Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 225 (2023) 3488–3497

Procedia Computer Science

www.elsevier.com/locate/procedia

27th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2023)

Expert-adapted language models improve the fit to reading times

Iza Škrjanec*, Frederik Yannick Broy, Vera Demberg

Department of Language Science and Technology, Informatics Campus, Saarland University, 66123 Saarbrücken, Germany

Abstract

The concept of surprisal refers to the predictability of a word based on its context. Surprisal is known to be predictive of human processing difficulty and is usually estimated by language models. However, because humans differ in their linguistic experience, they also differ in the actual processing difficulty they experience with a given word or sentence. We investigate whether models that are similar to the linguistic experience and background knowledge of a specific group of humans are better at predicting their reading times than a generic language model. We analyze reading times from the PoTeC corpus [15, 27] of eye movements from biology and physics experts reading biology and physics texts. We find experts read in-domain texts faster than novices, especially domain-specific terms. Next, we train language models adapted to the biology and physics domains and show that surprisal obtained from these specialized models improves the fit to expert reading times above and beyond a generic language model.

Keywords: Eye tracking; Background knowledge; Surprisal

1. Introduction

The concept of surprisal, or a word's predictability in context [10, 24], has been widely accepted as a measure of human processing difficulty [2, 41, 4, 6, 5, 38, 39]. However, the most effective way to estimate surprisal remains unclear, with research focusing on what language model architecture or size to use [3, 7, 29, 31, 23, 32].

Previous studies evaluating language models against measures of human processing difficulty have mostly predicted average reading times or those of a typical reader, overlooking the fact that individuals experience varying degrees of difficulty when reading a specific text. This variation is particularly significant when dealing with domain-specific texts, which may be challenging for novices but comparatively easy for experts. Models that can accurately predict the processing difficulty of a specific group of people have broad applications in readability research and language generation adaptation. The recently collected Potsdam Textbook Corpus (PoTeC) [15, 27], which includes eye-tracking measures over expository texts on biology and physics in German, provides empirical data with infor-

* Corresponding author.
E-mail address: skrjanec@lst.uni-saarland.de

1877-0509
10.1016/j.procs.2023.10.344

mation about readers, allowing researchers to analyze whether the reading times of novices and experts differ and whether domain-specific surprisal models are better predictors of expert reading times than generic surprisal models.

This study employs a German GPT-2 model, initially trained on generic text, and further trained on corpora of biology or physics Wikipedia articles to create two language models that simulate linguistic expectations of biology and physics experts, respectively. The findings reveal that specialized surprisal is a significant predictor of expert reading times, over and above a generic surprisal model, and this effect is particularly pronounced for domain-specific terms.

2. Background

2.1. Background knowledge and reading behavior

To comprehend a text and make accurate inferences about its content, readers must form an appropriate mental representation of the text. This representation is heavily influenced by the reader's background knowledge of the text topic (see [42] for overview).

According to the Construction-Integration model of reading, successful reading involves a dynamic interplay between the literal representation of the text and related schemata derived from the reader's background knowledge [20, 21]. Background knowledge is stored in long-term memory as a set of propositions, which are organized into various schemata. During reading, activated schemata aid in constructing a meaningful representation of the text [43]. When readers lack the required knowledge to integrate the text material properly, they struggle to construct an effective meaning representation and encounter comprehension difficulties [19, 18].

This processing difficulty is reflected in eye movement patterns during reading. Several eye-tracking studies have demonstrated that word frequency, length, and familiarity to the reader influence reading times. For instance, in their study of reading scientific texts, Just and Carpenter [16] identified seven words as completely unfamiliar to participants and found that the unfamiliar words were processed more slowly. Similar results were found in a study of real and pseudowords, with pseudowords receiving more refixations in the first pass and longer processing times, particularly for longer words [26].

Although eye-tracking studies have shown that less familiar words result in a greater processing effort and longer reading times, they have not examined background knowledge as a subject-related predictor, i.e. how the same text is read by subjects with varying levels of familiarity. In this paper, we examine the role of background knowledge in reading by comparing subjects who are experts in the text topic to those who are much less familiar with it. We aim to demonstrate how background knowledge influences reading behavior and whether we can better approximate this behavior with surprisal from expert-adapted language models.

2.2. Surprisal theory and pretrained language models

The way we process sentences depends on our expectations of what comes next, according to the expectation-based account [10, 24]. This theory suggests that the more surprising a word is given its preceding context, the more effort it takes to integrate it into the sentence. This effort is measured by surprisal, which is the negative log-probability of a word $surprisal(w) = -\log p(w|context)$. In other words, the less probable a word is, the higher its surprisal value. In several studies, surprisal has been linked to observed reading behavior, such as reading time, [2, 41, 4], the N400 amplitude [30] and language processing as measured with fMRI [40].

