.3 Research questions

The starting point of our work is Y. Zhang et al. (2021)'s finding of an earlier saturation effect (in terms of pre-trained tokens) for BLiMP as opposed to (Super)GLUE. If performance on BLiMP is already close to the optimum after pre-trained for 100M tokens, we suspect that a model with relatively small capacity is sufficient to reliably learn the required syntactic and semantic features. In contrast, the larger pre-trained token budget and model size needed for GLUE should also require a model with higher capacity.

Studies on reading time prediction generally use causal LMs trained on a next-word prediction task instead of masked LMs (Oh and Schuler, 2022; Arehalli et al., 2022; Jurayj et al., 2022) because of their closer similarity to human language processing. Although masked LMs such as BERT show some word order effects (Papadimitriou et al., 2022) and even garden-path effects (Irwin et al., 2023), they are cognitively implausible in the sense that they process all words in a sequence simultaneously when predicting a word at a masked position, rather than processing language sequentially. This *autoregressive* property mirrors human language processing, and is therefore desirable in studies with the primary goal of modeling human reading behaviour. We therefore employ decoder-only, GPT-like LMs (Radford et al., 2019) in our study, i.e., we want to answer the following research questions:

Research Question A: Are GPT-like models cognitively plausible in the sense that they are able to acquire (a degree of) formal and functional linguistic competence, while being also predictive of human processing effort?

Research Question B: Can such LMs be trained on the same data as a child has available during language acquisition (100M tokens)?

3.4 Previous work

Do we need transformers for cognitive plausibility? Despite promising findings by Hosseini et al. (2021), it has yet to be determined whether transformers, and decoder-only transformer LMs in particular, are cognitively plausible in the sense that they are data-efficient enough to acquire human-like[2] linguistic competence. Indeed, there are results that seem to partially contradict the necessity of LLMs with wide context windows in order for a model to exhibit human-like processing behaviour. Kuribayashi et al. (2022) showed that *reducing* context length of LLMs improves the fit of a LMM on gaze durations, with surprisal from a bigram GPT-2 model as a predictor yielding the largest log-likelihood reduction over the baseline model. Wilcox et al. (2020) failed to identify a relationship between psychometric predictive power (PPP) (Δ log-likelihood) and syntactic generalization, concluding that different models are needed for modeling human processing effort versus syntactic generalization.

Linguistic competence vs. psycholinguistic measures It has long been clear that LM capacity, and subsequently LM perplexity, does not necessarily correlate with human-likeness (Kuribayashi et al., 2021). LLMs such as GPT-3 in particular were found to have considerable disadvantages when it comes to predicting psycholinguistic measures from their next-word predictions: Oh and Schuler (2022) found an inverse relationship between both perplexity and LLM capacity, versus fit to human reading times. The authors of this study hypothesize that this is because transformers have access to the full sequence context, and are trained on large enough corpora to make use of the information that they contain. This relationship between model perplexity and reading times is however not intrinsic to transformer-based LMs: J. Hu et al. (2020) found a similar relationship for LSTM LMs, though small GPT-like models have an advantage over recurrent models.The impact of LM size on linguistic competence was investigated by Eldan and Y. Li (2023), who found that relatively small GPT2-like models ($<$10M parameters) manage to produce fluent English and can be trained on relatively small

2 Here, we do not use "human-like" to imply human-level performance, but rather that the model is *subject to similar processing constraints* as a human.

corpora with a reduced vocabulary. Their study also shows that the relationship still holds for small models, while also identifying trade-offs between model width (hidden size) and depth (number of decoder layers).As for training dataset size, Oh and Schuler (2023a) found that surprisal from transformer-based LLMs gives the best fit to reading times at about 2B train tokens, across a wide range of model sizes. The corpus used in their study is very large (300B tokens), allowing for extensive training of a model without repeating any data. Reaching the same number of update steps with the much smaller BabyLM corpus would require training for multiple epochs.

Single- vs. multi-epoch training Since the BabyLM training data is substantially smaller than the 2B tokens suggested by Oh and Schuler (2023a), training our models in a multi-epoch setting cannot be avoided. Previous research has shown that repeating the training data can have adverse effects: Xue et al. (2023) compared single-epoch vs. multi-epoch training in a limited data setting and show that multi-epoch training leads to overfitting, with little performance being gained after the first epoch. They also find that regularization can only partially alleviate the overfitting problem, with dropout having the largest effect. Not having to repeat the training data is advantageous for downstream tasks and psycholinguistic modeling, if a certain amount of training data is available: Oh and Schuler (2023a) found that reading time fit deteriorates after 2B tokens over a wide range of model sizes. However, it is not clear if repeating the training data would lead to an even stronger deterioration. If the corpus is substantially smaller than 2B tokens, repeating the training data could have a different effect, especially if the optimum of the reading time fit depends on the availability of the 2B tokens.

3.5 Methodology

Modeling We use the OPT architecture by S. Zhang et al. (2022a) with a language modeling head for pre-trained. Following our intuition that BLiMP should require much smaller model sizes than MSGS and GLUE, we train a series of OPT models of different sizes, varying only model width (hidden size) and model depth (number

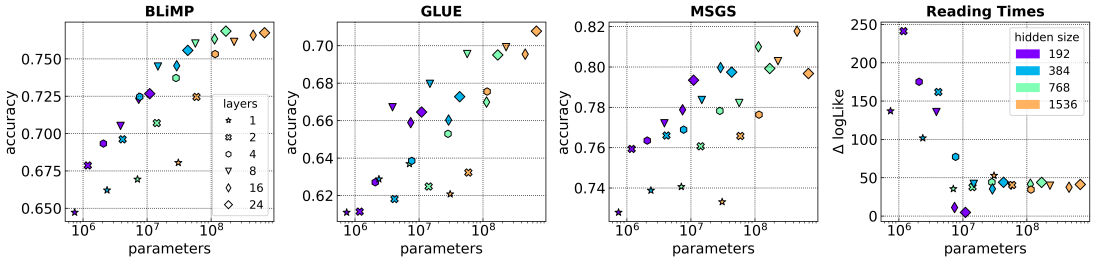


Figure 3.2: Task performance by model size (higher numbers are better). Baselines can be found in Appendix B.3.

of decoder layers). In total we train 24 models varying over 4 hidden sizes $l_{hidden} \in \{192, 384, 768, 1536\}$ and 6 numbers of decoder layers ($l_{decoder} \in \{1, 2, 4, 8, 16, 24\}$). We also adjust the dimension of the feedforward layers such that the size of the output vector $l_{forward} = 3 \times l_{hidden}$. Table B.1 in Appendix B shows the resulting model sizes. The models and all code for pre-trained are implemented with PyTorch (Paszke et al., 2019) and HuggingFace transformers (Wolf et al., 2020a), starting from their implementation of OPT. We also trained a new tokenizer on the training set of the BabyLM corpus, using the same vocabulary size $|V| = 50272$ as the original OPT tokenizer. We report all results as averages over 3 random seeds (see Appendix B for full results and standard error).

Training Following the Shortformer pipeline (O. Press et al., 2021), each model is trained for one epoch with an initial sequence length of 64, followed by 4 epochs with the full sequence length of 256. The full sequence length of 256 was chosen as a compromise between the relatively short test items in the challenge tasks (up to 128 tokens) In order to ensure that the model generalizes to longer sequences we use ALiBI (O. Press et al., 2022) instead of learned positional embeddings. This also ensures that our models generalize to the longer sequences in the Natural Stories corpus. We trained each model on a A100 GPU with 40 GB VRAM and an effective batch size of 128, using gradient accumulation for models that could not fit the full batch size. We used AdamW (Loshchilov and Hutter, 2019) as our optimizer with an initial learning rate of 0.001 and weight decay of 0.001 with 2000 linear warm-up steps. We use a dropout of 0.1 following the default HuggingFace transformers parameters for OPT.

Pre-training experiments We also experimented with changes to the pre-trained regime. We trained models on multiple permutations of the training dataset: ordering sequences according to length (number of words), word length (number of characters), sequence-level perplexity from a 3-gram LM trained on the same data, and different orderings of the subcorpora as in Mueller and Linzen (2023). None of these approaches resulted in significant performance gains in terms of perplexity and performance on the challenge tasks over a baseline model trained on the concatenated BabyLM corpus with subsequent shuffling of the sequences.

Evaluation We evaluated all models on the downstream tasks of the BabyLM challenge. While these three tasks test for the linguistic competence of a model, they do not quantify the cognitive effort associated with language processing. We therefore also evaluate all models on a reading time prediction task. For each model, we calculated surprisal on the items of the Natural Stories Corpus (Futrell et al., 2021). This corpus was chosen because its domain is close to at least one of the BabyLM subcorpora (Children's Stories). We fitted linear mixed model (LMM) models with random intercepts for subject, word and item (the id of the story); surprisal, word frequency, word length and sentence position as predictors and log-normalized reading times as the response variable. The exact formula is

$$\begin{aligned}&\text{lmer}(\log(\text{reading\_time}) \sim \text{word\_surprisal} + \text{len}(\text{word}) + \log(\text{word\_frequency}) + \\ &\quad \text{position} + (1|\text{word}) + (1|\text{subject}) + (1|\text{item}), \text{data} = .)\end{aligned}$$

For the reading time analysis we report the difference in log-likelihood between the models with surprisal as a predictor over a baseline model with only the control predictors. For all other tasks we report accuracy.

Code We used the evaluation code provided by the organizers of the BabyLM challenge[3], with some modifications to load custom models. The evaluation pipeline is based on the LM-Eval framework by Gao et al. (2021). Fine-tuning on GLUE and MSGS was done with the default hyperparameter settings, but we reduced the number of fine-tuning epochs to 3 as we did not observe any improvements after 3 epochs. The LMM models

3 https://github.com/BabyLM/evaluation-pipeline

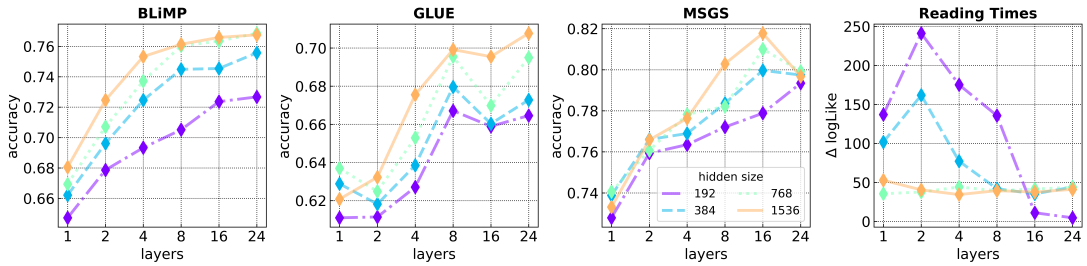


Figure 3.3: Task performance by hidden size, number of layers and task.

were fitted using the lmerTest R library (Kuznetsova et al., 2017) via the pymer4 Python package (Jolly, 2018). The code to pre-trained and evaluate all models is publicly available on GitHub[4]. The model with the highest BLiMP accuracy and detailed results for the LMM models are made available at the same location, alongside instructions on how to run the training and evaluation pipelines.

3.6 Results

Fine-tuning GLUE Fine-tuning on GLUE was overall very unstable and often failed to outperform the baseline. This was mainly due to the one-size-fits-all approach to the fine-tuning hyperparameters; we repeated several more fine-tuning runs with different hyperparameter settings on some of the GLUE tasks, and found that, e.g., RTE profited from a longer warm-up period (which is in line with the findings of Mosbach et al. (2021) for BERT-like models), but most other sub-tasks fine-tuned with the same hyperparameters showed a drop in performance. While we could have optimized hyperparameters for all sub-tasks, the main objective of the BabyLM challenge is to improve the pre-trained part of the NLP pipeline. Thus, we decided to fine-tune with the default hyperparameters, only adjusting the number of epochs as we found that the fine-tuning runs already converged after a few epochs.

Model size Figure 3.2 shows the relationship between model size and task performance: While GLUE (Spearman's $\rho = 0.7739$, $p < 1^{-4}$) and MSGS ($\rho = 0.7148$, $p < 1^{-4}$)

4 https://github.com/uds-lsv/babylm

performance scales with model size, BLiMP performance plateaus after reaching a model size of about 50M parameters ($\rho = 0.8835$, $p < 1^{-4}$). In contrast, reading time fit was negatively correlated with model size ($\rho = -51.39$, $p < 0.05$). All correlations are statistically significant with $p < 1^{-4}$. No single model performed best on all three challenge tasks, with large differences in the size of the best model. Figure 3.1 shows that similar correlations hold for model perplexity and task performance (BLiMP: $\rho = -0.9765$, $p < 1^{-4}$, GLUE: $\rho = -0.8287$, $p < 1^{-4}$, MSGS: $\rho = -0.8661$, $p < 1^{-4}$); negative correlations mean that lower perplexity leads to higher performance. We found strong positive correlations (pictured in Figure B.2 in Appendix B.1) between performance on the challenge tasks (BLiMP and GLUE ($\rho = 0.8784$), BLiMP and MSGS ($\rho = 0.9182$) and GLUE and MSGS ($\rho = 0.815$) generally with $p < 1^{-4}$).

Model width vs. depth While BLiMP performance was not found to be strongly correlated with either the number of decoder layers or hidden size, GLUE and MSGS showed some variability based on the number of layers. For GLUE the only configuration that showed a monotonic improvement in performance was a hidden size of 1536, with models with more decoder layers achieving higher accuracy in this setting. For MSGS we observed a drop in performance for the models with 24 decoder layers at the largest hidden sizes (384, 768). Overall, the effect of hidden size and number of layers was minor when compared to overall model size. In contrast, the best fit on the reading time data was achieved with the second smallest model with only 2 decoder layers and a hidden size of 192. Figure 3.3 illustrates this trend: for the challenge tasks, performance increases with the number of layers (though not monotonically), whereas Δ log-likelihood of the LMM models decreases with the number of layers at $l_{hidden} = 192$ and, to a lesser extent, at $l_{hidden} = 384$, while deeper models with more decoder layers and larger hidden sizes perform considerably worse.

Possible confounds The reading time analysis suffers from several potential confounding factors: Firstly, the domain of the training data differs considerably from the data in the Natural Stories corpus. While the training data also contains some longer texts (Wikipedia, Children's Stories), most of the corpora are more representative of spoken language (Open Subtitles, BNC Spoken, CHILDES). In addition, most sequences are

relatively short, with a median sequence length of 8 in the Open Subtitles corpus, which accounts for >50% of the training data. This is considerably less than the median sequence length of 22 in the Natural Stories corpus. Another confounding factor might be the difference in exposure to language data of the model and that of the participants of the original reading time study. Futrell et al. (2021) do not provide demographic data of their participants, but since data collection was done via Amazon Mechanical Turk we can safely assume that the mean age of the participants was higher than 13, meaning that they were exposed to considerably more language data than the 100M tokens in the BabyLM corpus. Although a recent study by Oh and Schuler (2023a) showed that reading time fit (in terms of Δ log-likelihood) from transformer models still profits from pre-trained data multiple orders of magnitude larger than our corpus, with an optimum at 2B tokens, this is partially alleviated in this study by the multiple-epoch training regime, totalling about 500M tokens seen by each of our models. Since Oh and Schuler (2023a) found that training on more *unseen* tokens after reaching the optimum leads to a quick deterioration of reading time fit proportional to model size, it is unclear what impact repeating the training data would have on the reading time fit.

3.7 Reading time prediction in a multi-epoch setting

Experimental setup In order to evaluate whether the negative correlation is an artifact of the domain mismatch between the BabyLM corpus and the items in the Natural Stories corpus or the repetition of the training data before reaching the optimal token budget, we conduct two additional experiments: First, we retrain all models on the BabyLM corpus for a single epoch, saving intermediate checkpoints at 100, 500, 1000, 2000 and 3000 training steps. Then, we use the intermediate models to fit LMM models to the reading time data, using the same formula as given in Section 3.5. Second, we replicate these experiments on Wikitext-103, a corpus of similar size that does not have the same limitations of the BabyLM corpus (i.e. an average sequence length and a domain closer to the Natural Stories corpus). The models trained on Wikitext-103 serve

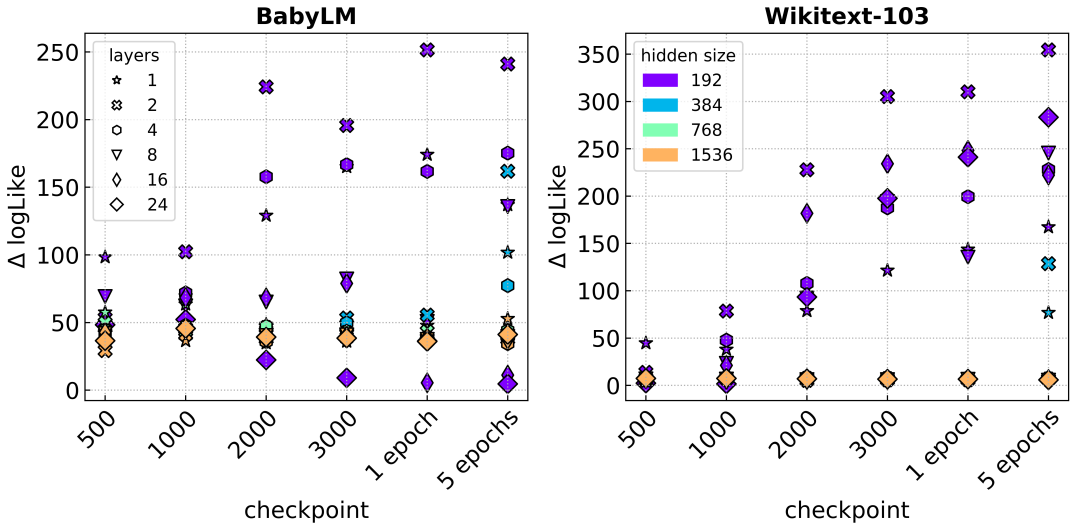


Figure 3.4: Reading time fit in terms of Δ log-likelihood over a base model without surprisal as a predictor, on the BabyLM and Wikitext-103 data after 500, 1000, 2000 and 3000 training steps (1/8, 1/4, 1/2 and 3/4 of an epoch) and 1 and 5 epochs.

as a control for the experiments on the BabyLM corpus and were not included in the final submission. Since the results indicate that larger models yield a worse reading time fit, we restrict the experiment to small models (1-4 layers, all hidden sizes) and larger models with the smallest and largest hidden size (192 and 1536). The models are trained with the same hyperparameter settings as the original models, but sequence length is not reduced in the first epoch.

Results Figure 3.4 shows a somewhat different picture for the models trained on Wikitext-103, with reading time fit of smaller models increasing over the whole pre-trained procedure, while models with $l_{hidden} > 192$ almost never improve over the baseline model. In contrast, the reading time fit of the LMs trained on the BabyLM data improves significantly over the baseline for shallower models (< 2 decoder layers), while staying roughly constant for deeper and wider models (16, 24 decoder layers). However, the relationship between the number of training steps and reading time fit is not monotonic, with a slight decrease after training for 4 more epochs for the best model. While the models trained on the Wikitext-103 dataset yield a better fit to reading times in terms of Δ log-likelihood, the basic finding on the BabyLM data is corroborated: exposing a transformer model to multiple repetitions of the training data before reaching

the optimal token budget does not lead to a decrease in reading time fit, but also does not improve over the single epoch setting in a meaningful way. The results also show that the improved reading time fit for $l_{hidden} = 192$ cannot be attributed to smaller model size alone, as the deepest model with that hidden size, 24*192 shows an improved fit over the baseline, while 1*384, a model with a comparable number of parameters, but a larger hidden size, does not. In conclusion, we did not find a degradation of reading time fit when repeating the training data, with similar effects of LM size on reading time fit for Wikitext-103 and the BabyLM corpus (see Table B.2 in Appendix B.3 for Spearman's ρ's and p-values). We also found The BabyLM corpus to be advantageous for this task in the sense that – in contrast to Wikitext-103 – reading time fit from all models improved over the baseline LMM model.

3.8 Discussion

Correlation between BLiMP, GLUE & MSGS The experiments presented in Section 3.6 provide evidence for a correlation between LM performance on BLiMP, GLUE and MSGS tasks when pre-training on the BabyLM corpus. This correlation is in accordance with established effects of training dataset size (Y. Zhang et al., 2021), and interactions of train corpus size and model capacity (Eldan and Y. Li, 2023, Kaplan et al., 2020). However, no single model achieves the highest score on all three tasks: BLiMP shows diminishing returns for model sizes larger than 50M tokens, while the best model on MSGS (16*1536) is substantially smaller than the best model on GLUE (24*1536). This discrepancy between the best model on the BabyLM challenge tasks and on the reading times prediction task is illustrated by Figure 3.5. The correlation between BLiMP/MSGS and GLUE may be an artifact of the sub-optimal fine-tuning on GLUE, failing to outperform the baseline model. It cannot be ruled out that the results would change when determining the optimal hyperparameters for each sub-task individually. However, even if the correlation were an artifact of the pre-training data, the findings of a negative correlation between model size and reading time fit would still hold.

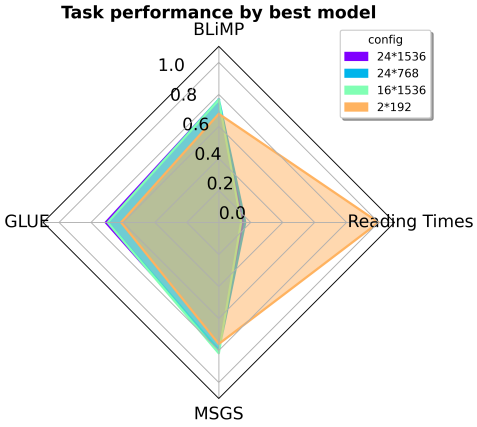


Figure 3.5: Performance of the best models by task. Reading times Δ log-likelihoods are normalized in the interval $[0, 1]$.

Cognitive plausibility The best fit on self-paced reading times from the Natural Stories corpus was obtained with the second-smallest model, with models with $l_{hidden} > 192$ only slightly improving over the baseline. The second suite of experiments in Section 3.7 confirms that this is not solely caused by the multi-epoch training regime necessitated by the small token budget. The reason for the mismatch between measures of cognitive plausibility (reading times) and measures of formal (BLiMP, MSGS) and functional linguistic competence (GLUE) is rooted in the interaction of pre-training regime and model size: While it is feasible to train a model that performs comparatively well on all four tasks on a budget of 100M tokens, the sweet spot for model size and dataset size is reached much earlier for the reading time prediction task than for the BabyLM challenge tasks. This problem could easily be resolved by using one model when modeling reading times (or any other psycholinguistic measure), and another model when either of the forms of linguistic competence is the aim. This might be a valid and promising approach in a situation where the understanding of the research object does not depend on the connectedness of its experimental analoga. In the case of our research object – the human language faculty – it may not be necessary to find a single analogon that accounts for all its components, but since we *know* that the human language faculty is part of a unified cognitive system (with specialized sub-units) performing the tasks which the modern language modeling pipeline of pre-training and fine-tuning splits up into individual

modules, it would be worthwhile to move in the direction of a unified approach that accounts for both forms of linguistic competence and empirical evidence of processing effort. This could be achieved through adjustments to the pre-training regime (in terms of data, modeling objective etc.), as suggested by the BabyLM challenge, or through adjustments to the model architecture.

Size of transformer models The results of the reading time prediction study on the BabyLM corpus indicate that it in fact has an *advantage* over Wikitext-103, although the LMs trained on the latter achieve larger Δ log-likelihoods on average: Since the largest models fail to improve over the baseline model if trained on Wikitext-103, it is possible that some properties of the language in the BabyLM corpus facilitate the learning mechanism that actuates the correlation of LM surprisal and reading times. The reason for the worse fit of surprisal from the larger models may be that both Wikitext-103 and the BabyLM corpus are not large enough to induce the learning bias needed to give good predictions of reading times in larger models, with Figure 3.4 showing that the results on the BabyLM corpus are much less stable than on Wikitext-103 and the improvements over the baseline much less sharply linear. In summary, our results lead to the following answers to our research questions:

Result: Research question A GPT-like LMs can be cognitively plausible and display formal and functional linguistic competence, although not both at the same time...

Result: Research question B ...under the constraint of a developmentally plausible training dataset.

3.9 Conclusion

Our study highlights the challenges of training a LM that performs well on tasks requiring some degree of formal and functional linguistic competence as defined by Mahowald et al. (2023), while also being predictive of the psycholinguistic measure of reading times. We find that small, shallow models of less than 5M parameters yield the best fit to the psycholinguistic measure, while performance on BLiMP, GLUE and

MSGS improves with increasing model size, although to a different degree for each of the tasks. This has implications for research on cognitively or developmentally plausible models of human language processing: in the case of a small, domain-specific training corpus it is not feasible to pre-train an LLM that displays formal linguistic competence and performs well on a reading time prediction tasks, a conclusion also drawn by Wingfield and Connell (2022). Consequently, research in this direction has concentrated on fine-tuning pre-traineded LLMs on domain-specific data, e.g., Skrjanec et al. (2023). A promising approach to a unified architecture could be relegating special tasks (such as classifying a sequence as in GLUE) to adapters (Houlsby et al., 2019), sub-networks within a pre-traineded LM. This approach is common in multilingual language modeling (Pfeiffer et al., 2022a; Alabi et al., 2022), where its success is partially attributed to its ability to separate general linguistic knowledge from language-specific information. A similar modeling decision may be necessary for cognitively plausible language models.

While this chapter considered LMs as models of the language faculty of a single human, Chapter 4 looks at LMs as models of a single or multiple languages, i.e., languages as abstract objects separate from their realization in any speaker. This includes a change in the scope of our research questions: While we have so far considered differences in linguistic competence and processing effort between speakers and language structures, we will now move on to differences in the organization of the pronominal systems between languages. This is still a synchronic comparison, but it makes use of surprisal in a very different way: Instead of using surprisal as an explanans of observed differences in language processing, we correlate surprisal with the properties of personal pronouns (such as gender and number) to recover general facts about the structure of a language.

3.10 Limitations

The results of the paper mainly hold for decoder-only transformer LMs. While these LMs are closer to human language processing in the sense that they process language incrementally, this has some disadvantages for reading time predictions, since humans do not attribute equal importance to each word, skipping some words in the process, and

typically integrate words from the left- *and* right-hand context of a fixated word. While the first point can be addressed by explicitly modelling skipping behaviour (Hahn and Keller, 2016), the second could require a solution closer to masked language models.

A second limitation is the focus on self-paced reading time as the psycholinguistic response variable. Since the setup of self-paced reading studies, with the participants observing a single word at a time, distorts the natural reading process, the measure itself may be not that cognitively plausible. This could be addressed by repeating the experiments on corpora from eye-tracking studies such as the Dundee corpus (Kennedy and Pynte, 2005). There is evidence that much larger models than those tested in the current study still improve the fit to total reading times in less restricted experimental settings (Varda and Marelli, 2023). The latter study also shows that the fit to psycholinguistic measures varies over languages and writing systems.

Another option is modelling brain activity patterns directly by predicting N400 and P600 EEG signals, which have the additional advantage of providing a means of decomposing LM surprisal without the proxy of linguistic structure, as shown by J. Li and Futrell (2023).

3.11 Future Work

As already hinted on in Section 3.9, there are multiple promising approaches to make LMs more human-like. While it has become clear that the limitation of exposure to linguistic data during pre-training is crucial, I will set out breifly other aspects of LM training data and architecture that promise to increase the human-likeness of the language a LM acquires, as measured by its formal and functional linguistic competence and PPP.

3.11.1 Syntax-aware language models

One option is the explicit inclusion representations of syntactic structure as inductive bias in the training data, e.g., by training an LM like Transformer Grammars (Sartran et al., 2022) on linear representations constituency-parsed data. This approach might suffer from data sparsity, as gold standard constituency parses are difficult to come by, at least for languages other than English. My expection is that a LM trained in this way could be more sensitive to syntactic ambiguities and therefore better predict reading times of, e.g., garden-path sentences, which are notoriously difficult to predict even for LMs jointly trained on a language modeling and parsing task (Arehalli et al., 2022; K.-J. Huang et al., 2023).

3.11.2 Classification vs. generation

When comparing a LM's PPP and its performance on classification tasks like CoLA (Warstadt et al., 2019), this comparison is usually made between two *instances* of the same model: PPP is evaluated on surprisal estimates obtained from the pre-trained model (that was only trained on a language modeling task). In contrast, classification tasks in their classical configuration require a new set of weights θ_{CL} called the *classification head* that projects a contextualized embedding (of class token in encoder-only models or

of the sequence-final token in decoder-only models) to the number of classes. This new set of weights is either trained jointly with the weights of the transformer θ_{TF}, modifying both in the process, or only θ_{CL} is trained while the weights of the transformer are frozen. In both cases, the different kinds of linguistic competence are distributed over disjoint parts of the same model: Task knowledge resides either in the classification head θ_{CL}, or jointly in the updated transformer weights θ'_{TF} and the classification head, and there is no model that would possess both types of linguistic competence as discussed in Section 3.1. While it would be possible to use the fine-tuned transformer θ'_{TF} with the language modeling head θ_{LM}, previous work has shown that full fine-tuning of a LM deteriorates its language modeling quality (Mosbach et al., 2020), which directly impacts formal linguistic competence, since benchmarks like BLiMP use surprisal to distinguish between grammatical and ungrammatical sequences. The deterioration of LM quality could be counteracted by reformulating a classification task as a sequence to sequence task, or by jointly training on the classification task and on a language modeling task.

3.11.3 Models with human-like memory limitations

An important limitation of the cognitive plausibility of transformer LMs is their large context window: Even the LMs in this study where trained on sequences of 128 tokens, and usually (maximum) context sizes are much larger. This is corroborated by the findings of Kuribayashi et al. (2022) and indirectly Oh et al. (2024), although this study did not explicitly model the effect of context size on reading time fit. However, even limiting the context size to a more human-like value (as was done in the study on personal pronouns described in Chapter A) is clearly far removed from how humans process language: In transformers the context is either present or absent, and never subject an actual memory limitation[5]. This could be addressed by combining a decoder with a small context window of n tokens with an embedding model that compresses

5 Of course any individual model might not be able to make use of all the information in its context window, e.g., because it is too small or was trained on insufficient data.

contexts longer than n tokens into a vector representation, which is then passed to the decoder alongside the fully attended context. This could also be achieved by applying different types of attention to the context, as is done in the Perceiver architecture (Jaegle et al., 2021).

3.11.4 Syntactic structure through inductive bias

Applying a specific LM architecture – be it of the transformer kind or other – to the study human language or, as in the case of BabyLM, human language acquisition, always has the disadvantage of requiring the researcher to subscribe to certain assumptions about the human language faculty that are forced upon them by that architecture. Thus, the incentive to use an architecture that makes as little assumptions about the data it sees as possible is easily explainable. It is maybe from this point of view that a position like that of Piantadosi (2023) or Futrell and Mahowald (2025) becomes understandable to people such as myself, who start out from linguistic theories and see LMs as a somewhat misused tool to (in)validate these theories. In the absence of a model-intrinsic syntactic (or other) bias, the actual language data a LM is exposed to during training becomes the focus of investigation. In this context questions of compositional generalization become relevant, and I would like to build on previous work in this direction (J. Hu et al., 2020; Mueller and Linzen, 2023; Leong and Linzen, 2023). I am also interested in the apparent disjunction of syntactic and semantic knowledge in LMs (see Weissweiler et al. (2023) for a treatment if English comparative correlatives), which also hearkens back to my point about the disjunction of formal and functional linguistic competence in LMs I raised in Section 3.11.2. It seems to me that the kind of statistical learning that is inherent to transformer LMs works differently for syntactic and semantic knowledge, and investigating both in the same model might not be justified.

4

They saw it, onu, 它, coming: An information theoretic study of cross-linguistic variation in personal pronouns

In this study, we applied a pre-trained multilingual LM, mGPT (Shliazhko et al., 2022), in a comparative study of personal pronouns in a parallel corpus (mini-CIEP+, Verkerk and Talamo 2024) of 17 languages from 8 language families. The reasons for choosing a pre-trained model instead of, e.g., training our own model on the parallel data are twofold: Firstly, most of the data in the corpus is still under copyright, and training a LM on it, strictly speaking, would be illegal. Secondly, using a model that was pre-trained on a large corpus ensures LM quality, which is particularly important when we consider longer context sizes. We complement the results for the multilingual model with results for multilingual model with monolingual models for 4 languages that we trained on wiki-40B (Guo et al., 2020). We aim to capture the predictability of personal pronouns using the information theoretic measure of surprisal, which characterizes the information value of a word given its preceding context. We compute the surprisal of personal pronouns in these languages at different context sizes. Then, Bayesian Regression models are fitted with surprisal-based response variables and a range of independent variables, including the frequency of the pronoun and various morpho-syntactic parameters (syntactic role, number, person, and others). These parameters are extracted from grammars and from mini-CIEP+, which is automatically annotated in the Universal Dependencies framework using pre-trained models.

The content in this chapter is based on:
Luigi Talamo et al. (2025a). “They saw it, onu, 它, coming: An information theoretic study of cross-linguistic variation in personal pronouns,” in: *Linguistics*, to appear

Luigi Talamo and Annemarie Verkerk conceptualized the research and led the paper writing. Luigi Talamo collected the dataset of relevant pronouns in miniciep+ and fitted the regression models. Julius Steuer selected the information-theoretic response measures and calculated surprisal values.

4.1 Introduction

4.1.1 Rationale

Cross-linguistic work on independent personal pronouns (you, me, she, them, ...) has focused on pronominal paradigms and associated universals (Joseph H. Greenberg, 1963; Cysouw, 1997; Mühlhäusler, 2001; Cysouw, 2009)[1], diachronic patterns regarding strong vs. weak forms and the development of pronominal clitics and bound affixes (also known as concord, Heine and Song (2011) and Konnerth and Sansò (2021)), their interaction with wider nominal morpho-syntax (Louagie and Verstraete, 2015) and usage-based accounts focusing on pragmatics (Langacker, 2007; Gardelle and Sorlin, 2015). These foci emerge from the deictic nature of pronouns, as they point towards referents previously established in discourse and change their referent depending on the context. Especially interesting from a typological perspective is how the usage of personal pronouns interacts with their function, both from the perspective of information structure as from the perspective of communicative efficiency. Arnold and Zerkle (2019) capture two main viewpoints on the "why" of personal pronouns: (1) The "pragmatic selection models", which work on the premise that personal pronouns are selected in discourse because of the nonlinguistic representation of the referent’s status as

1 Joseph H. Greenberg (1963) states: "Universal 42: All languages have pronominal categories involving at least three persons and two numbers."

"accessible", "salient", "prominent", or "in focus" (Arnold and Zerkle, 2019, p. 2), and (2), the "rational models", where the choice for personal pronouns is rooted in a need for efficient communication, being maximally informative while using efficient forms – personal pronouns are often very short. The main concern of Arnold and Zerkle (2019) is which model does a better job at explaining the the choice language users make between personal pronouns and other (types of) nominal items used for reference. This is highly interesting from a pragmatic and psycholinguistic perspective, but given the very limited work in these disciplines on non-European and non-WEIRD [2] languages, it is also Eurocentric. It is guided by the idea that a speaker has to encode the referent either using a nominal form or a pronoun, given that this is compulsory in English and other well-studied European languages. Many languages of the world manage referent tracking by indexing referents on the verb exclusively (Dryer, 2013a) and have no obligatory overt pronominal or nominal referents. Nevertheless - they tend to have personal pronouns (Joseph H. Greenberg, 1963; Cysouw, 1997), but simply use them far less frequently and for different purposes (Siewierska, 2012). In all likelihood, the usage of personal pronouns in these languages is nevertheless rooted in their inherent referential function, information structure, and communicative efficiency. We want to assess this both from a cross-linguistic as well as an information-theoretic perspective. While it would be optimal to conduct psycholinguistic experiments across a range of diverse languages, we instead conduct a token-based corpus study (see Levshina 2022b) across seventeen languages (given data availability, see below). We use the information theoretic metric surprisal (Hale, 2001b; Levy, 2008) to approximate accessibility in context. Surprisal measures how predictable a token is given its preceding context: when the token is unpredictable, surprisal is high; when it is predictable, surprisal is low. Predictability in context is directly related to cognitive effort; and correspondingly, a lot of work has gone into relating surprisal to various measures of cognitive effort in humans. Demberg and Keller (2008a) find that surprisal (lexical predictability) mediates self-paced reading time and Smith and Levy (2013) expand this finding to fixation time in eye-tracking experiments. Frank (2013) find that surprisal predicts the "strength" of

2 WEIRD stands for Western, Educated, Industrial, Rich, Democratic (Henrich et al., 2010)

the N400 effect; words with higher surprisal have more negative N400s. Arnold and Zerkle (2019)'s account of the choice between pronouns and full nominal referents is rooted in predictability: personal pronouns are expected to be highly predictable because they are given; their referent should be accessible in the (immediate) context. More recent work has shown that items that are more accessible from the context have lower surprisal (Zarcone et al., 2016; Aina et al., 2021), and that this effect is modulated by the distance of the pronoun from a context item, i.e., the size of the preceding context. Hence, we expect surprisal to correlate with accessibility. In this paper, we are concerned with linguistic predictors of surprisal, given that pronouns are used in highly divergent ways across the seventeen sampled languages.

