Katja I. Haeuser* and Jutta Kray

Not so SUBTLE(X): word frequency estimates and their fit to sentential reading times in interaction with predictability

https://doi.org/10.1515/ling-2024-0143
Received August 7, 2024; accepted May 20, 2025; published online July 28, 2025

Abstract: Frequency and predictability are two prominent psycholinguistic variables that determine the ease of word comprehension and have informed models of language processing. Here, we pooled the data from five self-paced reading studies to investigate (1) the usefulness of three well-known frequency databases of German in accounting for word reading times in context (i.e., the SUBTLEX-DE, CELEX, and dlexDB databases), and (2) whether frequency and predictability have additive or interactive effects on lexical processing. Regarding (1), goodness of fit comparisons between the three frequency measures showed that, in the majority of models, dlexDB frequencies performed best (in contrast to earlier investigations recommending to use SUBTLEX), even though nearly all frequency effects were statistically invariant and dwarfed by the contributions of other more potent variables such as predictability or trial number. Regarding (2), we found that, even though predictability influenced reading times, there was no evidence for interactive effects of frequency and predictability. Our results call into question the current default practice in many psycholinguistic studies to rely on subtitle norms when it comes to estimating lexical frequencies, but they also suggest that frequency effects may be negligible in paradigms which promote contextual word-by-word reading. Our findings are more in line with modular models of language comprehension in which lexical access operates independently from contextual predictability.

Keywords: frequency; predictability; self-paced reading; surprisal theory

1 Introduction

Psycholinguistic research is often concerned with the question of what makes words easier to process. Two important variables that impact word and sentence processing

***Corresponding author: Katja I. Haeuser**, Department of Psychology, Campus A 1.3., Room 2.13, Saarland University, 66123 Saarbrücken, Germany, E-mail: khaeuser@coli.uni-saarland.de
Jutta Kray, Department of Psychology, Saarland University, Saarbrücken, Germany; and Collaborative Research Center Information Density and Linguistic Encoding (SFB 1102), Saarland University, Saarbrücken, Germany

are frequency and predictability (e.g., Ehrlich and Rayner 1981; Inhoff and Rayner 1986; for reviews, see Kuperbergand Jaeger 2016; Staub 2015, 2024; Van Petten and Luka 2012). Whereas frequency reflects the exposure of a given language user to a particular word by measuring how frequently that word occurs in large collections of written or spoken speech, predictability indexes the likelihood of encountering a word in a particular context. For example, given a sentence context such as *The state of Bavaria in Germany is known for making excellent* …, a continuation such as *beer* is, to many people, more predictable than a continuation such as *pretzels* or *knodels*, even though all three words are relatively plausible continuations given the context. In this paper, we investigate the usefulness of two prominent word frequency databases of German in accounting for human word reading times in context, and we also investigate whether frequency and predictability have joint or separable effects on word processing.

2 Predictors of word comprehension

2.1 Frequency

One of the most important psycholinguistic variables to explain processing times of words is frequency, which is normally measured as the frequency of occurrence of a word given a particular corpus. The usual observation made across many, many studies is that high-frequency words are easier to process than low-frequency words (for review, see Brysbaert et al. 2018; Staub 2015, 2024), and this facilitatory effect of word frequency has been found for a variety of research methods, such as lexical decision (e.g., Brysbaert and New 2009; Brysbaert et al. 2011a; Chen et al. 2018; Heister and Kliegl 2012), word recognition and naming (Schilling et al. 1998), eye-tracking during reading where frequency influences even the earliest stages of word processing such as lexical access (e.g., Inhoff and Rayner 1986; Kliegl et al. 2004; Rayner et al. 2004; Whitford and Titone 2014), and self-paced reading, as in the present study (e.g., Kapteijns and Hintz 2021; Tremblay et al. 2011). For the German language, a variety of frequency measures have been published so far (for review, see Brysbaert et al. 2011a; Heister and Kliegl 2012); three that are critical in the present study are the CELEX, dlexDB, and SUBTLEX corpora. CELEX is a corpus dating back to the mid-1990s that consists of 5.4 million word tokens derived from written text (i.e., newspapers and books) and 600 thousand word tokens derived from transcribed speech (Baayen et al. 1995). Despite its age, CELEX continues to be used by state-of-the-art research (see e.g., Crossley et al. 2014; Mickan et al. 2024; Petilli and Marelli 2024), which is likely a result of the long-standing prominence of this corpus (see Heister et al. 2011, for discussion). DlexDB, in turn (see Heister et al. 2011), is a more recent compilation and relies on the corpus compiled by DWDS (i.e., *Digitales Wörterbuch der* *deutschen Sprache*; see Geyken 2007). DlexDB consists of 122,816,010 tokens and is, in comparison to CELEX, somewhat less restricted in terms of genre as it comprises texts from books, newspaper articles, transcribed speech from all over the 20th century. Finally, the SUBTLEX database is a corpus of 25.4 million words that was derived from the subtitles of 4,610 movies and TV series (i.e., German and non-German TV material; see Brysbaert et al. 2011a; also see Heister and Kliegl 2012, for discussion).

Given the importance of word frequency in language processing, an obvious question is which one of the three corpora – i.e., CELEX, dlexDB, or SUBTLEX – psycholinguists should rely on. To date, several studies have obtained independent evidence that the SUBTLEX frequencies tend to outperform other corpora when it comes to accounting for word processing times (e.g., Brysbaert et al. 2011a, 2018; Chen et al. 2018; Heister and Kliegl 2012). For example, Brysbaert et al. (2011a) compared the goodness of fit of the SUBTLEX and the CELEX corpora (among others) in predicting lexical decision times of German words presented in isolation. They relied on previously published data from three lexical decision experiments where participants were mainly psychology students from various German universities. Goodness of fit was operationalized as the magnitude of the correlation coefficient between the various frequency measures and lexical decision times, such that higher correlation coefficients were deemed better than lower ones, as they reflect a greater amount of explained variance. According to the results, SUBTLEX outperformed the other frequency databases across all three experiments. According to the authors, the superiority of the SUBTLEX corpus compared to other corpora might be related to the type of register that is normally used in movies and TV shows, compared to written sources such as newspapers or books. Psychology students may have had less exposure to this latter type of register, by watching more television and spending less time reading (also see Brysbaert et al. 2011b). Similar findings were obtained in a subsequent study by Heister and Kliegl (2012), who compared the goodness of fit of several corpora in accounting for lexical decision times, by additionally taking into account the emotional valence of words. The authors found that corpora that include large amount of words high in emotional valence (such as movie subtitle corpora or corpora derived from tabloid newspapers such as the German newspaper *Bild*) outperformed others in accounting for lexical decision times. The authors explain this result by arguing that emotional words are processed faster than neutral ones, yielding larger correlation coefficients with response times. However, another intuitive reason for the superiority of the subtitle corpus could simply lie in the fact that both studies relied on previously published lexical decision data that, at least partially, consisted of word lists high in emotional valence (see Brysbaert et al. 2011a: 417). It seems somewhat intuitive that a frequency corpus based on emotional and dramatic content (as in movies and TV shows) would explain more variance in emotional rather than neutral word lists (see Baayen et al. 2016;