To estimate surprisal, we need a language model (LM) that assigns probabilities to words given their context. Recently, large pretrained neural language models have been for that, as these models are a lot more accurate in language tasks than previous model types. Although neural LMs were initially designed for natural language processing tasks, there is active ongoing research on their cognitive plausibility and use in psycholinguistic research [22, 33, 44]. The models are trained on vast amounts of text from different domains, and they can be used as domain-general models or further trained and fine-tuned for specific domains and tasks to achieve higher performance [8]. However, not all pretrained neural LM simulate human reading behavior equally well and the literature shows there is no simple recipe. While attention has been shown as more predictive of human processing than the recurrent mechanism [29], studies show that smaller Transformer models fit reading times better [33] and so do models with limited context [22]. In this

Table 1. Excerpt from biology text B0 on gel electrophoresis. The words in italics are technical terms of level 1, while bold print marks technical terms of level 2. The remaining words are considered common (level 0).

Um das Vorhandensein der **Polymerasekettenreaktion-Produkte** feststellen zu können, verwendet man die **Gelelektrophorese** mit einer anschließenden Behandlung des Gels durch **Ethidiumbromid**. Bei dieser Methode wird die *negative Ladung* der *DNA-Fragmente* genutzt. Die *DNA-Fragmente* sind auf Grund der darin enthaltenen **Phosphatreste** *negativ* geladen. Anlegen eines *elektrischen Feldes* lässt die **Polymerasekettenreaktion-Produkte** durch ein gelartiges *Medium*, zumeist **Agarose**, wandern. Die *DNA-Fragmente* wandern durch das **"Agarose-Sieb"** vom *Minuspol* zum *Pluspol*. /.../

Table 2. Average accuracy (and standard deviation) on background knowledge and text comprehension questions. Unpaired t-tests showed that the two reader groups significantly differ in biology knowledge ($t(73) = 3.0$, $p = 0.004$) and text comprehension ($t(73) = 3.24$, $p = 0.002$), as well as physics knowledge ($t(73) = -3.57$, $p = 0.0006$) and text comprehension ($t(73) = -2.66$, $p = 0.01$).

		Biology texts		Physics Texts	
		Background knowledge	Text comprehension	Background knowledge	Text comprehension
Major	Biologists	75.71 ± 22.81	52.71 ± 20.87	50.65 ± 13.46	24.81 ± 10.38
	Physicists	60.59 ± 19.77	38.54 ± 15.45	62.85 ± 16.11	33.33 ± 17.28

work, we explore whether a LM can account for variation in background knowledge between readers. We use a small GPT-2 model, namely the German domain-general [13] as well as specialized GPT-2 models to investigate their fit to reading times of experts and novices in chosen domains.

3. Data and language models

To model the role of background knowledge in reading behavior, we use the Potsdam Textbook Corpus (Section 3.1) of eye-tracking measures over expository text. We also explore how expert-adapted language models contribute to predicting reading times. Section 3.2 presents the domain-specific Wikipedia corpora for adapting the German GPT-2 language model to the domains of biology and physics.

3.1. The Potsdam Textbook Corpus (PoTeC)

The Potsdam Textbook Corpus (PoTeC) [15, 27] is a collection of eye-tracking measures gathered over texts from German biology and physics text books. It comprises eye movement recordings from 75 subjects while reading 6 texts on biology and 6 on physics. Each text focuses on a single topic from the respective domain. All subjects were German native speakers studying either physics (N=32) or biology (N=43) at university.

After reading, participants were given three multiple-choice comprehension questions without the option of returning to the text. Additionally, three multiple-choice background knowledge questions were asked for each text. On average, students showed a higher background knowledge as well as more accurate text comprehension when they were reading texts from the domain of their major (i.e. biologists reading biology, and physicists reading physics) as shown in Table 2.

The stimuli texts have 158 words on average (minimally 126 and maximally 180). They are annotated for terminology. Common words that belong to general vocabulary are labeled as level 0, while technical terms that are not very specific are labeled as level 1, and highly specific terms are labeled as level 2. Table 1 shows an excerpt from a biology text with terminology annotation for the three levels. In total, there are 954 words in biology texts, 79% of these are of level 0, 11% are level 1, and 9% are technical terms of level 2. The distribution is similar in physics texts with 941 words in total, out of which 78% are of level 0, 15% of level 1, 7% of level 2. Figure 3.1 shows that in both domains, technical terms have longer reading times than common words. For common words, there is little difference between subjects reading in- and out-of-domain texts, while on technical words experts (biologists when reading biology,physicists when reading physics) are faster.