4.1.2 The 101 on personal pronouns

What makes personal pronouns so special is that, compared with other lexical items that refer to beings or things, their referent is not stable but rather changes throughout any sizable chunk of spoken discourse or text: they are deictic terms. In spoken but also in written language, personal pronouns are one of the main strategies through which speakers and writers enable hearers and readers to keep track of entities without using full or proper nouns whenever they need to refer to an entity (Arnold and Zerkle, 2019). This is especially true for third person pronouns, which are used to refer to participants that have been mentioned previously – they are used as anaphoras that refer back to antecedents. However, first and second person pronouns are used for referring to previously established referents and for perspective taking too, especially in written dialogues. In spoken discourse, the speaker may talk about themselves using first person pronouns such as *I* or *me*, whereas the listener would refer to the speaker using the second person pronoun *you* (and would in turn use *I* or *me* to refer to themselves). Both hearer and speaker may use third person pronouns such as *she* or *them* to refer to people that are not part of the speech situation. The distinction in what we call person, then, is rooted in the triadic nature of communication: deictic resolution of roles taken by the speaker, hearer, and relevant other third parties (Helmbrecht, 2004, p. 11). Pronouns

also frequently encode syntactic role, number, gender, and other categories. These are properties that flag useful information about the referent: are they agents or patients, was it one entity or several? Feminine or masculine? Animate or inanimate? But the usage of personal pronouns is not the only way to resolve deictic argument tracking and co-reference: the usage of independent pronouns interacts with bound pronominal affixes. Many languages have obligatory bound pronominal affixes on predicates that indicate person, number, as well as other categories. For subjects, Dryer (2013a) lists 437 languages that have "subject affixes on verbs", making this by far the most common strategy for expressing pronominal subjects. The other strategy for expressing pronominal subjects is by using independent pronouns, either obligatorily (in Dryer's sample, 82 languages), optionally (61 languages), or in a different place compared to full noun phrases (67 languages). This implies that many languages may exclusively use bound pronominal affixes for pronominal reference, a characteristic that has been called "pro-drop" (Helmbrecht, 2004, p. 55), or leave coreference resolution up to context, which has been termed "null-subject" (Camacho, 2013). It also implies that in such languages, independent subject pronouns are used in a very different way compared to languages with obligatory overt subject pronouns: in the later type, these will be far more frequent, while in the former type, independent pronouns may be used only for specific pragmatic reasons (see for example Länsisalmi (2001)). We will observe the consequences of this later on. Personal pronoun systems vary greatly when it comes to the parameters they encode; here we quickly mention a few very common ones, while in Section 4.2.1 we review all that matter for the sampled languages: alternative morpho-phonological forms (such as clitics or stressed pronouns), syntactic role, position, person, number, gender, and clusivity. As for person, third person pronouns are different from first and second pronouns: first and second pronouns (almost) exclusively have human or animate referents and are first on the person hierarchy, while third person pronouns can also refer to inanimates and are used as anaphors, they refer to participants that were mentioned earlier in discourse (Helmbrecht, 2004, pp. 28, 50). Concerning syntactic role, subject pronouns are more common than direct object pronouns (Helmbrecht, 2004, pp. 28–29). In addition, speech act type is an important determinant: in expressions of

internal mental states such as wishes, obviously first person pronouns are commonly used; while in speech acts that center on the hearer, such as questions, we frequently find second person pronouns. In this study, we do not further consider speech act types. Number has been discussed extensively in the literature in terms of markedness with reference to the number hierarchy (Joseph Harold Greenberg, 1966, p. 31): singular < plural < dual < trial, where singular is unmarked and markedness increases as one moves to the right. Note that this characterization is rooted in part in frequency (across nominal, pronominal and verbal marking of number), where singular is the most common category, plural the second-most, etc. Lastly, politeness, where some languages (64 out of 207 in Helmbrecht 2013) encode social hierarchies in their pronominal paradigms (see for example Levshina 2017). An example is Dutch, whose speakers use *jij* "2SG.FAMILIAR" vs. *u* "2SG.HONORIFIC" in different social contexts. In addition, Helmbrecht (2013) lists seven languages where pronouns are avoided for reasons of politeness; we sample two of these, Japanese and Indonesian. Again, we will observe the impact of this linguistic behavior later on; and we discuss further details in Section 4.2.

4.1.3 Theory, aims and expectations

Given cross-linguistic differences in the usage of pronouns, as well as expected differences in usage within pronominal paradigms, we wish to investigate if a general information-theoretic measure, surprisal, can capture aspects of information status, namely, givenness. While independent personal pronouns are always given (Helmbrecht, 2004, p. 29), we expect surprisal of a pronoun to fluctuate depending on both cross-linguistic and paradigmatic differences just discussed. In a way, then, we want to test whether surprisal reflects what we know from a typological and linguistic perspective. While surprisal in our study is used as a global, observed measure that is to be explained by typological properties of a language, in the psycholinguistic literature it is commonly employed as an explanans of processing effort on the level of the individual, see Section 4.1.1 and for example, the recent series of studies that aim to predict

reading times from language model surprisal (Oh and Schuler, 2023a; Wilcox et al., 2020; Arehalli et al., 2022; J. Huang et al., 2023). While these studies differ from our approach by employing surprisal as a predictor and not a measure of processing difficulty, the linking hypothesis between surprisal and processing difficulty remains the same: A linguistic structure receives high surprisal if it is difficult to process (e.g., a low-frequency lexeme, a syntactic structure with multiple levels of recursion, or a long dependency between a pronoun and its referent), and low surprisal if it is easy to process (see for further details Section 4.3). Surprisal is inherently linked to frequency; all other things being equal, lexical items that appear more frequent will have lower surprisal. Frequency, at least for independent personal pronouns, is inherently linked to all of the aspects shortly discussed above (person, number, etc.) and to a range of further dimensions, unpacked in Section 4.2.1 below. We predict a strong correlation between surprisal and frequency, and wish to observe if other, related dependencies also hold; these are further discussed in Section 4.3.3 on linear models.

Since frequency is so important, our data source is chosen such that it allows for direct comparison: a parallel corpus called mini-CIEP+ (Verkerk and Talamo 2024, see Section 4.2.2 below). This corpus is fully annotated in the Universal Dependencies (UD) (De Marneffe et al., 2021), and we rely on several layers ofUD annotation to extract the relevant linguistic information. We take ten languages from different branches of Indo-European, namely Irish (Celtic), Dutch (Germanic), Italian (Romance), Albanian (Albanian), Greek (Hellenic), Lithuanian (Baltic), Ukrainian (Slavic), Armenian (Armenian), Persian (Iranian), and Hindi (Indo-Aryan). Then we add languages from seven different families: Arabic (Semitic), Basque (Isolate), Hungarian (Uralic), Indonesian (Austronesian), Japanese (Japonic), Mandarin Chinese (henceforth: Mandarin, Sino-Tibetan), and Turkish (Turkic). This brings our sample to seventeen languages, related only at a deep level, but with a strong bias towards Europe. Unfortunately this bias cannot be avoided in this study, as mini-CIEP+ is built on translations and the version used does not feature any languages spoken outside of Eurasia/Papunesia. The premise of our study, then, is to do quantitative typology on the basis of well-established typological categories, in the same vein as Say (2014), Levshina (2017), Naranjo and

Becker (2018), Gerdes et al. (2019), Masini and Mattiola (2019), Levshina (2022a), Talamo and Verkerk (2022), Levshina et al. (2023), Jing et al. (2023), Talamo et al. (2025b), Tao et al. (2024), and many others. Similar to some of these studies, this study is highly exploratory in nature. While we have some general ideas on which pronouns are more surprising than others, and how this could be related to accessibility from an information processing perspective, we basically see this study as an exploration of possible patterns, not as a test of any function- or theory-driven hypotheses. We are aware of several limitations of our study. These include the language sample, which mostly features synthetic languages from Eurasia and a possible overestimation of rates of pronoun usage. Furthermore, there are a host of other factors that will interact with the study of surprisal of personal pronouns (some of these are reviewed in Section 4.2 and Section 4.6.4 on limitations) that we do not include here. As for the predictors of surprisal, we look here at the surprisal of a pronoun given the preceding seven words. In Appendix E, we additionally look at surprisal as a function of context size, from $n = 1$ to $n = 127$ preceding words, operationalized as the area under the curve (AUC). These are admittedly relatively simple metrics that may not capture well what surprisal could tell us about pronoun usage and the accessibility of referents. We hope, however, that this study can shed first light on the divergent use of personal pronouns across languages, and how that usage may be explained in information-theoretic terms. We leave for future work studies of languages whose type we were unable to sample altogether, such as polysynthetic languages.

4.2 Data

4.2.1 Data from grammars: typological parameters

Drawing on the literature on personal pronouns and person marking, we prepared a typology of personal pronouns in the relevant languages. The parameters that we found to be relevant are shown in Table 1 (for a more detailed overview on all sampled

languages, see Appendix A in the Supplementary Materials). We have extracted values for these parameters from up-to-date reference grammars and through consulting native speakers and language experts. We do not go over all parameters in full, for reasons of space, but highlight those we think merit discussion. Independent personal pronouns are distinguished from another personal marker strategy, bound pronominal markers or index marking, by the notion of independent form. Independent form is defined in Siewierska (2012, pp. 17–22) on a morphological and prosodic basis: independent forms are forms that are morphologically and prosodically independent. Semi-bound forms, or "clitics", stand in between independent personal pronouns and bound index markers, in that they form a "phonological unit with a word (their host)" (Siewierska, 2012, p. 26), but in certain constructions they are realized as independent forms. For the purpose of the present paper we consider both independent and semi-bound personal pronouns and we refer to them collectively as "personal pronouns". However, as (Siewierska, 2012, p. 26) points out, the distinction between bound index markers and semi-bound clitics is sometimes blurred (for a recent discussion, see Haspelmath (2023) on Romance object clitics). In five languages of our sample, a set of personal markers are clearly described as clitics in the consulted grammars for Albanian Camaj and Fox (1984, p. 94), Arabic Badawi (2016, pp. 51–53), Greek Holton (2012, pp. 114–115), Indonesian Sneddon and Adelaar (2010, pp. 165–167), and Italian Maiden and Robustelli (2014, pp. 93–100). Clitics are often described in grammars as "weak" or "unstressed forms"; however, since Siewierska (2012) definition of independent form does not include the phonological nature of the marker, stress or its absence is not a sufficient condition for qualifying a form as independent or dependent. Accordingly, weak or unstressed forms can be independent forms. In our sample, Dutch (Donaldson, 2008, pp. 66–69) and Irish (Stenson, 2019, pp. 157–161) have unstressed forms which cannot be defined as clitics. Albanian uses clitics for the dative and the accusative; these forms are indeed unstressed, but they are only used for three specific constructions: the double object, in preverbal position and for the object argument of an imperative (Camaj and Fox, 1984, pp. 94–95). In contrast, Irish has two sets of independent personal markers, a set of personal pronouns and a set of augmented personal pronouns,

which "are used whenever one would put stress on an English pronoun" (Stenson, 2019, p. 160). The set of non-augmented personal pronouns thus qualify as unstressed forms, but are not restricted to certain positions or constructions and the opposition between the two sets is of a contrastive nature. Aside from morphological and phonological form, there is a third aspect that may license an alternate set of pronouns, namely, its discourse function. This opposition is commonly referred to in the grammars as "emphatic vs. non-emphatic" or "intensive vs. non-intensive" pronouns. According to Siewierska (2012, pp. 67–68), the term "emphatic" should be restricted to pronouns performing exclusively an emphatic function. The only candidate in our sample with such pronouns might be Basque, whose intensive (emphatic) forms are "most typically used when the pronoun is in focus" and "may also be used as topicalized noun phrases [and] are frequently preceded by ordinary pronouns" (Trask, 2003, p. 152). Since the languages of our sample display a maximum of two sets of personal pronouns, we have decided to conflate the phonological, morphological and discourse aspect into a single parameter, "alternative form" in Table 4.1, and list there if such sets are attested. As for the encoding of the syntactic roles of personal pronouns, languages of the sample use two strategies: case marking and adpositions. Here we have used grammars to investigate which strategy is used, typically making use of a synoptic table giving a list of the pronouns along with their morpho-syntactic features. Since the distinction between syntactic roles is not always clear in grammars, most notably for indirect object and oblique pronouns, we have opted to consider only a binary opposition: subject vs. non-subject pronouns. For instance, since Italian uses three different independent forms for subject, object and indirect object and prepositions for oblique argument, we have coded its encoding of the syntactic roles as "case marking" for both subjects and non-subjects and as "adpositional marking" for non-subjects. Most languages have a pronominal case system, five do not (Indonesian, Irish, Japanese, Mandarin, Persian). Note that even for these five, subject and non-subject roles are annotated in the parallel corpus, making case a parameter available for all languages. In six languages of the sample, Armenian (colloquial varieties: Dum-Tragut (2009, p. 307)), Hindi (colloquial varieties), Hungarian, Indonesian, Irish

and Persian, the adposition may be fused with the personal pronouns, giving rise to inflectional adpositions. This happens under different morpho-syntactic conditions in different languages. In Armenian (colloquial varieties), Hindi (colloquial varieties), Hungarian, Indonesian and Persian bound personal forms are still transparent; for instance, colloquial Hindi, *mujha-ko* "to me" vs. *tuma-ko* "to you" (Aarushi Singhal, p.c.); Hungarian, *iránt-em* "towards me" vs. *iránt-ed* "towards you" (Rounds 2009: 145-150); Indonesian, *kepada-ku* "to me" vs. *kepada-mu* "to you" (Sneddon and Adelaar, 2010, pp. 166–167). In Irish the paradigm of the inflectional adpositions is "not entirely predictable, but some general patterns can be discerned" i.e., *chugam* "toward me" vs. *chugat* "toward you" (Stenson, 2019, p. 219). Another parameter that figures prominently in the relevant literature is the expression of pronominal subjects; we follow here the typology proposed by Dryer (2013a), which is reduced to four types relevant for our sample and is slightly adapted as follows: (1) independent forms in subject position that are normally if not obligatorily present; (2) dependent forms (affixes) on verbs; (3) optional independent forms in subject position; and (4) mixed: more than one of the above types with none dominant. Most of the languages of our sample belong to the first or to the second type, two languages of our sample, Mandarin and Japanese, belong to the third type, i.e., they have optional independent forms in subject position but without verb affixation (Jiang (2016, pp. 508–509); Kaiser et al. (2013, p. 137)), and another language, Irish, belongs to the fourth type, since verbal suffixes can sometimes replace pronouns in subject position (Stenson, 2019, pp. 26–37). The position of the pronouns with respect to its predicate is also relevant here; since existing typological datasets (Dryer, 2013b; Dryer and Gensler, 2013) and consulted grammars do not distinguish between pronominal and non-pronominal arguments, we have extracted from mini-CIEP+ the frequency of personal pronouns in pre-verbal and post-verbal positions; see Appendix A.3 for more information. Following Dryer (2013c), we have classified languages as "no dominant (order)" if "text counts reveal [...] the frequency of the two orders is such that the more frequent order is less than twice as common as the other" (see also Talamo and Verkerk (2022, p. 180) for a discussion). This classification in Table 4.1 is given as an approximation, in the language modeling

experiments we conduct, the ratio of preverbal versus postverbal pronouns is taken. Finally, Table 1 includes four morpho-syntactic parameters: person, number, gender and clusivity. Another feature that would have been relevant is politeness (Helmbrecht, 2004; Helmbrecht, 2013). We have decided to not consider it in this study as in several languages, polite forms are homophones with non-polite forms; for example, Italian *lei* can be both 3SG.F "she" or 2SG.POLITE "thou". These pronouns are included in our study, but the parameter "politeness" is not coded for. In addition, there are several languages in which terms of address like "sir", "chief" and the like are used as replacements of pronouns for reasons of politeness (Helmbrecht, 2013); see Luong (1990) for an example and Chao (1956) for an influential study that covers polite terms of address and pronouns in Mandarin). Including these and coding for polite vs. regular pronouns is left for future studies. Another relevant factor would have been the position the pronoun has in a coreference chain, whether the pronoun is an anaphor or a cataphor, and if the former, the length to its antecedent. A possible resource in order to conduct a study that would take coreference into account is CorefUD, a multilingual collection of corpora with a standardized format for coreference resolution that can be used to generate new annotations (Nedoluzhko et al., 2022). Again, we leave this for further study.

Language	Family	Alternative Form	Case Marking	Adp. Marking	Pronominal Subject	Position	Paradigm Size	Person	Number	Gender	Clusivity
Albanian	IE	clitics for non-subj	subj, non-subj	prepositions for non-subj	affixes on verbs	pre for subj, post for non-subj	43	1,2,3	Sg, Pl	M,F	-
Arabic	Afro-Asiatic	clitics for non-subj	subj, non-subj	prepositions and inflectional prepositions for non-subj	affixes on verbs	no data	53	1,2,3	Sg, Pl, Dual	M,F	-
Armenian	IE	-	subj, non-subj	pre/postpositions for non-subj	affixes on verbs	pre for subj no dominant for non-subj	25	1,2,3	Sg,Pl	-	-
Basque	Isolate	emphatic pronouns for subj, non-subj	subj, non-subj	postpositions for non-subj	affixes on verbs	pre for subj and non-subj	67	1,2,3	Sg,Pl	-	-
Dutch	IE	unstressed pronouns for subj, non-subj	subj, non-subj	prepositions for non-subj	obligatory pronouns	pre for subj, no dominant for non-subj	31	1,2,3	Sg,Pl	M,F,N	-
Greek	IE	clitics for subj, non-subj	subj, non-subj	prepositions for non-subj	affixes on verbs	pre for subj, post for non-subj	24	1,2,3	Sg,Pl	M,F,N	-
Hindi	IE	-	subj,non-subj	postpositions for non-subj	affixes on verbs	pre for subj and non-subj	72	1,2,3	Sg,Pl	M,F	-
Hungarian	Uralic	-	subj, non-subj	postpositions and inflectional postpositions for non-subj	affixes on verbs	pre for subj,post for non-subj	81	1,2,3	Sg,Pl	-	-
Indonesian	Austronesian	clitics for subj + non-subj	-	prepositions for non-subj	obligatory pronouns	pre for subj,post for non-subj	26	1,2,3	Sg,Pl	-	incl, excl
Irish	IE	stressed forms for subj+ non-subj	-	inflectional prepositions for non-subj	mixed	post for subj and non-subj	92	1,2,3	Sg,Pl	M,F	-
Italian	IE	clitics for non-subj	subj, non-subj	prepositions for non-subj	affixes on verbs	pre for subj, post for non-subj	24	1,2,3	Sg,Pl	M,F	-
Japanese	Japonic	-	-	postpositions for subj, non-subj	optional pronouns	pre for subj and non-subj	99	1,2,3	Sg,Pl	M,F	-
Lithuanian	IE	-	subj, non-subj	prepositions for non-subj	affixes on verbs	pre for subj, no dominant for non-subj	42	1,2,3	Sg,Pl, Dual	M,F	-
Mandarin	Sino-Tibetan	-	-	prepositions for non-subj	optional pronouns	pre for subj, post for non-subj	20	1,2,3	Sg,Pl	-	incl,excl
Persian	IE	clitics for non-subj	-	prepositions and inflectional prepositions for non-subj	affixes on verbs	pre for subj and non-subj	23	1,2,3	Sg,Pl	-	-
Turkish	Turkic	-	subj,non-subj	postpositions for non-subj	affixes on verbs	pre for subj and non-subj	43	1,2,3	Sg,Pl	-	-
Ukrainian	IE	-	subj,non-subj	prepositions for non-subj	mixed	pre for subj no dominant for non-subj	27	1,2,3	Sg,Pl	M,F,N	-

Table 4.1: Parameters extracted from grammar and mini-CIEP+ (position and paradigm size). A comma indicates an opposition between values, while a plus sign (+) indicates a lack thereof.

4.2.2 Extracting data from mini-CIEP+: the UD framework

The data source for the current study is mini-CIEP+ (full name: mini Corpus of Indo-European Prose Plus, Verkerk and Talamo 2024). This is a parallel corpus featuring the first 15% of ten texts of prosaic nature (nine novels and one diary). It is a derivative work of the Corpus of Indo-European Prose Plus (CIEP+), from which the authors sampled a legally-sharable portion. We have sampled from mini-CIEP+ seventeen languages from different families; thirteen languages have at least eight texts out of the ten parallel texts; two languages (Armenian, Japanese) have six texts and one language, Basque, has seven texts. Whereas languages with six texts already reached a comparable size, we supplemented one language, Irish, which has five texts, with five non-parallel texts (four texts originally written in Irish and one in English: Verkerk and Talamo 2024). mini-CIEP+ has been automatically parsed so that it has annotation in line with the Universal Dependencies (UD) framework (De Marneffe et al., 2021). This has been done using Stanford Stanza v. 1.8.0 (Qi et al., 2020) with models trained on UD v. 2.13 (Zeman et al., 2023). This annotation concerns: parts of speech (UPOS), language-specific parts of speech (XPOS), lemmatization, morpho-syntactic features (verbal and nominal categories: Universal Features) and syntactic dependencies (Universal Relations). All levels of UD annotation work at the token level, describing syntactic relations between tokens and providing lexical and morpho-syntactic information on each token. This has some consequences in the annotation of syntactic role given that there are two different strategies; case marking results in one token, while adpositional marking results in two tokens, the adposition and the pronoun. For example, the comitative/instrumental of the first person singular is *benimle* (case marking) in Turkish and *met mij* (adpositional marking) in Dutch. The UD annotation treats the first strategy as one single token and the second strategy as two tokens; in both cases, it is the personal pronoun form that receives annotation for the syntactic relation, the oblique object. In order to implement the parameters discussed in the previous section we have exploited all levels of UD annotation; this allowed us to reduce noise (such as non-pronouns mistaken for pronouns by the parser). The implementation of parameters varies from

language to language, basically based on how large the pronoun paradigm is considering inflecting adpositions; Appendix A.3 in the Supplementary Materials includes a full overview of our methodology. The default procedure, which is applied to roughly half of the languages, is to extract tokens matching a list of pronoun forms collected from grammars and other sources. We filter on their UPOS label and generate lists of tokens with their morpho-phonological form, syntactic role and morpho-syntactic features. Finally, forms are filtered on the basis of their syntactic role, keeping only verbal arguments. However, sometimes it was not possible to include all pronominal forms, more specifically, when the language has inflectional adpositions, leading to a proliferation of the pronominal forms. In this case, we conduct the corpus query on the basis of the UPOS annotation layer and manually check the extracted tokens in order to remove non-pronominal forms. As for the alternative form of the personal pronouns, we have distinguished independent forms from clitics using the annotation provided in the Universal Features. Where this information was not available, like for Albanian and Greek, we have used lists of clitics taken from grammars. Similarly, lists of forms have been used for distinguishing stressed and unstressed pronouns in Dutch and Irish, and emphatic from non-emphatic pronouns in Basque. For the reasons discussed in the previous section, we have merged these three parameters into a single parameter, marking independent, unstressed and non-emphatic forms as "standard" and clitic, stressed and emphatic forms as "alternative". As for syntactic roles, the relevant annotation level is the so-called Universal Relations. As stated in the guidelines, the UD framework assumes the following clause arguments: subject (nsubj), object (obj), indirect object (iobj) and oblique (obl). Furthermore, several languages use relation subtypes, which are a subset of syntactic relations providing further information; for instance, more than half of the languages in our sample use the label "obl:agent" to identify the agent of a passive construction. In order to ensure comparability across languages, we have conflated these subtypes with their relation type (for example, "obl:agent" with "obl"). Furthermore, for the reasons already expressed in the previous section, we have chosen to contrast subject with non-subject pronouns, conflating direct objects with indirect objects and obliques. As for the morpho-syntactic features, for most languages Universal Features

provided information on clusivity, gender, number and person; where this information is not available, as in the case of Japanese for number, person and gender, we have filled these parameters using a list of forms annotated for these categories. Finally, we have extracted the position of the personal pronoun with respect to its predicate, i.e., whether the token pronoun is preverbal or postverbal. A data row, to which we add the information-theoretic complexity measures discussed in the following section, looks like this for Italian lo "3SG.M.OBJ.ENCL":

token	position	altform	synrole	person	number	gender	clusivity
lo	pre	alt	non-subj	third	sing	Masc	non-clusive

4.3 Information-theoretic metrics and computational models

4.3.1 Importance of context for pronoun surprisal

As becomes clear from the examples in Chapter 2.3, the exact form a pronoun takes depends heavily on its context. This may be the words following the the pronoun (as in *L"homme est allé chercher ses enfants dans l"école maternelle* ("The man has picked up his children from kindergarten"), where the possessive pronoun *ses* agrees with its head *enfants* in number), or the words preceding it (in the English translation of our example, the pronoun his agrees with its antecedent man in gender and number). In our study, we consider only those differences in predictability that stem from information in the preceding context, in line with surprisal theory and the incremental, left-to-right nature of human language processing while reading. The degree to which the surrounding context is relevant for the predictivity of a pronoun will differ from language to language, depending on how many categories it marks on the pronoun and on modality, register, and position with a text or spoken discourse. While first and second person

pronouns do not as frequently agree with their antecedent as third person pronouns, they may be highly predictable in discourse, e.g., second person pronouns may have high predictability in a dialogical context, whereas first person pronouns may have high predictability in a first-person narrative. A third axis on which the predictability of a pronoun may vary is the amount of material intervening between it and its antecedent. This may be an arbitrary number of words (or sentences, or paragraphs), and surprisal theory stipulates that increasing context should on average decrease surprisal, that is, if we add context to a pronoun like *ihn*, we would expect surprisal to decrease by a certain amount when averaging over many contexts. We aim to capture a general surprisal as well as the decrease in surprisal with increasing size of the preceding context (see Section 4.3.1 and Appendix A.5). We measure surprisal different context sizes $t \in \{1, 3, 7, 15, 31, 63, 127\}$. Obviously, there will be some kind of relation between surprisal values calculated for the same token under different context sizes; and it would be rather tedious to discuss results for all seven different windows. Therefore we pick surprisal with $t = 7$ as the main analysis that we discuss. We believe this is appropriate given that this is the size of a short sentence, in line with human processing capabilities (see for example Bamman et al. 2020 on coreference resolution in pronouns in English literature). In addition, in we assess how surprisal changes as context size increases. We do this by plotting context sizes sequentially which results in the so-called surprisal reduction curve; we measure the AUC. This captures a summary measure of sensitivity of the pronoun to the increasing context size. Due to space constraints in the original paper, these analyses (methods and results) are included in Appendix A.5 and only shortly discussed in the following section.

4.3.2 Surprisal of personal pronouns

We calculate average surprisal $\hat{S}_t$ at context size t over all w_n, where w_n is an instance of a personal pronoun P, with N_P the number of instances of that pronoun in the corpus. P refers to the instances of a personal pronoun in the corpus that share a unique

combination of grammatical properties as described in Section 2.2. The general formula for the average surprisal of P at context size t is then given as:

$$\hat{S}_t(P) = \frac{1}{N_P} \sum_{w_n \in P} \text{surp}(w_n | w_{n-t}, ..., w_{n-1}) \tag{4.1}$$

In theory, mGPT could extend to much longer contexts than any of our context sizes $t \in \{1, 3, 7, 15, 31, 63, 127\}$. Since we want to control for the amount of information in the context, we restrict the number of context tokens to exactly t, and extract the probability distribution for the t^{th} token, that is the personal pronoun at position t of the corpus. We are aware of the fact that using a multilingual model such as mGPT to estimate these entropies may bias next-word probabilities for languages that are underrepresented in the training data in unpredictable ways. In order to evaluate whether our results hold for entropies estimated with monolingual models trained on a similar number of tokens, we conduct a series of controlled pre-training experiments using the Wiki-40B dataset (Guo et al., 2020) (see Appendix A.4 in the Supplementary Materials).

4.3.3 Linear mixed models

On the basis of the typological survey described in Section 4.2.1, we identified the following linguistic variables that are relevant for all sampled languages (see for further details below): Person, number, case, and position:

Person. This variable can take the values "first", "second", "third" in all sampled languages. In the statistical models, which will be further discussed below, we take "third" as the reference category. We do this while this enables us to compare first and second person pronouns to third person pronouns easily; as third person pronouns have been argued to be different in nature from first and second person pronouns (see Introduction).

Number. This variable can take the values "singular" and "plural" in all sampled languages. Additionally, Arabic and Lithuanian have a category "dual". For these languages we subsume "dual" and "plural" under a single variable "non-singular". We take "singular" as the reference category.

Case. This variable is taken from the syntactic dependencies annotation of the corpus. We only differentiate the subject forms ("subj") from the non-subject forms ("non-subj"), because the non-subject forms are annotated quite differently between languages in the UD framework (see above). Most languages in our sample have an accusative pronominal case marking system, which of course greatly overlaps in the sense that (for example) "subj" forms are nominative forms like I and not accusative forms like me (despite for mistakes introduced by dependency parsing, which are unfortunately unavoidable). Indonesian, Irish, Japanese, Mandarin, and Persian do not have a pronominal case marking system, but are treated exactly the same way as languages that do: by fully relying on the syntactic annotation of the pronouns as "subj" or "non-subj". We take "subj" as the reference category.

Position. To capture potential differences related to word order, we code whether the pronoun appears before the root verb of the clause ("pre") or after the verb ("post"). We take "pre" as the reference category. In addition, 4.1 listed if the pronominal subject is obligatorily expressed or not (Dryer's 2013 typology). This is captured by a continuous measure, namely:

Frequency. This variable corresponds to the frequency of the pronoun with the same combination of the variables relevant for that language (person, number, gender, etc.), so as to effectively deal with syncretism. For example, *zij* (and unstressed counterpart *ze*) in Dutch can refer to the 3SG.F form (*Zij is een ingenieur* "She is a engineer) and to the 3PL form (*Zij maken lawaai* "They are making noise"). For the first example; *zij* is annotated as "third person", "singular", and "feminine", and for this token, would be associated with the frequency of *zij* forms which are the same. Correspondingly in the second example, the frequency of *zij* forms that are "third person" and "plural" is taken. This variable effectively captures the frequency of pronouns, creating a continuous assessment of the obligatoriness of expressing both the subject and the non-subject pronouns. For a given corpus, we calculate relative frequencies by summing the absolute frequencies of each pronoun form with the same combination of all relevant variables and dividing these by the total corpus size in terms of pronoun tokens; then take the

natural logarithm. Then there are variables which matter only for a subset of the sampled languages:

Gender. Most sampled languages have a pronominal gender marking system. In Albanian, Arabic, Armenian, Hindi, Irish, Italian, Japanese, Lithuanian, Mandarin, and Ukrainian, the opposition is between "masculine" and "feminine". In Dutch and Greek, the opposition is between "masculine", "feminine", and "neuter". Basque, Hungarian, Indonesian, Persian and Turkish do not have a pronominal gender marking system. Gender is only relevant for third person pronouns in the languages of our sample; with the exception of Japanese, in which gender is relevant for all three persons (Kaiser et al., 2013, pp. 137–142). Accordingly, we add "ungendered" as a variable for pronouns that are not gendered and make this the reference category, so we can observe the behavior of gendered pronouns.

Clusivity. This is only relevant for very few languages: Indonesian and Mandarin have an "inclusive" vs. "exclusive" opposition in first person pronouns. Clusivity, like gender, is not relevant for all pronouns: in the languages of our sample that have clusivity, it is only relevant for first person pronouns. Therefore we add "not clusive" as a variable for pronouns that are neither "inclusive" nor "exclusive" and make this the reference category.

Alternative form. Some languages have pronominal clitics; these are Albanian, Arabic, Greek, Indonesian, Italian, and Persian. Dutch has a set of unstressed pronouns. Basque has emphatic pronouns and Irish has a set of stressed pronouns. As the reference category, we take independent forms for languages with pronominal clitics, the stressed pronouns for Dutch, the non-emphatic pronouns for Basque and the unstressed pronouns for Irish. Our statistical model takes into account all variables discussed so far, except for inflectional adpositions and politeness. Only a small set of languages have inflectional adpositions: Arabic, Hungarian, Irish, and Persian. At the time of writing, this is something that cannot be easily extracted from the UD-annotated data. Politeness is excluded for reasons explained in Section 4.2 We leave the analysis of these parameters to further work. We include both a multi-language model with the variables that are present in all languages (person, number, case, position, frequency) and language

individual models, where the relevant remaining variables are also included. Models are fitted using brms Bürkner (2017)'s Bayesian multilevel models, which uses STAN (Carpenter et al., 2017) for Bayesian inference in R (R core team, 2024). The main models have language-specific intercepts and slopes for frequency, which is the variable we suspect is most closely associated with surprisal:

```
surpisal7 <- brm(formula= surprisal_7 ~ synrole + number + person + rel_freq +
    position + (1+rel_freq|lang), family = gaussian(identity))
```

Aside from this multi-language model, we include individual models for each language in order to sensibly model the variables listed in (6-8) above. These cannot be included in the multi-language model as the variables are not relevant for certain languages, whose data would be deleted before model fitting. The models for Irish and Indonesian are given below in order to exemplify what these look like:

```
surprisal7_Irish <- brm(formula= surprisal_7 ~ synrole + number + person +
    rel_freq + gender + position + form, family = gaussian(identity))
```

```
surprisal7_Indonesian <- brm(formula= surprisal_7 ~ synrole + number + person
    + rel_freq + position + clusivity, family = gaussian(identity))
```

4.4 General results

4.4.1 Size of the paradigm

Rather than extracting the paradigm size from grammars, we have opted for a corpus-based approach in the computation of this parameter, which is given in Table 4.1 and plotted in Figure 4.1. In our approach, a distinct form of the paradigm is operationalized as a unique combination of values from some of the parameters discussed in Section 4.2.1: person, number, case, gender, clusivity and alternative form, plus the graphemic form of the personal pronoun.