Heister and Kliegl 2012, for discussion). As a case in point, Baayen et al. (2016) showed that subtitle frequencies perform worse in predicting outcomes of tasks other than lexical decision (in this case, eye-tracking during reading), simply because lexical decision promotes the impact of variables highlighted in subtitle frequencies.

An open question is therefore not only how various frequency databases (here CELEX, dlexDB, and SUBTLEX) perform when using neutral words as opposed to emotional ones, but also, how they perform in psycholinguistic paradigms other than lexical decision that likely promote different processing strategies. For example, lexical decision tasks present words in isolation, which may promote rather wholistic, one-on-one, recognition of words, where frequency might have a strong impact (e.g., Barton et al. 2014). Sentential reading paradigms such as used in the present study emphasize serial processing, due to the requirement of recognizing words in a broader linguistic context, which may down-regulate the impact of frequency effects (e.g., Kretzschmar et al. 2015).

2.2 Predictability

Predictability is a second prominent measure that explains the efficiency of language processing. Predictability is often assessed by means of the cloze procedure in which participants are asked to complete sentence frames that are truncated before the final word (e.g., Taylor 1953). The cloze probability of a word then reflects the proportion of participants who completed the sentence with that word. A robust body of research has shown that high-predictability words are processed more quickly and effortlessly than low-predictability words, and across a variety of research methods (for reviews, see Kuperberg and Jaeger 2016; Kutas and Federmeier 2011; Staub 2015; Van Petten and Luka 2012, among others).

An empirical question in psycholinguistic research is whether frequency and predictability affect word processing independently from one other or jointly (e.g., Kretzschmar et al. 2015; Shain 2019; Staub 2015; Whitford and Titone 2014). This question is not only relevant for models of eye-movement control during reading (e.g., Engbert et al. 2005; Reichle et al. 2009), which differ from one another in how strongly they advocate additive versus interactive effects of frequency and predictability. This question is also relevant to psycholinguistic models of word recognition, which emphasize either context-independent lexical access mechanisms (e.g., Coltheart et al. 2001; Seidenberg and McClelland 1989), advocating separable effects of frequency and predictability, or unified comprehension that depends entirely on the incremental probability of a linguistic structure, subsuming frequency and predictability in a single processing stage (e.g., Hale 2001; Levy 2008; Norris 2006; for dissemination, see Shain 2024; Staub 2024).

If predictability and frequency interact, an intuitive hypothesis could be that predictability effects should be more pronounced for low-frequency words, as high-frequency words should have strong levels of activation in mental lexicon anyways (e.g., Kretzschmar et al. 2015; Shain 2024; Staub 2024). Crucially, this hypothesis is not supported by empirical results obtained in previous eye-tracking studies (e.g., Altarriba et al. 1996; Ashby et al. 2005; Bélanger and Rayner 2013; Kliegl et al. 2004; Rayner et al. 2004; Whitford and Titone 2014; for review, see Staub 2015). For example, Rayner et al. (2004) investigated fixations made on high- and low-frequency words presented in sentential contexts that either rendered the words predictable or unpredictable. Across eye-movement measures, the authors found no statistical evidence for interactive effects of frequency and predictability: even though fixations were shorter for high-compared to low-frequency words, predictability reduced fixation durations to a similar extent for high- and low-frequency words. Even though the absence of a statistical effect does not constitute evidence for the null hypothesis (see Wagenmakers et al. 2007, for dissemination), it is striking to observe that, out of the numerous eye-tracking studies that have examined joint effects of predictability and frequency over the years, none have observed interaction effects between predictability and frequency, which would be expected based on Type I error probability alone (see Kretzschmar et al. 2015). Nevertheless, many of the eye-tracking studies that factorially manipulated predictability and frequency were likely underpowered: for conventional ANOVA analyses, a 2 × 2 design requires a total of 84 subjects to detect a statistical interaction with 80 % power. Few if none of the previous reading time studies included such a sample size.

3 The present study

Our contributions in the present study were twofold: our first goal was to investigate the usefulness of the three frequency databases (SUBTLEX, DELEX, and dlexDB) in predicting human processing times using a task that emphasizes serial word processing in context, a self-paced reading task. Previous studies had investigated the performance of the various frequency databases exclusively by means of lexical decision tasks, using sets of emotionally valenced words, and relying on measures obtained from psychology students. Our goal here was to test the validity of the three frequency databases using a different task, a more neutral set of words, and a more diverse sample.

Our second goal was to investigate potentially interactive versus additive effects of frequency and predictability. Additive effects would support models of language processing that argue for a staged processing architecture where lexical access proceeds independently of contextual predictability (e.g., Coltheart et al. 2001;

Seidenberg and McClelland 1989). Interactive effects of frequency and predictability would support unified processing models that subsume frequency and predictability in one single processing stage (e.g., Levy 2008; Norris 2006; see Shain 2024; Staub 2024).

4 Methods

4.1 Participants

We pooled the data from five self-paced reading studies conducted in our lab. The final sample of participants included 329 native German-speaking adults (*mean age* = 26 years, range = 18–41; 184 female, 139 male, 3 non-binary), who participated for financial compensation (participants recruited through Prolific, *n* = 171) or course credit (participants recruited through the university's study recruitment website, *n* = 155). Participants recruited through the university's study recruitment website were mostly first- and second-year psychology undergraduates. The majority of participants recruited over Prolific were not students but worked in part-time or full-time jobs (*n* = 87). Sixteen Prolific participants were unemployed/job-seeking; 46 Prolific participants had not provided information on employment status. Hence, our participants were not limited to psychology students, and occupational background varied among participants. All participants had normal or corrected-to-normal vision and reported no neuropsychiatric medication and/or a history of language impairments at the time of testing.