In total, PoTeC contains 142,125 data points (75 subjects × (954+941) words). We excluded data points that were not fixated in the first pass or were not fixated at all or belong to the first word in the text. After log-transforming the reading times, we remove data points that lie over 3 standard deviations from the grand mean of first-pass or from

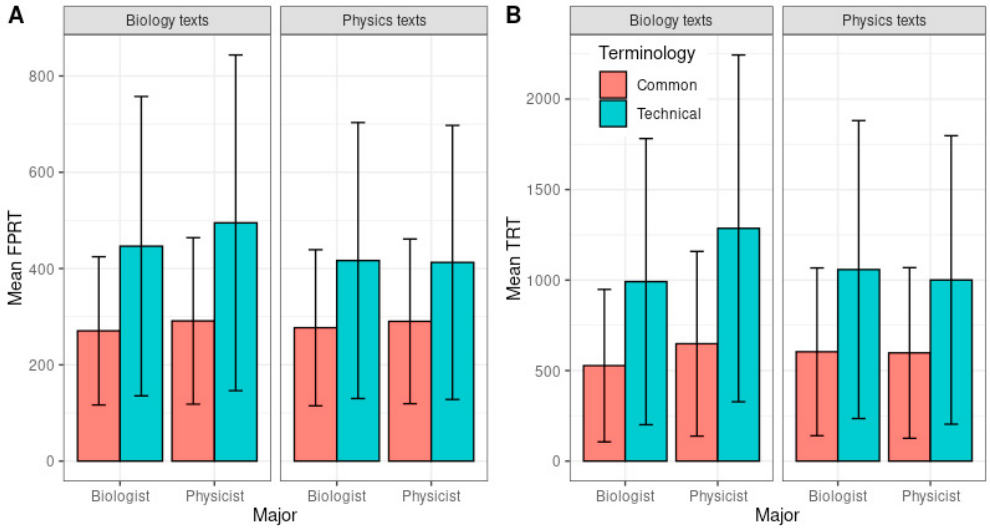

Fig. 1. Average first-pass (FPRT; A) and total reading time (TRT; B) in PoTeC after data cleaning. The error bars represent standard deviation.

the grand mean of total reading time. In total, 28.4% of observations were removed, resulting in 1,356 data points per subject on average (minimally 911 and maximally 1,621) and around 50.8k data points per each of domain.

3.2. Training of expert-adapted GPT-2 models

The expert-adapted GPT-2 language models were created by continuing training the domain-generic German GPT-2 [13]. To create a training set for each of the expert LMs, we scrapped domain-related articles from German Wikipedia. From the list of Wikipedia categories, we hand-picked 35 biology categories (e.g. botany, chronobiology, zoology) and 32 physics categories (e.g. astrophysics, quantum mechanics, thermodynamics). The `wikipediaapi` library was used to collect the names of subcategories of these categories. We used the `beautifulsoup` and `requests` libraries to scrape all articles belonging to the (sub)categories, yielding 15,883 articles (8.4 million words) for the Wiki-Bio, and 11,347 articles (5.8 million words) for the Wiki-Phys corpus.

The original German GPT-2 model is Generative Pre-trained Transformer [36] trained on the German Wikipedia, OpenLegalData and news. The vocabulary includes around 50k byte-level tokens. To continue training the model on the next-word-prediction objective, we use the `transformers` library and train the German GPT-2 model on Wiki-Bio to obtain the BioGPT-2, and on Wiki-Phys for the PhysGPT-2 model (for hyperparameters see Appendix A.1).

The general German GPT-2 as well as the expert-adapted BioGPT-2 and PhysGPT-2 models were evaluated intrinsically for their fit to specific domains. We calculate their perplexity over a set of held-out texts: domain-generic text (a recent news article), excerpts from MSc theses in biology and physics, and stimuli from the PoTeC corpus. Table 3 shows the expert language models achieve lower perplexity on the in-domain texts than out-of-domain. Similarly, the general GPT-2 model has the lowest perplexity on generic text, but it rises on domain-specific texts.

Table 3. Perplexity of three German language models (general GPT-2, BioGPT-2 and PhysGPT-2) on held-out test texts.

	Language model		
	General GPT-2	BioGPT-2	PhysGPT-2
Domain-generic text	22.79	22.95	24.36
MSc thesis biology	55.76	17.53	22.51
MSc thesis physics	66.18	29.09	22.29
Mean over PoTeC biology texts	27.19	25.86	31.69
Mean over PoTeC physics texts	24.35	26.40	21.04

4. The effect of background knowledge on reading time

In our first analysis, we explore the effect of the reader's background knowledge on reading times. We are interested in reading measures that are associated with lexical access (i.e. early measures) as well as those that index higher order processes like semantic integration and revision (i.e. late measures) [37]. We choose first-pass reading time as our early, and total reading time as our late measure of interest. We assume that, upon reading a word, lexical retrieval is easier for experts as they are most likely familiar with the word. Similarly, integration of the word's meaning into the preceding context should be less difficult for experts as this process is supported by their background knowledge on the text's topic. We thus expect longer first-pass and longer total reading times for novices relative to experts. We hypothesize this effect will be even more pronounced for technical terminology, as technical terms are typically more complex than common vocabulary: they are domain-specific, less frequent and often composed of foreign language morphemes, which makes it harder to decode their meaning without sufficient prior knowledge.