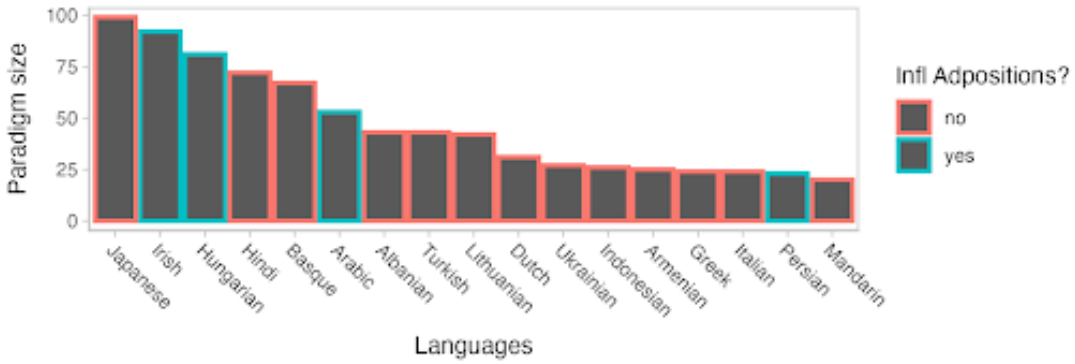
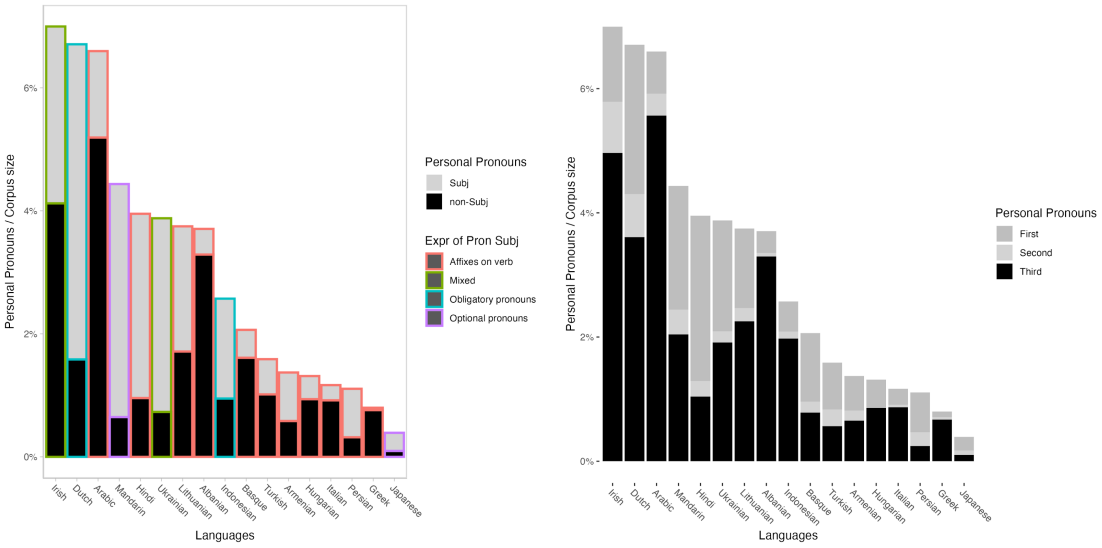


Figure 4.1: Size of the paradigm of the personal pronouns, as operationalized by our corpus-based approach. Languages with inflectional adpositions are highlighted in blue.

As shown in Figure 4.1, languages vary quite considerably in the size of the personal pronoun paradigm; the highest number of forms is found in languages with inflectional adpositions. As expected, languages with a high number of morphological cases i.e., at least six cases, also appear in the top part of the list, while languages with few or no morphological cases are placed at the bottom. A notable exception is Persian, a language with inflectional adpositions; however, unlike the other three languages with inflectional adpositions, the pre-trained UD model for Persian is able to split the token into the adposition and the clitic object pronoun, thus drastically reducing the number of forms.

4.4.2 Frequency of personal pronouns

The bar plots in Figures 4.2a and 4.2b show the relative frequencies of personal pronouns in the sampled languages. For a better visualization, we calculate percentages by taking the number of pronoun tokens in the corpus, and dividing that by the total size of the corpus in tokens, and multiplying that number by 100. We start with Figure 4.2a, which shows relative frequencies of the personal pronouns according to syntactic role and in relation to the typology of Dryer (2013a). With respect to corpus size, Irish has the highest percentage of personal pronouns (7%), while Japanese has the lowest (0.4%). The ranking reflects to a certain extent this typology: a language with obligatory subject



(a) Frequency of the personal pronouns with respect to the mini-CIEP+ corpus size, with the ratio between subject and non-subject pronouns and indicating types for the expression of pronominal subject taken from Dryer (2013a).

(b) Frequency of the personal pronouns with respect to the mini-CIEP+ corpus size, with the ratio between first, second and third person.

pronouns (Dutch) ranks second; with the exception of Arabic, languages with subject indexing on the verb (the "affixes on verb" type) occupy the lower ranks and languages of the mixed and optional type show, as expected, a mixed behavior. Dutch has the highest relative frequency of subject pronouns (5.1%) and Greek, a language of the "affixes on verb" type, the lowest (0.04%). As for the ratio of subject and non-subject pronouns, seven languages of the sample have more non-subject pronouns than subject pronouns, and nine languages more subject pronouns than non-subject. The greatest discrepancy is found in Greek, where non-subject pronouns are twenty times more frequent than subject pronouns and in Mandarin, where subject pronouns are five times more than non-subject pronouns. Two languages, Armenian and Lithuanian, have a quasi 1:1 ratio, with a slight predominance of subject pronouns. Figure 4.2b shows the ratio between the different values for person; across languages, the third personal pronoun (mean across languages: 1.8%) is more frequent than the first personal pronoun (mean: 1%), which in turn is more frequent than the second personal pronoun (mean: 0.3%). Individual languages support the 3>1>2 hierarchy, with the alternative hierarchy

1>3>2 in Basque, Hindi, Japanese, Persian and Turkish. Accordingly, the first personal pronoun is always more frequent than the second personal pronoun. The lowest relative frequency for the third personal pronoun is found in Japanese (0.1%) and the highest in Arabic (5.5%), while Hindi has the highest relative frequency for the first personal pronoun (2.7%) and Greek the lowest (0.09%). Finally, Irish has the highest relative frequency for the second personal pronoun (0.83%) and Hungarian the lowest (0.01%).

4.4.3 Personal pronoun length: surprisal

It is well-known since Zipf's seminal studies (Zipf, 1935; Zipf, 1949) that word frequency negatively correlates with word length i.e., shorter words tend to be more frequent. Levshina (2022a) additionally finds a correlation between word length and bigram surprisal, i.e., words with higher bigram surprisal tend to be longer. In Arnold and Zerkle (2019)'s review of rational models, the choice for pronouns is rooted in part in their length, as they are typically shorter than full nominal referents. We test this by performing a Pearson's correlation test with personal pronoun length as the predictor. We have included in the analysis all languages of our sample with the exception of Japanese and Mandarin, as these languages are written in logographic and syllabic scripts, not alphabetic ones. Pronoun length consists of the number of the characters encoded in Unicode format and computed using the R (R Core Team 2024) function nchar(). Pearson's correlations are performed using the R function cor(), checking the significance with the cor.test() function. Across all languages i.e., using a multi-language correlation, the correlation between average surprisal based on the preceding seven tokens and personal pronoun length is significant ($p < 0.05$) but quite low (Pearson"'s $\rho = 0.229$, df $= 1269$). It is positive, such that pronouns with a longer form as measured in number of characters also have a higher surprisal. For individual languages, this general pattern holds, with the exception of six languages: Armenian, Basque, Dutch, Greek, Indonesian and Persian. There is quite a lot of variation in the strength of the correlations; only one language, Arabic, has a negative correlation between surprisal

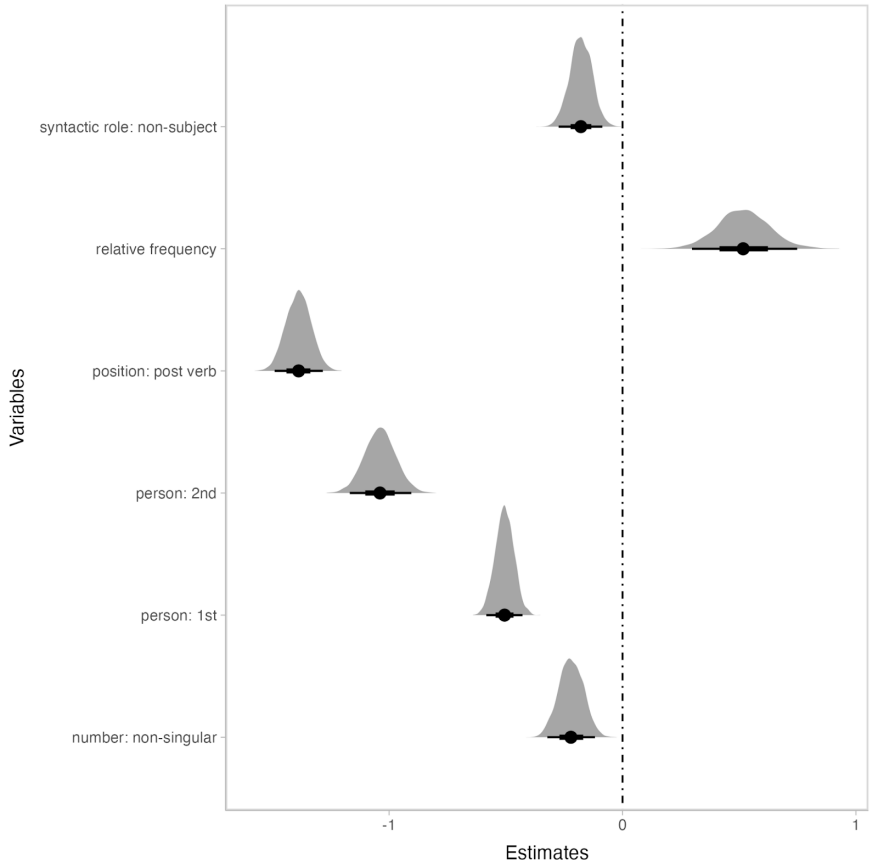


Figure 4.3: Posterior densities of the fixed effects of the multi-language LMM. Main intercept is excluded as they are removed much further from 0 (Estimate 3.23; CI: 2.37, 4.49).

and pronoun length. For further information, see Appendix A.6 in the Supplementary Materials.

4.5 Multi-language LMM

We first discuss the fixed effects from the models introduced in Section 4.2.1, and then the random effects. We take surprisal based on the preceding seven tokens as the response variable and determine relevance of predictors by asserting whether the 95% confidence interval (CI) excludes zero using Figure 4.3 (and similar figures below). Relative frequency has relevant coefficient estimates, meaning that less relative frequent

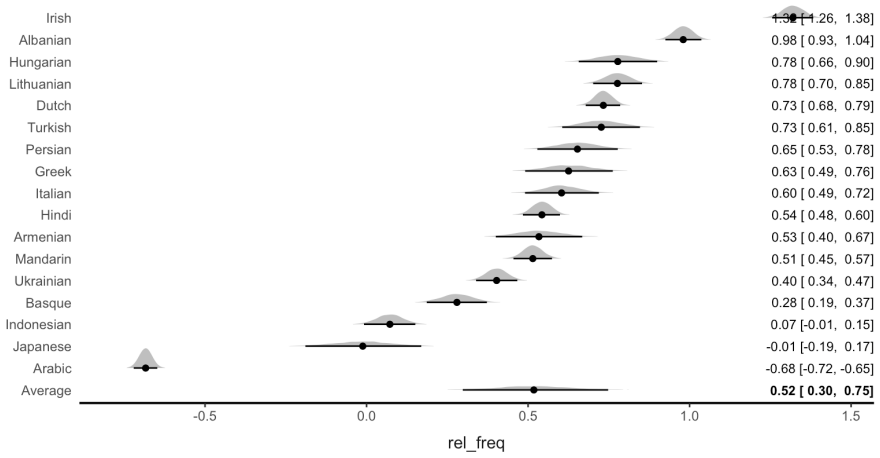


Figure 4.4: Posterior densities of the random slopes of the multi-language LMM.

pronouns (as relative frequency is logged) have greater surprisal (Estimate: 0.52; CI: 0.30, 0.75), as expected. As for the position of the pronoun with respect to the verb, postverbal pronouns have lower surprisal when compared to preverbal pronouns (Estimate: -1.39; CI: -1.49, -1.28). We observe across languages relevant negative coefficient estimates of the first (Estimate: -0.51; CI: -0.58, -0.43) and second (Estimate: -1.04; CI -1.17, -0.90) person pronouns, implying that both first and second person pronouns are less surprising than third person pronouns. Furthermore, when compared to subject pronouns, non-subject pronouns have lower surprisal (Estimate: -0.18; CI: -0.27, -0.09). Lastly, non-singular pronouns show lower surprisal than singular pronouns (Estimate: -0.22; CI: -0.32, -0.12).

Since the frequency of personal pronouns is so varied across languages (see Figure 4.2a), we estimated both random intercepts and slopes on relative frequency. The estimates for the random slopes are given in Figure 4.4. The estimates from the model are offsets from the fixed estimate also included in the model. Summing over the fixed estimates (see "Average" in Figure 4.4) and the random slope offsets, we see that languages differ quite widely in the effect relative frequency has on surprisal. In all languages with the exception of Arabic (negative correlation) the effect is positive. Since the predictor is logged, this implies that less frequent pronouns are more surprising, all other things being equal. But in Irish, with the highest random slope estimates, the effect is much more pronounced than in Indonesian, which has the lowest. There is no

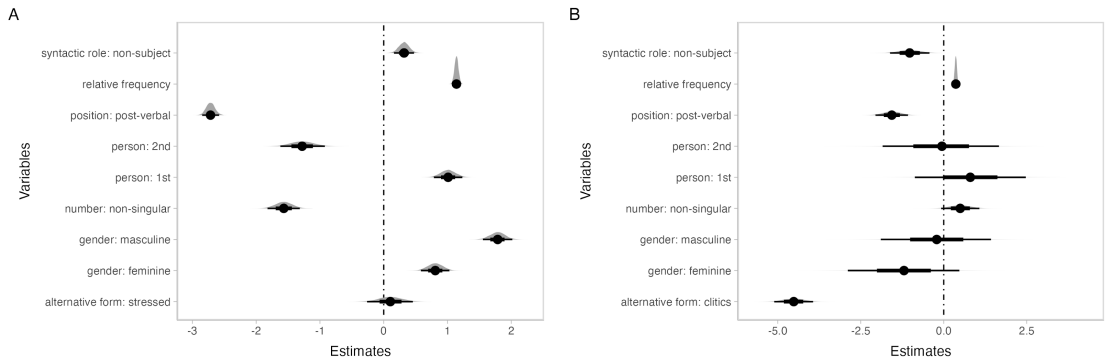
obvious relationship between Figure 4.2a-4.2b and Figure 4.4, except that languages with fewer pronouns overall (Basque, Hungarian and Japanese), tend to have more pronounced effects of frequency. Interestingly, as shown in Figure 4.1 and discussed in Section 4.4.1, Basque, Hungarian and Japanese are among the languages with the highest number of forms per paradigm: 67, 81 and 99, respectively. The languages with the lowest slope estimates (Indonesian, Ukrainian, Mandarin) have (in our sample), medium pronoun use and small paradigm size, but belong to different types.

4.5.1 Individual language LMM

We now turn to the results of the individual language models. For the predictors shared by all languages (person, number, syntactic role, position, relative frequency), we ask if individual languages show differences as compared with the multi-language LMM Gender, clusivity and alternative form are discussed in separate sections. As observed for the multi-language models, surprisal is generally well-explained by the linguistic predictors. Nevertheless, for nearly half of the languages of the sample we observe only few relevant coefficient estimates; these languages are Hindi, Hungarian, Indonesian, Japanese, Lithuanian, Persian and Turkish.

4.5.2 Relative frequency, person, number, syntactic role, position

We find that relative frequency is a relevant variable with positive coefficients across 14 languages out of 17, confirming the negative correlation (as relative frequency is logged) between relative frequency and surprisal also observed in the multi-language model. Exceptions are Arabic (Estimate: -0.69; CI: -0.77, -0.60) and Indonesian (Estimate: -0.11; CI: -0.18, -0.03), for which we find relevant negative coefficients. We do not find relevant coefficients in Japanese. The cross-linguistic differences in strength of this relationship (see Figure 4.4) are also observable in the language individual LMMs. In Appendix A.7 (see the Supplementary Materials) this is clearly illustrated. Appendix A.7

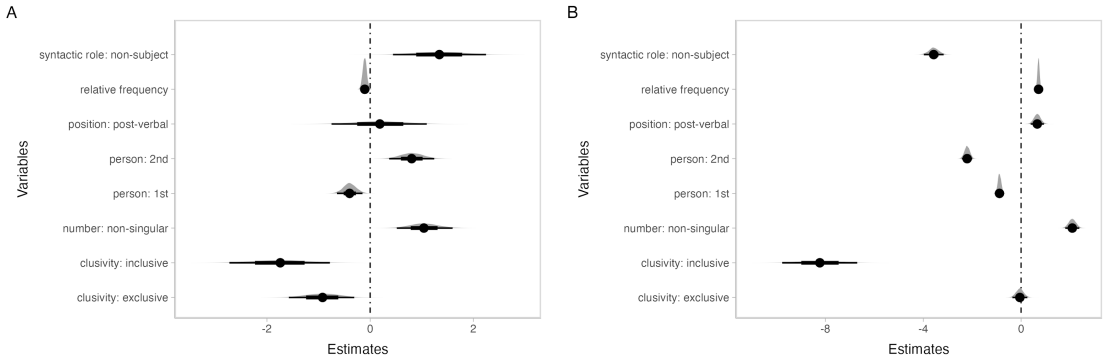


(a) Dutch **(b)** Albanian

Figure 4.5: Posterior densities of the fixed effects of two LMMs; A: Dutch; B: Albanian

presents an overview of the relationship between surprisal and frequency in individual languages. It allows the reader to assess the mean surprisal of individual pronouns in each language and its relationship with frequency. In addition, in Figure 4.5, the estimates of the predictors for Dutch (4.5a) and Albanian (4.5b) are plotted. Here, we can observe that the effect for relative frequency in Dutch (Estimate: 1.14; CI: 1.08, 1.20) is much bigger than that in the Albanian model (Estimate: 0.36; CI: 0.30, 0.43).

All languages have relevant CIs for the **position** of the pronoun, with the exception of Indonesian and Japanese; in the latter language, this variable is not included in the multi-linear model, as pronouns always precede. Opposite to the multi-language model, in the individual models the effect is negative for all languages (postverbal pronouns have less surprisal than preverbal ones) but for Irish (Estimate: 6.35; CI: 5.22, 7.49), Italian (Estimate: 0.94; CI: 0.56, 1.32) and Mandarin (Estimate: 0.66; CI: 0.39, 0.94), meaning that, especially in Irish, preverbal pronouns have lower surprisal than postverbal pronouns. In Irish, the "runaway" estimates for position probably have to do with the fact that almost all pronouns are post-verbal, Irish being a predominantly verb-initial language. Regarding both first and second **person** pronouns, the multi-language model had negative coefficient estimates, with first person pronouns having a more negative coefficient than second person pronouns. In general, then, third person pronouns (the reference category) have higher surprisal than first or second person pronouns. Across languages, however, this does not hold at all. We find non-relevant posterior estimate distributions for both the first and the second person in Albanian,



(a) Indonesian **(b)** Mandarin

Figure 4.6: Posterior densities of the fixed effects of two LMMs; A: Indonesian; B: Mandarin.

Arabic, Italian, Japanese, Lithuanian, and Ukrainian. For the remaining languages, we observe several types:

- Both first and second person pronouns have positive coefficients (only Irish).
- Positive coefficient estimates for second person pronouns, negative coefficients for first person pronouns: Indonesian (see Figure 4.6); this is also the case for Hungarian (the estimates for the first person pronoun include zero in Hungarian).
- Positive coefficient estimates for first person pronouns, negative coefficients for second person pronoun: Armenian (the estimates for the second person pronoun include zero in this language), Dutch (see Figure 4.5).
- Both first and second person pronouns have negative coefficients. This is almost true for Basque, Japanese, and possibly Turkish; however, estimates for the first person pronoun include zero in all of the three languages. Hindi and Mandarin (see Figure 4.6 B), and Persian are good examples of this.

With the exception of Albanian, Armenian, Hungarian, Turkish and Ukrainian, the number predictor is relevant in the individual LMMs of eleven languages; we can observe here the following two types:

- Non-singular pronouns are more surprising than singular pronouns. This is line with the multi-language model (see Figure 4.4), and occurs in three languages:

Indonesian (Estimate: 1.05; CI: 0.52, 1.60, see Figure 4.6a), Irish (Estimate: 4.87; CI: 4.11, 5.63) and Mandarin (Estimate: 2.09; CI: 1.81, 2.38, see Figure 4.6b).

- Non-singular pronouns have less surprisal than singular pronouns. This happens in seven languages, but with estimates that are less far removed from zero: Basque (Estimate: -2.76; CI: -3.59, -1.96), Dutch (Estimate: -1.57; CI: -1.82, -1.32, see Figure 4.5a), Greek (Estimate: -0.41; CI: -0.77, -0.04), Hindi (Estimate: -0.53; CI: -0.84, -0.23), Italian (Estimate: -1.56; CI: -2.09, -1.02), Lithuanian (Estimate: -0.40; CI: -0.63, -0.17) and Persian (Estimate: -0.58; CI: -1.00, -0.15).

As for syntactic role, its estimates are not relevant in the individual LMMs for Armenian, Greek, Hindi and Japanese; we have again two types here:

- Non-subjects pronouns have less surprisal, as in Albanian (Estimate: -1.03; CI: -1.62, -0.43; see Figure 4.5b), Arabic (Estimate: -1.32; CI: -1.86, -0.81), Basque (Estimate: -1.64; CI: -2.34, -0.95), Italian (Estimate: -2.95; CI: -3.89, -2.00), Mandarin (Estimate: -3.57; CI: -3.98, -3.17, see Figure 4.6b), Persian (Estimate: -1.08; CI: -1.36, -0.80) and Ukrainian (Estimate: -0.59; CI: -0.90, -0.27). This group of languages follow the trend evident in the multi-language model (see Figure 4).

- Non-subject pronouns have greater surprisal. Languages belonging to this group are Dutch (Estimate: 0.32; CI: 0.16, 0.48, see Figure 4.5a), Hungarian (Estimate: 0.56; CI: 0.04, 1.06), Indonesian (Estimate: 1.34; CI: 0.44, 2.25, see Figure 4.6a), Irish (Estimate: 2.24; CI: 1.80, 2.68), Lithuanian (Estimate: 0.50; CI: 0.29, 0.71) and Turkish (Estimate: 1.77; CI: 1.51, 2.0).

4.5.3 Gender

With respect to gender, we have relevant coefficients in three languages out of nine; exceptions are Arabic, Greek, Hindi, Italian, Japanese and Lithuanian. These three languages can be organized into the following types:

- Positive coefficient estimates for the feminine and the masculine gender: Dutch (feminine: Estimate: 0.81; CI: 0.58, 1.03; masculine: Estimate: 1.79; CI: 1.55, 2.02; see Figure 4.5a) and Ukrainian (feminine: Estimate: 0.58; CI: 0.14, 1.03; masculine: Estimate: 0.49; CI: 0.10, 0.89).
- Positive coefficient estimates for the neuter gender: Ukrainian (Estimate: 1.33; CI: 0.62, 2.06).
- Negative coefficient estimates for the neuter gender: Dutch (Estimate: -1.24; CI: -1.50, -0.98; see Figure 4.5a).
- Positive coefficient estimates for the feminine gender. This happens only in Irish (Estimate: 5.54; CI: 4.48, 6.55).

4.5.4 Clusivity

Only two languages of our sample have clusivity, Indonesian and Mandarin. We observe a single type, with negative coefficients for both the inclusive and exclusive personal pronouns. In Mandarin, the coefficients for the exclusive personal pronouns encompasses zero, but they are far removed from zero for inclusive personal pronouns (Estimate: -8.23; CI: -9.77, -6.70, see Figure 4.6b). In Indonesian, the coefficients are less far removed from zero (see Figure 4.6a), but they are relevant for both the inclusive (Estimate: -1.75; CI: -2.73, -0.78) and the exclusive (Estimate: -0.93; CI: -1.58, -0.31) personal pronouns.

4.5.5 Alternative form

Among the languages with alternative form, the individual LMM for Dutch (stressed vs. unstressed forms) does not have relevant coefficient estimates. With respect to the clitic pronouns we have two types:

- Pronominal clitics have negative coefficient estimates: Albanian (Estimate: -4.52; -5.11, -3.93; see Figure 4.5b) and Greek (Estimate: -1.93; -3.18, -0.68).
- Pronominal clitics have positive coefficient estimates: Italian (Estimate: 2.47; CI: 1.53, 3.38) and Persian (Estimate: 1.09; CI: 0.28, 1.90).

For stressed vs. unstressed forms we find that stressed forms have positive coefficient estimates in Irish (Estimate: 4.24; CI: 3.07, 5.48). Finally, for emphatic vs. non-emphatic forms we find that emphatic forms have positive coefficient estimates in Basque (Estimate: 4.24; CI: 1.45, 7.06).

4.5.6 Some possible patterns and their tentative explanations

Here, we speculate on (explanations for) certain cross-linguistic patterns that we find; a comprehensive summary of the results is presented in Section 4.4-4.5. First, Dutch and Ukrainian are the only languages of the sample to show relevant coefficient estimates for gender; together with Greek, they are also the only languages to feature a third value for gender, neuter. The four-value system for gender (feminine, masculine, neuter, non-gendered) of these languages perhaps leads to more uncertainty as compared to the three-value system (feminine, masculine, non-gendered) of the other gendered languages. This uncertainty may translate into an effect of gender on surprisal. While this effect is always positive for the feminine and masculine pronouns, raising their surprisal with respect to the non-gendered pronouns, it can be either positive or negative for the neuter pronouns; in Dutch, neuter pronouns are less surprising than non-gendered pronouns, while in Ukrainian we observe the opposite effect. Clusivity is negatively correlated with surprisal in both Mandarin and Indonesian, which could be explained by the fact that the inclusion/exclusion of the listener in the discourse is highly predictable from the context. We also observe that inclusive pronouns are less surprising than exclusive pronouns, but unfortunately this is limited to Indonesian, as there are no relevant coefficients for exclusive pronouns in Mandarin. As for languages with alternative forms (clitic, stressed and emphatic forms), we suspect that the positive correlation with surprisal

might be a by-product of the manipulation of information structure. Although we have posited a distinction between stressed (phonetic) and emphatic pronouns, stressed pronouns in Irish and emphatic pronouns in Basque are found in strategies that modify the information structure of the sentence. Stressed pronouns are used in Irish with the "same effect as contrastive stress in English" and are found in cleft sentences (Stenson, 2019, p. 160), while emphatic pronouns in Basque are used when the pronoun is in focus or in topicalized noun phrases (see Section 4.2.1). The same reasoning applies to clitic forms. Coefficient estimates for surprisal and clitic forms are far removed from zero in Albanian, Greek, and Italian, with a substantial difference: estimates are negative for Albanian and Greek but positive for Italian and Italian. These three languages have clitic doubling, as in the following examples:

(1a) Albanian (IE, Albanian: Camaj and Fox 1984, p. 95)

Zefi	më	e	dha	librin.
Zefi	DAT.3SG	ACC.3SG.M	give.PST	book.ACC.SG.M

"Zefi gave me the book"

(1b) Italian (IE, Romance: Maiden and Robustelli 2014, p. 359)

Un	caffè	lo	prenderei	proprio	volentieri!
A	coffee(M.SG)	3SG.M	take.COND	just	love

A coffee I"d just love!

(1c) Greek (IE, Greek: Holton 2012, p. 511)

Ton	sunántisa	ton	adelphó	tis	prochthís.
CC.3SG.M	meet.PST	the	brother.ACC.SG.M	her	yesterday.

"I met her brother yesterday."

This strategy is widespread in Albanian Camaj and Fox (1984, pp. 94–95), where the nominal object of a ditransitive construction is obligatorily indexed on a clitic, as in Example (1a) ; note also that the clitic forms are frequently combined, resulting in a single form, for instance in Example (1a) *më e* > *ma*. The Albanian parser is able to decompose the single form into two clitic forms, so the computational model has access to both forms. The obligatoriness of the clitic doubling strategy for the ditransitive

construction in Albanian is a possible explanation for the negative correlation with surprisal, which is further reinforced by the frequent combination of the two clitics: not only the reader or hearer speaker expects two clitics, but the recipient clitic "prompts" the appearance of the object clitic later on in the sentence. In Greek and in Italian the clitic doubling strategy is restricted to constructions changing the information structure of a declarative sentence, that is, to cleft sentences, right and left dislocations and hanging topics (Holton 2012, pp. 255–266; Maiden and Robustelli 2014, pp. 359–365). However, in Greek clitics are less surprising than non-clitic pronouns, while in Italian they are more surprising. Comparing Greek with Italian, Greek has a much more flexible word order (Holton 2012, p. 518; see also Lascaratou 1998, Talamo and Verkerk 2022), which favors the above-mentioned constructions; possibly, Greek clitics are less surprising than independent forms for this reason. By contrast, strategies manipulating the information structure are less common in Italian and, consequently, clitic pronouns, as the alternative forms in Basque and in Irish, are perceived as surprising, still with coefficients not as far removed from zero.

4.6 Discussion and conclusion

4.6.1 Summary

We start our discussion by addressing, from a quantitative perspective, one of the few typological proposals on personal pronouns: Dryer (2013a)'s typology on the expression of pronominal subject. Our data support the relevance of two of the four categories described there, "affixes on verb" and "obligatory pronouns", while the other two categories, "mixed" and "optional pronouns", seem problematic. In the "mixed" category, we encounter a language like Irish, which uses subject pronouns as frequently as languages with "obligatory pronouns", such as Dutch and Indonesian. At the same time, Ukrainian, another "mixed" language, has very few subject pronouns, more or less the same percentage as observed in languages with "affixes on verb". As for

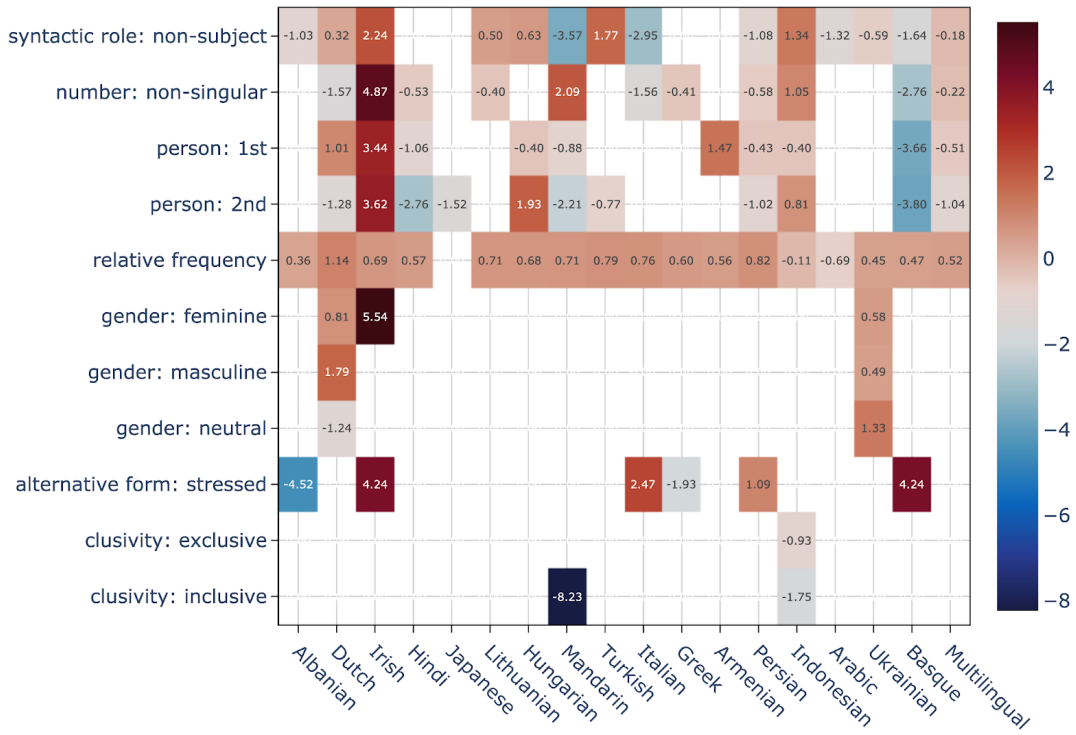


Figure 4.7: Overview of the estimates of the effect of all variables on the surprisal of personal pronouns from single-language linear mixed-effects models and the multilingual model. Only effects of interest are shown, i.e., where the 95 % CI of the coefficient estimates does not include zero.

the languages with "optional pronouns", the two languages in our sample with this classification (Mandarin and Japanese) behave quite differently: Mandarin has the highest percentage of subject pronouns in the languages of the sample, while Japanese has a percentage of subject pronouns similar to languages with "affixes on verbs". While Dryer (2013a)'s typology may make sense from a morphosyntactic perspective, it does not necessarily make sense from a token-based perspective (see Levshina 2022b). We call out for a systematic cross-linguistic study of the use of personal pronouns, considering the impact of verbal person-number agreement, politeness strategies, and other factors.

Beyond the crude data count and focusing on independent and semi-bound personal forms, we have tried to offer a first insight into the behavior of personal pronouns on the information theoretic measure surprisal across different languages. We employ two key

response variables, surprisal as calculated with a preceding context size of seven tokens and AUC (see Appendix A.5). For surprisal with context size seven, the multi-language model has relevant but weakly supported predictors, which are not always matched by results from the individual languages models. As shown in Figure 4.7, the only coefficient estimate which is relevant for both the multi-language and the individual language models is relative frequency: pronouns with higher relative frequency are less surprising. This relation is supported in 14 out of 17 languages of the sample; at the same time, there are pronounced differences in the strength of this relation (see Figure 4.5 and Appendix A.6). In most languages (13/17), postverbal pronouns have lower surprisal than preverbal ones. It would be interesting to find out if this set of pronouns are also non-subject pronouns, where the reduction in surprisal could be rooted in cues on whether a pronoun is coming follows from the valency of the verb. In the individual language models, we mostly find splits between groups of languages such that, for example, first and second person pronouns are more surprising than third person pronouns in Irish, but less surprising in Hindi, Mandarin and Persian (plus borderline cases). This is the case for all linguistic predictors: we do not observe any effect which is consistently present and of the same polarity across all languages. We find more interesting coefficient estimates when looking at features which are shared only by a subset of the sample; Dutch and Ukrainian have relevant coefficients for gender, and clusivity has a strong effect on surprisal in Indonesian and Mandarin. The three types of alternative forms have an effect on surprisal: phonological forms in Irish, morphological forms in Albanian, Greek and Italian and focus forms in Basque. In the remainder of this section we address some of the limitations of the present study, such as the impersonal usage of personal pronouns, their speech act types and the presence of verbal indexing (Section 4.6.4); in Section 4.6.2, we conclude the paper by going beyond surprisal and addressing venues for further research.

4.6.2 Conclusions: moving from surprisal to accessibility

Typological work on personal pronouns tells us that pronouns are accessible in context, and that the usage of pronouns differs radically across languages. The usage of personal pronouns is rooted in grammar, but at the same time, different pronouns abide by different types of rules: subject pronouns may be "dropped" in languages with verb agreement, but object pronouns may be obligatorily present. In addition, usage of pronouns may be entirely different across speaker communities: in some cultures, they are avoided entirely. Rational models (see overview in Arnold and Zerkle 2019) posit that speakers and writers make choices on how to encode referents, and that these choices are rooted in efficiency. In this paper, we have not compared personal pronouns to other types of (nominal) referents, but rather we have tried to assess variation between and within languages when it comes to the surprisal of personal pronouns. What we can say is that frequency is a universal predictor of surprisal, and position (pre- or postverbal) is another relevant predictor for most languages. However, we did not find other predictors that are consistently relevant across languages. The linguistic predictors we find evidence for point at local effects, patterns that are found only in a few or a single language. Many parameters are involved here – we have tried to capture the most salient ones, but we leave to further study the impact of those we have not included as well as their interactions. In short: our findings speak towards the highly varied use of pronouns across these different seventeen languages, with respect to using personal pronouns per se, but also with respect to different pronouns within a single language. While pronouns might always be accessible in context for readers and other language users, they do not behave unitedly when we study their predictability in terms of surprisal. Especially low-frequency pronouns may be very surprising. It will be worth unpacking variation between and within languages further in future studies, above all by focusing on other strategies for encoding referents (full noun phrases, terms of address, see Helmbrecht 2013). Pairing quantitative studies with experimental and qualitative ones will shed further light on the interaction between frequency and usage patterns. Ultimately, we hope to establish more exactly what such patterns of usage imply for accessibility in

terms of human processing, and if and how we can relate these to predictability in computational terms.