4.2 Materials

Materials consisted of 46 constraining German sentence frames (e.g., *Unser freundlicher Nachbar mähte für uns kürzlich den…*, English translation: 'Our friendly neighbor mowed for us recently the…'), which were continued with a predictable or unpredictable noun (e.g., *Rasen*, 'lawn'; vs. *Hof*, 'courtyard'). Each sentence was concluded by a sentence continuation (identical for predictable and unpredictable versions), which was inserted to allow for spill-over effects after the noun, common to self-paced reading (Keating and Jegerski 2015; Witzel et al. 2012; e.g., *auf dem Grundstück nebenan*, 'of the property next door'). Examples of experimental sentences and critical regions are presented in Table 1.

Noun predictability was assessed by means of a cloze pre-test in which 40 psychology students who did not participate in the main experiment (age range: 18–26, 26 female, 14 male), were asked to complete each sentence frame with the first word that came to mind. According to the results of the cloze test, predictable nouns had a

Table 1: Examples of stimuli and critical-region words.

Example sentence	Critical region		
	Noun	Spill 1	Spill 2
Da Anne Angst vor Spinnen hat, geht sie bei sich zuhause nur ungern nach unten in den	*Keller/Garten*	*ihrer*	*Eltern*
Since Anne is afraid of spiders, she doesn't like going down into the	basement/garden	(by) her	parents
Unser freundlicher Nachbar mähte für uns kürzlich den	*Rasen/Hof*	*auf*	*dem*
Our friendly neighbor mowed for us recently the	lawn/yard	on	the

cloze probability of 0.78 (*Range* = 0.29–1.00). Unpredictable nouns had a cloze probability of zero throughout (i.e., no unpredictable noun was ever produced as a response in the cloze test). All sentential materials are accessible under this paper's OSF project using the link, https://osf.io/zn4am/?view_only=1e69f68799e143be9b40dc1f2f83b719.

4.2.1 Frequency estimates

Frequency estimates for each word in the critical region were taken from three corpora: SUBTLEX-DE, CELEX, and dlexDB.[1] The critical region consisted of the predictable/unpredictable noun, as well as the two words following the noun (i.e., the spill-over region; see Table 1). For each word, we added to the data the word form frequency estimates from all three corpora. Common problems when comparing frequency estimates from different corpora are differences in corpus size and missing values. In order to address these, we converted the per-million frequency estimates to Zipf values (e.g., Brysbaert and Diependale 2013; Carroll 1970; Chen et al. 2018; Heuven et al. 2014; Zipf 1949; see). Zipf values are useful because they quantify frequency estimates on a common scale (i.e., 1–7, with 1 being very highly infrequent and 7 being extremely frequent), while also taking into account missing observations in the corpus. To convert to Zipf values, we used the formula, log10[$(f + 1)/N$] + 3, where f refers to the raw frequency count of a word in a corpus and N to the size of the corpus (in millions).[2] Figure 1 shows histograms of Zipf frequencies for each word

[1] We note that the dlexDB database interface has been archived and is no longer available. Therefore, we extracted the dlexDB frequencies from the DWDS website using the link, https://www.dwds.de/r/lexdb.

[2] Given the small size of the CELEX corpus, the Zipf frequency estimates for this corpus may be noisier than the estimates from the other two corpora. We thank our reviewer Harald Baayen for pointing this out. Note that we replicated all findings reported below when using log-per-million estimates instead of Zipf values.

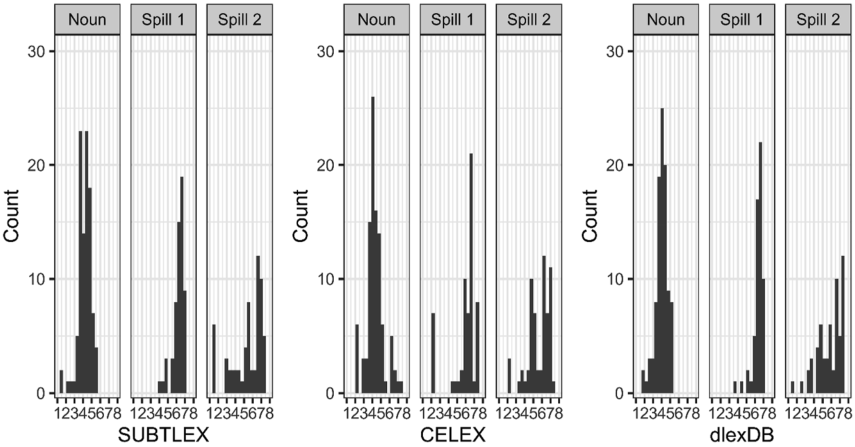

Figure 1: Histograms of Zipf frequency values in SUBTLEX-DE, CELEX, and dlexDB, split out by each word in the critical region.

in the critical region, split out by SUBTLEX-DE, CELEX, and dlexDB. As can be gleaned from the figure, the frequency estimates in the three corpora were similar, but not identical, which underscores the relevance of our research question.

4.3 Procedure

We pooled the data from five online self-paced reading experiments which we ran in the past; all experiments were run using the experimental presentation software *LabVanced* (Finger et al. 2017). The main task was a word-by-word self-paced reading task (no moving-window reading; no mask). Comprehension questions that followed after one third of the sentences ensured that participants were reading for comprehension. Participants saw one word appear in the center of the screen at a time and pressed the space bar to reveal the next word. They were instructed to read all sentences as quickly and accurately as possible, and to answer all true/false comprehension questions as accurately as possible by pushing the "J" (Yes, correct) and "N" (No, incorrect) keys on the keyboard. Sentences were separated by a 500 ms fixation cross. Each experiment also included a minimum of 30 moderately predictable filler sentences (taken from the Potsdam sentence corpus or earlier studies in our lab). The filler sentences were added to make sure that participants continued to generate predictions in the course of the experiment, despite having predictions disconfirmed multiple times (e.g., Delaney-Busch et al. 2019; Fine et al. 2013; Kukona and Hasshim 2024).

5 Results

Five subjects were excluded from further analysis because their accuracy rate on the comprehension questions was below 70 %. The remaining participants understood the sentences very accurately (*Mean* = 92 %; *Range* = 73–100 %). We take this to mean that participants were reading the sentences attentively. Prior to analysis, and based on visual inspection of the data, we trimmed the reading time (RT) data minimally by excluding reading times faster than 100 ms and slower than 3,000 ms (for similar outlier criteria, see e.g., Linzen and Jaeger 2016). This procedure affected less than 1 % of all data points.