4.1. Analysis

We use a linear-mixed effects regression model (LMER) and backward selection for each of the two reading measures. First-pass reading time is the sum of all fixation durations that occur on the word in the first pass, i.e. the amount of time that passes when the gaze first enters the word region to when it first leaves this region. Total reading time is the sum of all fixation durations on the word. Both are measured in milliseconds.

For both of the measures, we fit an LMER model using the log-transformed reading times as the response variable. Model predictors include the control variables of word length, log-transformed word frequency (from dLexDB [11]), surprisal from general German GPT-2 and word index in sentence, as well as predictors of interest: terminology, readers' expertise and an interaction between terminology and expertise. The terminology variable was encoded using two levels, "technical" (which includes levels 1 and 2, see Section 3.1) and "common" (level 0). The expertise variable contains the levels "expert" (when a physicist is reading physics, or a biologist reading biology texts) and "novice" (when the participant is reading a text out of their domain). Additionally, a by-subject random intercept as well as a by-word random intercept and by-word random slope for expertise level are included.[1] We conducted our analyses using linear-mixed effects model in R [35] with the `lme4` package [1], version 1.1-32.

Word-level surprisal values were extracted from the German GPT-2 model by feeding it each text separately. Each text fit in the maximal context size limitation of 1,024 tokens. The texts were tokenized using the model's byte-pair encoding tokenizer. In case a single word was tokenized into multiple subword tokens, the negative log probabilities of subword tokens were summed to calculate the surprisal of the word. For the regression models, all continuous predictors were centered around their respective means. Both terminology and expertise were sum-coded (Terminology: "common" as -1, "technical" as 1. Expertise: "novice" as -1, "expert" as 1.)

4.2. Results

The final models for log first-pass times and log total reading times are displayed in Table 4.[2] We find significant main effects of word length, word index in sentence and surprisal on first-pass as well total reading time: words that are longer, appear at the sentence beginning or have a higher surprisal have longer reading times. This confirms the findings from previous studies. There is a significant effect of word frequency on total reading time, but not in the first pass. The models also reveal a main effect of expertise and terminology as well as their interaction: technical terms were generally read more slowly than common words, and experts required less time for reading. The significant interaction shows that technical terms in particular are read faster by experts than novices. A subset analysis confirms this interpretation: when we split the data according to terminology type, we find that expertise has a significant main effect on both splits: for technical terms (β=-0.037, SE=0.004, t=-8.86) and for common words (β=-0.006, SE=0.002, t=-3.74), with the effect being larger for technical terms.

[1] Linear mixed-effects regression model: RT ~ WordLen + LogFreqLemma + WordIndexInSentence + GenSurprisal + Expertise*Terminology + (1 | SubjectID) + (1 + Expertise | WordID). RT is either first-pass or total reading time.

[2] Code is available at https://github.com/izaskr/expertLM_reading_time.

Table 4. Regression coefficients and test statistics from the linear mixed-effects model for each reading measure. The asterisk indicates the p-value: *** ($p < 0.001$), ** ($p < 0.01$), * ($p < 0.05$), no asterisk ($p > 0.05$).

	FPRT				TRT			
Variable	β	SE	t		β	SE	t	
(Intercept)	5.62	0.02	348.12	***	6.31	0.03	190.06	***
Word length	0.04	0.001	29.96	***	0.06	0.002	27.61	***
Word freq					-0.01	0.003	-2.94	**
Word index in sentence	-0.002	0.001	-3.73	***	-0.005	0.001	-7.39	***
GenSurprisal	0.02	0.001	11.59	***	0.03	0.002	16.54	***
Expertise	-0.02	0.002	-11.34	***	-0.08	0.002	-34.09	***
Terminology	0.05	0.01	7.08	***	0.08	0.01	8.95	***
Expertise:Terminology	-0.01	0.002	-7.98	***	-0.02	0.002	-9.49	***

5. Modelling reading behavior with expert-adapted language models

In our second analysis, we investigate the relation between reading times and surprisal based on expert language models. Our previous analysis confirmed surprisal as a measure of processing load: words that were less predictable had longer reading times. We also demonstrated that experts read the texts faster, and that this difference in reading times is particularly large for technical terms. This means that technical words induced a smaller processing load on experts than on novices. Surprisal in the first analysis was estimated based on a general LM. To find whether linguistic expectations of expert readers are better represented by an expert-adapted language model, we repeat the analysis but include surprisal from expert LMs as an additional predictor. We hypothesize that surprisal from an expert LM will provide a better fit for reading times of experts, especially for technical terms.