4.6.3 Conclusions: Multilingual LMs in typoligical research

As to the applicability of multilingual LMs to our research question, we have shown that by relating surprisal from mGPT to properties of personal pronouns, we can recover much of the differences between pronominal systems in the languages in our corpus, which was possible precisely because of the multilingual nature of the LM. Thus, we conclude that multilingual LMs trained on a large amounts of text data hold some promise for synchronic or typological studies, although future work should evaluate the impact of the distribution of languages in the training data on cross-lingual similarities (see also the next section). The two following Chapters 5-6, in contrast, will deal with a diachronic research question that requires careful control of the pre-training data and tokenization, which we have left out in this chapter.

4.6.4 Limitations

Relevant factors that were not considered for annotation were already discussed in part earlier; these include politeness and coreference resolution. Another factor that is likely to impact personal pronoun usage, at least for some languages, is impersonal usage (Kitagawa and Lehrer 1990; also see the title of the paper). For example, De Hoop and Tarenskeen (2015) show that in Dutch declaratives, a little over 40% of the usage of *je* "2SG, unstressed" is generic (impersonal) in the same sense as you in the English sentence "You/One would think that he would have understood by now." It is unclear how much the impersonal use of pronouns has affected our analysis. Siewierska (2012, Ch. 5.5) notes that the impersonal use of second person pronouns is well-spread throughout Europe and also attested in Mandarin; however, as we have reported in Section 4.4.2, first and third person pronouns are more frequent than second person

pronouns in most languages of the sample, and first person pronouns are more frequent than second person pronouns in all languages of the sample. Speech act type – whether the sentence or clause that the personal pronoun finds itself in is commissive, directive, assertive or part of other illocutionary acts – is another variable that highly impacts referent choice. As stated in the Introduction, in expressions of internal mental states such as wishes (expressive acts), obviously first person pronouns are used; while in speech acts that center on the hearer, such as requests or commands (directive speech acts), we find second person pronouns. Here we should also reflect on the genre and stylistic choices made by the authors and translators of the source material in mini-CIEP+. More specifically, two texts of mini-CIEP+ are narrated using almost exclusively the first person, namely, Anne Frank's diary and Coelho's The Zahir; in contrast, novels such as Süskind's Perfume, Eco's The Name of the Rose and Kazantzakis's Zorba the Greek are told in the third person and contain a great deal of dialogue. Dialogue in novels may be closer to spoken discourse. However, rational use of third person pronouns is often overruled in prose: proper names like Alice are still used instead of she or her, even if it is clear that the sentence described her actions, state of being, or things that happen to her. It would be highly interesting to see future cross-linguistic work on personal pronouns that takes into account medium, genre, and ultimately, illocutionary act. With the exception of Indonesian, Japanese and Mandarin, all languages in our sample use verbal affixes to signal the morpho-syntactic properties of the subject to a certain extent. As we have discussed in the Introduction, the most obvious limitation of a token-based study such as ours is that it cannot take into account morphological marking. Furthermore, languages heavily vary with respect to the information encoded by verbal affixes (see Berdicevskis et al. 2020 on the trade-off in Slavic). The verbal affixes of Ukrainian index referent for number and gender in the past tense:

(2a) Ukrainian (IE, Slavic: Pugh and I. Press 1999, p. 225)
Boris chitav.
Boris read.PST.M.SG
"Boris read, was reading"

(2b) Ukrainian (IE, Slavic: Pugh and I. Press 1999, p. 225)
Vona umylasya.
She wash.PST.F.SG
"She washed (herself)"

while the other languages of the sample index person and number with different numbers of forms in their paradigm. For instance, the present tense in Dutch has three different verbal affixes: a zero form for 1SG, *-t* for 2SG and 3SG, and *-en* for 1PL, 3PL and possibly for the 2SG (Donaldson, 2008, p. 170). In contrast, the present tense in Italian has different forms for each combination of person and number in the paradigm, for a total of six forms: *-o* for 1SG, *-i* for 2SG, a zero form for 3SG, *-iamo* for 1PL, *-te* for 2PL and *-no* for 3PL, plus three distinct verbal thematic vowels between the verbal root and the verbal affixes, with their selection dependent on the specific verb Maiden and Robustelli (2014, pp. 219–221). The extent to which verbal indexing is informative about person, number, gender and other features of the referent is thus highly language specific. Furthermore, in many languages verb paradigms are suppletive or irregular, their number ranging from very few verbs (Albanian: Camaj and Fox 1984, pp. 229–232; Hindi: Kachru 2006, pp. 79–82; Persian: Yousef 2018, pp. 159, 182) to several verbs (Armenian: Dum-Tragut 2009, pp. 277–284; Dutch: Donaldson 2008, pp. 191–205; Greek: Holton 2012, pp. 196–206; Hungarian: Rounds 2001, pp. 269–278; Irish: Stenson 2019, pp. 83–91; Italian: Maiden and Robustelli 2014, pp. 224–240; Ukrainian: Pugh and I. Press 1999, pp. 209, 219–230); this suggests that referent resolution is not always straightforward in languages with sizable paradigms. Aside from the lack of annotation for possibly relevant variables and unresolved cross-linguistic differences in the informativity of verbal indexing, a methodological note should be made on the statistical modeling. In order to avoid problems with multiple testing, we opted for two relatively simple models that center on explaining surprisal and AUC in terms of linguistic predictors. In the multi-language models, we only added random intercepts and slopes for relative frequency, because of a priori beliefs that this is one of the strongest predictors of surprisal. But it would have been possible to add random intercepts (and slopes) for other variables, interactions between predictors

(person and number, for example), and possibly predictors not discussed above. We have refrained from doing this given the explorative nature of this study, and leave further exploration of both cross-linguistic and language individual patterns to further study A further limitation of our study is that personal pronouns are studied in isolation, i.e., without reference to average surprisal for all tokens in the studied corpora, or with respect to other strategies for reference to entities or things: nominal phrases. Given that the interaction between personal pronouns and verbal number-person agreement impacts information structure, it is highly likely that there is a correlation with the (frequency of) use of overt noun phrases to establish reference. This is yet another worthy topic of further research. Finally, while our sample was chosen with the goal of maximizing geographic and genealogical diversity, we want to point out that it does not allow for the evaluation of geographical or genetic effects on pronoun surprisal. Although 10 out of 17 languages belong to the Indo-European language family, they all represent primary branches of that family, and the geographical distance between them varies considerably. For future work, we would expect that languages with a low geographic and genealogical distance are similar in their use of personal pronouns, resulting in similar effects of pronoun features on surprisal.

4.7 Future Work

Apart from the directions already set out throughout this chapter, I see two more general research questions that arise from our decision to use a multilingual corpus and a multilingual transformer LM.

4.7.1 Impact of source language & translation

The first relates to the nature of our corpus. From the 10 books considered in this study, one was originally written in Spanish (*A Hundred Years of Solitude*), two in Portuguese (*The Alchemist*, *Zahir*), one in Italian (*The Name of the Rose*), two in English (*Alice in Wonderland*, *Through the Looking Glass*), one in Dutch (*The Diary of a Young Girl*), one in French (*The Little Prince*), one in German (*Perfume*), and one in Greek (*Zorba the Greek*). To what degree the original language of a book impacts pronoun use in the translation is a question that we have evaded, under the tacit assumption that this effect would "average out" over the 8 different original languages. The primary avenue to address the effect of the original language is to adapt our statistical analysis by grouping translated books with the same original language together and contrast them with untranslated books that were originally written in that language, and fitting LMMs on both groups.

4.7.2 In what sense are "multilingual" LMs multilingual?

However, this only addresses the discrepancy between original and translated *evaluation* data. A related effect stems from the multilingual nature of mGPT: All 61 languages in the pre-training data share the parameter space, and it is reasonable to assume that the dominant languages in the pre-training data shape the LM's linguistic capabilites. This was noted early on in the history of multilingual LMs, e.g., for mBERT (Wu and Dredze, 2020), although to my knowledge there is little research into the exact

makeup of the parameter space of a multilingual model. While there are LMs that explicitly separate parameter spaces of individual languages, e.g., by using adapters (Pfeiffer et al., 2022b)[3] or language-specific tokenizers (Rust et al., 2021), it is not *a priori* clear if this approach is more cognitively plausible than a shared parameter space. A related but divergent research question is how transformer LMs recover linguistic information throughout the forward pass; this could be answered by probing the hidden representations of intermediate layers, e.g., for the properties of a personal pronoun that we considered in this study.

4.7.3 Morphologically adequate tokenization

Previous work from multiple research directions indicates that the default tokenization strategy for (multilingual) LMs does not leverage much of the information contained in the syntactic and morphological structure of the languages in the training data: Nair and Resnik (2023) pointed out that reading time fit to surprisal of high-fertility words[4] (that is, words that are split into multiple subword tokens by the tokenizer) is on average worse than for words that are not split, and find that morphologically aware tokenizers potentially provide better surprisal estimates for high-fertility words. This result has been corroborated outside of the language modeling context by Beinborn and Pinter (2023), who similar to Hofmann et al. (2021) indicate that a broader coverage of derivational morphology correlates with psycholinguistic measures; see also Batsuren et al. (2024) for a more recent work. While Zouhar et al. (2023) showed that BPE tokenization is a fairly good approximation for at least some languages, work on morphologically complex languages like Finnish and Turkish (Ismayilzada et al., 2024) shows that the default tokenization strategy is not optimal if one desires the LM to display compositional generalization at least on the morphological level - a capacity that humans surely possess. This has implications for linguistic research: If a researcher is interested in a specific linguistic phenomenon, say, gender agreement

3 But see Alabi et al. (2024) for the effect of the dominant language of the "adapted" model

4 I use "word" here to refer to whitespace-delimited character sequences in an alphabetical script.

between adjectives and nouns, the way in which this task is solved by a LM will in part depend on its tokenizer: If, in an extreme case, all adjectival forms in the experimental stimuli are part of the tokenizer's vocabulary (i.e., they are not split into subwords), the LM need have learnt the gendered property of each form during training. In the opposite case, the tokenizer always splits adjectival forms into a stem (consisting of any number of subwords) and the gender affixes. In this case the LM has to rely entirely on a learnt compositional generalization mechanism (apart from heuristics, e.g., frequent associations of specific adjective stems and affixes in the training data). While in practice most tokenizers will lie in between these extremes, it is crucial to know which strategy a LM learns, and whether that strategy aligns with the cognitive theory that is supposed to be tested. It would be worthwhile to evaluate which tokenization strategy best represents human language in this regard, or indeed for any other linguistic phenomenon.

5

Modeling Diachronic Change in English Scientific Writing over 300+ Years with Transformer-based Language Model Surprisal

This work began as an attempt to addresses two challenges to applying surprisal theory to diachronic development of scientific English in the Royal Society Corpus (Fischer et al., 2020a). Firstly, surprisal is strongly correlated to model and training data size, as evidenced by, i.a., Kaplan et al. (2020) and Scao et al. (2022) or more recently and on smaller scales Wal et al. (2025)[1], as surprisal is exactly the function optimized by most LM training algorithms. Since the amount of available text data increases exponentially over time, surprisal from periods further in the past will necessarily be higher than surprisal on periods closer to the present. The second factor, tied to first, is the drastic expansion of the scientific vocabulary, which biases the quality of a tokenizer trained naively on the entire corpus in question towards later periods. We address these by using a shared vocabulary and employing a robust training strategy that includes initial uniform sampling from the corpus and continuing pre-training on specific temporal segments. We then conduct an empirical analysis highlighting the predictive power of surprisal from transformer-based LMs, particularly in analyzing complex linguistic structures like relative clauses.

1 These scaling laws do not necessarily hold true when it comes to modeling human linguistic competence; this is discussed in more detailed in Chapters 2.3.4 and 3.

5.1 Introduction

Language models, particularly those rooted in machine learning and neural networks, have revolutionized the way we analyze and understand the intricacies of linguistic change (Kim et al., 2014; Hamilton et al., 2016; Dubossarsky et al., 2019). Different models, such as n-gram, LSTM or transformer models, offer diverse possibilities for analyzing language variation and change due to their underlying architectures. N-gram models are usually based on rather small and fixed-size context windows, excelling in capturing local patterns of variation. Transformer models, instead, employ attention mechanisms and deep neural networks capturing long-range dependencies and global context in language data. While they are less efficient in terms of training compared to n-gram models, they excel at capturing complex syntactic and semantic relationships, making them well-suited for analyzing possibly broader and more complex linguistic trends.

Various studies, especially concerned with lexical semantic change, already employ transformer-based models successfully (Giulianelli et al., 2020). However, comparability of the models over time is not a trivial task as the data sets often vary greatly in terms of corpus and vocabulary size, especially for historical material where the data cannot be extended. In this paper, we apply transformer-based models to explore diachronic linguistic change in 300 years of English scientific writing. In particular, we create models of surprisal (the predictability of a word given its previous context, Shannon, 1948), which are comparable over time. Surprisal models allow us to investigate how change in language use is possibly driven by optimization effects, given that surprisal is proportional to cognitive effort (Hale, 2001a; Levy, 2008). A major assumption for the evolution of scientific writing is that it becomes more informationally dense over time (Biber and Gray, 2016) due to the increasing specialization and diversification of scientific disciplines. On the other hand, conventionalization effects are at play which modulate the informational load. This balance between highly informative content and conventionalized ways of expression allows for an optimal code for expert-to-expert communication (Degaetano-Ortlieb and Teich, 2019). Our overarching aim

is to create robust diachronic language models capable of capturing and quantifying optimization effects in language use. We begin by outlining previous research on changes in English scientific writing, emphasizing the use of information-theoretic notions (specifically surprisal) to capture changes related to efficiency in communication. We continue by elaborating on the challenges in using models with restricted window sizes (n-gram models) and the motivation to apply transformer-based models as well as the challenges associated with the implementation of these models to diachronic data. Next, we describe the dataset used in our study and discuss the methods we adopt for modeling changes over time using transformer-based models. Working with historical linguistic data presents unique challenges, and we address some of these in detail (vocabulary shifts and train set bias). In our analysis section, we assess our transformer-based surprisal models, comparing surprisal trends with those found in earlier studies. We supplement this with a focused study on relative clauses, which require understanding long-range dependencies. These dependencies can be effectively captured by transformer-based models as they consider larger context windows.

The contributions of this study lie in addressing key modeling challenges inherent in historical data analysis, however, our approach has broader applications, extending to other areas where resources are limited.

5.2 Previous Work and Rationale

Diachronic change in the English scientific register has received ample attention in previous work. Earlier, descriptive (Halliday, 1988; Halliday and Martin, 1993) as well as corpus-based studies (e.g., Biber et al., 1999; Biber and Gray, 2011; Biber and Gray, 2016) report on a central mechanism in scientific language which shifts grammatical complexity from the clausal level (subordination and coordination) to the phrasal level (see also Hundt et al., 2012) leading to an increasingly nominal instead of verbal style. Another central development in the scientific register is the conventionalization of lexico-grammatical features, which has been detected to be a necessary condition for innovation on the one hand and grammaticalization on the other (Bybee, 2010; Schmid,

1994). Innovation is probably the most obvious mechanism, as a natural reaction to the need to create new vocabulary for newly arising concepts. Furthermore, diversification of certain features in increasingly distinct contexts has been observed to be at play in the course of the creation of new scientific disciplines and the formation of their respective sublanguages (Halliday, 1988; Harris Sabbettai, 1991).

While the mentioned studies are either qualitative in nature or at most frequency-based, more recent studies have employed information-theoretic measures such as n-gram-based surprisal to detect diachronic changes in the register (Degaetano-Ortlieb and Teich, 2016; Degaetano-Ortlieb and Teich, 2018; Teich et al., 2021). Surprisal is formalized as the negative log probability of a unit in context $Surprisal(unit_i) = -log_2 P(unit_i|Context)$, which results in bits of information (Shannon, 1948). The motivation to abandon a mere frequency-based approach in favor of n-gram-based surprisal is the assumption that linguistic change underlies the rational strive for communicative efficiency. Since surprisal is a widely-used measure of information, which has been shown to be correlated with cognitive effort in online language processing (e.g., Levy, 2008; Demberg and Keller, 2008b) it is well suited to giving a communicative explanation for changes in the lexical as well as grammatical level. Degaetano-Ortlieb and Teich (2019), for instance, show that in scientific writing certain grammatical patterns become less surprising over time, i.e., increasingly conventionalized in their contexts, while specific lexical items show a trend toward "innovation and increase in expressivity" (Degaetano-Ortlieb and Teich, 2019, p. 26) indicated by phases of high surprisal when new concepts enter the language and phases of stability/consolidation when items of lexical usage become conventionalized in their contexts. An example of the interplay between lexical innovation and grammatical consolidation is the noun – preposition – noun pattern (e.g., *oxide of iron*) becoming extremely predictable as a grammatical pattern while serving as a "habitual host for lexical innovation and terminology formation" Degaetano-Ortlieb and Teich (2019, p. 26). The mentioned studies give a plausible explanation for the underlying motives of language change, however, their underlying 4-gram language models (i.e., surprisal based on a word's predictability given its previous three words as context) are fairly restricted in terms of context size.

While the narrow context of only three preceding words is well suited for detecting optimization of shorter linguistic units such as the above-mentioned noun-preposition-noun pattern, it is less well suited for drawing conclusions about the diachronic development of linguistic structures exceeding this window size. A possible solution to this is to replace the n-gram with a model covering a larger context window such as an LSTM (Hochreiter and Schmidhuber, 1997) or Transformer (Radford et al., 2018), which is the approach we present in the present paper. However, applying such models, especially transformer-based ones, poses several challenges (e.g., varying corpus and vocabulary sizes or selection of data for modeling). In this paper, we work towards addressing some of these challenges (cf. Section 6.5). While there is a growing body of research on which language model architecture to choose if restricted to a limited compute budget (e.g., Scaboro et al., 2021) or a specific dataset size (e.g., Hoffmann et al., 2022), it is less clear what approach one should take if either one's compute budget or dataset size is very small. This becomes especially pressing in the case of historical corpora as there are only limited ways of extending the data. While it is possible to train large language models on historical English data (e.g., Hosseini et al., 2021), doing so might be undesirable for a number of reasons. When the reason for training a language model is to create a computational model which serves as an approximation – ideally a cognitively plausible one – of a speaker of a specific time period, the option of training the language model on large-scale text data is not available, since the training data should, for example, be restricted to a time period *preceding* any text from that time period, with the basic assumption that a speaker did not have access to future text productions. This is of course a simplified assumption; in the case of written text, it is plausible to assume that *every* reader has access to some amount of data that lie in the future from the perspective of any given text. However, there are domains in which this amount can plausibly be assumed to be small, such as scientific English writing.

5.3 Data and Methods

5.3.1 The Royal Society Corpus

We use the Royal Society Corpus (RSC) Fischer et al. (2020b) and Kermes et al. (2016) as a data set. The RSC is based on the *Philosophical Transactions* and the *Proceedings* of the Royal Society of London. In total, it comprises 295 895 749 tokens and 47 837 texts, which were published between 1665 and 1996. The RSC incorporates a comprehensive set of metadata such as text categories (e.g.,, articles, abstracts), authorship, title, publication date, and historical periods (ranging from decades to half-centuries), along with linguistic annotations at multiple layers including tokens (featuring to some extent both normalized and original forms), lemmas, and parts of speech, utilizing TreeTagger (Schmid, 1994), and Universal Dependency parsing achieved by Stanza (Qi et al., 2020) with combined models. Given that the texts underwent OCR, several preliminary procedures were employed to counteract OCR inaccuracies to the greatest extent feasible (for an in-depth explanation, refer to Kermes et al., 2016; Menzel et al., 2021).

5.3.2 Problems for diachronic language modeling

Previous research on the diachronic linguistic development of English scientific writing (Degaetano-Ortlieb et al., 2018; Degaetano-Ortlieb and Teich, 2018; Degaetano-Ortlieb and Teich, 2019) has employed surprisal derived from n-gram language models as a proxy of a model's linguistic knowledge. The hypothesis is that as syntactic structures are conventionalized (i.e., become more predictable) over time, surprisal from n-gram language models fitted to texts from successive time slices of the RSC decreases. While previous work indeed found such an effect (Krielke, 2021), n-gram language models are only a good approximation for local effects and ideally, a more cognitively plausible model should have access to the full sentence-level context. A possible solution is to

replace the n-gram by a neural language model with a larger context window (LSTM or transformer). However, for historical data such as the RSC, this is far from trivial as the number of texts and the number of tokens per year decrease exponentially as we go back in time. In such a setting, it becomes increasingly difficult to train and compare language models for two reasons:

1. **Vocabulary Shift.** When sampling from periods t and t^{+1}, the sets of word types or vocabularies V^t, V^{t+1} will be partially disjoint, i.e., $V_t \cap V^{t+1} \neq \emptyset$. When training language models M^t, M^{t+1} on t and $t+1$, the probability distributions P_{M^t}, $P_{M^{t+1}}$ cannot be directly compared since they are defined over different sets of events.

2. **Train Set Bias.** Let C^t be the set of texts from period t. Since $|C^t| << |C^{t+1}|$, M^{t+1} will see much more training data than M^t when naively sampling from the corpus. The probability estimates derived from M^{t+1} will be tighter than those derived from M^t as a function of the train set size, if the vocabulary stays constant, i.e., for an identical prefix $w_{0...i-1}$ we expect that $P_{M^{t+1}} w_i | w_{0...i-1} \geq P_{M^t} w_i | w_{0...i-1}$.

5.3.3 Approach

Continuous Pre-training While vocabulary shifts can be addressed by sharing a unified vocabulary over all models, train set bias requires sampling the train set such that M^t and M^{t+1} are trained on a similar number of tokens, which is problematic because for earlier periods we may have only very little data. In order to alleviate the effects of trains set bias, we make use of the default NLP pipeline of pre-training a transformer model on a more general dataset D^0_{PT} sampled uniformly from the each time period C^t, and then continue pre-training on the documents of a specific year C^t. In our experiments we use the the smallest version decoder-only OPT architecture (S. Zhang et al., 2022b), with randomly initialized weights.

Pre-training Dataset We sample a pre-training dataset D^0_{PT} from all documents in the corpus such that an equal number of tokens is sampled from the documents C^t. Sampling 10^5 tokens yields $|D^0_{PT}| \approx 2 \times 10^6$. We derive a unified vocabulary by training a BPE tokenizer with $|V| = 5 \times 10^4$ on D_{PT}. We then pre-train on D^0_{PT}, obtaining a set of pre-trained weights θ_{PT}. Surprisal for words that are split into subwords by the tokenizer is calculated by summing their respective log probabilities.

Pre-training on Individual Years We sample datasets D^t_{PT} for each year t in the corpus. Each D^t_{PT} consists of the documents from a period of k years prior to t such that starting from $t_0 = t - k$:

$$D^t_{PT} = \bigcup_{t'=t_0}^{t-1} C^{t'} \tag{5.1}$$

We then initialize the model with θ_{PT} and fine-tune on D^t_{FT} until validation loss converges. Hyperparameters for continuous pre-training can be found in Table 5.1. This results in a similar number of training steps (300-400) on each D^t_{PT}, independent of $|D^t_{PT}|$.

Hyperparam	Value
Batch Size	128
Learning Rate	1^{-3}
Warmup	Linear, 10%
Optimizer	AdamW

Table 5.1: Pre-training hyperparameters

Implementation We used HuggingFace Transformers (Wolf et al., 2020b) to train the BPE tokenizer and to pre-train and fine-tune the OPT model. The source code will be made available on GitHub alongside instructions to replicate the result upon publication.

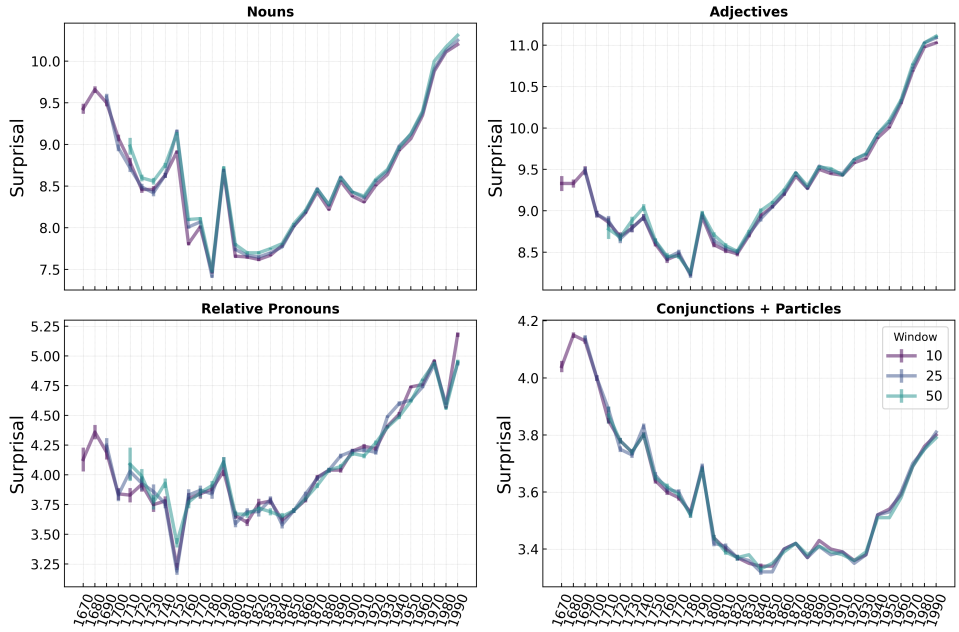


Figure 5.1: Surprisal on nouns, adjectives, relative pronouns ('which'+'that' with XPOS tag 'WDT') and conjunctions/particles as defined by the UPOS annotations of the RSC. Error bars show standard error.

5.4 Analyzing Linguistic Change

One main assumption regarding the development of scientific English is a balance between informationally dense scientific content and conventionalized scientific style (cf. Degaetano-Ortlieb and Teich, 2019). In fact, it has been shown that scientific English changes from a verbal involved style with embedded clausal structure (see Example (1)) toward a heavy nominal style with long nominal phrases (see Example (2)) and predictable grammatical structures (Biber and Gray, 2016; Degaetano-Ortlieb and Teich, 2019).

(1) *And if the greatest part of these Vessels are Arteries, or other Vessels,* ***that*** *immediately* ***receive*** *liquors from them;* ***I may prove***, ***I think****, from another Experiment, made by Injection into a part of the Arteria praeparans, before* ***I began*** *to expand the Body of the Testis;* ***whereupon*** *opening the part,* ***which I saw*** *discoloured,* ***I found****, that many*

of these Tubes ***had received*** *some of the fine particles of that matter,* ***which I tinged*** *my injected Spirit with.* (King and de Grieff, 1669)

(2) *On the other hand,* ***a clear red-green stripe pattern of predominantly positive or negative response*** *emerges in* ***the vertical motion signal****.* (Zanker, 1996)

We start by testing whether the results obtained by transformer-based surprisal models are in line with previous findings. We employ the Mann-Kendall trend test (in the Python implementation by Hussain and Mahmud (2019)) to confirm visually salient increases or decreases of surprisal (either on specific words or averaged over POS tags). In particular, we report the direction of the trend (increasing, decreasing, or no trend), its slope s and the p-value p associated with it. Figure 5.1 shows surprisal of the lexical word classes nouns and adjectives as well as the grammatical function words relative pronouns (e.g., *which*, *that*) and conjunctions/ particles (e.g., *and*, *to*). The surprisal values are calculated as the mean surprisal of all words belonging to a class per decade, determined by the word's POS tag in the CoNLLU. From the 1800s onward we can see an increase in surprisal for word classes associated with nominal style, nouns ($s = 0.0181$, $p < 0.001$) and adjectives ($s = 0.174$, $p < 0.001$). Function words, instead, show a steady decrease during the 17$^{\text{th}}$ up to the 1840s ($s = -0.01$, $p < 0.001$) followed by a slight but not significant increase from the 1930s onward ($s = 0.0093$, $p = 0.2097$). Thus, while nouns and adjectives in general carry more information (higher surprisal, between 7.5 - 12 bits) on average, function words carry much less information (between 3-5 bits). While these findings are in line with the general trend observed in previous work (cf. Kermes and Teich, 2017), we should state that considering average lexical surprisal of words belonging to specific word classes only shows a very aggregated picture as a result of the confluence of different factors, e.g., word frequency, vocabulary diversity, collocational distributions of words per decade.

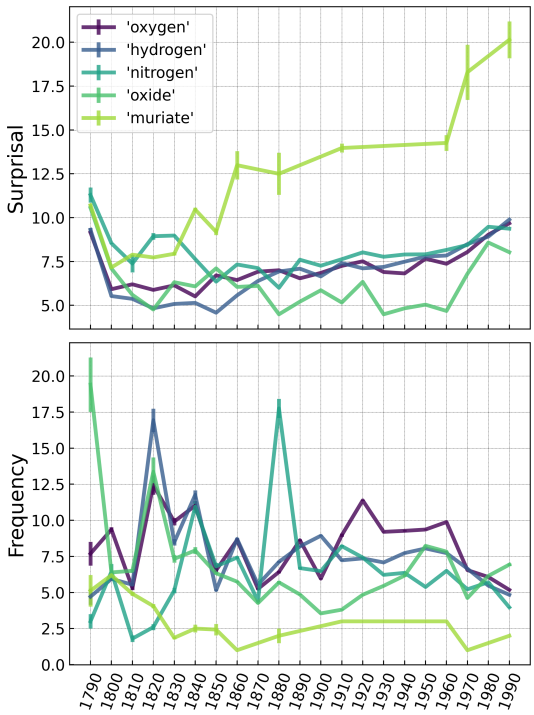


Figure 5.2: Average surprisal and frequency of nouns contributing to the surprisal increase and related to the chemical revolution. Error bars show standard error over documents.

5.4.1 Modeling Convention and Innovation with Surprisal

A more thorough inspection at the lexical level (nouns and adjectives in Figure 5.1) seems to indicate a wave-like tendency with periods of alternating peaks (e.g., 1680, 1750, 1790, 1890, 1990) and troughs (e.g., 1730, 1780, 1820, 1910) in surprisal. Considering that peaks in surprisal indicate an increase in the use of unpredictable words given their previous context and troughs stand for more predictable usages, the observed changes could be related to discoveries triggering new vocabulary in the corpus at hand, especially at points with abrupt changes such as in the 1790s for nouns. In fact, this peak coincides with the chemical revolution where the 100-year-old phlogiston theory was replaced by the evidence around the discovery of oxygen and hydrogen. Troughs, on the other hand, reflect periods of consolidation, where new vocabulary is integrated into language use possibly becoming established terminology (cf. Degaetano-Ortlieb and Teich, 2019). To test this assumption, we further inspect the nouns contributing most

to the surprisal increase in the 1790s (see Figure 5.2 showing surprisal and frequency). At their first mentioning in the 1790s, these nouns are relatively high in surprisal, but strongly decrease in the decade 1800, when their frequency increases, stabilizing at a mid-surprisal range in the coming decades (1800-1840). This is clearly an indication of a point in time (1790s) of innovation in terms of the use of new lexemes followed by a period of conventionalization, where new terminology was established in the new chemistry field. Thereafter, the chemical elements oxygen ($s = 0.031$, $p < 0.001$), hydrogen ($s = 0.0424$, $p < 0.001$), and nitrogen ($s = 0.0188$, $p < 0.05$) show a continuous increase in surprisal, with a clear peak from the 1970s to the 1990s. This tendency seems to indicate two distinct mechanisms that might have an impact on the nouns' surprisal: (1) given that the frequency is not decreasing until the 1960s, the nouns might be used in more diverse contexts which would explain their increase in surprisal (e.g., *thought that/permeable to/the ketonic oxygen* with high surprisal of *oxygen* >10 bits), (2) in the period of the 1970s to the 1990s, their frequency decreases, which might explain their even stronger increase in surprisal in that later period.

5.4.2 Modeling Surprisal for Long-Range Dependencies

In this second analysis, we focus on relative clause (RC), which inherently involve long dependencies (see Example (3)), necessitating models that can appropriately handle such complexities. In this regard, transformer-based models are arguably more effective than n-gram or LSTM models as they can make use of very long contexts, offering deeper, context-sensitive analysis that is crucial for accurately capturing the nuanced aspects of these linguistic structures.

(3) *The **protein**, for which the detailed biochemical pathway analysis conducted by the researchers identified several novel interaction partners, **exhibits** properties consistent with increased metabolic resistance.*

Surprisal of Relativizers: We start by considering the surprisal of the relative pronouns in Figure 5.1, which show a clear turn in 1920, while conjunctions and particles seem to stabilize in surprisal for more than 100 years between 1800 and 1920. Interestingly for relativizers, the development until the 19th century is not exactly in line with previous research on the diachronic development of relativizer informativity (cf. Krielke, 2021), which reports on an overall slightly increasing surprisal of relativizers (*which* and *that*). Krielke (2021) explains the upward trend with the frequency decrease of relativizers in scientific writing. However, they report on an increasing conventionalization of *which* as the preferred relativizer in scientific writing occurring in increasingly uniform contexts accounting for some very low surprisal values. An explanation for the differences in our results might be the different modeling of surprisal since Krielke (2021) uses a 4-gram surprisal model based on probability estimates of relativizers within 50-year periods. As our estimates are based on a dynamic context size and models are more precise in terms of surprisal prediction in different time slices due to the balanced vocabulary size of each slice, we can assume that our results reflect changes in relativizer informativity more reliably. Moreover, the increase in surprisal in the 20th century in our data may have two mutually non-exclusive explanations: (a) relativizers become less frequent in the 20th century, and (b) they occur in increasingly diverse contexts as a reaction to specialization, e.g., they might follow a higher diversity of nouns, which are also less predictable (cf. Figure 5.1).

Relativizers in Context: We furthermore inspect more thoroughly the reasons for the increasing surprisal values of relativizers in the 20th century by examining a) the frequency distributions of relativizers over time (Figure 5.3) as well as the immediate lexical contexts of relativizers (*that* and *which*, Table 5.2). The frequencies show that the overall strongest decrease in relativizers over time happens roughly in the 19th century. This shows that the sudden increase in surprisal of relativizers in the 20th century does not seem to be motivated by a major frequency drop but rather by a specialization of contexts that relativizers tend to occur in. The most frequent part-of-speech 3-grams preceding the relativizers *that* and *which* reflect this: *that* in the decade 1990 mostly occurs after complex noun phrases (Table 5.2), which can be assumed to decrease the

predictability of the relativizer. The preceding contexts of *which* are more predictive since they introduce either prepositional RSCs (e.g., *the way in which*) or restrictive RSCs separated from the matrix clause with a comma (e.g., *the experiment, which*). What is interesting here is the fact that the more predictable *which* steadily drops in frequency while *that* steadily increases from 1900 onward. This could explain the overall increase in surprisal since the strongly conventionalized *which* becomes less influential in the average surprisal while *that* becomes less predictable in context and more frequent.

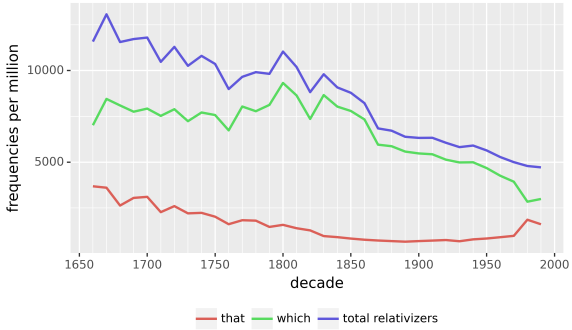


Figure 5.3: Distribution of relativizers in the RSC.

While this kind of pattern-based context analysis would also be possible using 4-grams, our surprisal model should also account for larger contexts and better surprisal estimates compared to the 50-year-based 4-gram surprisal model used in previous studies allowing more reliable interpretations of diachronic trends.