We report our results in two separate sections. The first section addresses research question one, i.e., RQ (1) Which one of the three frequency corpora performs best in predicting reading times in context? The second section addresses research question two, i.e., RQ (2) Is there evidence for interactive effects of frequency and predictability? All analysis scripts are accessible under this paper's OSF project using the link, https://osf.io/zn4am/?view_only=1e69f68799e143be9b40dc1f2f83b719.

5.1 Research Question 1: Which frequency corpus performs best in predicting reading times in context?

5.1.1 Approach to analysis

To address research question one, we compared – for each word in the critical region, i.e., the noun, as well as the two words in the spill-over region – the goodness of fit of three critical models: a first model that specified the SUBTLEX frequency estimates, a second model that specified the CELEX frequency estimates, and a third model that specified the dlexDB frequency estimates. The idea of this approach was to keep the models as similar as possible but vary only the frequency estimates.

To analyze the data statistically, we used generalized additive mixed models (GAMMs; e.g., Baayen et al. 2017; Chuang et al. 2021; Wieling 2018; Wood 2017) in R (R Core Team 2021). The advantage of using GAMMs over canonical LMERs is that they allow one to account for non-linear effects (i.e., smooth terms) and trial-by-trial dependencies through autocorrelation parameters (i.e., accounting for the fact that the RTs at time *t* cannot be completely independent from the RTs at time *t-1*). Note that, here, we used *bam()* instead of *gam()*, because bam works more efficiently with large data sets at minimum cost of accuracy. All models were run with the directive discrete set to TRUE, such that covariates were binned in a mathematically principled way to enable faster estimation of model coefficients (e.g., Chuang et al. 2021; Wieling 2018).

In all models, the outcome variable was word-by-word reading times, transformed using the optimal transformation to approximate normality of residuals (see below). The predictor variable consisted of scaled Zipf frequency (i.e., SUBTLEX, CELEX, or dlexDB frequency; see above for conversion to Zipf values). Each model additionally specified a control predictor for predictability[3] (two levels, predictable and unpredictable, dummy-coded with 1 and 0), scaled word length, and orthographic neighborhood size. For other control predictors, we conducted initial model comparisons to check whether the inclusion of variables such as target sample (Prolific vs. student) and scaled word position was warranted. The results across models showed robust improvements in model fit for the inclusion of target sample (as a main effect, and, whenever necessary, as an interaction term with frequency; see model reports below). For word position, no improvements in model fit were found, so we dropped this variable. Note that we could not include valence as additional predictor here, as only one third of all lexical items in our investigation had equivalents in a German valence database (i.e., Kanske and Kotz 2010; we return to this point in the discussion). Random effects consisted of random intercepts for subjects and items, as well as a by-subject random-slope adjustment for smoothed trial number. We also added a smooth term for trial by target sample, in order to allow for the possibility that trial effects may vary in a non-linear fashion depending on sample. The autocorrelation parameter was set to *rho* = 0.07. For example, the syntax for the best-fitting model for nouns was,

```
bam(

boxcox(RT) ~ scale(frequency) + predictability + scale(length) + sample + scale(frequency):sample + s(trial, by = sample) + s(subject,

…trial, bs = “re”) + s(subject, bs = “re”) + s(item,

…bs = “re”), AR.start = AR.start, rho = 0.07, data = dat

).
```

Table 2 shows beta-coefficients and *t*-values for the parametric (i.e., linear) predictors in each model; for full model outputs, including s-tables, the reader is referred to the OSF site of this article.

Goodness of model fit was assessed by means of Akaike information criterion (AIC). AIC is a measure of fit that penalizes a model for having more variables. Larger

[3] Note that we ran follow-up models that included a non-linear effect of frequency by condition, but found that these models did not consistently improve model fit. We therefore dropped this term from the models.

Table 2: Parametric coefficients and *t*-values for SUBTLEX, CELEX, and dlexDB models.

	SUBTLEX		CELEX		DlexDB	
	b	***t***	***b***	***t***	***b***	***t***
Noun						
Frequency	0.009679	0.066	0.14761	1.279	−0.00523	−0.032
Predictability	−0.523889	−3.079	−0.57747	−3.442	−0.54403	−3.222
Group	5.245676	3.623	5.24794	3.625	5.24920	3.626
Length	0.869562	5.077	0.78256	4.215	0.88161	5.160
N-Size	0.214548	1.381	0.07297	0.457	0.21633	1.348
Frequency:group	−0.360929	−2.104	–	−0.28920	−1.587	
Noun+1						
Frequency	−0.1929	−1.128	−0.09737	−0.815	−0.05115	−0.276
Predictability	−0.7026	−6.133	−0.70089	−6.118	−0.70170	−6.125
Group	3.3878	3.653	3.38648	3.651	3.38748	3.653
Length	0.4143	1.534	0.53543	2.284	0.54377	2.070
N-Size	0.4485	1.957	0.48443	2.106	0.46425	2.016
Frequency:group	0.2101	1.814	–	0.22838	1.985	
Noun+2						
Frequency	−1.1193	−3.485	0.4560	1.356	−2.2160	−6.554
Predictability	−0.9499	−8.174	−0.9492	−8.164	−0.9448	−8.140
Group	3.6685	3.951	3.6810	3.963	3.66	3.95
N-Size	0.4425	1.461	−0.7986	−3.054	1.03	3.80
Length	–	–	–			
Frequency:group	–	–	–			

values indicate worse fit, corrected for the number of variables (Winter 2019). The critical AIC values per model for each word in the critical region are presented in Table 3; best-fitting models are highlighted in bold. Note that, in order to complement the AIC model comparisons with another measure of goodness of fit (i.e., log likelihood), we ran follow-up models in which we aligned the number of predictors per model. The results of these other models converged with our main findings and are not reported here.

5.1.2 Results

Noun. Based on visual inspection of the data, noun reading times were boxcox-transformed and multiplied by –1,000, to improve legibility of model outputs (Baayen and Milin 2010).

Table 3: AIC of best-fitting models using SUBTLEX, CELEX, and dlexDB frequencies for words in the critical region.

	Noun	Noun+1	Noun+2
SUBTLEX	**75,061.45**	68,547.82	64,120.80
CELEX	75,066.21	68,549.04	64,129.79
DlexDB	75,064.8	**68,546.75**	**64,094.54**

Best-fitting models appear in bold.