5.1. Analysis

We again use linear mixed-effects regression models with log-transformed first-pass and log-transformed total reading times as the response variable, modeling each reading measure separately. We extend the set of predictors from the first analysis by adding surprisal from expert LMs (henceforth, specialized surprisal). To create the variable of specialized surprisal, the surprisal from BioGPT-2 is used when the reader is an expert in biology, and the surprisal estimates from PhysGPT-2 when the reader is a physics experts. We note that specialized surprisal is highly correlated with surprisal from the general LM ($r = 0.97$). To study the contribution of specialized surprisal, we therefore residualize the specialized surprisal as follows: we use the PoTeC stimulus texts and fit a linear regression model with specialized surprisal as the outcome, and general surprisal as the predictor.[3] We term the residuals of this model residualized specialized surprisal and use it as an additional continuous predictor *ResidSpecSurprisal* to represent the reader's linguistic and expectations beyond the general reading skill (as operationalized by general surprisal).

We present two analyses that include residualized specialized surprisal. In the first analysis, we aim to answer the question whether residualized surprisal can predict reading times of experts better than the generic surprisal model. We therefore include it in the model as a main effect[4] and perform model comparison in order to test whether the inclusion of this predictor significantly improves model fit. Secondly, we would like to test whether specialized surprisal estimates are particularly strongly weighted for experts reading terminology from their own field. We therefore fit a more complex model which includes a three-way interaction between residualized specialized surprisal, expertise and terminology. We regress reading time onto the same covariate predictors as before.[5] We also included a three-way interaction between general surprisal, expertise and terminology.

[3] Linear regression model: SpecSurprisal ~ GenSurprisal.

[4] Linear mixed-effects regression model: RT ~ WordLen + WordIndexInSentence + Expertise * Terminology + GenSurprisal + ResidSpecSurprisal + (1 | SubjectID) + (1 + Expertise | WordID).

[5] Linear mixed-effects regression model: RT ~ WordLen + WordIndexInSentence + Expertise * Terminology * GenSurprisal + Expertise * Terminology * ResidSpecSurprisal + (1 | SubjectID) + (1 + Expertise | WordID).

Table 5. Regression coefficients and test statistics from the linear mixed-effects including residualized specialized surprisal as a main effect. For significance notation, see Table 4.

	FPRT				TRT			
Variable	β	SE	t		β	SE	t	
(Intercept)	5.62	0.02	348.12	***	6.31	0.03	190.03	***
Word length	0.04	0.001	29.95	***	0.06	0.002	27.62	***
Word freq					-0.008	0.003	-2.95	***
Word index in sentence	-0.002	0.001	-3.84	***	-0.005	0.001	-7.55	***
Expertise	-0.02	0.002	-9.48	***	-0.08	0.003	-30.72	***
Terminology	0.05	0.01	7.17	***	0.08	0.01	9.07	***
GenSurprisal	0.02	0.001	11.62	***	0.03	0.002	16.45	***
ResidSpecSurprisal	0.01	0.002	3.82	***	0.01	0.003	4.81	***
Expertise:Terminology	-0.01	0.002	-6.84	***	-0.02	0.002	-8.06	***

5.2. Results

In our first analysis, we found that residualized specifialized surprisal significantly improves model fit: for first-pass time ($\chi^2(1) = 14.46$, $p < 0.001$) and for total reading time ($\chi^2(1) = 23.00$, $p < 0.0001$). Table 5 presents the coefficients, standard error and test statistics of the LMER models that include residualized specialized surprisal. Residualized specialized surprisal has a significant positive effect on both reading measures. The effects of terminology, expertise and the two-way interaction between expertise and terminology is still significant in this model, indicating that the specialized surprisal does not perfectly account for all of the reading time differences between experts and non-experts reading these texts.

In our second analysis, we added the three-way interaction term between expertise, terminology and residualized specialized surprisal, as well as the three-way interaction between expertise, terminology and general surprisal. The results of these models are presented in Table 6 and show that the three-way interaction with residualized specialized surprisal is statistically significant for both the first-pass and the total reading time model. This indicates that residualized specialized surprisal estimates are weighted more into the model for the data points where experts are reading terminology from their own domain. We also found a significant positive interaction between terminology and general surprisal, indicating that for domain-specific words, the reading time estimate should be increased in proportion to the general surprisal of the word. To isolate the relation of surprisal sources and expertise given the terminology type, we conduct a follow-up analysis on data split by terminology type (either only "common" or "technical").

For **common words**, we have no strong expectations regarding whether specialized surprisal should help in the model. In the first-pass times of common words, both general (β =0.01, SE=0.002, t=8.42, p<2e-16) and residualized specialized surprisal (β =0.008, SE=0.003, t=2.5, p=0.0128) have significant main effects, but only general surprisal is in significant interaction with expertise (β =-0.002, SE=0.0005, t=-3.00, p=0.003). For total reading times, the results on common words are similar (main effect of general surprisal: β =0.03, SE=0.002, t=11.88, p<2e-16; main effect of residualized specialized surprisal: β =0.01, SE=0.004, t=2.86, p= 0.004). The interactions of both surprisal estimates with expertise are also significant (general surprisal and expertise: β =-0.002, SE=0.0007, t=-2.67, p=0.008; residualized specialized surprisal and expertise β =-0.008, SE=0.003, t=-2.25, p=0.03). The negative interactions indicate that for experts, the reading times of the surprisal models should be less strong than predicted by the main effects, i.e., experts read common words faster in their own domain.