Surprisal in Long-Range Dependencies; Here we ask whether syntactic context has an influence on the predictability of syntactic elements relying on longer distances to their syntactic heads. As a plausible measure to define syntactic context, we make use of the well-established metric dependency length (DL) describing the distance from an element X to its syntactic head (Heringer et al., 1980; Hudson, 1995). We apply this metric to measure the distance between the head noun of an RSC and its embedded verb to find out whether the distance correlates with the predictability of the embedded verb. One assumption could be that the closer the relevant information (head noun) is located

freq.	trigram	example
3034	IN DT NN	of the acid *that*
2893	DT JJ NN	the muriatic acid *that*
1627	NN IN NNS	number of experiments *that*
1473	IN JJ NNS	in various instances *that*
1415	DT NN NN	the iron particles *that*
6392	DT NN IN	the way in *which*
3123	JJ NN ,	unique solution, *which*
2866	JJ NN IN	special case in *which*
2818	NN NN ,	length scale *which*
2722	DT JJ NN	the complex plane *which*
1767	(CD)	(3.1) *which*

Table 5.2: POS trigrams preceding *that* and *which*

to the upcoming dependent (embedded verb) the lower the surprisal of the upcoming word:

(4) *The **woman** who **ate** the sandwich was hungry.* ($DL = 2$)

(5) *The **woman** whom the manager of operations wanted to talk to **was** upset.* ($DL = 10$)

We start by inspecting the diachronic development ofDL and the surprisal of embedded verbs in RSCs by decades (see Figure 5.4).DL and surprisal are negatively correlated, i.e. the opposite of our assumption is the case: in decades where the embedded verb on average is further away from its syntactic head the verb's average surprisal is lower, and in decades where the verb is closer to its syntactic head the verb is more surprising:

(6) *The **questions** that we might **ask** are not easy to answer.* ($DL = 4$)

(7) *The **questions** that **lie** on the table are not easy to answer.* ($DL = 2$)

One explanation is that at the point of encountering the relativizer the entropy (i.e. uncertainty about the rest of the sentence, Hale, 2006) is higher than at a later point in the relative clause.

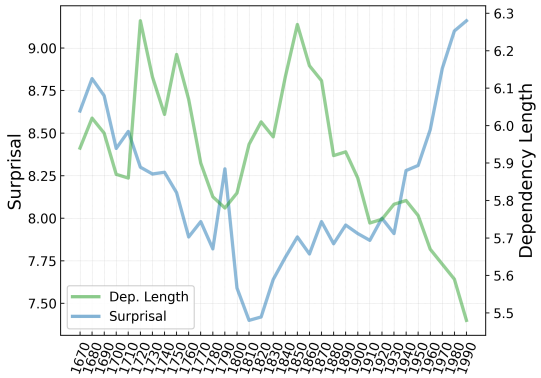


Figure 5.4: Average surprisal on the verb in the RSC and dependency length to its head noun in the RSC. Negative correlation for surprisal and dependency length (Spearman's $\rho = -0.35, p < 0.05$).

To get a better intuition about the reasons for the lower predictability in shorter syntactic contexts and the higher predictability in longer syntactic contexts, we extract the most surprising contexts of embedded verbs in RSCs in 1990 plus the least surprising contexts in 1820. Example (8) reflects two aspects we have mentioned so far. First, the conventionalized pattern of prepositional RSCs (i.e. *determiner noun preposition*) contexts not only seem to increase the predictability of the upcoming relativizer but also that of the embedded verb. Second, surprisal at the main verb (participle) in the RSC might be reduced due to being in a highly predictable passive construction (*has been + participle*).

(8) *the first **case** in which a quantitative attempt has been **made*** ($Surprisal = 0.0142$)

Conversely, the high-surprisal contexts are extremely short where the head noun is directly followed by the relativizer and the embedded verb (Example (9)).

(9) *recordings in the region of the pontine **nuclei** that **VERB*** ($Surprisal = 9.9998$)

The latter, high-surprisal cases syntactically belong to the subject RC type, while the first, low surprisal cases belong to the oblique relative clause type. The negative correlation between surprisal andDL can thus also be explained with Hale's Entropy Reduction Hypothesis (ERH, Hale, 2006) – uncertainty about the rest of the sentence tends to decrease as new words are introduced, and the degree of this reduction aligns with the information that the word conveys within the context of the current sentence (cf. Frank, 2013, p. 476). Thus, a high surprisal value at the verb of a subject RC directly following the relativizer is equivalent to a strong reduction of uncertainty about the rest of the sentence since at this point the relativized grammatical relation can be resolved. The low surprisal at the embedded verbs of RCs extracted from other positions (e.g., oblique) implies that entropy reduction here is much lower since a lot of information for disambiguation about the rest of the sentence has been given before.

For comparison, we consider the 1820s whereDL is comparatively long, while surprisal is fairly low. Example (10-a) shows a particularly long dependency relation between the head noun (*bundles*) and the embedded verb (*composed*) with extremely low surprisal.

(10) *denote the* ***bundles*** *of fibres of which the brain is* ***composed*** ($Surprisal = 0.1132$)

Since the immediate left context (i.e. *the brain is*) of *composed* is not particularly predictive, while the head noun *bundles* is much more so, our surprisal estimates seem to rely on a context beyond the immediately preceding 3-gram. Thus, our modeling approach allows us to capture long-range syntactic relations. Together with the fact that our results align with observations from psycho-linguistic studies (e.g., Frank, 2013), we conclude that our methodology for generating diachronically comparable surprisal estimates provides plausible metrics to investigate the interplay between information content and syntactic context in diachronic language change.

5.5 Conclusion

We demonstrated the efficacy of transformer-based surprisal models in analyzing diachronic linguistic change, highlighting their capacity to accommodate long-range dependencies and global context. The application of these models to historical data is not without challenges given the exponential increase in texts and tokens over time, leading to partially disjoint vocabularies and non-comparable probability distributions across different periods. Significant disparities in training set sizes between time periods further complicate the modeling process. To address this, we implement two key strategies: (1) sharing a unified vocabulary over all models, (2) pre-training on a more general dataset sampled uniformly from the whole corpus and then continue pre-training on documents of a specific year. Our models also uniquely incorporate a temporal aspect, restricting training data to texts published before the target period (which can be adapted for other research questions).

Our empirical analysis, compared against prior studies using n-gram models on the Royal Society Corpus, revealed both corroborative and novel insights. Specifically, the examination of linguistic phenomena, such as relative clauses, underscored the superiority of transformer-based models in predicting changes not solely based on changes related to frequency distributions. These models, with their broader contextual awareness, facilitated a more nuanced exploration of communicative efficiency, aligning with theoretical frameworks like Hale's Entropy Reduction Hypothesis. The inclusion of DL annotations further enriched our analysis, allowing for a more granular examination of syntactic structures. This provided deeper insights into the adaptive mechanisms of language, reflecting shifts in complexity and efficiency within scientific discourse.

This study should be considered to be the groundwork for Chapter 6, where we use the methodology laid out here to estimate the memory-surprisal trade-off (Hahn et al., 2021) of different periods of the RSC.

6

Modeling the Memory-Surpisal Trade-Off over Time: Communicative Efficiency Decreases with Lexico-Grammatical Change in Scientific English

In this paper, we apply the methodology developed in the previous chapter to explore the influence of diachronic variation of the memory-surprisal trade-off (MST) in the scientific English of the Royal Society Corpus (RSC), spanning from the 18th century to modern time. In order to assess the impact of intra-linguistic variation (register), we compare scientific English with "general language" using parts of the Corpus of Historical American English (COHA). We observe a clear diachronic effect for scientific English towards decreased efficiency as scientific texts shift from verbal to nominal style and the lexicon in the scientific domain expands, while in general language the effect is less pronounced.

The content in this chapter is based on:
Julius Steuer et al. (under review). "Modeling the Memory-Surpisal Trade-Off over Time: Communicative Efficiency Decreases with Lexico-Grammatical Change in Scientific English". In: *Submitted to ACL Rolling Review*

Julius Steuer and Marie-Pauline Krielke conceptualized the research and led the paper writing. Julius Steuer trained the LMs and ran the evaluation. Stefania Degaetano-

Ortlieb, Elke Teich and Dietrich Klakow helped set out the research agenda, suggested new directions and gave feedback during the experimental phase.

6.1 Introduction

The development of scientific English over the last 300 years was characterized by a shift from more intricate sentence structure with a high degree of clausal embedding towards increasingly informationally packed noun phrases, shorter sentences, and decreasing dependency length (DL). These changes have led to the conclusion that scientific English has become syntactically less complex at sentence level and more complex at noun phrase level over time (Juzek et al., 2020; Krielke et al., 2022; Krielke, 2024). At the same time, scientific English has expanded its vocabulary drastically from ca. 1900 onward, as seen, e.g., in the exponential increase of noun types in the RSC (Fischer et al., 2020a) (see Figure 6.3).

In this paper, we set out to model the impact of lexical expansion and syntactic change on communicative efficiency in terms of the memory-surprisal trade-off (MST, Hahn et al., 2021). The MST unifies two competing approaches to communicative efficiency: Surprisal theory (Levy, 2008), focusing on expectation-based efficiency, assumes that a word w_t becomes easier to predict the more context information (e.g., preceding words $w_1, ..., w_{t-1}$) is available. Dependency Locality Theory (Gibson, 2000) assumes that memory-based efficiency is optimized if words that are close to each other in a dependency tree (see Figure 6.1) are also close to each other in the surface form of a sentence, i.e., if dependency lengths between words on average are small. The MST combines these approaches by positing that for a given language, the actual word order in a sufficiently large corpus balances the requirements of predictive processing (surprisal theory) and communicative efficiency (dependency locality) by optimizing the amount of information that needs to be stored in memory to reach an average surprisal level. In their seminal paper, Hahn et al. (2021) showed that the syntax of a typologically diverse set of languages is optimized with respect to this trade-off.

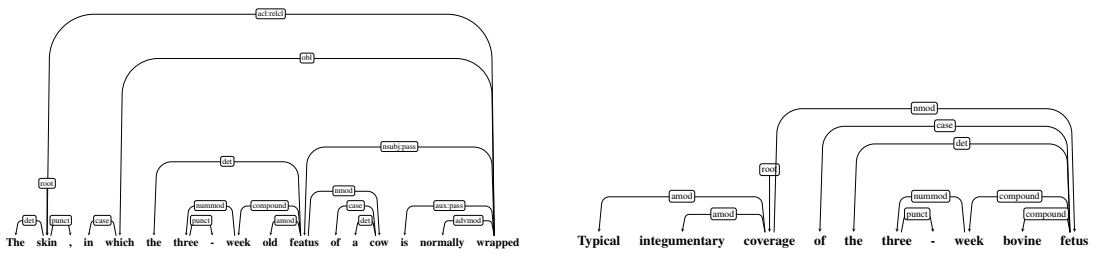


Figure 6.1: Dependency structures of (a) noun phrase with relative clause postmodification (15 words) and (b) noun phrase with multiple premodication (9 words).

We ask whether the communicative efficiency of scientific English as measured by the MST changes over time due to linguistic change, and if so, whether it becomes better or worse and which factors are most influential. We also ask whether any changes in MST are register-specific, i.e., whether scientific English is affected more than general English, and whether the language of different scientific disciplines reacts differently (due to domain-specific lexical expansion).

6.2 Related Work

6.2.1 Diachronic development of scientific English

In the past 300 years, scientific English has undergone substantial changes on the lexical and grammatical levels (e.g., Banks, 2003; Halliday, 1988). Lexis is continuously expanded with new technical terms (Halliday and Martin, 1993; G. Wang et al., 2023), and due to the increasing shared background knowledge within individual scientific disciplines, grammatically explicit constructions such as clausal subordination (Example in Figure 6.1a) become less frequent (Hundt et al., 2012; Krielke, 2024) in favor of a dense, implicit nominal style with heavy noun phrase constructions (Biber and Gray, 2011; Biber and Clark, 2002) (cf. Example in Figure 6.1b).

According to rational communication, diachronic change is a continuous process of adapting the linguistic system to emerging communicative needs while holding process-

ing effort stable. Information-theoretic approaches (Degaetano-Ortlieb and Teich, 2022) have shown that periods of *lexical expansion* are associated with increased surprisal (and thus increased processing load), (e.g., Steuer et al., 2024b), while *grammatical conventionalization* leads to optimization of expectation-based processing for increasingly predictable grammatical constructions (Degaetano-Ortlieb and Teich, 2022; Degaetano-Ortlieb et al., 2019; Teich et al., 2021; Bizzoni et al., 2020).

To cognitively assess syntactic phenomena, dependency locality (the distance between syntactically related words) has been used to approximate the processing difficulty of working memory (Gibson, 1998; Gibson, 2000; Lewis and Vasishth, 2005). While overall, languages tend to minimize the length of their syntactic dependencies (Futrell et al., 2015; Liu, 2008) compared to random baseline word orders, this also applies diachronically (Gulordava and Merlo, 2015; Lei and Wen, 2020) and in specific registers (Juzek et al., 2020; Krielke, 2024). In the present paper, we set out to measure communicative efficiency by applying the MST over time as well as by register (scientific vs. non-scientific language).

6.2.2 Memory-surprisal models

Hahn and Futrell (2020) extend expectation-based processing models (Levy, 2008) and lossy compression theory (Cover and Thomas, 2006) to propose an information-theoretic framework for memory efficiency in language. They define memory efficiency as a trade-off between surprisal and memory usage where reducing average surprisal per word requires storing more information about past context. Applying this to 54 languages, they find that word order optimizes processing efficiency under memory constraints, supporting the idea that syntax facilitates efficient online processing. Hahn et al. (2021) extend the notion of the MST proposing the Efficient Tradeoff Hypothesis, which suggests that word order in natural language is shaped by pressures to optimize this tradeoff. They further derive that languages achieve more efficient tradeoffs when they exhibit information locality, i.e. predictive information about a word is concentrated in its immediate preceding linguistic context. While these approaches have proven a

cross-linguistic tendency to order words and morphemes to achieve a maximally efficient tradeoff between memory and surprisal, to date, the approach has not been applied to intralinguistic or diachronic studies.

6.2.3 Intuition

Figure 6.2 illustrates the relationship between MST curves and area under the curve (AUC) for five years from the RSC: the curve for the last year (1900) starts at the highest unigram surprisal and then converges to the highest surprisal level after the maximum amount of memory bits, resulting in the highest AUC value. Conversely, the first year (1820) starts at the lowest unigram surprisal and reaches the lowest surprisal level overall after the maximum number of memory bits, resulting in the lowest AUC and thus a more optimal MST. Between those years, 1950 still follows the (expected) trend of increasing unigram surprisal, but intersects the MST curve of 1850, leading to a *lower* AUC and a temporary increase in optimality w.r.t. the MST.

6.3 Rationale and Hypotheses

Over time, scientific English has shifted toward a nominal style with high lexical density and syntactic conventionalization, promoting local but implicit dependencies. Simultaneously, vocabulary growth increases lexical variability, suggesting an interaction between lexis and grammar. Nominal constructions reduce the efficiency of memory-based prediction due to implicit dependencies, whereas verbal constructions support explicit, less local dependencies and benefit more from memory. As vocabulary expands, the average lexical surprisal rises, implying that the minimum achievable surprisal in later periods exceeds that of earlier ones.

Specifically, we expect that changing preferences for specific syntactic constructions will lead to different shapes of the MST. For instance, a language variety (e.g., register and/or period) with a high usage of subordinate constructions leads to longer

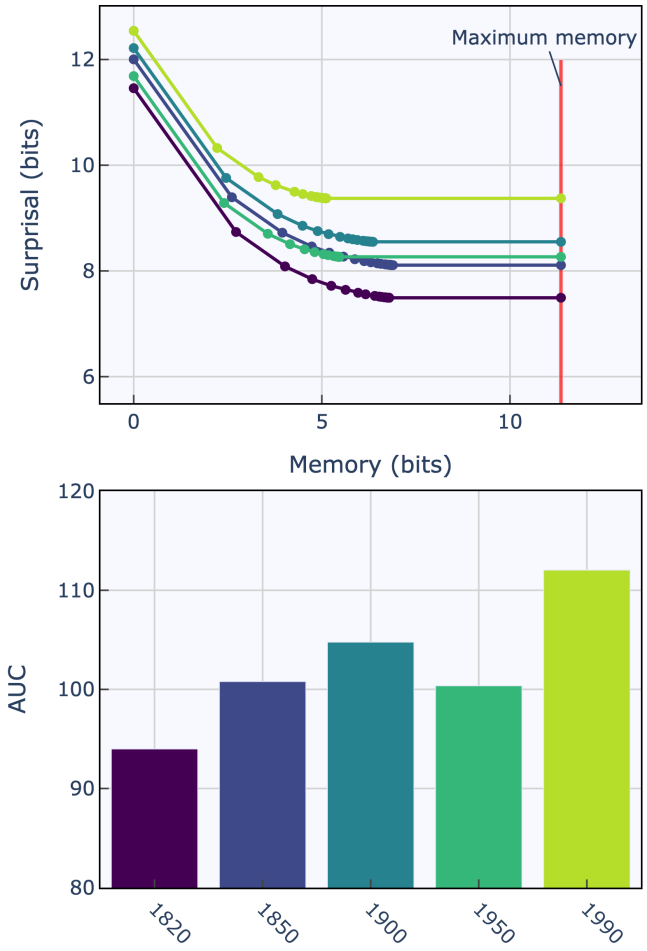


Figure 6.2: MST curves (1820-1990) and their respective areas under the curves (AUC). MST curves where extended to the maximum per-document memory in the RSC.

dependencies generating longer predictive contexts (e.g., Figure 6.1a). Such constructions benefit from higher memory usage to predict the next word, since more memory helps to reduce surprisal. In contrast, varieties using highly dense constructions (e.g., Figure 6.1b), less memory should be enough on average to predict the next word, while more memory should not necessarily improve the prediction. To quantify the quality of the MST over time, we calculate the AUC of the memory–surprisal graph per decade in scientific and general English. We compare the AUCs calculated for both corpora per 50-year periods to find out if optimization on memory efficiency develops differently in scientific vs. general English. For a more fine-grained analysis, we calculate the MST for word classes (nouns, verbs, other) and compare the AUCs respectively. Since the

AUC can only give us a reduced picture of the actual shape of the MST, we also consider the actual MST graphs and interpret their slopes. If a graph flattens at a low memory budget, this means, more memory does not contribute to improving the prediction of the next word. If a graph decreases steadily, this means that every further token held in memory improves the prediction of the next word further. @Based on the attested developments in scientific English, we form the following hypotheses:

H1.1: Impact of average surprisal We expect the MST to deteriorate (i.e. increasing AUC) in both RSC and COHA (scientific vs. "general" English) due to the general increase in surprisal through vocabulary expansion.

H1.2: Impact of register We expect the MST to deteriorate (i.e. increasing AUC) more strongly in the RSC than for COHA due to a stronger vocabulary increase in scientific English.

H2: Difference between POS The vocabulary expansion affects the MST of nouns (i.e. increasing AUC) more than other POS, especially in scientific English.

H3: Effects of nominal style on shape of the MST curves Over time, we expect to find a weaker surprisal reduction with more bits of memory, especially in the RSC.

6.4 Data

6.4.1 Royal Society Corpus

We use two English diachronic corpora covering the time between 1750 and 2000. For scientific English we use the Royal Society Corpus (RSC) (Fischer et al., 2020a), consisting of the publications of the Royal Society of London with 47K documents and 300M tokens. We evaluate the evolution of the MST in 3 sub-samples, given that the RSC was split into subjournals around the 1900: (1) **RSC** encompasses all documents from 1665 to 1900, and from 1900 onward (2) **RSC-A** includes the Proceedings and Transactions of the Mathematical, Physical and Engineering Sciences, and (3) **RSC-B** containing publications of the Proceedings and Transactions of the Biological Sciences.

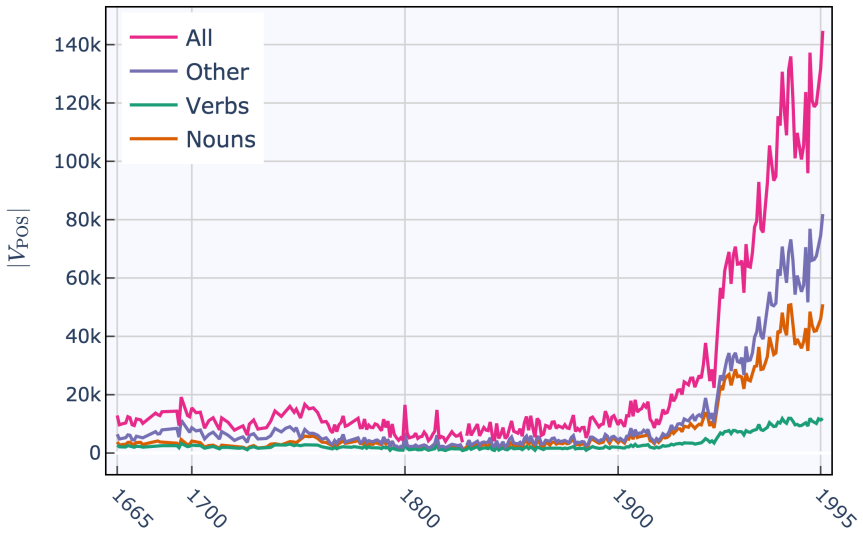


Figure 6.3: Increase of types per part-of-speech (POS) in the RSC over time; "Other" contains all POS except nouns and verbs.

Documents from a forth category containing, e.g., obituaries were excluded from the analysis.

6.4.2 Corpus of Historical American English

For general English, we use a reduced version (masked words) of the multi-genre, diachronic COHA corpus (Davies, 2021). The full COHA comprises over 475 million words spanning the 1820s to 2010s. To make the linguistic annotation comparable in both corpora, we parse and POS-tag the corpus with the Stanza software package (Qi et al., 2020), using the default English parser.

6.4.3 Corpus subsampling

We follow the diachronic language modeling approach introduced by Steuer et al. (2024b) by subsampling train sets of approximately identical size for each year in a corpus (see Table C.1 in Appendix C.1). For the tokenization methods not based on

subwords, we apply a post-processing step that reduces the number of vocabulary items to obtain approximately similar vocabulary sizes of ≈ 80.000. For each tokenization method, we choose a separate threshold frequency t_{REPL} that any token in the train set must exceed to be included in the tokenizer's vocabulary. We split the train set by white spaces and replace all words that occur only t_{REPL} times in the train set by an "unknown" token that corresponds to its POS tag as given by the UPOS column in the conllu file. Then, we replace all out-of-vocabulary (OOV) items in the validation and test sets in the same way.

6.5 Methods

6.5.1 Tokenization

We tested several tokenization methods for both corpora. These methods are described in detail in Appendix C.1. For the results in the main paper, we used a *lempos-based* tokenization: We first split the train corpus by whitespaces, and then replace each word with a concatenation of its lemma form as given by the UPOS tag as given by the respective columns of the conllu file. In case the absolute frequency of a word did not exceed the threshold value t_{REPL} it is replaced by its UPOS tag. We then replace OOV items in the validation and test set in the same way. The final tokenizer (used for all models trained on that corpus) is trained on the concatenated train sets of each corpus. This dampens the effect of the exponential increase of noun types in the RSC, and allows a closed, word-based vocabulary sampled equally from all years of the corpus.

6.5.2 Language models

For each tokenization method, we use Hugging Face transformers (Wolf et al., 2020a) to train the base version of the OPT architecture (S. Zhang et al., 2022a) on each subset of the training corpus (i.e., the train set of pretainging to a single year in either RSC or

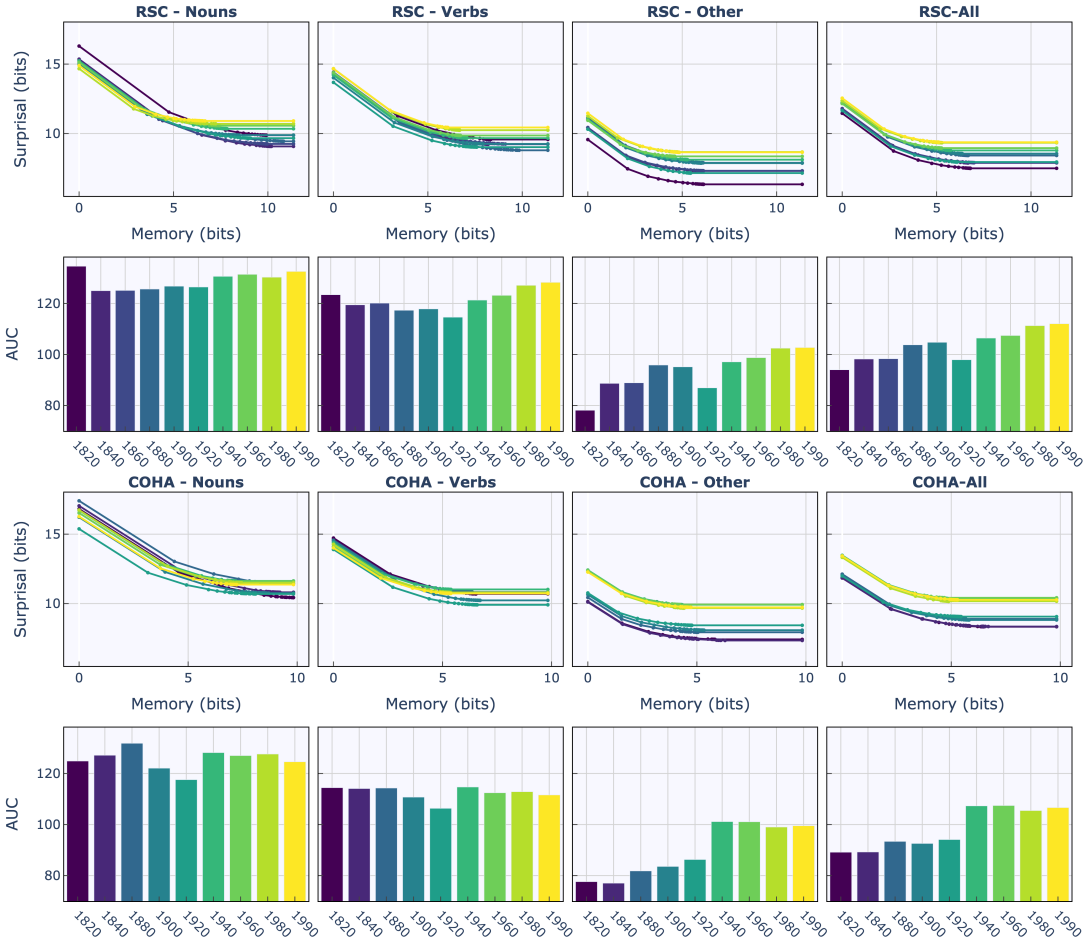


Figure 6.4: Memory-surprisal trade-off curves for 10 years from the RSC. Surprisal was averaged over documents and cross-validation folds. Each dot on a curve corresponds to a surprisal - memory pair, starting with unigram surprisal and no memory. All curves where extended to the maximal amount of memory available.

COHA) for 10 epochs with a batch size of 256, a learning rate of 5×10^{-5} and a linear learning rate warmup over 50% of training steps. Word-level models were trained with a context window of 32, and the BPE model with a context window of 64. Training was done on a cluster of 8 Nvidia A100 GPUs with 40GB of memory and took about 2 hours per model. We then used the LMs to estimate surprisal values on all documents from each test year of the two corpora. The code

6.5.3 Surprisal estimation

For each context size T ranging from $T = 0$ (unigram surprisal) to $T_{Max} = 20$, we estimate average surprisal $\hat{S}_T$ on a document D of $|D|$ words following Hahn et al. (2021):

$$\hat{S}_T = \frac{1}{|D| - T} \sum_{t=T}^{|D|} -\log_2 p(w_t | w_{t-T}, ..., w_{t-1}) \tag{6.1}$$

We estimate $p(w_t|w_{t-T}, ..., w_{t-1})$ directly from a transformer model averaging $\hat{S}_T$ on the documents from a single year over 5 models trained on different cross-validation splits as described in Section 6.4. Since the model may overfit for larger values of T due to data sparsity, we stop estimating $\hat{S}_T$ if $\hat{S}_T > \hat{S}_{T-1}$ and substitute $\hat{S}_{T-1}$ for $\hat{S}_T$. Since we want to compare the MST of different POS tags, we calculate $\hat{S}_T$ for a a given set of POS tags $P = \{p_1, ..., p_{|P|}\}$ and a subset of words $D_P \subseteq D$ as:

$$\hat{S}^P_T = \frac{1}{|D_P| - T} \sum_{t=T}^{|D_P|} -\log_2 p(w_t | w_{t-T}, ..., w_{t-1}) \tag{6.2}$$

6.5.4 AUC calculation

We then use surprisal estimates $\hat{S}^P_T$ to calculate mutual information I^P_T for each context size T as $I^P_T = \hat{S}^P_{T-1} - \hat{S}^P_T$, and memories M^P_T as $\sum_{t=0}^{T} t I^P_T$. We chose the following POS tag sets: Nouns (UPOS = "NOUN"), verbs (UPOS = "VERB") and other (all other POS). After estimating $\hat{S}^P_T$s and I^P_T, we calculate the area under the memory-surprisal trade-off curve (AUC) by applying the trapezoidal rule using the corresponding function of the scikit-learn Python package (Pedregosa et al., 2011).

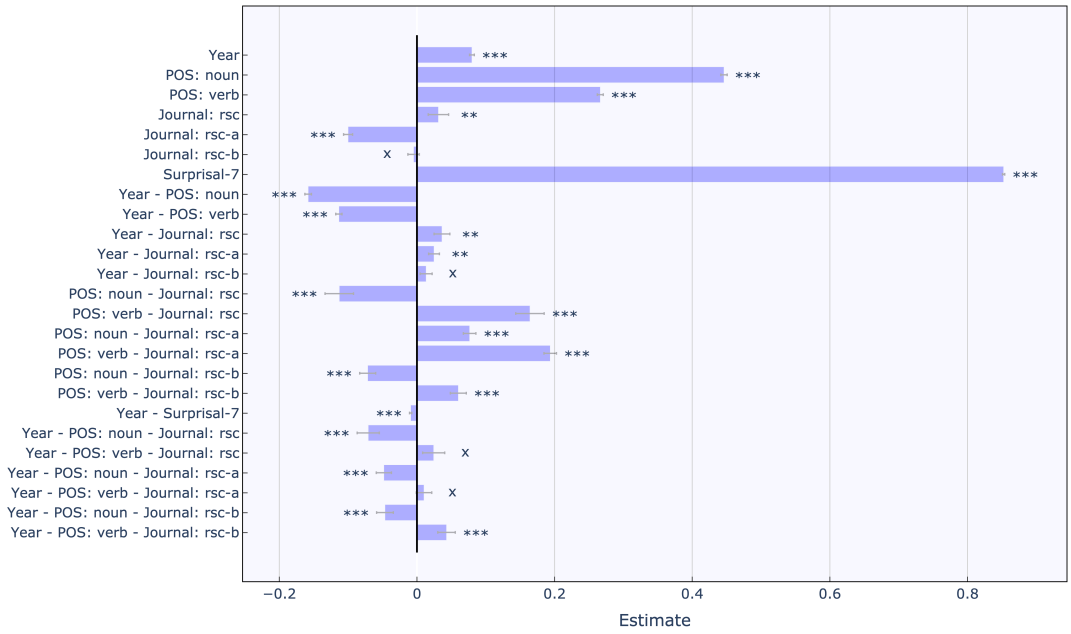


Figure 6.5: Effects of part of speech, journal and time on AUC for the period from 1820 to 1996. Reference levels for factor variables are "other" (POS) and "coha" (journal). Significance levels: '***' $p < 0.001$, '**' $p < 0.01$, '*' $p < 0.05$, 'x' $p >= 0.05$. Error bars show standard error of the coefficient estimate.

6.5.5 Statistical modeling

To assess the temporal development of the MST in the two corpora, we fit linear mixed-effects models (LMEs) via the lmerTest R package with AUC as response variable and average per-document 7-gram surprisal (surprisal estimated from the transformer model with 6 words in the context), journal, period, and POS as dependent variables. As we calculate the AUC for each document in the corpus, we include the document ID as a random effect nested in the corpus variable. We fit a separate LME for each tokenization method. We used the following formula to fit all regression models:

$$\text{lmer}(\text{auc} \sim \text{year} * \text{pos} * \text{journal} + \text{surprisal-7} * \text{year} + (1|\text{corpus}/\text{doc\_id}), \text{data} = .)$$

We normalized auc, year and surprisal-7 to the interval $[0, 1]$. We chose "coha" (that is, the whole COHA corpus) as the base level of the journal variable, and "other" (not noun or verb) as the base level of the POS variable.

6.6 Analysis

6.6.1 Effect of surprisal

Overall, the observed effects are in line with our expectations. We find the strongest effect for 7-gram surprisal (Estimate: 0.8525; confidence interval (CI): 0.8492, 0.8557; $t = 522.77$). A positive estimate corresponds to an increase in AUC, i.e., a worse MST, while a negative estimate corresponds to a decrease in AUC and a better MST compared to the base level of the variable. Given the fact that AUC is correlated with the surprisal values at different memory budgets, the strong association between the predictor and response is plausible. Figure 6.5 is a detailed overview of all effects of interest, besides surprisal. We also see main effects of POS, with both nouns (Estimate: 0.44; CI: 0.43, 0.45; $t = 96, 17$) and verbs (Estimate: 0.27, CI: 0.26,0.28; $t = 61.47$) having on average higher AUCs than other POS, which is in line with their generally higher surprisal.

6.6.2 Effect of RSC subjournals

Comparing the language of the three subjournals of the RSC to COHA, we find a mixed picture. AUC is higher for the early RSC (up until 1900), though this effect is small (Estimate: 0.03; CI: 0.002,0.06; $t = 2.11$), while we find no significant effect for RSC-B. For RSC-A we find a large negative effect (Estimate: -0.1, CI: -0.11, -0.09; $t = -15.41$), showing that the language of this subjournal is optimized w.r.t. the MST compared to general English.

6.6.3 Effect of time

AUC increases gradually over time (main effect of the "Year" variable; Estimate: 0.0798, CI: 0.0731, 0.0866; $t = 23.26$). We find significant interactions of time and POS, with

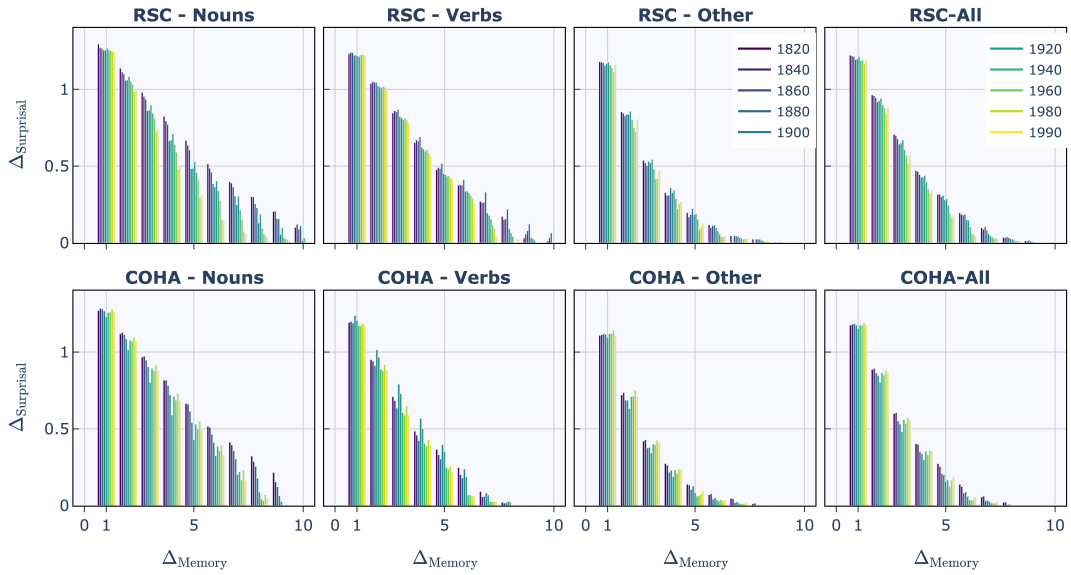


Figure 6.6: Average slope of the MST curve at equidistant memory intervals of 1 bit.

both nouns (Estimate: -0.15; CI: -0.17, -0.14; $t = -32.94$) and verbs (Estimate: -0.11; CI: -0.12, -0.1; $t = -24.96$) showing a markedly slower increase in AUC than other POS. This effect is stronger in the RSC than in COHA, with triple interactions between time, POS and journal indicating a slower increase for nouns compared to other POS in RSC (Estimate: -0.7; CI: -0.10, -0.04, $t = -4,41$), RSC-A (Estimate: -0.04; CI: -0.07, -0.03; $t = -4.37$), and RSC-B (Estimate: -0.05; CI: -0.07, -0.03; $t = -3.82$).