The SUBTLEX model fit the data best (see Table 3). In none of the three models, the frequency predictor explained a substantial amount of variance (all *p*'s > 0.05; see Table 2). The predictability effect was significant in all models and suggested slowed reading for unpredictable compared to predictable words (all *p*'s < 0.01; see Table 2).

Noun+1. Based on visual inspection of the data, the reading time data of the first spill-over word were square-root transformed and multiplied by –1,000.

The dlexDB model fit the data best (see Table 3). None of the three models suggested that frequency explains a substantial amount of variance in the data (all *p*'s > 0.1; see Table 2). Unpredictable words slowed reading down, compared to predictable ones, as was suggested by significant effects of predictability across models (all *p*'s < 0.001, see Table 2).

Noun+2. The reading time data of the second spill-over word were square-root transformed and multiplied by –1,000. The predictor *length* was dropped from this model, as it led to multicollinearity with frequency.

The dlexDB model fit the data best (see Table 3). Unlike in the noun and noun+1 models, the models for this region showed significant effects of frequency throughout (see Table 2; all *p*'s < 0.05). Again, predictability had a substantial effect on reading times, in that unpredictable words slowed down reading comprehension (all *p*'s < 0.1).

5.2 Research Question 2: Is there evidence for interactive effects of frequency and predictability?

5.2.1 Approach to analysis

To address the question if frequency and predictability interact, we assessed, for each word in the critical region, whether adding to the model the interaction between frequency and predictability significantly improved model fit, compared to a base model that did not include the interaction. To this end, we chose to proceed with the best-fitting models of word frequency, i.e., the SUBTLEX model for the noun region, and the dlexDB models for the noun+1 and noun+2 regions. However, we note that we obtained qualitatively similar findings when running models that

Table 4: AIC for base and interaction models.

	Noun	Noun+1	Noun+2
Base	**75,061.45**	**68,546.75**	**64,094.54**
Interaction	75,063.15	68,546.78	64,097.58

Better-fitting models appear in bold.

included the other frequency predictors. Table 4 details AIC values for base and interaction models; best-fitting models are highlighted in bold.

5.2.2 Results

The results are easily summarized. In none of the models did we obtain evidence supporting the hypothesis that adding the interaction between predictability and frequency substantially improves model fit (see Table 4). Similarly, in none of the models did the interaction between frequency and predictability reach statistical significance (all *p*'s > 0.10).

6 Discussion

Frequency estimates are the bread and butter of psycholinguistic experiments. Even paradigms that do not directly investigate frequency effects in language processing need to control, or at least statistically account for, basic processing differences based on word frequency. In research question 1 of this paper (RQ 1), we investigated the usefulness of three prominent frequency databases of German in accounting for reading times of words in context: the SUBTLEX, CELEX, and dlexDB databases (Baayen et al. 1995; Brysbaert et al. 2011a). Previous studies had reported that subtitle frequencies (as in SUBTLEX; e.g., Brysbaert et al. 2011a; Chen et al. 2018; Heister and Kliegl 2012) perform best in estimating human processing times, but these studies relied on previously published lexical decision time data of psychology students and used non-canonical stimuli. Here, we aimed to include a more norm-typical set of words and a more diverse participant sample. Our second research goal in this study was to garner empirical evidence for additive versus interactive effects of frequency and predictability during word-by-word reading. This question is relevant for models of language processing, which diverge in their predictions regarding whether lexical access of words proceeds context-independently (e.g., Coltheart et al. 2001; Seidenberg and McClelland 1989), or is modulated by both frequency and predictability as they both determine the processing difficulty of a word (e.g., Hale 2001; Levy 2008).

In order to address RQ (1), we compared the goodness of fit for models that differed in the frequency measure they specified, i.e., the SUBTLEX-DE, dlexDB, and CELEX frequency. Our results are less consistent than previous research but they show that, in the majority of models, dlexDB – and not SUBTLEX – performs best in estimating reading times in context. This aspect of our results is not in line with previous research utilizing lexical decision tasks to show that SUBTLEX frequencies consistently outperform other frequency measures. That being said, in none of the models presented in our analysis (except for one model, the noun+2 model), word frequency explained a substantial amount of variance in the data. In addition, supplementary analyses presented in the appendix showed that, compared to a base model which did not contain frequency, adding frequency as predictor did not consistently improve model fit for either SUBTLEX, CELEX, or dlexDB. This could indicate that predictors other than word frequency are more potent in explaining word reading times in context, for example predictability, which had substantial facilitatory effect on reading times in all models (see Table 2).

In order to address RQ (2), we investigated whether predictability and frequency interact in modulating reading times. The answer is a clear "No". In none of the models did we find evidence to support the idea that adding the interaction between frequency and predictability significantly improved the goodness of fit of the model. This aspect of our findings mirrors past research showing that predictability and frequency likely impact reading times at distinct processing stages, and independently from one another. Hence, frequency and predictability seem to have additive, rather than interactive, effects, which supports modular models of language processing arguing that lexical access and contextual integration proceed independently from one another. We discuss our findings in greater detail below.

6.1 Estimating frequency effects in psycholinguistic research

The results of the present study could not substantiate the superiority of subtitle frequencies that previous research has highlighted (e.g., Brysbaert et al. 2011a, 2018; Chen et al. 2018; Heister and Kliegl 2012). Instead, in the majority of our analyses on critical-region words (i.e., the noun+1 and noun+2 analyses, but not the noun analysis, see below), models that included the dlexDB frequencies had the best fit to the data. We believe that the primary reason for these discrepant findings lies in the methodology that was used here. The present study relied on self-paced reading of words in context, whereas most previous investigations had used lexical decision tasks. We know that lexical decision likely promotes wholistic rather than serial processing of words (e.g., Barton et al. 2014), and that wholistic processing is more likely to be impacted by form-related variables such as lexical frequency, but also valence, a variable the present study could not include (see Method section). It seems trivial that frequency databases such as SUBTLEX, which are known for overusing sad and happy words rich in such form-related variables (e.g., Baayen et al. 2016), would have the best fit to lexical decision data. Ultimately, the results from our study call into question the current practice of many, many psycholinguistic studies that use subtitle frequencies as their *de facto* default.