For **technical terms**, out of the two sources of surprisal, only the general one has a significant main effect (β =0.02, SE=0.003, t=6.41, p<2.94e-10) on first reading times. However, there is a significant interaction between residualized specialized surprisal and expertise (β =0.01, SE=0.004, t=2.8, p=0.003); while the interaction between general surprisal and expertise was not found to be statistically significant. For the total times model, general surprisal has a significant main effect (β =0.03, SE=0.003, t=10.88, p<2e-16), but no significant interaction with expertise. The model revealed a significant main effect of residualized specialized surprisal (β =0.02, SE=0.005, t=3.7, p=0002) and its interaction with expertise (β =0.01, SE=0.004, t=2.97, p=0.003092). These models thus indicate that the predictive effect of the surprisal estimates is particularly strong for experts within their own domain.

Table 6. Regression coefficients and test statistics from the linear mixed-effects model including interactions between residualized specialized surprisal, expertise and terminology type. For significance notation, see Table 4.

	FPRT				TRT			
Variable	β	SE	t		β	SE	t	
(Intercept)	5.62	0.02	345.03	***	6.31	0.03	189.02	***
Word length	0.04	0.001	29.89	***	0.06	0.002	26.34	***
Word freq					-0.010	0.003	-3.38	***
Word index in sentence	-0.002	0.001	-3.75	***	-0.005	0.001	-7.53	***
Expertise	-0.02	0.002	-9.68	***	-0.08	0.003	-29.19	***
Terminology	0.05	0.01	6.94	***	0.08	0.01	8.39	***
GenSurprisal	0.02	0.001	11.98	***	0.03	0.002	16.41	***
ResidSpecSurprisal	0.01	0.002	3.63	***	0.01	0.003	4.80	***
Expertise:Terminology	-0.01	0.002	-6.44	***	-0.02	0.003	-6.89	***
Expertise:GenSurprisal	-0.0001	0.0004	-0.30		-0.001	0.001	-1.52	
Terminology:GenSurprisal	0.004	0.001	3.52	***	0.003	0.002	1.98	*
Expertise:ResidSpecSurprisal	0.004	0.002	2.06	*	0.002	0.003	0.66	
Terminology:ResidSpecSurprisal	0.001	0.002	0.38		0.002	0.003	0.71	
Expertise:Terminology:GenSurprisal	0.001	0.0004	3.17	**	0.001	0.001	1.77	
Expertise:Terminology:ResidSpecSurprisal	0.005	0.002	2.29	*	0.009	0.003	3.57	***

6. Discussion

Numerous studies have found empirical evidence that a word's predictability in context (operationalized as surprisal from a LM) indexes language processing load. Our study focuses on how surprisal relates to reading times of readers who vary in their level of background knowledge and familiarity with the text domain.

We first investigated the effect of background knowledge on reading. Literature shows that in the process of reading, the reader's prior knowledge contributes to creating the meaning representation of the text. In the present study of reading expository texts, we found that the readers' background knowledge significantly modulates reading behavior: readers who are more familiar with the text domain read faster. This holds for reading in the first pass, as well as total reading time of a word. Considering terminology, technical terms are generally difficult to read and our study shows that they are especially difficult for readers who are less familiar with the domain.

We also address the question whether surprisal can account for the differences in reading times. We use surprisal estimates from the German GPT-2 model as well as residualized surprisal from models that are domain-adapted on biology and physics text to represent expert readers in the PoTeC dataset. While both sources of surprisal have a main effect on reading times, their contribution with respect to terminology and expertise differ. For the observations of the second analysis, we can conclude that general surprisal relates more to the linguistic features (whether the word is common or technical is indicated by its surprisal). In contrast, residualized specialized surprisal is connected to the reader, their level of familiarity with the domain and how this affects their reading times, and is particularly helpful for explaining reading times of a specialist in their own domain.

While our results do demonstrate that adapting a language model to specific reader properties (such as their domain knowledge) improves the fit to reading times, our findings also open up a lot of questions: how exactly should a model best be adapted to model a group of readers with a specific property (or how should a model best be trained to model a specific reader)? How exactly should the texts for training a model be chosen, and how should we model text comprehension in a specific domain, i.e., how should the general language model be mixed with a purely domain-specific model, i.e. for how long should the model be trained on more domain-specific contents?

In our approach, we continued training of the pre-trained German model strictly on the domain-specific data, but other options could be suitable too, such as interleaving training on domain-specific as well as generic text. Another question opened by our analysis is to what extent do both types of language models fit the reading times of individual readers, i.e., are there expert readers who are particularly well or poorly represented by the surprisal from the specialized model?