6.6.4 Effect of POS

Apart from the main effect of POS, we also find significant interactions of POS and journal: Verbs generally have a higher AUC than other POS in RSC (Estimate: 0.16; CI: 0.12, 0.2; $t = 7.95$), RSC-A (Estimate: 0.19, CI: 0.17, 0.21; $t = 21.33$), and RSC-B (Estimate: 0.06; CI: 0.04, 0.08; $t = 5.19$) compared to COHA, while nouns are overall associated with lower AUC. This is in line with our findings for the interaction of time, POS and journal: Not only do nouns in the RSC generally have a lower AUC (see Section 6.6.1), but the increase in AUC over time is not as large as may be expected based on the overall increase. Thus, while the number of nominal vocabulary items

increases drastically over time in the RSC, the syntax of scientific English is still in some sense optimized w.r.t. the MST for nouns and verbs.

6.7 From AUC to Shape of the MST curves

6.7.1 Effect of nominal style

In the previous section, we have analyzed the overall development of optimality in the two corpora as measured by the AUC. However, the AUC is only an approximation of optimality, given that MST curves whose AUC is compared are parallel in time. Furthermore, even when two curves do not cross, the degree to which more bits of memory reduce surprisal is not covered by the AUC. We will therefore analyze in more detail the individual shapes of the MST curves as well as the surprisal reduction rate per every additional bit of memory.

Looking at Figure 6.4, we see that the MST curves show different shapes in different years. Especially for nouns in the RSC, MST curves show an interesting picture: While surprisal in the first 100 years (1820 - 1920) continuously drops per additional bit of memory, in the last 60 years, surprisal shows very little reduction with less than 5 bits of memory. A similar trend can be observed for verbs in the RSC and other POS, however, not as pronounced as for nouns. At the same time, nouns show a decreasing unigram surprisal, which is surprising given the fact that the number of nominal vocabulary items increases over time. It shows, however, that in the case of nouns, the increase in AUC over time is not owed to increasing surprisal but instead to the decreasing surprisal reduction per bit of information held in memory. A meta-interpretation of this would be that increasingly dense structures as typical for nominal style lead to a decreasing information gain through additional memory, or in other words: If information is packed in dense constructions, only locally placed information helps reduce surprisal of the next word, while with less dense constructions, longer context windows are beneficial for prediction of the next word.

6.7.2 Effect of NP density

This interpretation is backed by the calculation of the average slope of the memory-surprisal curves at equidistant memory intervals of one bit (see Figure 6.6). For nouns, less and less information is gained (or surprisal reduced) per additional bit of memory for each step of 20 years. Compared to verbs and other POS, this is especially pronounced. Comparing RSC and COHA, the slope of the MST curves levels out faster for COHA than for the RSC, i.e., the LMs trained on the RSC data can make use of more bits of memory. This difference may be a result of generally longer sentence lengths in scientific English than in general English. Comparing POS, in both corpora, the temporal effect is strongest for nouns and especially pronounced in the RSC. For verbs, the slope is fairly similar across time in the RSC, indicating that there has been less change in predictive contexts of verbs than for nouns. This is plausible given that most changes in scientific English are known to have affected the structure of noun phrases, which have become increasingly dense over time.

6.8 Conclusion

We examined the communicative efficiency of scientific vs. general English over time, as measured by the Memory-Surprisal Tradeoff (MST). Our central question was whether MST optimality has changed diachronically and, if so, whether such changes vary across registers. This inquiry was motivated by the well-documented shift in English toward nominal rather than verbal style, manifested in complex, informationally dense noun phrases. While the Efficient Tradeoff Hypothesis predicts that more optimal orderings with respect to locality should yield more efficient MSTs, our findings indicate the opposite: denser encodings and vocabulary expansion over time appear to reduce optimality. Specifically, we identified two key factors: growing vocabulary size leads to higher average lexical surprisal, and less predictive contexts result in less efficient memory usage. The observed trends in AUC values suggest a general decline

in optimality over time. However, this interpretation must be qualified, as vocabulary growth is an inherent feature of language evolution. Although our subsampling strategy was designed to mitigate the influence of vocabulary size on surprisal estimates, the overall trend persists. This raises important questions regarding the comparability of surprisal values across historical stages. To address this, we also analyzed the average slope of the MST curves, capturing the information gained per bit of memory independently of absolute surprisal levels. This analysis revealed that for short memory contexts (1 – 3 bits), the tradeoff remains relatively stable over time, suggesting that efficiency has declined primarily for longer contexts. We interpret this as evidence that English has become less optimal in terms of long-range predictability, consistent with a broader shift toward shorter, denser encodings.

6.9 Limitations & future work

There are several limitations to our study. First, our analysis of scientific English distinguishes between three journals within the RSC (RSC, Journal A and Journal B). It is important to note that these journals reflect both different disciplines (biology and mathematics) and represent different time periods (Journals A and B are only published from 1900 onward, RSC contains all earlier publications). A more detailed analysis of the three journals could reveal variation among scientific disciplines. Furthermore, more fine-grained distinctions by topic, author, or subfield could give additional insights into how efficiency varies along these lines. Second, our comparison contrasts these journals with the entirety of COHA, rather than with more carefully matched subsets of general English. A more nuanced comparison might better isolate register-specific effects. Third, we did not investigate which specific documents or genres are driving the observed increase in surprisal over time, nor did we examine which texts could be considered particularly (non-)optimal w.r.t. the MST. Addressing these points in future work would provide a more detailed understanding of the interaction between register, vocabulary growth, and communicative efficiency. Finally, although Scientific English may appear less optimal, in our surprisal models, we have not accounted for

the factors of specialization and background knowledge. This is because our modeling is based on the entire corpus, which may mask discipline-specific effects. These effects could become apparent if we were to model each discipline separately. Additionally, psycholinguistic studies on expert text processing would be necessary to draw more definitive conclusions.

7

Conclusion

Contents

Here follows a short summary of my research objectives set out in Chapter 1.2. I will first discuss the contributions to each of the research objectives, and close with some general points on the application of LMs to linguistic research questions.

7.1 Cognitive modeling on small scales

Running counter to the general finding in the world of (large) LMs that an increase in model and data size improves performance in textual downstream tasks, this trend generally does not hold for PPP, and particularly and perhaps surprisingly also not in case of a developmentally plausible pre-training corpus as in Chapter 3. In the light of the findings of Oh et al. (2024), I suspect that the effect of low-frequency tokens is exacerbated by the small pre-training corpus and the multi-epoch training regime, a hypothesis that should be tested in future work. I take this to mean that smaller,

syntactically informed models should be used to shed light on the actual processing mechanism in the human brain, which indeed has become more frequent in recent years (Oh et al., 2022; Timkey and Linzen, 2023).

7.2 Using multilingual LLMs for typological research

In our study on cross-lingual pronoun surprisal in Chapter 4, we were able to recover properties of personal pronouns in 14 languages by correlating LM surprisal with the pronouns' properties. This shows that multilingual models can be useful for typologists, especially when surprisal on sequences from low-resource languages is required. In contrast to the study in Chapter 3, this displayed some of the benefits of scaling model and data size. However, the translated nature of our multilingual corpus and the predominance of English in the training data of the LM is a potential double confound: Since most of the translations have English as their source language, it is conceivable that in combination with the predominantly English LM some of the syntactic features of languages other than English are skewed into this direction. This could be evaluated in two ways: Firstly by including English in our analysis and extending our comparison of for multi- and monolingual models, and secondly by inspecting the parameter space of the multilingual LM for shared representations between languages.

7.3 Diachronic language modeling reveals optimization processes

In Chapter 5, I established a framework for diachronic language modeling that accounts for the increase of training data and lexical items over time, and makes surprisal values comparable. In Chapter 6, I then compared the MST of texts from different time periods of the RSC based on surprisal values estimated in the way described in Chapter 5. On the surface, the language in the RSC becomes *less* optimal over time (after controlling

for the increase in surprisal). A closer look on different POS reveals that the decrease in "optimality" is slowed down by changes involving noun phrases, pointing to ongoing optimization processes that go beyond the MST. However, the assumption of an average reader is too simplistic and needs to be complemented by LMs trained on subsets of the corpus that represent different scientific audiences, similar to Skrjanec et al. (2023).

7.4 The importance of choosing the right LM

To sum up, for any linguistic research question that can be tackled with LMs, it is important to know the requirements of the task/the system one wants to model. This requires a conscious choice of pre-training data, model size and architecture in order to obtain good surprisal estimates. Sometimes one can get surprisingly far with big off-the-shelf models, as with mGPT in the typological study in Chapter 4, underscoring the utility of multilingual models for research in low-resource languages. If one is interested in comparing LM performance on syntax (or any other task) to humans, one should probably train their own model on an appropriate corpus (e.g., BabyLM), ideally with some kind of syntactic bias (Linzen and Baroni, 2021; Mueller and Linzen, 2023), a memory capacity closer to humans (Kuribayashi et al., 2022; Timkey and Linzen, 2023) or careful manipulations of the training data that enforce a LM to generalize on the syntactical level (Leong and Linzen, 2024). Finally, this thesis showed that settings like modeling diachronic changes indeed require careful experimental designs and strict control of the training data and tokenization. As computational linguists, our main interest should be in language, and we should not mistake a LM for a model of language based on its perceived competence in language tasks (or rather, tasks that humans usually solve with language). Instead we should choose the LM as represented by a combination of architecture, training data and training method according to our hypothesis, and then use the LM to test that hypothesis. This is not to say that the reverse procedure is invalid: Its limitations lie in the statistical nature of the learning algorithm and the (purposefully) general architecture of transformers. Properties of language discovered in this way are necessarily confined to what is recoverable from

language data via statistical learning, and we should keep in mind that language data, for all its richness, is not the mirror image of the human language faculty.

List of Figures

List of Tables

List of Acronyms

1PL first person plural 80

1SG first person singular 80

2PL second person plural 80

2SG second person singular 46, 52, 78, 80

3PL third person plural 59, 80

3SG third person singular 52, 56, 59, 73, 80

ACC accusative case 73

AUC area under the curve 48, 57, 76, 80, 107–109, 114–118, 127

CC clause-chain marker 73

CI confidence interval 65–68, 70–72, 115, 116, 126

COND conditional mood 73

DAT dative case 73

DL dependency length 98, 99, 101, 102, 104

ENCL clitic form 56

F feminine gender 52, 59, 80

LM language model 1–9, 11, 14–17, 19–25, 28, 32, 33, 35–41, 78, 82–85, 121–123, 125

LMM linear mixed model 12, 13, 25, 28–31, 33, 65–71, 82, 126, 127, 188, 189

M masculine gender 56, 73, 79

MST memory-surprisal trade-off 103–109, 113–119, 122, 123, 127

OBJ grammatical object 56

OOV out-of-vocabulary 111, 187, 188

PPP psychometric predictive power 3, 12, 14, 17, 25, 38, 121

PST past tense 73, 79, 80

RC relative clause 96

RSC Royal Society Corpus 90, 91, 93, 98–100, 102–104, 107–112, 115–119, 122, 126, 127

SG singular number 73, 79, 80

UD Universal Dependencies 47, 54, 55, 59, 60, 62

UID uniform information density 12, 17

Bibliography

Aina, Laura, Xixian Liao, Gemma Boleda, and Matthijs Westera (2021). "Does referent predictability affect the choice of referential form? A computational approach using masked coreference resolution". en. In: *Proceedings of the 25th Conference on Computational Natural Language Learning*. Online: Association for Computational Linguistics, pp. 454–469. DOI: 10.18653/v1/2021.conll-1.36. URL: https://aclanthology.org/2021.conll-1.36 (visited on 06/10/2025) (cit. on p. 44).

Alabi, Jesujoba, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow (Oct. 2022). "Adapting Pre-trained Language Models to African Languages via Multilingual Adaptive Fine-Tuning". In: *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, pp. 4336–4349. URL: https://aclanthology.org/2022.coling-1.382 (cit. on p. 36).

Alabi, Jesujoba, Marius Mosbach, Matan Eyal, Dietrich Klakow, and Mor Geva (Aug. 2024). "The Hidden Space of Transformer Language Adapters". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, pp. 6588–6607. DOI: 10.18653/v1/2024.acl-long.356. URL: https://aclanthology.org/2024.acl-long.356/ (cit. on p. 83).

Annepaka, Yadagiri and Partha Pakray (Mar. 2025). "Large language models: a survey of their development, capabilities, and applications". en. In: *Knowledge and Information Systems* 67.3, pp. 2967–3022. ISSN: 0219-3116. DOI: 10.1007/s10115-024-02310-4. URL: https://doi.org/10.1007/s10115-024-02310-4 (visited on 08/21/2025) (cit. on p. 2).

Aoyama, Tatsuya and Ethan Gotlieb Wilcox (July 2025). "Language Models Grow Less Humanlike beyond Phase Transition". In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar. Vienna, Austria: Association for Computational Linguistics, pp. 24938–24958. ISBN: 979-8-89176-251-0. DOI: 10.18653/v1/2025.acl-long.1214. URL: https://aclanthology.org/2025.acl-long.1214/ (cit. on p. 14).

Arehalli, Suhas, Brian Dillon, and Tal Linzen (Dec. 2022). "Syntactic Surprisal From Neural Models Predicts, But Underestimates, Human Processing Difficulty From Syntactic Ambiguities". In: *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, pp. 301–313. URL: https://aclanthology.org/2022.conll-1.20 (visited on 02/13/2023) (cit. on pp. 11, 13, 14, 22, 24, 38, 47).

Arnold, Jennifer E. and Sandra A. Zerkle (Oct. 2019). "Why do people produce pronouns? Pragmatic selection vs. rational models". en. In: *Language, Cognition and Neuroscience* 34.9, pp. 1152–1175. ISSN: 2327-3798, 2327-3801. DOI: 10.1080/23273798.2019.1636103. URL: https://www.tandfonline.com/doi/full/10.1080/23273798.2019.1636103 (visited on 06/10/2025) (cit. on pp. 42–44, 64, 77).

Badawi, El-Said M. (2016). *Modern written Arabic: a comprehensive grammar*. eng. Second Revision edition. Routledge comprehensive grammars. London New York: Routledge. ISBN: 978-0-415-66748-7 978-1-315-85615-5 978-1-317-92982-6 978-1-317-92983-3 978-1-317-92984-0. DOI: 10.4324/9781315856155 (cit. on p. 49).

Bamman, David, Olivia Lewke, and Anya Mansoor (May 2020). "An Annotated Dataset of Coreference in English Literature". eng. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Marseille, France:

European Language Resources Association, pp. 44–54. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.6/ (visited on 06/10/2025) (cit. on p. 57).

Banks, David (2003). "The evolution of grammatical metaphor in scientific writing". In: *Amsterdam Studies in the Theory and History of Linguistic Science Series 4*, pp. 127–148. DOI: 10.1075/cilt.236.07ban (cit. on p. 105).

Batsuren, Khuyagbaatar, Ekaterina Vylomova, Verna Dankers, Tsetsuukhei Delgerbaatar, Omri Uzan, Yuval Pinter, and Gábor Bella (Apr. 2024). *Evaluating Subword Tokenization: Alien Subword Composition and OOV Generalization Challenge*. arXiv:2404.13292. DOI: 10.48550/arXiv.2404.13292. URL: http://arxiv.org/abs/2404.13292 (visited on 11/06/2024) (cit. on p. 83).

Beinborn, Lisa and Yuval Pinter (Dec. 2023). "Analyzing Cognitive Plausibility of Subword Tokenization". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 4478–4486. DOI: 10.18653/v1/2023.emnlp-main.272. URL: https://aclanthology.org/2023.emnlp-main.272 (visited on 12/04/2024) (cit. on p. 83).

Bender, Emily M. and Alexander Koller (July 2020). "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 5185–5198. DOI: 10.18653/v1/2020.acl-main.463. URL: https://aclanthology.org/2020.acl-main.463 (cit. on p. 16).

Berdicevskis, Aleksandrs, Karsten Schmidtke-Bode, and Ilja Serant (2020). "Subjects tend to be coded only once: Corpus-based and grammar-based evidence for an efficiency-driven trade-off". en. In: *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*. Düsseldorf, Germany: Association for Computational Linguistics, pp. 79–92. DOI: 10.18653/v1/2020.tlt-1.8. URL:

https://www.aclweb.org/anthology/2020.tlt-1.8 (visited on 06/10/2025) (cit. on p. 79).

Biber, Douglas and Victoria Clark (2002). "Historical shifts in modification patterns with complex noun phrase structures". In: *English Historical Morphology. Selected Papers from 11 ICEHL, Santiago de Compostela, 7 – 11 September 2000* 11. Ed. by Maria López—Couso Teresa Fanego and Javier Perez—Guerra, pp. 43–66. DOI: 10.1075/cilt.223.06bib (cit. on p. 105).

Biber, Douglas and Bethany Gray (2011). "Grammatical change in the noun phrase: The influence of written language use". In: *English Language and Linguistics* 15.2, pp. 223–250. ISSN: 1360-6743, 1469-4379. DOI: 10.1017/S1360674311000025. URL: https://www.cambridge.org/core/product/identifier/S1360674311000025/type/journal_article (visited on 10/06/2022) (cit. on pp. 87, 105).

— (2016). *Grammatical Complexity in Academic English: Linguistic Change in Writing*. Studies in English Language. Cambridge: Cambridge University Press. ISBN: 978-1-107-00926-4. DOI: 10.1017/CBO9780511920776. URL: https://www.cambridge.org/core/books/grammatical-complexity-in-academic-english/0986B2B0F882B68A7427909C6A8C428D (visited on 10/21/2021) (cit. on pp. 86, 87, 93).

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan (1999). *Longman grammar of spoken and written English*. Longman (cit. on p. 87).

Biderman, Stella, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal (2023). "Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling". In: *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*. Ed. by Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett. Vol. 202. Proceedings of Machine Learning Research. PMLR, pp. 2397–

2430. URL: https://proceedings.mlr.press/v202/biderman23a.html (cit. on p. 8).

Bisk, Yonatan, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian (Nov. 2020). "Experience Grounds Language". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 8718–8735. DOI: 10.18653/v1/2020.emnlp-main.703. URL: https://aclanthology.org/2020.emnlp-main.703 (cit. on p. 16).

Bizzoni, Yuri, Stefania Degaetano-Ortlieb, Peter Fankhauser, and Elke Teich (2020). "Linguistic Variation and Change in 250 years of English Scientific Writing: A Data-driven Approach". In: *Frontiers in Artificial Intelligence, section Language and Computation*. Ed. by David Jurgens. DOI: https://doi.org/https://doi.org/10.3389/frai.2020.00073. URL: https://www.frontiersin.org/articles/10.3389/frai.2020.00073/full (cit. on p. 106).

Bürkner, Paul-Christian (2017). "**brms** : An *R* Package for Bayesian Multilevel Models Using *Stan*". en. In: *Journal of Statistical Software* 80.1. ISSN: 1548-7660. DOI: 10.18637/jss.v080.i01. URL: http://www.jstatsoft.org/v80/i01/ (visited on 06/10/2025) (cit. on p. 61).

Bybee, Joan (2010). *Language, Usage and Cognition*. Cambridge: Cambridge University Press. ISBN: 978-0-521-85140-4. DOI: 10.1017/CBO9780511750526. URL: https://www.cambridge.org/core/books/language-usage-and-cognition/46BF7213957AF53492A7B03A9BCE9DA0 (visited on 08/27/2025) (cit. on p. 87).

Camacho, José (2013). *Null subjects*. eng. Cambridge studies in linguistics. Cambridge: Cambridge University Press. ISBN: 978-1-139-52440-7 (cit. on p. 45).

Camaj, Martin and Leonard Fox (1984). *Albanian grammar: with exercises, chrestomathy and glossaries*. eng. Wiesbaden: Harrassowitz. ISBN: 978-3-447-02467-9 (cit. on pp. 49, 73, 80).

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell (2017). "*Stan* : A Probabilistic Programming Language". en. In: *Journal of Statistical Software* 76.1. ISSN: 1548-7660. DOI: 10.18637/jss.v076.i01. URL: http://www.jstatsoft.org/v76/i01/ (visited on 06/10/2025) (cit. on p. 61).

Chao, Yuen Ren (Jan. 1956). "Chinese Terms of Address". In: *Language* 32.1, p. 217. ISSN: 00978507. DOI: 10.2307/410666. URL: https://www.jstor.org/stable/410666?origin=crossref (visited on 06/10/2025) (cit. on p. 52).

Chen, Angelica, Ravid Schwartz-Ziv, Kyunghyun Cho, Matthew Leavitt, and Naomi Saphra (2024). "Sudden Drops in the Loss: Syntax Acquisition, Phase Transitions, and Simplicity Bias in MLMs". In: *International Conference on Learning Representations (ICLR)*. Spotlighted. URL: https://arxiv.org/abs/2309.07311 (cit. on p. 14).

Chomsky, Noam (2023). "The False Promise of ChatGPT". In: *The New York Times*. URL: https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html (cit. on pp. 2, 15).

Cover, TM and Joy A Thomas (2006). *Elements of information theory*. Hoboken, NJ: Wiley-Interscience. DOI: 10.1002/047174882X (cit. on p. 106).

Cysouw, Michael (1997). *3rd Person Personal Pronoun: A Universal Category?* URL: http://www.cysouw.de/home/manuscripts_files/cysouwTIN97.pdf (visited on 03/21/2025) (cit. on pp. 42, 43).

— (2009). *The Paradigmatic Structure of Person Marking*. eng. Oxford Studies in Typology and Linguistic Theory Ser. Oxford: Oxford University Press. ISBN: 978-0-19-955426-3 978-0-19-156499-4 (cit. on p. 42).

Davies, Mark (2021). *Corpus of Historical American English (COHA)*. Version V12. DOI: 10.25346/S6/6I8JL1. URL: https://doi.org/10.25346/S6/6I8JL1 (cit. on p. 110).

De Hoop, Helen and Sammie Tarenskeen (Oct. 2015). "It's all about you in Dutch". en. In: *Journal of Pragmatics* 88, pp. 163–175. ISSN: 03782166. DOI: 10.1016/j.pragma.2014.07.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S0378216614001362 (visited on 06/10/2025) (cit. on p. 78).

De Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre, and Daniel Zeman (May 2021). "Universal Dependencies". en. In: *Computational Linguistics*, pp. 1–54. ISSN: 0891-2017, 1530-9312. DOI: 10.1162/coli_a_00402. URL: https://direct.mit.edu/coli/article/doi/10.1162/coli_a_00402/98516/Universal-Dependencies (visited on 06/10/2025) (cit. on pp. 47, 54).

Degaetano-Ortlieb, Stefania, Hannah Kermes, Ashraf Khamis, and Elke Teich (Dec. 2018). *An Information-Theoretic Approach to Modeling Diachronic Change in Scientific English*. en. Pages: 258-281 Section: From Data to Evidence in English Language Research. Brill. ISBN: 978-90-04-39065-2. DOI: 10.1163/9789004390652_012. URL: https://brill.com/display/book/edcoll/9789004390652/BP000014.xml (visited on 12/20/2022) (cit. on p. 90).

— (2019). "An Information-Theoretic Approach to Modeling Diachronic Change in Scientific English". In: *From Data to Evidence in English Language Research*. Ed. by Carla Suhr, Terttu Nevalainen, and Irma Taavitsainen. Language and Computers. Brill, pp. 258–281. DOI: 10.1163/9789004390652 (cit. on p. 106).

Degaetano-Ortlieb, Stefania and Elke Teich (2016). "Information-based modeling of diachronic linguistic change: from typicality to productivity". In: *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pp. 165–173 (cit. on p. 88).

— (Aug. 2018). "Using relative entropy for detection and analysis of periods of diachronic linguistic change". In: *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Santa Fe, New Mexico: Association for Computational Linguistics, pp. 22–33. URL: https://aclanthology.org/W18-4503 (visited on 12/20/2022) (cit. on pp. 88, 90).

Degaetano-Ortlieb, Stefania and Elke Teich (Feb. 2019). "Toward an optimal code for communication: The case of scientific English". en. In: *Corpus Linguistics and Linguistic Theory* 18.1, pp. 175–207. ISSN: 1613-7027, 1613-7035. DOI: 10.1515/cllt-2018-0088. URL: https://www.degruyter.com/document/doi/10.1515/cllt-2018-0088/html (visited on 12/20/2022) (cit. on pp. 86, 88, 90, 93, 95).

— (Feb. 2022). "Toward an optimal code for communication: The case of scientific English". en. In: *Corpus Linguistics and Linguistic Theory* 18.1, pp. 175–207. ISSN: 1613-7027, 1613-7035. DOI: 10.1515/cllt-2018-0088. URL: https://www.degruyter.com/document/doi/10.1515/cllt-2018-0088/html (visited on 12/20/2022) (cit. on p. 106).

Demberg, Vera and Frank Keller (Nov. 2008a). "Data from eye-tracking corpora as evidence for theories of syntactic processing complexity". en. In: *Cognition* 109.2, pp. 193–210. ISSN: 00100277. DOI: 10.1016/j.cognition.2008.07.008. URL: https://linkinghub.elsevier.com/retrieve/pii/S0010027708001741 (visited on 06/10/2025) (cit. on p. 43).

— (2008b). "Data from eye-tracking corpora as evidence for theories of syntactic processing complexity". In: *Cognition* 109.2, pp. 193–210. ISSN: 00100277. DOI: 10.1016/j.cognition.2008.07.008. URL: https://linkinghub.elsevier.com/retrieve/pii/S0010027708001741 (visited on 12/06/2022) (cit. on p. 88).

Dentella, Vittoria, Fritz Günther, and Evelina Leivada (Dec. 2023). "Systematic testing of three Language Models reveals low language accuracy, absence of response stability, and a yes-response bias". en. In: *Proceedings of the National Academy of Sciences* 120.51, e2309583120. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2309583120. URL: https://pnas.org/doi/10.1073/pnas.2309583120 (visited on 07/04/2025) (cit. on p. 15).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume*

1 (Long and Short Papers). Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423 (cit. on p. 7).

Donaldson, Bruce C. (2008). *Dutch: a comprehensive grammar*. eng dut. 2nd ed. Routledge comprehensive grammars. Abingdon, Ox New York, N.Y: Routledge. ISBN: 978-0-415-43231-3 978-0-415-43230-6 978-0-203-89532-0 (cit. on pp. 49, 80).

Dryer, Matthew S. (2013a). "Expression of Pronominal Subjects (v2020.4)". In: *The World Atlas of Language Structures Online*. Ed. by Matthew S. Dryer and Martin Haspelmath. Type: Data set. Zenodo. DOI: 10.5281/zenodo.13950591. URL: https://doi.org/10.5281/zenodo.13950591 (cit. on pp. 43, 45, 51, 62, 63, 74, 75).

— (2013b). "Order of Subject and Verb (v2020.4)". In: *The World Atlas of Language Structures Online*. Ed. by Matthew S. Dryer and Martin Haspelmath. Type: Data set. Zenodo. DOI: 10.5281/zenodo.13950591. URL: https://doi.org/10.5281/zenodo.13950591 (cit. on p. 51).

— (2013c). "Order of Subject, Object and Verb (v2020.4)". In: *The World Atlas of Language Structures Online*. Ed. by Matthew S. Dryer and Martin Haspelmath. Type: Data set. Zenodo. DOI: 10.5281/zenodo.13950591. URL: https://doi.org/10.5281/zenodo.13950591 (cit. on p. 51).

Dryer, Matthew S. and Orin D. Gensler (2013). "Order of Object, Oblique, and Verb (v2020.4)". In: *The World Atlas of Language Structures Online*. Ed. by Matthew S. Dryer and Martin Haspelmath. Type: Data set. Zenodo. DOI: 10.5281/zenodo.13950591. URL: https://doi.org/10.5281/zenodo.13950591 (cit. on p. 51).

Dubossarsky, Haim, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg (July 2019). "Time-Out: Temporal Referencing for Robust Modeling of Lexical Semantic Change". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational

Linguistics, pp. 457–470. DOI: 10.18653/v1/P19-1044. URL: https://aclanthology.org/P19-1044 (cit. on p. 86).

Dum-Tragut, Jasmine (2009). *Armenian: modern Eastern Armenian*. eng. London Oriental and African language library volume 14. Amsterdam: John Benjamins. ISBN: 978-90-272-3814-6 978-90-272-8879-0 (cit. on pp. 50, 80).

Eldan, Ronen and Yuanzhi Li (2023). "TinyStories: How Small Can Language Models Be and Still Speak Coherent English?" In: DOI: 10.48550/ARXIV.2305.07759. URL: https://arxiv.org/abs/2305.07759 (visited on 05/21/2023) (cit. on pp. 25, 33).

Elman, Jeffrey L. (Mar. 1990). "Finding Structure in Time". en. In: *Cognitive Science* 14.2, pp. 179–211. ISSN: 0364-0213, 1551-6709. DOI: 10.1207/s15516709cog1402_1. URL: https://onlinelibrary.wiley.com/doi/10.1207/s15516709cog1402_1 (visited on 06/25/2025) (cit. on p. 2).

Fernandez Monsalve, Irene, Stefan L. Frank, and Gabriella Vigliocco (Apr. 2012). "Lexical surprisal as a general predictor of reading time". In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France: Association for Computational Linguistics, pp. 398–408. URL: https://aclanthology.org/E12-1041 (cit. on p. 20).

Fischer, Stefan, Jörg Knappen, Katrin Menzel, and Elke Teich (May 2020a). "The Royal Society Corpus 6.0: Providing 300+ Years of Scientific Writing for Humanistic Study". English. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 794–802. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.99 (cit. on pp. 85, 104, 109).

— (May 2020b). *The Royal Society Corpus 6.0: Providing 300+ Years of Scientific Writing for Humanistic Study*. English. Marseille, France. URL: https://aclanthology.org/2020.lrec-1.99 (cit. on p. 90).

Frank, Stefan L. (2013). "Uncertainty Reduction as a Measure of Cognitive Load in Sentence Comprehension". In: *Topics in Cognitive Science* 5.3, pp. 475–494. ISSN: 1756-8765. (Visited on 10/20/2023) (cit. on pp. 43, 101).

Frank, Stefan L. and Rens Bod (June 2011). "Insensitivity of the human sentence-processing system to hierarchical structure". eng. In: *Psychological Science* 22.6, pp. 829–834. ISSN: 1467-9280. DOI: 10.1177/0956797611409589 (cit. on p. 12).

Futrell, Richard, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven T. Piantadosi, and Evelina Fedorenko (2021). "The Natural Stories corpus: A Reading-Time Corpus of English Texts Containing Rare Yyntactic Constructions". In: *Language Resources and Evaluation* 55.1, pp. 63–77. URL: https://aclanthology.org/L18-1012/ (cit. on pp. 21, 28, 31, 125).

Futrell, Richard and Kyle Mahowald (May 2025). *How Linguistics Learned to Stop Worrying and Love the Language Models*. arXiv:2501.17047 [cs]. DOI: 10.48550/arXiv.2501.17047. URL: http://arxiv.org/abs/2501.17047 (visited on 07/04/2025) (cit. on pp. 2, 15, 40).

Futrell, Richard, Kyle Mahowald, and Edward Gibson (2015). "Large-scale evidence of dependency length minimization in 37 languages". In: *Proceedings of the National Academy of Sciences* 112.33, pp. 10336–10341. DOI: 10.1073/pnas.1502134112. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1502134112. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1502134112 (cit. on p. 106).

Gao, Leo, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou (Sept. 2021). *A framework for few-shot language model evaluation*. Version v0.0.1. DOI: 10.5281/zenodo.5371628. URL: https://doi.org/10.5281/zenodo.5371628 (cit. on p. 28).

Gardelle, Laure and Sandrine Sorlin, eds. (2015). *The pragmatics of personal pronouns*. eng. Studies in language companion series 171. Amsterdam Philadelphia: John Benjamins Publishing Company. ISBN: 978-90-272-5936-3 (cit. on p. 42).

Gerdes, Kim, Sylvain Kahane, and Xinying Chen (2019). "Rediscovering Greenberg' s Word Order Universals in UD". en. In: *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*. Paris, France: Association for Computational Linguistics, pp. 124–131. DOI: 10.18653/v1/W19-8015. URL: https://www.aclweb.org/anthology/W19-8015 (visited on 06/10/2025) (cit. on p. 48).

Gibson, Edward (Aug. 1998). "Linguistic complexity: locality of syntactic dependencies". In: *Cognition* 68.1, pp. 1–76. ISSN: 0010-0277. DOI: 10.1016/S0010-0277(98)00034-1. URL: https://www.sciencedirect.com/science/article/pii/S0010027798000341 (visited on 08/27/2025) (cit. on p. 106).

— (2000). "The dependency locality theory: a distance-based theory of linguistic complexity." In: *Image, language, brain: Papers from the first Mind Articulation Project Symposium*. Ed. by Yasushi Miyashita Alec Marantz and Wayne O' Neil. Cambridge, MA: MIT Press, pp. 95–126. DOI: 10.7551/mitpress/3654.003.0008 (cit. on pp. 104, 106).

Gilkerson, Jill, Jeffrey A. Richards, Steven F. Warren, Judith K. Montgomery, Charles R. Greenwood, D. Kimbrough Oller, John H. L. Hansen, and Terrance D. Paul (May 2017). "Mapping the Early Language Environment Using All-Day Recordings and Automated Analysis". en. In: *American Journal of Speech-Language Pathology* 26.2, pp. 248–265. ISSN: 1058-0360, 1558-9110. DOI: 10.1044/2016_AJSLP-15-0169. URL: http://pubs.asha.org/doi/10.1044/2016_AJSLP-15-0169 (visited on 07/31/2023) (cit. on p. 22).

Giulianelli, Mario, Marco Del Tredici, and Raquel Fernández (2020). "Analysing Lexical Semantic Change with Contextualised Word Representations". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Lin-*

guistics. Association for Computational Linguistics, pp. 446–457. URL: https://pure.uva.nl/ws/files/63250055/2020.acl_main.365.pdf (cit. on p. 86).

Goodkind, Adam and Klinton Bicknell (2018). "Predictive power of word surprisal for reading times is a linear function of language model quality". en. In: *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*. Salt Lake City, Utah: Association for Computational Linguistics, pp. 10–18. DOI: 10.18653/v1/W18-0102. URL: http://aclweb.org/anthology/W18-0102 (visited on 09/01/2023) (cit. on p. 13).

Grattafiori, Aaron, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla,

Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie

Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko,

Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma (2024). *The Llama 3 Herd of Models*. arXiv: 2407.21783 [cs.AI]. URL: https://arxiv.org/abs/2407.21783 (cit. on p. 8).

Greenberg, Joseph H. (1963). "Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements". In: *Universals of Human Language*. Ed. by Joseph H. Greenberg. Cambridge, Mass: MIT Press, pp. 73–113 (cit. on pp. 42, 43).

Greenberg, Joseph Harold, ed. (1966). *Universals of language: report of a conference held at Dobbs Ferry, New York, April 13-15, 1961*. eng. 2. ed., 7. pr. The M.I.T. Press paperback series 37. Meeting Name: International Conference on Language Universals. Cambridge, Mass.: M.I.T Pr. ISBN: 978-0-262-57008-4 (cit. on p. 46).