In the same vein, the superior performance of the SUBTLEX frequencies for specifically the noun region could be an artifact of the way the SUBTLEX corpus is built. The SUBTLEX corpus is built on large amounts of emotionally valenced words, and the frequency estimates that pertain to these words may simply work best for nouns, as most nouns are concrete, refer to objects in the world, and give the language user some kind of visual or tactile experience. Instead, most words in the N+1 and N+2 region in the present study were prepositions or pronouns – i.e., words for which such haptic experiences do not apply. Indeed, when we ran follow-up models on a subset of our data that only included normal nouns (irrespective of region), we found that SUBTLEX frequencies had the best fit.

The fact that, across models, CELEX performed worst as frequency predictor could be a consequence of the sheer age of this corpus, but it could also be an artifact of its size: CELEX is the smallest corpus in our investigation and, despite the Zipf-normalization approach used here (which attempts to minimize differences in corpus size), the CELEX frequencies may have the noisiest estimates.

That being said, our results also show that, in the majority of our statistical models, word frequency did not account for a substantial amount of variance in the data. In contrast, model predictors that made much more sizable contributions to reading time were contextual predictability or non-parametric variables such as trial number. This aspect of our findings is noteworthy as it mirrors previous work using event-related potentials (ERPs)which has often shown that, even though frequency explains a sizable portion of N400 amplitude when words are presented in isolation (e.g., Grainger et al. 2012), these frequency effects tend to become smaller or even disappear for words processed in sentential context (e.g., Osterhout et al. 1997; Van Petten and Kutas 1990). This shows that a linguistic context can diminish or even override frequency effects during language processing (see Kretzschmar et al. 2015; Staub 2024).

A question that remains open in the context of our statistical analyses is why frequency effects seemed to differ depending on population (i.e., Prolific workers vs. psychology students), at least numerically. This question was not of primary interest to the present study, but some of the models in our analysis showed clear improvements in data fit once the interaction between frequency and population was included. Nevertheless, the results are somewhat puzzling, as they are inconsistent: for nouns, psychology students showed numerically *smaller* frequency effects than

Prolific workers, but on the spill-over region (N+1 word), the reversed pattern surfaced, such that psychology students showed numerically *larger* frequency effects. We would like to refrain from over-interpreting these interactions, as most of them remained statistically non-significant. However, there is now a growing body of research showing that individual experience and expertise shape frequency effects in lexical processing (e.g., Baayen et al. 1996; Halteren et al. 2005; for discussion, see Baayen et al. 2016). This could potentially indicate that some of the interactions that surfaced in our analysis were, nevertheless, meaningful and could be driven by genuine individual differences between these two populations (e.g., educational attainment, socio-economic status, political views, but also internet usage or social media activity and many more). It will take more research with more consistently counterbalanced subject groups to further substantiate these findings. Nevertheless, this aspect of our results underscores the necessity of conducting research on target populations other than psychology students.

6.2 Additive versus interactive effects of frequency and predictability

Results for our second research question showed that frequency and predictability likely make independent contributions to language processing. Our statistical models yielded no improvement in model fit when specifying the interaction between frequency and predictability over and above taking into account both main effects. In the present study, this result is unlikely to be driven by low power, since we used a sufficient amount of both subjects and items (i.e., single words) in order to find an interaction between the two variables. However, our frequency estimates from all three corpora showed relatively low rates of variability relative to the full range of the frequency metric (see Figure 1). Therefore, we cannot fully exclude the possibility that different results would obtain with a more widespread set of frequency distributions. Nevertheless, the additive effects of predictability and frequency found here do match with a large body of previous research manipulating frequency and predictability factorially to show that there are no joint effects of predictability and frequency in eye-tracking during reading; rather, frequency and predictability have additive effects (e.g., Rayner et al. 2004). Additive effects of these variables support models of language processing that assume a staged processing architecture in which lexical properties of incoming input (such as frequency) are evaluated independently of prior context (such as predictability). An avenue for future research will be to pin down the exact time course with which these processes unfold. Classic psycholinguistic models of word processing such as the DRC (dual route cascaded) or the Logogen models (e.g., Coltheart et al. 2001; Morton 1969) assume a strictly serial architecture, in which lexical access precedes contextual or semantic integration, with the latter only happening during later stages of processing. Strictly speaking, this would imply that frequency and predictability effects cannot occur together, i.e., at a single processing stage, in line with the present results. In contrast to this, recent findings from eye-tracking challenge this strictly serial processing view by showing context effects even in very early measures of eye-movement control during reading (e.g., Staub 2015; Staub and Goddard 2019). In the same vein, large language models challenge the modularity view of classic psycholinguistic research as they no longer have a clear distinction between lexicon and syntax. It is difficult to put these results together and draw firm conclusions as to the underlying processing architecture of the cognitive system, but tentatively, they are consistent with a view in which frequency and predictability affect processing in potentially encapsulated modules, but they do so at overlapping time points (i.e., the seriality assumption would need to be revised). This is no doubt an exciting avenue for future research.

6.3 Limitations and conclusions

In the introduction of this paper, we asked a simple question: which frequency database should psycholinguistic research in German predominantly rely on? Given that our results largely show statistically invariant effects of word frequency, and since other predictors were nearly ten times as potent in explaining variability, one may feel tempted to argue that, for word processing in context, the choice of frequency database does not matter much. However, that is a stronger statement than we wish to make.

First, as we argued above, our results may reflect a lack of variance in frequency estimates rather than a genuine indifference between the frequency databases, which means that it will take more studies to answer this question definitively. Second, even if that conclusion is valid, our results are strictly limited to the research method used here, i.e., sentential self-paced reading. Above, we have reviewed evidence suggesting that different research methodologies (i.e., ERPs, eye-tracking) promote diverging processing strategies, i.e., wholistic versus serial processing, which emphasize and down-regulate effects of frequency (e.g., Barton et al. 2014). The self-paced reading paradigm used here likely reflects a hybrid between these two, as it presented single words to subjects at a time, but nevertheless required some degree of serial processing, as words needed to be comprehended within a sentential context. A fruitful avenue for further investigation will therefore be to investigate frequency effects on the same materials using research methodologies that promote more wholistic versus more serial processing (e.g., lexical decision vs. eye-tracking during reading, respectively). Only then will we be able to isolate genuine effects of frequency from those of the research paradigm at hand.