7. Conclusion

This study sheds light on the role of background knowledge in reading behavior and on the computational modeling of this role using surprisal theory. Our results show that experts read words faster than novices, especially in case of more difficult, technical terminology. We demonstrate that domain knowledge is an important factor of individual differences in reading behavior. Incorporating background knowledge of subjects informs not only theories of reading, but also supports computational modeling of cognitive load, which can be further applied in assessing text readability and automatic text adaptation given the reader's familiarity with the topic.

There has been little investigation of the similarity between pretrained language models and humans with respect to domain-specific background knowledge. We found that specialized surprisal from a domain-adapted LM improves the fit to reading times, showing that the relation of surprisal and processing load extends beyond the general to specific domains.

On the modeling level, a reader's command of one domain is just one aspect of their linguistic experience, as readers are typically acquainted with multiple domains at different levels. Furthermore, it is pertinent to include factors that have shown to affect how an individual might experience reading difficulty, for example working memory capacity [17, 9] or general reading skill [34].

In future work, we want to continue investigating the impact of background knowledge on eye movement beyond word-level analysis using clustering methods over scanpaths (see [14, 28]). In terms of broad-coverage simulation of human reading behavior, fine-tuning a LM on a single domain is not a very sustainable design choice. In the future, we plan to explore other techniques where the base LM stays in tact, while relevant weights in final layers are adapted to the given domain or task [12, 25].

Acknowledgements

The work reported in this paper has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102. The authors thank Lena A. Jäger and David R. Reich for providing additional information about the PoTeC corpus. The authors also thank Marius Mosbach for technical discussion on GPT-2 training, and Merel C. J. Scholman, Marian Marchal and Marjolein van Os for their useful comments on the analysis and results.

Appendix A. Training of specialized language models

A.1. Hyperparameter settings

The German GPT-2 model was further trained using the causal language modeling objective. For training both models the following hyperparameter setting were used: training and evaluation batch size of 8, learning rate of 1e-05, Adam optimizer with betas set to 0.9 and 0.999 and epsilon to 1e-08, and a linear learning rate scheduler. BioGPT-2 was trained with 10 warmup and 80k training steps. PhysGPT-2 was trained with 100 warmup and 50k training steps.

References

[1] Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67, 1–48.
[2] Demberg, V., Keller, F., 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. Cognition 109, 193–210.
[3] Demberg, V., Keller, F., Koller, A., 2013. Incremental, predictive parsing with psycholinguistically motivated Tree-Adjoining Grammar. Computational Linguistics 39, 1025–1066.
[4] Fernandez Monsalve, I., Frank, S.L., Vigliocco, G., 2012. Lexical surprisal as a general predictor of reading time, in: Proceedings of the 13th EACL, ACL, Avignon, France. pp. 398–408.
[5] Frank, S.L., Bod, R., 2011. Insensitivity of the human sentence-processing system to hierarchical structure. Psychological Science 22, 829–834.
[6] Frank, S.L., Otten, L.J., Galli, G., Vigliocco, G., 2013. Word surprisal predicts n400 amplitude during reading, in: Proceedings of the 51st ACL, ACL, Sofia, Bulgaria. pp. 878–883.