Gulordava, Kristina and Paola Merlo (Aug. 2015). "Diachronic Trends in Word Order Freedom and Dependency Length in Dependency-Annotated Corpora of Latin and Ancient Greek". In: *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*. Ed. by Joakim Nivre and Eva Hajiová. Uppsala, Sweden: Uppsala University, Uppsala, Sweden, pp. 121–130. URL: https://aclanthology.org/W15-2115/ (visited on 08/27/2025) (cit. on p. 106).

Guo, Mandy, Zihang Dai, Denny Vrandei, and Rami Al-Rfou (May 2020). "Wiki-40B: Multilingual Language Model Dataset". English. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Marseille, France: European Language Resources Association, pp. 2440–2452. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.297 (cit. on pp. 41, 58).

Hahn, Michael, Judith Degen, and Richard Futrell (July 2021). "Modeling word and morpheme order in natural language as an efficient trade-off of memory and surprisal." en. In: *Psychological Review* 128.4, pp. 726–756. ISSN: 1939-1471, 0033-295X. DOI: 10.1037/rev0000269. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/rev0000269 (visited on 11/07/2022) (cit. on pp. 7, 102, 104, 106, 113).

Hahn, Michael and Richard Futrell (2020). "Crosslinguistic Word Orders Enable an Efficient Tradeoff of Memory and Surprisal". In: *Society for Computation in Linguistics* 3.1 (cit. on p. 106).

Hahn, Michael and Frank Keller (Nov. 2016). "Modeling Human Reading with Neural Attention". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 85–95. DOI: 10.18653/v1/D16-1009. URL: https://aclanthology.org/D16-1009 (cit. on p. 37).

Hale, John (2001a). "A Probabilistic Earley Parser as a Psycholinguistic Model". In: *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*. NAACL '01. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 1–8 (cit. on pp. 2, 86).

— (2001b). "A Probabilistic Earley Parser as a Psycholinguistic Model". In: *Second Meeting of the North American Chapter of the Association for Computational Linguistics*. NAACL 2001. URL: https://aclanthology.org/N01-1021 (visited on 11/30/2022) (cit. on pp. 10, 43).

— (2006). "Uncertainty About the Rest of the Sentence". In: *Cognitive Science* 30.4, pp. 643–672. ISSN: 1551-6709. DOI: 10.1207/s15516709cog0000_64. URL: https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog0000_64 (visited on 12/16/2021) (cit. on pp. 100, 101).

Halliday, M.A.K. (1988). "On the language of physical science". In: *Registers of written English: Situational factors and linguistic features*. Ed. by Mohsen Ghadessy. tex.date-added: 2010-03-09 11:19:11 +0100 tex.date-modified: 2010-03-30 15:16:26 +0200. London: Pinter, pp. 162–177 (cit. on pp. 87, 88, 105).

Halliday, M.A.K. and James R. Martin (1993). *Writing science: Literacy and discursive power*. London: Falmer Press (cit. on pp. 87, 105).

Hamilton, William L., J. Leskovec, and Dan Jurafsky (2016). "Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. DOI: 10.18653/v1/D16-1229. URL: https://doi.org/10.18653/v1/d16-1229 (cit. on p. 86).

Harris Sabbettai, Zellig (1991). *A Theory of Language and Information: A Mathematical Approach*. Oxford, New York: Oxford University Press. ISBN: 978-0-19-824224-6 (cit. on p. 88).

Haspelmath, Martin (July 2023). "Defining the word". en. In: *WORD* 69.3, pp. 283–297. ISSN: 0043-7956, 2373-5112. DOI: 10.1080/00437956.2023.2237272. URL: https://www.tandfonline.com/doi/full/10.1080/00437956.2023.2237272 (visited on 06/10/2025) (cit. on p. 49).

Heilbron, Micha, Benedikt Ehinger, Peter Hagoort, and Floris de Lange (2019). "Tracking Naturalistic Linguistic Predictions with Deep Neural Language Models". In: *2019 Conference on Cognitive Computational Neuroscience*. Berlin, Germany: Cognitive Computational Neuroscience. DOI: 10.32470/CCN.2019.1096-0. URL: https://ccneuro.org/2019/Papers/ViewPapers.asp?PaperNum=1096 (visited on 02/07/2023) (cit. on p. 20).

Heine, Bernd and Kyung-An Song (Nov. 2011). "On the grammaticalization of personal pronouns". en. In: *Journal of Linguistics* 47.3, pp. 587–630. ISSN: 0022-2267, 1469-7742. DOI: 10.1017/S0022226711000016. URL: https://www.cambridge.org/core/product/identifier/S0022226711000016/type/journal_article (visited on 06/10/2025) (cit. on p. 42).

Helmbrecht, Johannes (2004). *Personal Pronouns - Form, Function and Grammaticalization*. Place: Erfurt Type: habil (cit. on pp. 44–46, 52).

— (2013). "Politeness Distinctions in Pronouns (v2020.4)". In: *The World Atlas of Language Structures Online*. Ed. by Matthew S. Dryer and Martin Haspelmath.

Type: Data set. Zenodo. DOI: 10.5281/zenodo.13950591. URL: https://doi.org/10.5281/zenodo.13950591 (cit. on pp. 46, 52, 77).

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan (June 2010). "The weirdest people in the world?" en. In: *Behavioral and Brain Sciences* 33.2-3, pp. 61–83. ISSN: 0140-525X, 1469-1825. DOI: 10.1017/S0140525X0999152X. URL: https://www.cambridge.org/core/product/identifier/S0140525X0999152X/type/journal_article (visited on 06/10/2025) (cit. on p. 43).

Heringer, Hans Jürgen, Bruno Strecker, and Rainer Wimmer (1980). *Syntax: Fragen, Lösungen, Alternativen*. Fink (cit. on p. 98).

Hochreiter, Sepp and Jürgen Schmidhuber (Nov. 1997). "Long Short-Term Memory". In: *Neural Computation* 9.8, pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. eprint: https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf. URL: https://doi.org/10.1162/neco.1997.9.8.1735 (cit. on p. 89).

Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre (2022). "An empirical analysis of compute-optimal large language model training". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., pp. 30016–30030. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/c1e2faff6f588870935f114ebe04a3e5-Paper-Conference.pdf (cit. on p. 89).

Hofmann, Valentin, Janet Pierrehumbert, and Hinrich Schütze (Aug. 2021). "Superbizarre Is Not Superb: Derivational Morphology Improves BERT's Interpretation of Complex Words". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Chengqing

Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Online: Association for Computational Linguistics, pp. 3594–3608. DOI: 10.18653/v1/2021.acl-long.279. URL: https://aclanthology.org/2021.acl-long.279 (visited on 11/26/2024) (cit. on p. 83).

Holton, David (2012). *Greek: a comprehensive grammar*. eng. Second edition. Routledge comprehensive grammars. Milton Park, Abingdon, Oxon New York, N.Y: Routledge. ISBN: 978-0-415-59202-4 978-0-203-80238-0 978-1-136-62633-3 978-1-136-62640-1. DOI: 10.4324/9780203802380 (cit. on pp. 49, 73, 74, 80).

Hosseini, Kasra, Kaspar Beelen, Giovanni Colavizza, and Mariona Coll Ardanuy (May 2021). *Neural Language Models for Nineteenth-Century English*. arXiv:2105.11321 [cs]. URL: http://arxiv.org/abs/2105.11321 (visited on 05/12/2023) (cit. on pp. 25, 89).

Houlsby, Neil, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly (2019). *Parameter-Efficient Transfer Learning for NLP*. arXiv: 1902.00751 [cs.LG] (cit. on p. 36).

Hu, Jennifer, Jon Gauthier, Peng Qian, Ethan Gotlieb Wilcox, and Roger P. Levy (2020). "A Systematic Assessment of Syntactic Generalization in Neural Language Models". In: Publisher: arXiv Version Number: 2. DOI: 10.48550/ARXIV.2005.03692. URL: https://arxiv.org/abs/2005.03692 (visited on 01/23/2023) (cit. on pp. 22, 25, 40).

Hu, Jennifer, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy (Sept. 2024). "Language models align with human judgments on key grammatical constructions". en. In: *Proceedings of the National Academy of Sciences* 121.36, e2400917121. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2400917121. URL: https://pnas.org/doi/10.1073/pnas.2400917121 (visited on 07/04/2025) (cit. on p. 15).

Hu, Michael Y., Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Leshem Choshen, Ryan Cotterell, Alex Warstadt, and Ethan Gotlieb

Wilcox, eds. (Nov. 2024). *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*. Miami, FL, USA: Association for Computational Linguistics. URL: https://aclanthology.org/2024.conll-babylm.0/ (visited on 08/11/2025) (cit. on p. 19).

Huang, Jiaxin, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han (Dec. 2023). "Large Language Models Can Self-Improve". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 1051–1068. DOI: 10.18653/v1/2023.emnlp-main.67. URL: https://aclanthology.org/2023.emnlp-main.67/ (visited on 06/10/2025) (cit. on pp. 11, 47).

Huang, Kuan-Jung, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen (Apr. 2023). *Surprisal does not explain syntactic disambiguation difficulty: evidence from a large-scale benchmark*. preprint. PsyArXiv. DOI: 10.31234/osf.io/z38u6. URL: https://osf.io/z38u6 (visited on 05/09/2023) (cit. on pp. 14, 22, 38).

Hudson, Richard (1995). "Measuring syntactic difficulty". In: *Manuscript, University College, London* (cit. on p. 98).

Hundt, Marianne, David Denison, and Gerold Schneider (2012). "Relative complexity in scientific discourse". In: *English Language and Linguistics* 16.2, pp. 209–240. ISSN: 1360-6743, 1469-4379. DOI: 10.1017/S1360674312000032. URL: https://www.cambridge.org/core/product/identifier/S1360674312000032/type/journal_article (visited on 10/05/2022) (cit. on pp. 87, 105).

Hussain, Md. and Ishtiak Mahmud (July 25, 2019). "pyMannKendall: a python package for non-parametric Mann Kendall family of trend tests." In: *Journal of Open Source Software* 4.39, p. 1556. ISSN: 2475-9066. DOI: 10.21105/joss.01556. URL: http://dx.doi.org/10.21105/joss.01556 (cit. on p. 94).

Irwin, Tovah, Kyra Wilson, and Alec Marantz (May 2023). "BERT Shows Garden Path Effects". In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 3220–3232. URL: https://aclanthology.org/2023.eacl-main.235 (visited on 05/08/2023) (cit. on pp. 22, 24).

Ismayilzada, Mete, Defne Circi, Jonne Sälevä, Hale Sirin, Abdullatif Köksal, Bhuwan Dhingra, Antoine Bosselut, Lonneke van der Plas, and Duygu Ataman (Oct. 2024). *Evaluating Morphological Compositional Generalization in Large Language Models*. arXiv:2410.12656 [cs]. URL: http://arxiv.org/abs/2410.12656 (visited on 10/30/2024) (cit. on p. 83).

Jaegle, Andrew, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira (June 2021). "Perceiver: General Perception with Iterative Attention". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 4651–4664. URL: https://proceedings.mlr.press/v139/jaegle21a.html (cit. on p. 40).

Jiang, Yan (2016). "Deixis and anaphora". In: *A Reference Grammar of Chinese* 中文参考□法. Ed. by Chu-Ren Huang and Dingxu Shi. Cambridge: Cambridge University Press, pp. 484–517. ISBN: 978-1-139-02846-2. DOI: 10.1017/CBO9781139028462.017. URL: https://www.cambridge.org/core/product/identifier/CBO9781139028462A117/type/book_part (visited on 06/10/2025) (cit. on p. 51).

Jing, Yingqi, Paul Widmer, and Balthasar Bickel (Dec. 2023). "Word order evolves at similar rates in main and subordinate clauses: Corpus-based evidence from Indo-European". en. In: *Diachronica* 40.4, pp. 532–556. ISSN: 0176-4225, 1569-9714. DOI: 10.1075/dia.20035.jin. URL: http://www.jbe-platform.com/content/journals/10.1075/dia.20035.jin (visited on 06/10/2025) (cit. on p. 48).

Jolly, Eshin (Nov. 2018). "Pymer4: Connecting R and Python for Linear Mixed Modeling". In: *Journal of Open Source Software* 3.31, p. 862. ISSN: 2475-9066. DOI:

10.21105/joss.00862. URL: http://joss.theoj.org/papers/10.21105/joss.00862 (visited on 07/29/2023) (cit. on p. 29).

Jurayj, William, William Rudman, and Carsten Eickhoff (Oct. 2022). *Garden-Path Traversal in GPT-2*. arXiv:2205.12302 [cs]. URL: http://arxiv.org/abs/2205.12302 (visited on 04/24/2023) (cit. on pp. 22, 24).

Juzek, Tom S, Marie-Pauline Krielke, and Elke Teich (Dec. 2020). "Exploring diachronic syntactic shifts with dependency length: the case of scientific English". In: *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*. Barcelona, Spain (Online): Association for Computational Linguistics, pp. 109–119. URL: https://aclanthology.org/2020.udw-1.13 (cit. on pp. 104, 106).

Kachru, Yamuna (Oct. 2006). *Hindi*. en. Vol. 12. London Oriental and African Language Library. Amsterdam: John Benjamins Publishing Company. ISBN: 978-90-272-3812-2 978-90-272-3821-4 978-90-272-9314-5. DOI: 10.1075/loall.12. URL: http://www.jbe-platform.com/content/books/9789027293145 (visited on 06/10/2025) (cit. on p. 80).

Kaiser, Stefan, Yasuko Ichikawa, Noriko Kobayashi, and Hilofumi Yamamoto (2013). *Japanese: a comprehensive grammar*. eng jpn. 2. edition. Routledge comprehensive grammars. London: Routledge. ISBN: 978-0-415-68737-9 978-0-415-68739-3 978-0-203-08519-6 (cit. on pp. 51, 60).

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei (Jan. 2020). *Scaling Laws for Neural Language Models*. arXiv:2001.08361 [cs, stat]. URL: http://arxiv.org/abs/2001.08361 (visited on 05/16/2023) (cit. on pp. 33, 85).

Katzir, Roni (Apr. 2023). *Why large language models are poor theories of human linguistic cognition. A reply to Piantadosi (2023)*. LingBuzz Published In: Ms., Tel Aviv University. URL: https://lingbuzz.net/lingbuzz/007190 (visited on 11/09/2023) (cit. on pp. 2, 15).

Kennedy, Alan and Joël Pynte (Jan. 2005). "Parafoveal-on-foveal effects in normal reading". en. In: *Vision Research* 45.2, pp. 153–168. ISSN: 00426989. DOI: 10.1016/j.visres.2004.07.037. URL: https://linkinghub.elsevier.com/retrieve/pii/S0042698904003979 (visited on 08/01/2023) (cit. on p. 37).

Kermes, Hannah, Stefania Degaetano-Ortlieb, Ashraf Khamis, Jörg Knappen, and Elke Teich (May 2016). *The Royal Society Corpus: From Uncharted Data to Corpus*. Portoro, Slovenia. URL: https://aclanthology.org/L16-1305 (cit. on p. 90).

Kermes, Hannah and Elke Teich (2017). "Average Surprisal of Parts-of-(S)peech". In: Birmingham, UK: Corpus Linguistics 2017. URL: https://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2017/general/paper207.pdf. published (cit. on p. 94).

Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov (June 2014). "Temporal Analysis of Language through Neural Language Models". In: *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. Baltimore, MD, USA: Association for Computational Linguistics, pp. 61–65. DOI: 10.3115/v1/W14-2517. URL: https://aclanthology.org/W14-2517 (cit. on p. 86).

Kitagawa, Chisato and Adrienne Lehrer (Oct. 1990). "Impersonal uses of personal pronouns". en. In: *Journal of Pragmatics* 14.5, pp. 739–759. ISSN: 03782166. DOI: 10.1016/0378-2166(90)90004-W. URL: https://linkinghub.elsevier.com/retrieve/pii/037821669090004W (visited on 06/10/2025) (cit. on p. 78).

Kodner, Jordan, Sarah Payne, and Jeffrey Heinz (2023). *Why Linguistics Will Thrive in the 21st Century: A Reply to Piantadosi (2023)*. arXiv: 2308.03228 [cs.CL]. URL: https://arxiv.org/abs/2308.03228 (cit. on p. 2).

Konnerth, Linda and Andrea Sansò (Aug. 2021). "Towards a diachronic typology of individual person markers". en. In: *Folia Linguistica* 55.s42-s1, pp. 1–24. ISSN: 0165-4004, 1614-7308. DOI: 10.1515/flin-2021-2012. URL: https://www.

degruyter.com/document/doi/10.1515/flin-2021-2012/html (visited on 06/10/2025) (cit. on p. 42).

Krielke, Marie-Pauline (2021). "Relativizers as markers of grammatical complexity: A diachronic, cross-register study of English and German". In: *Bergen Language and Linguistics Studies* 11.1, pp. 91–120. DOI: 10.15845/bells.v11i1.3440. URL: https://bells.uib.no/index.php/bells/article/view/3440 (cit. on pp. 90, 97).

— (2024). "Cross-linguistic Dependency Length Minimization in scientific language: Syntactic complexity reduction in English and German in the Late Modern period". In: *Languages in Contrast* 24.1, pp. 133–163. ISSN: 1387-6759, 1569-9897. DOI: 10.1075/lic.00038.kri. URL: https://www.jbe-platform.com/content/journals/10.1075/lic.00038.kri (visited on 02/20/2024) (cit. on pp. 104–106).

Krielke, Marie-Pauline, Luigi Talamo, Mahmoud Fawzi, and Jörg Knappen (June 2022). "Tracing Syntactic Change in the Scientific Genre: Two Universal Dependency-parsed Diachronic Corpora of Scientific English and German". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 4808–4816. URL: https://aclanthology.org/2022.lrec-1.514 (visited on 11/18/2022) (cit. on p. 104).

Kudo, Taku and John Richardson (Aug. 2018). *SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing*. arXiv:1808.06226 [cs]. DOI: 10.48550/arXiv.1808.06226. URL: http://arxiv.org/abs/1808.06226 (visited on 08/06/2025) (cit. on p. 1).

Kuribayashi, Tatsuki, Yohei Oseki, Ana Brassard, and Kentaro Inui (Dec. 2022). "Context Limitations Make Neural Language Models More Human-Like". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 10421–10436. URL: https://aclanthology.org/2022.emnlp-main.712 (cit. on pp. 14, 25, 39, 123).

Kuribayashi, Tatsuki, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui (Aug. 2021). “Lower Perplexity is Not Always Human-Like”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 5203–5217. DOI: 10.18653/v1/2021.acl-long.405. URL: https://aclanthology.org/2021.acl-long.405 (cit. on p. 25).

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H. B. Christensen (2017). “lmerTest Package: Tests in Linear Mixed Effects Models”. In: *Journal of Statistical Software* 82.13, pp. 1–26. DOI: 10.18637/jss.v082.i13 (cit. on pp. 12, 29).

Langacker, Ronald W. (Apr. 2007). “Constructing the meanings of personal pronouns”. en. In: *Aspects of Meaning Construction*. Ed. by Günter Radden, Klaus-Michael Köpcke, Thomas Berg, and Peter Siemund. Amsterdam: John Benjamins Publishing Company, pp. 171–187. ISBN: 978-90-272-3242-7 978-90-272-9255-1. DOI: 10.1075/z.136.12lan. URL: https://benjamins.com/catalog/z.136.12lan (visited on 06/10/2025) (cit. on p. 42).

Länsisalmi, Riikka (2001). “You and I in Japanese: What do "personal pronouns" do in Japanese discourse?” eng. In: *Finnish Journal of Linguistics*. -. Articles, pp. 121–149. ISSN: 2814-4376. - (cit. on p. 45).

Lascaratou, Chryssoula (Dec. 1998). “Basic characteristics of Modern Greek word order”. In: *Constituent Order in the Languages of Europe*. Ed. by Anna Siewierska. De Gruyter Mouton, pp. 151–172. ISBN: 978-3-11-015152-7. DOI: 10.1515/9783110812206.151. URL: https://www.degruyter.com/document/doi/10.1515/9783110812206.151/html (visited on 06/10/2025) (cit. on p. 74).

LeCun, Yann (1988). “A theoretical framework for back-propagation”. In: *Proceedings of the 1988 Connectionist Models Summer School, CMU, Pittsburg, PA*. Ed. by D. Touretzky, G. Hinton, and T. Sejnowski. Publisher: Morgan Kaufmann, pp. 21–28 (cit. on p. 10).

Lei, Lei and Ju Wen (May 2020). "Is dependency distance experiencing a process of minimization? A diachronic study based on *the State of the Union* addresses". In: *Lingua* 239, p. 102762. ISSN: 0024-3841. DOI: 10.1016/j.lingua.2019.102762. URL: https://www.sciencedirect.com/science/article/pii/S002438411930511X (visited on 08/27/2025) (cit. on p. 106).

Leivada, Evelina, Fritz Günther, and Vittoria Dentella (Sept. 2024). "Reply to Hu et al.: Applying different evaluation standards to humans vs. Large Language Models overestimates AI performance". en. In: *Proceedings of the National Academy of Sciences* 121.36, e2406752121. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2406752121. URL: https://pnas.org/doi/10.1073/pnas.2406752121 (visited on 07/04/2025) (cit. on p. 15).

Leong, Cara Su-Yi and Tal Linzen (June 2023). "Language Models Can Learn Exceptions to Syntactic Rules". In: *Proceedings of the Society for Computation in Linguistics 2023*. Ed. by Tim Hunter and Brandon Prickett. Amherst, MA: Association for Computational Linguistics, pp. 133–144. URL: https://aclanthology.org/2023.scil-1.11/ (cit. on p. 40).

— (2024). *Testing learning hypotheses using neural networks by manipulating learning data*. Version Number: 1. DOI: 10.48550/ARXIV.2407.04593. URL: https://arxiv.org/abs/2407.04593 (visited on 08/06/2025) (cit. on p. 123).

Levshina, Natalia (June 2017). "A Multivariate Study of T/V Forms in European Languages Based on a Parallel Corpus of Film Subtitles". In: *Research in Language* 15.2, pp. 153–172. ISSN: 1731-7533. DOI: 10.1515/rela-2017-0010. URL: https://czasopisma.uni.lodz.pl/research/article/view/2974 (visited on 06/10/2025) (cit. on pp. 46, 47).

— (May 2022a). "Corpus-based typology: applications, challenges and some solutions". en. In: *Linguistic Typology* 26.1, pp. 129–160. ISSN: 1430-0532, 1613-415X. DOI: 10.1515/lingty-2020-0118. URL: https://www.degruyter.com/document/doi/10.1515/lingty-2020-0118/html (visited on 06/10/2025) (cit. on pp. 48, 64).

— (Feb. 2022b). “Frequency, Informativity and Word Length: Insights from Typologically Diverse Corpora”. en. In: *Entropy* 24.2, p. 280. ISSN: 1099-4300. DOI: 10.3390/e24020280. URL: https://www.mdpi.com/1099-4300/24/2/280 (visited on 06/10/2025) (cit. on pp. 43, 75).

Levshina, Natalia, Savithry Namboodiripad, Marc Allassonnière-Tang, Mathew Kramer, Luigi Talamo, Annemarie Verkerk, Sasha Wilmoth, Gabriela Garrido Rodriguez, Timothy Michael Gupton, Evan Kidd, Zoey Liu, Chiara Naccarato, Rachel Nordlinger, Anastasia Panova, and Natalia Stoynova (July 2023). “Why we need a gradient approach to word order”. en. In: *Linguistics* 61.4, pp. 825–883. ISSN: 0024-3949, 1613-396X. DOI: 10.1515/ling-2021-0098. URL: https://www.degruyter.com/document/doi/10.1515/ling-2021-0098/html (visited on 06/10/2025) (cit. on p. 48).

Levy, Roger (2008). “Expectation-based syntactic comprehension”. In: *Cognition* 106.3, pp. 1126–1177 (cit. on pp. 2, 10, 43, 86, 88, 104, 106).

Lewis, Richard L. and Shravan Vasishth (2005). “An Activation-Based Model of Sentence Processing as Skilled Memory Retrieval”. In: *Cognitive Science* 29.3, pp. 375–419. DOI: 10.1207/s15516709cog0000_25. URL: https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog0000_25 (visited on 12/09/2021) (cit. on p. 106).

Li, Hang (July 2022). “Language models: past, present, and future”. en. In: *Communications of the ACM* 65.7, pp. 56–63. ISSN: 0001-0782, 1557-7317. DOI: 10.1145/3490443. URL: https://dl.acm.org/doi/10.1145/3490443 (visited on 08/06/2025) (cit. on p. 1).

Li, Jiaxuan and Richard Futrell (2023). “A decomposition of surprisal tracks the N400 and P600 brain potentials”. en. In: *Proceedings of the Annual Meeting of the Cognitive Science Society* 45.45. URL: https://escholarship.org/uc/item/75c569dr (visited on 07/31/2023) (cit. on pp. 20, 37).

Linzen, Tal and Marco Baroni (Jan. 2021). "Syntactic Structure from Deep Learning". en. In: *Annual Review of Linguistics* 7.1, pp. 195–212. ISSN: 2333-9683, 2333-9691. DOI: 10.1146/annurev-linguistics-032020-051035. URL: https://www.annualreviews.org/doi/10.1146/annurev-linguistics-032020-051035 (visited on 01/23/2023) (cit. on p. 123).

Liu, Haitao (2008). "Dependency Distance as a Metric of Language Comprehension Difficulty". In: *Journal of Cognitive Science* 9.2, pp. 159–191. DOI: 10.17791/JCS.2008.9.2.159. URL: https://doi.org/10.17791/JCS.2008.9.2.159 (visited on 08/27/2025) (cit. on p. 106).

Loshchilov, Ilya and Frank Hutter (2019). *Decoupled Weight Decay Regularization*. arXiv: 1711.05101 [cs.LG] (cit. on pp. 10, 27).

Louagie, Dana and Jean-Christophe Verstraete (June 2015). "Personal pronouns with determining functions in Australian languages". en. In: *Studies in Language* 39.1, pp. 159–198. ISSN: 0378-4177, 1569-9978. DOI: 10.1075/sl.39.1.06lou. URL: http://www.jbe-platform.com/content/journals/10.1075/sl.39.1.06lou (visited on 06/10/2025) (cit. on p. 42).

Luong, Hy V. (Jan. 1990). *Discursive Practices and Linguistic Meanings: The Vietnamese system of person reference*. en. Vol. 11. Pragmatics & Beyond New Series. Amsterdam: John Benjamins Publishing Company. ISBN: 978-90-272-5021-6 978-1-55619-277-7 978-90-272-8335-1. DOI: 10.1075/pbns.11. URL: http://www.jbe-platform.com/content/books/9789027283351 (visited on 06/10/2025) (cit. on p. 52).

Mahowald, Kyle, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko (2023). *Dissociating language and thought in large language models: a cognitive perspective*. arXiv: 2301.06627 [cs.CL] (cit. on pp. 16, 17, 19, 20, 35).

Maiden, Martin and Cecilia Robustelli (Feb. 2014). *A Reference Grammar of Modern Italian*. en. 0th ed. Routledge. ISBN: 978-1-4441-1678-6. DOI: 10.4324/978020378

3504. URL: https://www.taylorfrancis.com/books/9781444116786 (visited on 06/10/2025) (cit. on pp. 49, 73, 74, 80).

Masini, Francesca and Simone Mattiola (2019). "Come fare tipologia con categorie non tradizionali?" ita. In: *CLUB Working Papers on Linguistics*. Ed. by Gianollo and Mauri. Vol. 3. Bologna: AMS Acta, pp. 282–294. URL: https://cris.unibo.it/handle/11585/712260?mode=simple (visited on 07/22/2025) (cit. on p. 48).

Meister, Clara, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy (Nov. 2021). "Revisiting the Uniform Information Density Hypothesis". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 963–980. DOI: 10.18653/v1/2021.emnlp-main.74. URL: https://aclanthology.org/2021.emnlp-main.74/ (visited on 05/26/2025) (cit. on pp. 11–13, 17).

Menzel, Katrin, Jörg Knappen, and Elke Teich (2021). "Generating Linguistically Relevant Metadata for the Royal Society Corpus". In: (cit. on p. 90).

Mosbach, Marius, Maksym Andriushchenko, and Dietrich Klakow (2021). *On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines*. arXiv: 2006.04884 [cs.LG] (cit. on p. 29).

Mosbach, Marius, Anna Khokhlova, Michael A. Hedderich, and Dietrich Klakow (Nov. 2020). "On the Interplay Between Fine-tuning and Sentence-Level Probing for Linguistic Knowledge in Pre-Trained Transformers". In: *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Ed. by Afra Alishahi, Yonatan Belinkov, Grzegorz Chrupaa, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad. Online: Association for Computational Linguistics, pp. 68–82. DOI: 10.18653/v1/2020.blackboxnlp-1.7. URL: https://aclanthology.org/2020.blackboxnlp-1.7/ (cit. on p. 39).

Mueller, Aaron and Tal Linzen (2023). "How to Plant Trees in Language Models: Data and Architectural Effects on the Emergence of Syntactic Inductive Biases". In: Publisher: arXiv Version Number: 1. DOI: 10.48550/ARXIV.2305.19905. URL: https://arxiv.org/abs/2305.19905 (visited on 06/29/2023) (cit. on pp. 2, 28, 40, 123).

Mühlhäusler, Peter (2001). "Personal pronouns." In: *Language typology and language universals* 1. Ed. by Martin Haspelmath, Ekkehard König, Wulf Oesterreicher, and Wolfgang Raible, pp. 741–747 (cit. on p. 42).

Nair, Sathvik 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 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 11251–11260. DOI: 10.18653/v1/2023.findings-emnlp.752. URL: https://aclanthology.org/2023.findings-emnlp.752/ (visited on 06/10/2025) (cit. on p. 83).

Naranjo, Matías Guzmán and Laura Becker (2018). "Quantitative word order typology with UD". en. In: *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*. Vol. 155. Linköping Electronic Conference Proceedings, pp. 91–104 (cit. on p. 47).

Nedoluzhko, Anna, Michal Novák, Martin Popel, Zdenk abokrtský, Amir Zeldes, and Daniel Zeman (June 2022). "CorefUD 1.0: Coreference Meets Universal Dependencies". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis. Marseille, France: European Language Resources Association, pp. 4859–4872. URL: https://aclanthology.org/2022.lrec-1.520/ (visited on 08/07/2025) (cit. on p. 52).

Oh, Byung-Doh, Christian Clark, and William Schuler (Mar. 2022). "Comparison of Structural Parsers and Neural Language Models as Surprisal Estimators". English.

In: *Frontiers in Artificial Intelligence* 5. Publisher: Frontiers. ISSN: 2624-8212. DOI: 10.3389/frai.2022.777963. URL: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2022.777963/full (visited on 06/20/2025) (cit. on pp. 13, 122).

Oh, Byung-Doh and William Schuler (2022). "Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times?" In: Publisher: arXiv Version Number: 1. DOI: 10.48550/ARXIV.2212.12131. URL: https://arxiv.org/abs/2212.12131 (visited on 01/11/2023) (cit. on pp. 22, 24, 25).

— (Apr. 2023a). *Transformer-Based LM Surprisal Predicts Human Reading Times Best with About Two Billion Training Tokens*. arXiv:2304.11389 [cs]. URL: http://arxiv.org/abs/2304.11389 (visited on 07/27/2023) (cit. on pp. 11, 14, 23, 26, 31, 47).

— (Mar. 2023b). "Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times?" In: *Transactions of the Association for Computational Linguistics* 11, pp. 336–350. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00548. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00548/2075940/tacl_a_00548.pdf. URL: https://doi.org/10.1162/tacl%5C_a%5C_00548 (cit. on p. 14).

Oh, Byung-Doh, Shisen Yue, and William Schuler (Mar. 2024). "Frequency Explains the Inverse Correlation of Large Language Models' Size, Training Data Amount, and Surprisal's Fit to Reading Times". In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Yvette Graham and Matthew Purver. St. Julian's, Malta: Association for Computational Linguistics, pp. 2644–2663. URL: https://aclanthology.org/2024.eacl-long.162/ (cit. on pp. 14, 39, 121).

Papadimitriou, Isabel, Richard Futrell, and Kyle Mahowald (Mar. 2022). *When classifying grammatical role, BERT doesn't care about word order... except when it*

matters. arXiv:2203.06204 [cs]. URL: http://arxiv.org/abs/2203.06204 (visited on 04/03/2023) (cit. on p. 24).

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf (cit. on p. 27).

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830 (cit. on p. 113).

Pfeiffer, Jonas, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe (2022a). "Lifting the Curse of Multilinguality by Pre-training Modular Transformers". en. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, pp. 3479–3495. DOI: 10.18653/v1/2022.naacl-main.255. URL: https://aclanthology.org/2022.naacl-main.255 (visited on 03/24/2023) (cit. on p. 36).

— (July 2022b). "Lifting the Curse of Multilinguality by Pre-training Modular Transformers". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, pp. 3479–3495.

DOI: 10.18653/v1/2022.naacl-main.255. URL: https://aclanthology.org/2022.naacl-main.255/ (cit. on p. 83).

Piantadosi, Steven (2023). "Modern language models refute Chomsky' s approach to language". In: *Lingbuzz Preprint, lingbuzz* 7180. URL: https://lingbuzz.net/lingbuzz/007180/current.pdf (cit. on pp. 2, 15, 40).

Press, Ofir, Noah A. Smith, and Mike Lewis (2021). "Shortformer: Better Language Modeling using Shorter Inputs". en. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 5493–5505. DOI: 10.18653/v1/2021.acl-long.427. URL: https://aclanthology.org/2021.acl-long.427 (visited on 05/30/2023) (cit. on p. 27).

— (Apr. 2022). *Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation*. arXiv:2108.12409 [cs]. DOI: 10.48550/arXiv.2108.12409. URL: http://arxiv.org/abs/2108.12409 (visited on 04/25/2023) (cit. on p. 27).

Pugh, Stefan M. and Ian Press (1999). *Ukrainian: a comprehensive grammar*. eng ukr. Routledge grammars. London New York: Routledge, Taylor & Francis Group. ISBN: 978-0-415-15030-9 978-0-203-21558-6. DOI: 10.4324/9780203215586 (cit. on pp. 79, 80).

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning (Apr. 2020). *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages*. arXiv:2003.07082 [cs]. DOI: 10.48550/arXiv.2003.07082. URL: http://arxiv.org/abs/2003.07082 (visited on 05/07/2025) (cit. on pp. 54, 90, 110).

R core team, test (2024). *R: A Language and Environment for Statistical Computing*. Vienna, Austria (cit. on p. 61).

Radford, Alec, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. (2018). *Improving language understanding by generative pre-training* (cit. on p. 89).

Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever (2019). "Language Models are Unsupervised Multitask Learners". In: (cit. on pp. 14, 24).

Roark, Brian, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier (Aug. 2009). "Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing". In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, pp. 324–333. URL: https://aclanthology.org/D09-1034 (cit. on p. 20).

Rounds, Carol (2001). *Hungarian: an essential grammar*. eng. 2nd ed. Routledge essential grammars. London New York: Routledge. ISBN: 978-0-415-77737-7 978-0-203-88619-9 (cit. on p. 80).

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams (Oct. 1986). "Learning representations by back-propagating errors". en. In: *Nature* 323.6088, pp. 533–536. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/323533a0. URL: http://www.nature.com/articles/323533a0 (visited on 01/25/2023) (cit. on p. 10).

Rust, Phillip, Jonas Pfeiffer, Ivan Vuli, Sebastian Ruder, and Iryna Gurevych (Aug. 2021). "How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Online: Association for Computational Linguistics, pp. 3118–3135. DOI: 10.18653/v1/2021.acl-long.243. URL: https://aclanthology.org/2021.acl-long.243/ (cit. on p. 83).

Sartran, Laurent, Samuel Barrett, Adhiguna Kuncoro, Milo Stanojevi, Phil Blunsom, and Chris Dyer (Dec. 2022). "Transformer Grammars: Augmenting Transformer Language Models with Syntactic Inductive Biases at Scale". en. In: *Transactions of the Association for Computational Linguistics* 10, pp. 1423–1439. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00526. URL: https://direct.mit.edu/tacl/article/

doi/10.1162/tacl_a_00526/114315/Transformer-Grammars-Augmenting-Transformer (visited on 04/29/2024) (cit. on p. 38).