Regarding the choice of frequency database, our results are not in line with previous investigations showing that subtitle frequencies perform better than text-based frequency estimates (Brysbaert et al. 2011a; Chen et al. 2018). In fact, in the majority of our analyses, the dlexDB – and not subtitle – frequencies had the best fit to the data. Above, we have suggested possible explanations for these discrepant findings, e.g., differences in methodology. Ultimately, the findings reported here highlight that recommendations as to the use of one frequency database over another are only warranted when they are supported by a multitude of research methods. Hence, we concur with previous studies (e.g., Baayen et al. 2016; Heister and Kliegl 2012): issuing a blind recommendation to use subtitle frequencies as a default is, indeed, premature.

Acknowledgments: Thanks to Marjolein van Os and Emilia Ellsiepen for extracting the CELEX frequencies, and to Geraldine Müller and Jannick Reugels for helping with reference formatting. Special thanks to our two reviewers and to all members of the PLEAD (Psycholinguistic Experiments and Design) group at Saarland University, for many discussions and helpful feedback that informed the analysis presented in this article.

Author contributions: Conceptualization and data curation: KH. Formal analysis: KH. Funding acquisition: KH. Investigation and methodology: KH. Project administration, resources, software, supervision, validation, visualization: KH. Writing – original draft: KH, Writing – review and editing: KH and JK.

Research funding: This work was funded by the *Deutsche Forschungsgesellschaft* DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102, project A5.

Data availability: All data files, materials, and analysis scripts can be found on this paper's project page using the link, https://osf.io/zn4am/?view_only=1e69f68799e143be9b40dc1f2f83b719.

References

Altarriba, Jeanette, Judith F. Kroll, Alexandra Sholl & Keith Rayner. 1996. The influence of lexical and conceptual constraints on reading mixed-language sentences: Evidence from eye fixations and naming times. *Memory & Cognition* 24(4). 477–492.

Ashby, Jane, Keith Rayner & Clifton Charles. 2005. Eye movements of highly skilled and average readers: Differential effects of frequency and predictability. *The Quarterly Journal of Experimental Psychology Section A* 58(6). 1065–1086.

Baayen, Harald & Petar Milin. 2010. Analyzing reaction times. *International Journal of Psychological Research* 3(2). 12–28.

Baayen, Harald, Petar Milin & Michael Ramscar. 2016. Frequency in lexical processing. *Aphasiology* 30(11). 1174–1220.

Baayen, Harald, Richard Piepenbrock & Huib van Rijn. 1995. *The CELEX lexical database [CD-ROM]*. Philadelphia, PA: Linguistic Data Consortium.

Baayen, Harald, Hans van Halteren & Tweedie Fiona. 1996. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing* 11. 121–131.

Baayen, Harald, Shravan Vasishth, Reinhold Kliegl & Bates Douglas. 2017. The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language* 94. 206–234.

Barton, Jason J. S., Hashim M. Hanif, Laura Eklinder Björnström & Charlotte Hills. 2014. The word-length effect in reading: A review. *Cognitive Neuropsychology* 31(5–6). 378–412.

Bélanger, Nathalie N. & Keith Rayner. 2013. Frequency and predictability effects in eye fixations for skilled and less-skilled deaf readers. *Visual Cognition* 21(4). 477–497.

Brysbaert, Marc, Matthias Buchmeier, Marcus Conrad, Arthur M. Jacobs, Jens Bölte & Andrea Böhl. 2011a. The word frequency effect. *Experimental Psychology* 58(5). 412–424.

Brysbaert, Marc & Kevin Diependaele. 2013. Dealing with zero word frequencies: A review of the existing rules of thumb and a suggestion for an evidence-based choice. *Behavior Research Methods* 45. 422–430.

Brysbaert, Marc, Emmanuel Keuleers & Boris New. 2011b. Assessing the usefulness of Google Books' word frequencies for psycholinguistic research on word processing. *Frontiers in Psychology* 2. 27.

Brysbaert, Marc, Pawel Mandera & Emmanuel Keuleers. 2018. The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science* 27(1). 45–50.

Brysbaert, Marc & Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods* 41(4). 977–990.

Carroll, John B. 1970. An alternative to Juilland's usage coefficient for lexical frequencies, and a proposal for a standard frequency index (SFI). *ETS Research Bulletin Series* 3. 61–65.

Chen, Xiaocong, Yanping Dong & Xiufen Yu. 2018. On the predictive validity of various corpus-based frequency norms in L2 English lexical processing. *Behavior Research Methods* 50. 1–25.

Chuang, Yu-Ying, Janice Fon, Ioannis Papakyritsis & Harald Baayen. 2021. Analyzing phonetic data with generalized additive mixed models. In Martin Ball (ed.), *Manual of clinical phonetics*, 108–138. London: Routledge.

Coltheart, M., Kathleen Rastle, Conrad Perry, Robyn Langdon & Johannes Ziegler. 2001. DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review* 108(1). 204–256.

Crossley, Scott, Tom Salsbury, Ashley Titak & Danielle McNamara. 2014. Frequency effects and second language lexical acquisition: Word types, word tokens, and word production. *International Journal of Corpus Linguistics* 19(3). 301–332.

Delaney-Busch, Nathaniel, Emily Morgan, Ellen Lau & Gina R. Kuperberg. 2019. Neural evidence for Bayesian trial-by-trial adaptation on the N400 during semantic priming. *Cognition (The Hague)* 187. 10–20.

Ehrlich, Susan F. & Keith Rayner. 1981. Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior* 20(6). 641–655.

Engbert, Ralf, Antje Nuthmann, Eike M. Richter & Reinhold Kliegl. 2005. SWIFT: A dynamical model of saccade generation during reading. *Psychological Review* 112. 777–813.

Fine, Alex B., T. Florian Jaeger, Thomas A. Farmer & Ting Qian. 2013. Rapid expectation adaptation during syntactic comprehension. *PLoS One* 8(10). e77661.

Finger, Holger, Caspar Goeke, Dorena Diekamp, Kai Standvoß & Peter König. 2017. LabVanced: A unified JavaScript framework for online studies. In *International conference on computational social science (Cologne)*, 1–3. Cologne: University of Osnabrück.

Geyken, Alexander. 2007. The DWDS corpus: A reference corpus for the German language of the 20th century. In C. Fellbaum (ed.), *Idioms and collocations: Corpus-based linguistic, lexicographic studies*, 23–40. London: Continuum Press.

Grainger, Jonathan, Danielle Lopez, Marianna Eddy, Stéphane Dufau & Phillip J. Holcomb. 2012. How word frequency modulates masked repetition priming: An ERP investigation. *Psychophysiology* 49. 604–616.