[7] Goodkind, A., Bicknell, K., 2018. Predictive power of word surprisal for reading times is a linear function of language model quality, in: Proceedings of the 8th CMCL 2018, ACL, Salt Lake City, Utah. pp. 10–18.
[8] Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A., 2020. Don't stop pretraining: Adapt language models to domains and tasks, in: Proceedings of the 58th ACL, ACL, Online. pp. 8342–8360.
[9] Hahn, M., Degen, J., Futrell, R., 2021. Modeling word and morpheme order in natural language as an efficient tradeoff of memory and surprisal. Psychological Review 128, 726–756.
[10] Hale, J., 2001. A probabilistic Earley parser as a psycholinguistic model, in: Second Meeting of the NACL.
[11] Heister, J., Würzner, K., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., Kliegl, R., 2011. dlexDB. Psychologische Rundschau 62, 10–20.
[12] Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S., 2019. Parameter-efficient transfer learning for nlp., in: Chaudhuri, K., Salakhutdinov, R. (Eds.), ICML, PMLR. pp. 2790–2799.
[13] HuggingFace, 2021. German GPT-2 model. URL: https://huggingface.co/dbmdz/german-gpt2. Accessed: 2022-12-18.
[14] Hyönä, J., Lorch, R., Kaakinen, J., 2002. Individual differences in reading to summarize expository text: Evidence from eye fixation patterns. Journal of Educational Psychology - J EDUC PSYCHOL 94.
[15] Jäger, L.A., Kern, T., Patrick, H., 2021. Potsdam Textbook Corpus (PoTeC): Eye tracking data from experts and non-experts reading scientific texts. URL: https://osf.io/dn5hp/.
[16] Just, M.A., Carpenter, P.A., 1980. A theory of reading: from eye fixations to comprehension. Psychological review 87 4, 329–54.
[17] Kaakinen, J.K., Hyönä, J., Keenan, J.M., 2003. How prior knowledge, wmc, and relevance of information affect eye fixations in expository text. Journal of experimental psychology. Learning, memory, and cognition 29 3, 447–57.
[18] Kendeou, P., van den Broek, P., 2007. The effects of prior knowledge and text structure on comprehension processes during reading of scientific texts. Memory & Cognition 35, 1567–1577.
[19] Kendeou, P., Rapp, D.N., van den Broek, P., 2004. The influence of reader's prior knowledge on text comprehension and learning from text.
[20] Kintsch, W., 1998. Comprehension: A paradigm for cognition.
[21] Kintsch, W., van Dijk, T.A., 1978. Toward a model of text comprehension and production. Psychological Review 85, 363–394.
[22] Kuribayashi, T., Oseki, Y., Brassard, A., Inui, K., 2022. Context limitations make neural language models more human-like, in: Proceedings of the 2022 EMNLP, ACL, Abu Dhabi, United Arab Emirates. pp. 10421–10436.
[23] Kuribayashi, T., Oseki, Y., Ito, T., Yoshida, R., Asahara, M., Inui, K., 2021. Lower perplexity is not always human-like, in: Proceedings of the 59th ACL and the 11th International Joint Conference on Natural Language Processing, ACL, Online. pp. 5203–5217.
[24] Levy, R., 2008. Expectation-based syntactic comprehension. Cognition 106, 1126–1177.
[25] Li, M., Gururangan, S., Dettmers, T., Lewis, M., Althoff, T., Smith, N.A., Zettlemoyer, L., 2023. Branch-train-merge: Embarrassingly parallel training of expert language models.
[26] Lowell, R., Morris, R.K., 2014. Word length effects on novel words: Evidence from eye movements. Attention, Perception, & Psychophysics 76, 179–189.
[27] Makowski, S., Jäger, L.A., Abdelwahab, A., Landwehr, N., Scheffer, T., 2019. A discriminative model for identifying readers and assessing text comprehension from eye movements, in: Machine Learning and Knowledge Discovery in Databases, Springer. pp. 209–225.
[28] von der Malsburg, T., Kliegl, R., Vasishth, S., 2015. Determinants of scanpath regularity in reading. Cognitive science 39 7, 1675–703.
[29] Merkx, D., Frank, S.L., 2021. Human sentence processing: Recurrence or attention?, in: Proceedings of CMC, ACL. pp. 12–22.
[30] Michaelov, J.A., Bardolph, M.D., Van Petten, C.K., Bergen, B.K., Coulson, S., 2023. Strong Prediction: Language model surprisal explains multiple N400 effects. Neurobiology of Language , 1–71.
[31] Oh, B.D., Clark, C., Schuler, W., 2021. Surprisal estimators for human reading times need character models, in: Proceedings of the 59th ACL, ACL, Online. pp. 3746–3757.
[32] Oh, B.D., Clark, C., Schuler, W., 2022. Comparison of structural parsers and neural language models as surprisal estimators. Frontiers in AI 5.
[33] Oh, B.D., Schuler, W., 2023. Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times? TACL 11, 336–350.
[34] Ozuru, Y., Dempsey, K.B., McNamara, D.S., 2009. Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. Learning and Instruction 19, 228–242.
[35] R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.
[36] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., 2019. Language models are unsupervised multitask learners.
[37] Rayner, K., 1998. Eye movements in reading and information processing: 20 years of research. Psychological bulletin 124 3, 372–422.
[38] van Schijndel, M., Schuler, W., 2015. Hierarchic syntax improves reading time prediction, in: Proceedings of the 2015 NACL, ACL, Colorado.
[39] Shain, C., 2019. A large-scale study of the effects of word frequency and predictability in naturalistic reading, in: Proceedings of the 2019 Conference of NACL, ACL, Minneapolis, Minnesota. pp. 4086–4094.
[40] Shain, C., Blank, I.A., van Schijndel, M., Fedorenko, E., Schuler, W., 2019. fmri reveals language-specific predictive coding during naturalistic sentence comprehension. Neuropsychologia 138.
[41] Smith, N.J., Levy, R., 2013. The effect of word predictability on reading time is logarithmic. Cognition 128, 302–319.
[42] Smith, R.J., Snow, P.C., Serry, T.A., Hammond, L., 2021. The role of background knowledge in reading comprehension: A critical review. Reading Psychology 42, 214 – 240.
[43] Tapiero, I., 2007. Situation models and levels of coherence: Toward a definition of comprehension.
[44] Wilcox, E.G., Gauthier, J., Hu, J., Qian, P., Levy, R.P., 2020. On the predictive power of neural language models for human real-time comprehension behavior. ArXiv abs/2006.01912.