Say, Sergey (Jan. 2014). "Bivalent Verb Classes in the Languages of Europe". de. In: Publisher: Brill. DOI: 10.1163/22105832-00401003. URL: https://brill.com/view/journals/ldc/4/1/article-p116_4.xml (visited on 06/10/2025) (cit. on p. 47).

Scaboro, Simone, Beatrice Portelli, Emmanuele Chersoni, Enrico Santus, and Giuseppe Serra (Nov. 2021). "NADE: A Benchmark for Robust Adverse Drug Events Extraction in Face of Negations". In: *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*. Online: Association for Computational Linguistics, pp. 230–237. DOI: 10.18653/v1/2021.wnut-1.26. URL: https://aclanthology.org/2021.wnut-1.26 (cit. on p. 89).

Scao, Teven Le, Thomas Wang, Daniel Hesslow, Lucile Saulnier, Stas Bekman, M. Saiful Bari, Stella Biderman, Hady Elsahar, Niklas Muennighoff, Jason Phang, Ofir Press, Colin Raffel, Victor Sanh, Sheng Shen, Lintang Sutawika, Jaesung Tae, Zheng Xin Yong, Julien Launay, and Iz Beltagy (Nov. 2022). *What Language Model to Train if You Have One Million GPU Hours?* arXiv:2210.15424 [cs]. URL: http://arxiv.org/abs/2210.15424 (visited on 04/25/2023) (cit. on p. 85).

Schmid, Helmut (1994). "Probabilistic part-of-speech tagging using decision trees". In: *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK (cit. on pp. 87, 90).

Schrimpf, Martin, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko (Nov. 2021). "The neural architecture of language: Integrative modeling converges on predictive processing". en. In: *Proceedings of the National Academy of Sciences* 118.45, e2105646118. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2105646118. URL: https://pnas.org/doi/full/10.1073/pnas.2105646118 (visited on 07/28/2023) (cit. on p. 20).

Shain, Cory, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Philip Levy (Nov. 2022). *Large-Scale Evidence for Logarithmic Effects of Word Predictability on Reading Time*. preprint. PsyArXiv. DOI: 10.31234/osf.io/4hyna. URL: https://osf.io/4hyna (visited on 04/24/2023) (cit. on pp. 11, 13, 21).

Shannon, C. E. (July 1948). "A Mathematical Theory of Communication". en. In: *Bell System Technical Journal* 27.3, pp. 379–423. ISSN: 00058580. DOI: 10.1002/j.1538-7305.1948.tb01338.x. URL: https://ieeexplore.ieee.org/document/6773024 (visited on 12/08/2021) (cit. on pp. 86, 88).

Shliazhko, Oleh, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina (2022). "mGPT: Few-Shot Learners Go Multilingual". In: Publisher: arXiv Version Number: 1. DOI: 10.48550/ARXIV.2204.07580. URL: https://arxiv.org/abs/2204.07580 (visited on 07/13/2023) (cit. on pp. 8, 41).

Siewierska, Anna (2012). *Person*. eng. Cambridge textbooks in linguistics. Cambridge: Cambridge University Press. ISBN: 978-0-511-81272-9. DOI: 10.1017/CBO9780511812729 (cit. on pp. 43, 49, 50, 78).

Skrjanec, Iza, Frederik Y. Broy, and Vera Demberg (2023). *Expert-adapted language models improve the fit to reading times*. DOI: psyarxiv.com/dc8y6. URL: https://psyarxiv.com/dc8y6 (cit. on pp. 36, 123).

Smith, Nathaniel J. and Roger Levy (Sept. 2013). "The effect of word predictability on reading time is logarithmic". en. In: *Cognition* 128.3, pp. 302–319. ISSN: 00100277. DOI: 10.1016/j.cognition.2013.02.013. URL: https://linkinghub.elsevier.com/retrieve/pii/S0010027713000413 (visited on 06/10/2025) (cit. on p. 43).

Sneddon, James N. and Karl Alexander Adelaar (2010). *Indonesian: a comprehensive grammar*. eng ind. 2. ed. Routledge comprehensive grammars. London New York: Roultedge. ISBN: 978-0-415-58154-7 (cit. on pp. 49, 51).

Stenson, Nancy (2019). *Modern Irish: a comprehensive grammar*. eng. Routledge comprehensive grammars. London New York: Taylor & Francis. ISBN: 978-1-138-23652-3 978-1-138-23651-6 (cit. on pp. 49–51, 73, 80).

Steuer, Julius, Marie-Pauline Krielke, Stefania Degaetano-Ortlieb, Elke Teich, and Dietrich Klakow (under review). "Modeling the Memory-Surpisal Trade-Off over Time: Communicative Efficiency Decreases with Lexico-Grammatical Change in Scientific English". In: *Submitted to ACL Rolling Review* (cit. on pp. 4, 103).

Steuer, Julius, Marie-Pauline Krielke, Stefan Fischer, Stefania Degaetano-Ortlieb, Marius Mosbach, and Dietrich Klakow (May 2024a). "Modeling Diachronic Change in English Scientific Writing over 300+ Years with Transformer-based Language Model Surprisal". In: *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC-COLING 2024*. Ed. by Pierre Zweigenbaum, Reinhard Rapp, and Serge Sharoff. Torino, Italia: ELRA and ICCL, pp. 12–23. URL: https://aclanthology.org/2024.bucc-1.2 (cit. on p. 4).

— (2024b). "Modeling Diachronic Change in English Scientific Writing over 300+ Years with Transformer-based Language Model Surprisal". In: *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC)@ LREC-COLING 2024*, pp. 12–23. URL: https://aclanthology.org/2024.bucc-1.2/ (cit. on pp. 106, 110).

Steuer, Julius, Marius Mosbach, and Dietrich Klakow (Dec. 2023). "Large GPT-like Models are Bad Babies: A Closer Look at the Relationship between Linguistic Competence and Psycholinguistic Measures". In: *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*. Ed. by Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. Singapore: Association for Computational Linguistics, pp. 142–157. DOI: 10.18653/v1/2023.conll-babylm.12. URL: https://aclanthology.org/2023.conll-babylm.12 (cit. on pp. 4, 17, 19).

Talamo, Luigi, Julius Steuer, and Annemarie Verkerk (2025a). "They saw it, onu, 它, coming: An information theoretic study of cross-linguistic variation in personal pronouns," in: *Linguistics*, to appear (cit. on pp. 4, 42).

Talamo, Luigi and Annemarie Verkerk (2022). "A new methodology for an old problem: A corpus-based typology of adnominal word order in European languages". en. In: *Italian Journal of Linguistics* 2.34, pp. 171–226 (cit. on pp. 48, 51, 74).

Talamo, Luigi, Annemarie Verkerk, and Iker Salaberri (2025b). "A quantitative approach to clause type and syntactic change in two Indo-European corpora". en. In: *Italian Journal of Linguistics* 36.2, pp. 53–82. ISSN: 2499-8117, 2499-8125. DOI: 10.26346/1120-2726-225. URL: https://doi.org/10.26346/1120-2726-225 (visited on 06/10/2025) (cit. on p. 48).

Tao, Siyu, Lucia Donatelli, and Michael Hahn (Aug. 2024). "More frequent verbs are associated with more diverse valency frames: Efficient principles at the lexicon-grammar interface". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, pp. 11795–11810. DOI: 10.18653/v1/2024.acl-long.635. URL: https://aclanthology.org/2024.acl-long.635/ (visited on 06/10/2025) (cit. on p. 48).

Teich, Elke, Peter Fankhauser, Stefania Degaetano-Ortlieb, and Yuri Bizzoni (2021). "Less is More/More Diverse: On The Communicative Utility of Linguistic Conventionalization". In: *Frontiers in Communication* 5, p. 142. ISSN: 2297-900X. DOI: 10.3389/fcomm.2020.620275. URL: https://www.frontiersin.org/article/10.3389/fcomm.2020.620275 (visited on 10/29/2021) (cit. on pp. 88, 106).

Timkey, William and Tal Linzen (2023). "A Language Model with Limited Memory Capacity Captures Interference in Human Sentence Processing". en. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics, pp. 8705–8720. DOI: 10.18653/v1/2023.findings-

emnlp.582. URL: https://aclanthology.org/2023.findings-emnlp.582 (visited on 04/29/2024) (cit. on pp. 2, 122, 123).

Trask, Robert Lawrence (Dec. 2003). "The Noun Phrase: nouns, determiners and modifiers; pronouns and names". In: *A Grammar of Basque*. Ed. by José Ignacio Hualde and Jon Ortiz De Urbina. De Gruyter Mouton, pp. 113–170. ISBN: 978-3-11-089528-5. DOI: 10.1515/9783110895285.113. URL: https://www.degruyter.com/document/doi/10.1515/9783110895285.113/html (visited on 06/10/2025) (cit. on p. 50).

Varda, Andrea de and Marco Marelli (July 2023). "Scaling in Cognitive Modelling: a Multilingual Approach to Human Reading Times". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 139–149. DOI: 10.18653/v1/2023.acl-short.14. URL: https://aclanthology.org/2023.acl-short.14 (cit. on p. 37).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (cit. on pp. 1, 2, 7).

Verkerk, Annemarie and Luigi Talamo (May 2024). "mini-CIEP+ : A Shareable Parallel Corpus of Prose". In: *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC-COLING 2024*. Ed. by Pierre Zweigenbaum, Reinhard Rapp, and Serge Sharoff. Torino, Italia: ELRA and ICCL, pp. 135–143. URL: https://aclanthology.org/2024.bucc-1.15/ (visited on 06/10/2025) (cit. on pp. 41, 47, 54).

Wal, Oskar van der, Pietro Lesci, Max Muller-Eberstein, Naomi Saphra, Hailey Schoelkopf, Willem Zuidema, and Stella Biderman (Mar. 2025). *PolyPythias: Stability and Outliers across Fifty Language Model Pre-Training Runs*. arXiv:2503.09543

[cs]. DOI: 10.48550/arXiv.2503.09543. URL: http://arxiv.org/abs/2503.09543 (visited on 04/10/2025) (cit. on p. 85).

Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman (Feb. 2019). *GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding*. arXiv:1804.07461 [cs]. URL: http://arxiv.org/abs/1804.07461 (visited on 05/22/2023) (cit. on p. 20).

Wang, Gui, Hui Wang, Xinyi Sun, Nan Wang, and Li Wang (2023). "Linguistic complexity in scientific writing: A large-scale diachronic study from 1821 to 1920". In: *Scientometrics* 128.1, pp. 441–460. ISSN: 1588-2861. DOI: 10.1007/s11192-022-04550-z. URL: https://doi.org/10.1007/s11192-022-04550-z (visited on 07/21/2023) (cit. on p. 105).

Warstadt, Alex, Aaron Mueller, Leshem Chohen, Ethan Gotlieb Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Adina Williams, Bhargavi Paranjabe, Tal Linzen, and Ryan Cotterell (Dec. 2023a). "Findings of the 2023 BabyLM Challenge: Sample-efficient pretraining on developmentally plausible corpora". In: *Proceedings of the 2023 BabyLM Challenge*. Association for Computational Linguistics (ACL) (cit. on pp. 15, 22).

Warstadt, Alex, Aaron Mueller, Leshem Choshen, Ethan Gotlieb Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell, eds. (Dec. 2023b). *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*. Singapore: Association for Computational Linguistics. URL: https://aclanthology.org/2023.conll-babylm.0/ (visited on 08/11/2025) (cit. on p. 19).

Warstadt, Alex, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman (Dec. 2020a). "BLiMP: The Benchmark of Linguistic Minimal Pairs for English". en. In: *Transactions of the Association for Computational Linguistics* 8, pp. 377–392. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00321. URL: https://direct.mit.edu/tacl/article/96452 (visited on 01/23/2023) (cit. on pp. 14, 15, 22).

Warstadt, Alex, Amanpreet Singh, and Samuel R. Bowman (2019). *Neural Network Acceptability Judgments*. arXiv: 1805.12471 [cs.CL] (cit. on pp. 16, 38).

Warstadt, Alex, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman (2020b). "Learning Which Features Matter: RoBERTa Acquires a Preference for Linguistic Generalizations (Eventually)". en. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 217–235. DOI: 10.18653/v1/2020.emnlp-main.16. URL: https://www.aclweb.org/anthology/2020.emnlp-main.16 (visited on 06/26/2023) (cit. on p. 22).

Weissweiler, Leonie, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze (Oct. 2023). "Explaining pretrained language models' understanding of linguistic structures using construction grammar". English. In: *Frontiers in Artificial Intelligence* 6. Publisher: Frontiers. ISSN: 2624-8212. DOI: 10.3389/frai.2023.1225791. URL: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2023.1225791/full (visited on 08/19/2025) (cit. on p. 40).

Wilcox, Ethan Gotlieb, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger P. Levy (2020). "On the Predictive Power of Neural Language Models for Human Real-TimeComprehension Behavior". en. In: *Proceedings of the Annual Meeting of the Cognitive Science Society* 42.0. URL: https://escholarship.org/uc/item/738338tm (visited on 06/24/2025) (cit. on pp. 11, 13, 14, 20, 25, 47).

Wilcox, Ethan Gotlieb, Clara Isabel Meister, Ryan Cotterell, and Tiago Pimentel (Dec. 2023a). "Language Model Quality Correlates with Psychometric Predictive Power in Multiple Languages". en. In: Medium: application/pdf Publisher: ETH Zurich. DOI: 10.3929/ETHZ-B-000650659. URL: http://hdl.handle.net/20.500.11850/650659 (visited on 06/23/2025) (cit. on pp. 13, 14).

Wilcox, Ethan Gotlieb, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy (2023b). "Testing the Predictions of Surprisal Theory in 11 Languages". In: *Transactions of the Association for Computational Linguistics* 11. Place: Cambridge, MA Publisher: MIT Press, pp. 1451–1470. DOI: 10.1162/tacl_a_00612. URL:

https://aclanthology.org/2023.tacl-1.82/ (visited on 01/14/2025) (cit. on pp. 11, 13).

Wingfield, Cai and Louise Connell (2022). "Understanding the role of linguistic distributional knowledge in cognition". In: vol. 37. 10. Routledge, pp. 1220–1270. DOI: 10.1080/23273798.2022.2069278. eprint: https://doi.org/10.1080/23273798.2022.2069278. URL: https://doi.org/10.1080/23273798.2022.2069278 (cit. on p. 36).

Wittgenstein, Ludwig (1953). *Philosophische Untersuchungen*. Frankfurt am Main: Suhrkamp Verlag (cit. on p. 16).

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush (Oct. 2020a). "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Ed. by Qun Liu and David Schlangen. Online: Association for Computational Linguistics, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. URL: https://aclanthology.org/2020.emnlp-demos.6 (cit. on pp. 27, 111).

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush (2020b). *HuggingFace's Transformers: State-of-the-art Natural Language Processing*. arXiv: 1910.03771 [cs.CL] (cit. on p. 92).

Wu, Shijie and Mark Dredze (July 2020). "Are All Languages Created Equal in Multilingual BERT?" In: *Proceedings of the 5th Workshop on Representation Learning for NLP*. Ed. by Spandana Gella, Johannes Welbl, Marek Rei, Fabio Petroni, Patrick Lewis, Emma Strubell, Minjoon Seo, and Hannaneh Hajishirzi.

Online: Association for Computational Linguistics, pp. 120–130. DOI: 10.18653/v1/2020.repl4nlp-1.16. URL: https://aclanthology.org/2020.repl4nlp-1.16/ (cit. on p. 82).

Xue, Fuzhao, Yao Fu, Wangchunshu Zhou, Zangwei Zheng, and Yang You (2023). "To Repeat or Not To Repeat: Insights from Scaling LLM under Token-Crisis". In: DOI: 10.48550/ARXIV.2305.13230. URL: https://arxiv.org/abs/2305.13230 (visited on 05/31/2023) (cit. on p. 26).

Yousef, Saeed (2018). *Persian: a comprehensive grammar*. eng. Routledge comprehensive grammars. London New York: Routledge. ISBN: 978-1-138-92703-2 978-1-138-92702-5 (cit. on p. 80).

Zarcone, Alessandra, Marten Van Schijndel, Jorrig Vogels, and Vera Demberg (June 2016). "Salience and Attention in Surprisal-Based Accounts of Language Processing". In: *Frontiers in Psychology* 7. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2016.00844. URL: http://journal.frontiersin.org/Article/10.3389/fpsyg.2016.00844/abstract (visited on 06/10/2025) (cit. on p. 44).

Zeman, Daniel, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, eljko Agi, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Salih Furkan Akkurt, Gabriel Aleksandraviit, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Erik Andersen, Lene Antonsen, Tatsuya Aoyama, Katya Aplonova, Angelina Aquino, Carolina Aragon, Glyd Aranes, Maria Jesus Aranzabe, Bilge Nas Arcan, Hórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Katla Ásgeirsdóttir, Deniz Baran Aslan, Cengiz Asmazolu, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Mariana Avelãs, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Starkaður Barkarson, Rodolfo Basile, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Shabnam Behzad, Juan Belieni, Kepa Bengoetxea, brahim Benli, Yifat Ben Moshe, Gözde Berk, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agn Bielinskien, Kristín Bjarnadóttir, Rogier Blokland, Victoria

Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaar, António Branco, Kristina Brokait, Aljoscha Burchardt, Marisa Campos, Marie Candito, Bernard Caron, Gauthier Caron, Catarina Carvalheiro, Rita Carvalho, Lauren Cassidy, Maria Clara Castro, Sérgio Castro, Tatiana Cavalcanti, Gülen Cebirolu Eryiit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír éplö, Neslihan Cesur, Savas Cetin, Özlem Çetinolu, Fabricio Chalub, Liyanage Chamila, Shweta Chauhan, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Jayeol Chun, Juyeon Chung, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çar Çöltekin, Miriam Connor, Claudia Corbetta, Daniela Corbetta, Francisco Costa, Marine Courtin, Benoît Crabbé, Mihaela Cristescu, Vladimir Cvetkoski, Ingerid Løyning Dale, Philemon Daniel, Elizabeth Davidson, Leonel Figueiredo de Alencar, Mathieu Dehouck, Martina de Laurentiis, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Adrian Doyle, Timothy Dozat, Kira Droganova, Magali Sanches Duran, Puneet Dwivedi, Christian Ebert, Hanne Eckhoff, Masaki Eguchi, Sandra Eiche, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Toma Erjavec, Farah Essaidi, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Federica Favero, Jannatul Ferdaousi, Marília Fernanda, Hector Fernandez Alcalde, Amal Fethi, Jennifer Foster, Theodorus Fransen, Cláudia Freitas, Kazunori Fujita, Katarína Gajdoová, Daniel Galbraith, Federica Gamba, Marcos Garcia, Moa Gärdenfors, Fabrício Ferraz Gerardi, Kim Gerdes, Luke Gessler, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökrmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Gricit, Matias Grioni, Loïc Grobol, Normunds Gruztis, Bruno Guillaume, Kirian Guiller, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Haji, Jan Haji jr., Mika Hämäläinen, Linh Hà M, Na-Rae Han, Muhammad Yudistira Hanifmuti, Takahiro Harada, Sam Hardwick, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváová,

Florinel Hociung, Petter Hohle, Yidi Huang, Marivel Huerta Mendez, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, lájídé Ishola, Artan Islamaj, Kaoru Ito, Sandra Jagodziska, Siratun Jannat, Tomá Jelínek, Apoorva Jha, Katharine Jiang, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Hüner Kakara, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Ritván Karahóa, Andre Kåsen, Tolga Kayadelen, Sarveswaran Kengatharaiyer, Václava Kettnerová, Lilit Kharatyan, Jesse Kirchner, Elena Klementieva, Elena Klyachko, Petr Kocharov, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Mehmet Köse, Alexey Koshevoy, Natalia Kotsyba, Jolanta Kovalevskait, Simon Krek, Parameswari Krishnamurthy, Sandra Kübler, Adrian Kuqi, Ouzhan Kuyrukçu, Asl Kuzgun, Sookyoung Kwak, Kris Kyle, Käbi Laan, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phng Lê Hông, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Lauren Levine, Cheuk Ying Li, Josie Li, Keying Li, Yixuan Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Yi-Ju Jessica Lin, Krister Lindén, Yang Janet Liu, Nikola Ljubei, Irina Lobzhanidze, Olga Loginova, Lucelene Lopes, Stefano Lusito, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Menel Mahamdi, Jean Maillard, Ilya Makarchuk, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Büra Maran, Ctlina Mrnduc, David Mareek, Katrin Marheinecke, Stella Markantonatou, Héctor Martínez Alonso, Lorena Martín Rodríguez, André Martins, Cláudia Martins, Jan Maek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Tatiana Merzhevich, Niko Miekka, Aaron Miller, Karina Mischenkova, Anna Missilä, Ctlin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nepore-

Brzkalne, Manuela Nevaci, Lng Nguy ~ ên Th, Huyên Nguy ~ ên Th Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Alireza Nourian, Maria das Graças Volpe Nunes, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Hulda Óladóttir, Adéday Olúòkun, Mai Omura, Emeka Onwuegbuzia, Noam Ordan, Petya Osenova, Robert Östling, Lilja Øvrelid, aziye Betül Özate, Merve Özçelik, Arzucan Özgür, Balkz Öztürk Baaran, Teresa Paccosi, Alessio Palmero Aprosio, Anastasia Panova, Thiago Alexandre Salgueiro Pardo, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Giulia Pedonese, Angelika Peljak-apiska, Siyao Peng, Siyao Logan Peng, Rita Pereira, Sílvia Pereira, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Andrea Peverelli, Jason Phelan, Claudel Pierre-Louis, Jussi Piitulainen, Yuval Pinter, Clara Pinto, Rodrigo Pintucci, Tommi A Pirinen, Emily Pitler, Magdalena Plamada, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalnia, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Robert Pugh, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andreia Querido, Andriela Rääbis, Alexandre Rademaker, Mizanur Rahoman, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Joana Ramos, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Mathilde Regnault, Georg Rehm, Arij Riabi, Ivan Riabov, Michael RieSSler, Erika Rimkut, Larissa Rinaldi, Laura Rituma, Putri Rizqiyah, Luisa Rocha, Eiríkur Rögnvaldsson, Ivan Roksandic, Mykhailo Romanenko, Rudolf Rosa, Valentin Roca, Davide Rovati, Ben Rozonoyer, Olga Rudina, Jack Rueter, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardi, Stephanie Samson, Manuela Sanguinetti, Ezgi Sanyar, Dage Särg, Marta Sartor, Mitsuya Sasaki, Baiba Saulte, Agata Savary, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Emmanuel Schang, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Syeda Shahzadi, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Maria Shvedova, Janine Siewert, Einar Freyr Sigurðsson, João Silva, Aline Silveira, Natalia Silveira, Sara Silveira, Maria Simi,

Radu Simionescu, Katalin Simkó, Mária imková, Haukur Barri Símonarson, Kiril Simov, Dmitri Sitchinava, Ted Sither, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Per Erik Solberg, Barbara Sonnenhauser, Shafi Sourov, Rachele Sprugnoli, Vivian Stamou, Steinthór Steingrímsson, Antonio Stella, Abishek Stephen, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Daniel Swanson, Zsolt Szántó, Chihiro Taguchi, Dima Taji, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Dipta Tanaya, Mirko Tavoni, Samson Tella, Isabelle Tellier, Marinella Testori, Guillaume Thomas, Sara Tonelli, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sveinbjörn Hórðarson, Vilhjálmur Horsteinsson, Sumire Uematsu, Roman Untilov, Zdeka Ureová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Elena Vagnoni, Sowmya Vajjala, Socrates Vak, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Uliana Vedenina, Giulia Venturi, Eric Villemonte de la Clergerie, Veronika Vincze, Natalia Vlasova, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Shira Wigderson, Sri Hartati Wijono, Vanessa Berwanger Wille, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Qishen Wu, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Arife Betül Yenice, Olcay Taner Yldz, Zhuoran Yu, Arlisa Yuliawati, Zdenk abokrtský, Shorouq Zahra, Amir Zeldes, He Zhou, Hanzhi Zhu, Yilun Zhu, Anna Zhuravleva, and Rayan Ziane (2023). *Universal Dependencies 2.13*. URL: http://hdl.handle.net/11234/1-5287 (cit. on p. 54).

Zhang, Susan, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer (2022a). "OPT: Open Pre-trained Transformer Language Models". In: Publisher: arXiv Version Number: 4. DOI:

10.48550/ARXIV.2205.01068. URL: https://arxiv.org/abs/2205.01068 (visited on 04/24/2023) (cit. on pp. 26, 111).

Zhang, Susan, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer (2022b). "OPT: Open Pre-trained Transformer Language Models". In: *CoRR* abs/2 205.01068v4. DOI: 10.48550/ARXIV.2205.01068. arXiv: 2205.01068. URL: https://arxiv.org/abs/2205.01068v4 (cit. on p. 91).

Zhang, Yian, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman (Aug. 2021). "When Do You Need Billions of Words of Pretraining Data?" In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 1112–1125. DOI: 10.18653/v1/2021.acl-long.90. URL: https://aclanthology.org/2021.acl-long.90 (cit. on pp. 23, 24, 33).

Zipf, George (1935). *The Psycho-Biology of Language. An Introdution to Dynamic Philology*. Cambridge, Mass: MIT Press. ISBN: 978-0-262-74002-9. URL: https://mitpress.mit.edu/9780262740029/the-psycho-biology-of-language/ (cit. on p. 64).

— (1949). *Human behavior and the principle of least effort*. Cambridge, Mass.: Addison-Wesley Press. (cit. on p. 64).

Zouhar, Vilém, Clara Meister, Juan Gastaldi, Li Du, Tim Vieira, Mrinmaya Sachan, and Ryan Cotterell (July 2023). "A Formal Perspective on Byte-Pair Encoding". In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 598–614. DOI: 10.18653/v1/2023.findings-acl.38. URL: https://aclanthology.org/2023.findings-acl.38/ (cit. on p. 83).

A

Pronoun Surprisal

A.1 Appendix A

The pronominal systems: data from grammars. An extended version of Table 1, providing details on alternative forms, case and adpositional marking, expression of pronominal subject, pre/post verbal position, person, number, gender, clusivity and additional notes. https://zenodo.org/records/15064869

A.2 Appendix B

Pre/post verbal position: data from mini-CIEP+. Frequency tables for the pre/post verbal positions of personal pronouns in the mini-CIEP+ corpus. https://zenodo.org/records/15064883

A.3 Appendix C

UD implementation of the pronominal systems. The document describes the implementation of the personal pronominal system and its features using several layers of annotation from the Universal Dependencies framework: universal parts of speech,

language-specific parts of speech, universal features, syntactic dependencies as well as lists of word forms and lemmas. https://zenodo.org/records/15064946

A.4 Appendix D

Comparison of surprisal from mGPT and surprisal from monolingual models trained on the Greek, Italian, Turkish and Ukrainian sections of theWiki-40b dataset. https://zenodo.org/records/12633172

A.5 Appendix E

Data analysis using the area under surprisal curve. As an alternative measure, we introduce the area under surprisal curve (AUC). We perform a multi-linear regression model using AUC and the data discussed for surprisal with context 7. https://zenodo.org/records/15131405

A.6 Appendix F

Data files containing Rdata from the brm analysis and results, pronoun frequency with respect to the size of individual language mini-CIEP+ corpora, surprisal/AUC correlations with word length, coefficient tables, heatmap with estimates and coefficients for the brms model. https://zenodo.org/records/15064756

A.7 Appendix G

Interactive plot displaying the interaction between surprisal and frequency for each language of the sample. https://zenodo.org/records/15064733

B

Babylm

B.1 Validation perplexity

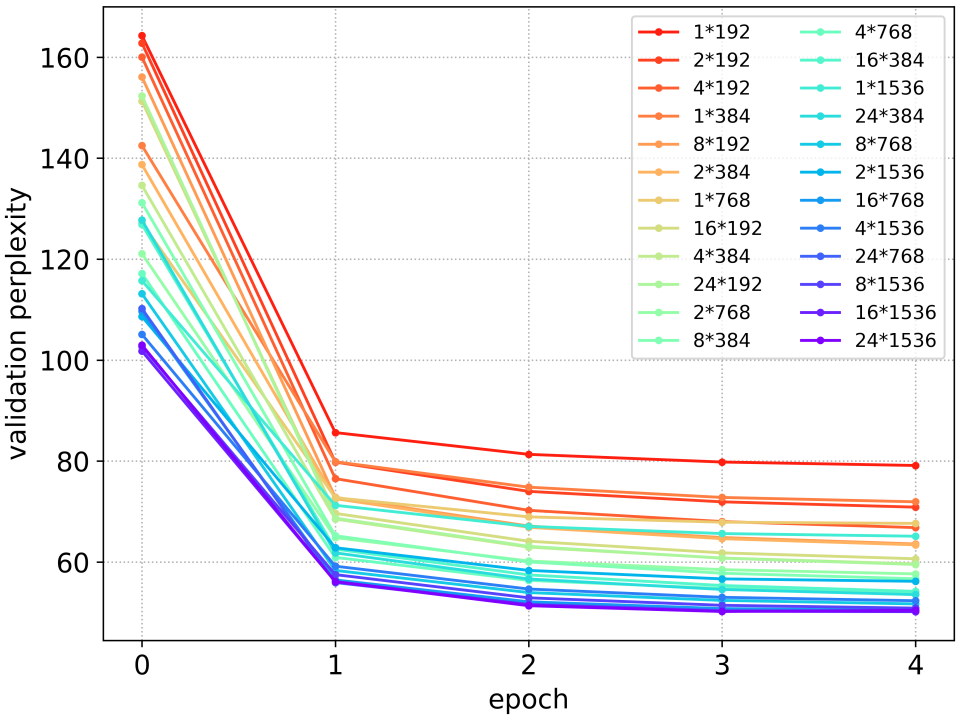


Figure B.1: Validation perplexity by configuration and epoch on the development set of the BabyLM corpus.

B.2 OPT models

$l_{decoder}$	l_{hidden}	**Parameters (non-embedding)**
1	192	0.74
2	192	1.19
4	192	2.07
8	192	3.85
16	192	7.41
24	192	10.9
1	384	2.37
2	384	4.14
4	384	7.69
8	384	14.79
16	384	28.99
24	384	43.18
1	768	7.09
2	768	14.18
4	768	28.35
8	768	56.70
16	768	113.41
24	768	170.11
1	1536	30.69
2	1536	59.00
4	1536	115.69
8	1536	229.01
16	1536	455.67
24	1536	682.32

Table B.1: OPT models sizes in million parameters by hidden size and number of decoder layers. The number of parameters does not include the embedding table, which is always of the size $l_{emb} \times |V| = 768 \times 50272 = 38.608.896$, as in OPT-128m.

B.3 Detailed resutls: Reading time experiments

Corpus	Step	Spearman's ρ	p-value
babylm	500	-0.5913	0.0097
babylm	1000	-0.6285	0.0052
babylm	2000	-0.7833	0.0001
babylm	3000	-0.7874	0.0001
babylm	1	-0.7915	0.0001
babylm	5	-0.614	0.0067
wikitext-103	500	0.0815	0.7478
wikitext-103	1000	-0.4241	0.0794
wikitext-103	2000	-0.7482	0.0004
wikitext-103	3000	-0.7441	0.0004
wikitext-103	1	-0.7172	0.0008
wikitext-103	5	-0.7523	0.0003

Table B.2: Spearman's ρ of model size (in terms of number of parameters) and Δ log-likelihood over the baseline LME model. Steps 1 and 5 refer to the first and fifth epoch.

B.4 Detailed results: BabyLM challenge tasks

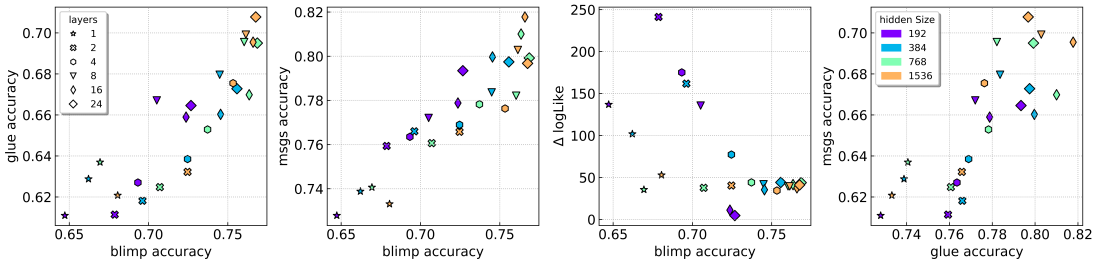


Figure B.2: Correlation of LM performance on BLiMP vs. GLUE, BLiMP vs. MSGS, GLUE vs. MSGS.

Corpus	Tokenizer	Tokens	t_{REPL}	$\lvert V \rvert$
RSC	Lempos	2.5M	1	79K
	Word		1	74K
	BPE		0	100K
COHA	Lempos	3.5M	1	83K
	Word		3	98K
	BPE		0	100K

Table C.1: Corpus sizes and data preprocessing parameters. 10% of the sampled tokens were used as a development set.

C

RSC MST

C.1 Tokenization Methods

In order to mitigate the problem of vocabulary expansion, we employ and independently evaluate three tokenization strategies, which all drastically reduce the number of tokens in the vocabulary and do not require a model whose parameters are mostly in the embedding layer (which would happen in case of a vocabulary of about 500K tokens, as in COHA).

Word-level tokenization: This is the simplest tokenization approach and requires a few tweaks to work. We use word-level tokenization with replacement of out-of-vocabulary (OOV) items instead of a subword tokenization method because words that

are split into many subtokens due to high tokenizer fertility would be assigned higher surprisal values by default. The surprisal of these de-facto OOV items would artificially inflate our AUC measure and obscure the impact of word order on AUC.

Lempos tokenization: This tokenization approach is derived from word-level tokenization, but reduces the size of the unigram vocabulary even further by replacing word forms with a combination of the corresponding lemma and UPOS tag.

BPE tokenization: We use the default implementation of the BPE algorithm in the Hugging Face tokenizers Python package to train a tokenizer with a vocabulary size of 100K on the subsampled version of each corpus. We did not replace OOV words, as those are handled by the tokenization algorithm. An overview of the tokenization methods, thresholds and examples of a tokenized sentence can be found in Table C.1.

C.2 Consistency across tokenization methods

We re-fitted all LMMs with surprisals and AUCs from language models trained on word-level and BPE-tokenized versions of the subsampled corpora. We found that, while effect sizes vary greatly between tokenizations, the direction of the effects is consistent. See Figure C.1 for a detailed overview of the LMM coefficients for all three tokenization methods.

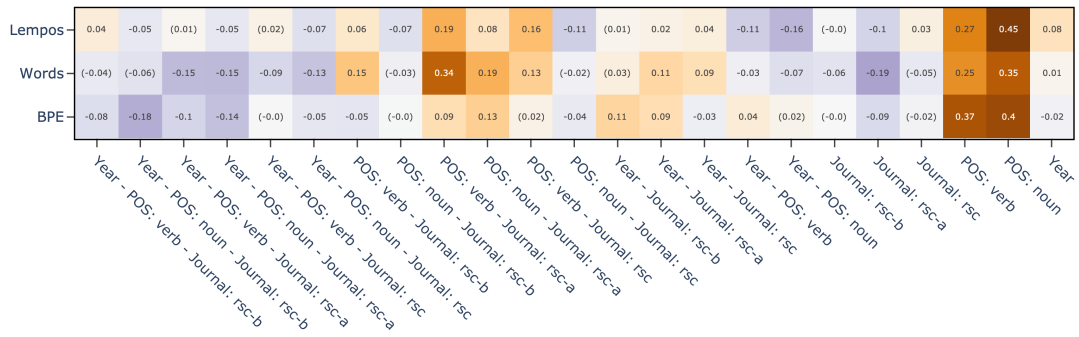


Figure C.1: LMM coefficients for AUC, surprisal from language models trained on lempos, word-level and BPE tokenizations. Non-significant effects in parentheses.