Hale, John. 2001. A probabilistic Earley parser as a psycholinguistic model. In *NAACL '01: Proceedings of the second meeting of the North American chapter of the Association for Computational Linguistics*, 159–166. Association for Computational Linguistics.

Halteren, Hans van, Harald Baayen, Fiona Tweedie, Marco Haverkort & Neijt Anneke. 2005. New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics* 12. 65–77.

Heister, Julian & Reinhold Kliegl. 2012. Comparing word frequencies from different German text corpora. In K.-M. Würzner & E. Pohl (eds.), *Lexical resources in psycholinguistic research*, 27–44. Potsdam: Universitätsverlag Potsdam.

Heister, Julian, Kay-Michael Würzner, Johannes Bubenzer, Edmund Pohl, Thomas Hanneforth, Alexander Geyken & Reinhold Kliegl. 2011. dlexDB–eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychologische Rundschau* 62(1). 10–20.

Heuven, Walter J. B. van, Pawel Mandera, Emmanuel Keuleers & Brysbaert Marc. 2014. SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology* 67. 1176–1190.

Inhoff, Albrecht W. & Keith Rayner. 1986. Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & Psychophysics* 40(6). 431–439.

Kanske, Philipp & Sonja A. Kotz. 2010. Leipzig affective norms for German: A reliability study. *Behavior Research Methods* 42. 987–991.

Kapteijns, Bob & Florian Hintz. 2021. Comparing predictors of sentence self-paced reading times: Syntactic complexity versus transitional probability metrics. *PLoS One* 16(7). e0254546.

Keating, Gregory D. & Jill Jegerski. 2015. Experimental designs in sentence processing research: A methodological review and user's guide. *Studies in Second Language Acquisition* 37(1). 1–32.

Kliegl, Reinhold, Ellen Grabner, Martin Rolfs & Ralf Engbert. 2004. Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology* 16(1–2). 262–284.

Kretzschmar, Franziska, Matthias Schlesewsky & Adrian Staub. 2015. Dissociating word frequency and predictability effects in reading: Evidence from coregistration of eye movements and EEG. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 41(6). 1648–1662.

Kukona, Anuenue & Nabil Hasshim. 2024. Mouse cursor trajectories capture the flexible adaptivity of predictive sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 50(10). 1650–1661.

Kuperberg, Gina R. & T. Florian Jaeger. 2016. What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience* 31. 32–59.

Kutas, Marta & Kara D. Federmeier. 2011. Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology* 62. 621–647.

Levy, Roger. 2008. Expectation-based syntactic comprehension. *Cognition (The Hague)* 106(3). 1126–1177.

Linzen, Tal & Tim F. Jaeger. 2016. Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive Science* 40(6). 1382–1411.

Mickan, Anne, Ekaterina Slesareva, James M. McQueen & Kristin Lemhöfer. 2024. New in, old out: Does learning a new language make you forget previously learned foreign languages? *Quarterly Journal of Experimental Psychology* 77(3). 530–550.

Morton, John. 1969. Interaction of information in word recognition. *Psychological Review* 76(2). 165–178.

Norris, Dennis. 2006. The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review* 113(2). 327–357.

Osterhout, Lee, Michael Bersick & Richard McKinnon. 1997. Brain potentials elicited by words: Word length and frequency predict the latency of an early negativity. *Biological Psychology* 46. 143–168.

Petilli, Marco A. & Marco Marelli. 2024. Visual intuitions in the absence of visual experience: The role of direct experience in concreteness and imageability judgements. *Journal of Cognition* 7(1). 3.

R Core Team. 2021. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rayner, Keith, Jane Ashby, Pollatsek Alexander & Erik D. Reichle. 2004. The effects of frequency and predictability on eye fixations in reading: Implications for the EZ Reader model. *Journal of Experimental Psychology: Human Perception and Performance* 30(4). 720–732.

Reichle, Erik D., Tessa Warren & Kerry McConnell. 2009. Using E-Z Reader to model the effects of higher-level language processing on eye movements during reading. *Psychonomic Bulletin & Review* 16. 1–21.

Schilling, Hildur E., Keith Rayner & James I. Chumbley. 1998. Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory & Cognition* 26(6). 1270–1281.

Seidenberg, Mark S. & James L. McClelland. 1989. A distributed, developmental model of word recognition and naming. *Psychological Review* 96(4). 523–568.

Shain, Cory. 2019. A large-scale study of the effects of word frequency and predictability in naturalistic reading. In Jill Burstein, Christy Doran & Thamar Solorio (eds.), *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies, vol. 1, Long and short papers*, 4086–4094. Minneapolis: Association for Computational Linguistics.

Shain, Cory. 2024. Word frequency and predictability dissociate in naturalistic reading. *Open Mind* 8. 177–201.

Staub, Adrian. 2015. The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Language and Linguistics Compass* 9(8). 311–327.

Staub, Adrian. 2024. Predictability in language comprehension: Prospects and problems for surprisal. *Annual Review of Linguistics* 11. https://doi.org/10.1146/annurev-linguistics-011724-121517.

Staub, Adrian & Kirk Goddard. 2019. The role of preview validity in predictability and frequency effects on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 45(1). 110–127.

Taylor, Wilson L. 1953. “Cloze procedure”: A new tool for measuring readability. *Journalism Quarterly* 30(4). 415–433.

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

Van Petten, Cyma & Marta Kutas. 1990. Interactions between sentence context and word frequency in event-related brain potentials. *Memory & Cognition* 18. 380–393.

Van Petten, Cyma & Barbara J. Luka. 2012. Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology* 83(2). 176–190.

Wagenmakers, Eric-Jan, Han L. J. van Der Maas & Raoul P. P. P. Grasman. 2007. An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review* 14(1). 3–22.

Whitford, Veronica & Debra Titone. 2014. The effects of reading comprehension and launch site on frequency-predictability interactions during paragraph reading. *Quarterly Journal of Experimental Psychology* 67(6). 1151–1165.

Wieling, Martijn. 2018. Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics* 70. 86–116.

Winter, Bodo. 2019. *Statistics for linguists: An introduction using R*. New York: Routledge.

Witzel, Naoko, Jeffrey Witzel & Kenneth Forster. 2012. Comparisons of online reading paradigms: Eye tracking, moving-window, and maze. *Journal of Psycholinguistic Research* 41. 105–128.

Wood, Simon N. 2017. *Generalized additive models: An introduction with R*, 2nd edn. New York: Chapman and Hall/CRC.

Zipf, George Kingsley. 1949. *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley.