The Effects of Surprisal across Languages: Results from Native and Non-native Reading

Andrea Gregor de Varda
Department of Psychology
University of Milano – Bicocca
a.devarda@campus.unimib.it

Marco Marelli
Department of Psychology
University of Milano – Bicocca
marco.marelli@unimib.it

Abstract

It is well known that the *surprisal* of an upcoming word, as estimated by language models, is a solid predictor of reading times (Smith and Levy, 2013). However, most of the studies that support this view are based on English and few other Germanic languages, leaving an open question as to the cross-lingual generalizability of such findings. Moreover, they tend to consider only the best-performing eye-tracking measure, which might conflate the effects of predictive and integrative processing. Furthermore, it is not clear whether prediction plays a role in non-native language processing in bilingual individuals (Grüter et al., 2014). We approach these problems at large scale, extracting surprisal estimates from mBERT, and assessing their psychometric predictive power on the MECO corpus, a cross-linguistic dataset of eye movement behavior in reading (Siegelman et al., 2022; Kuperman et al., 2020). We show that surprisal is a strong predictor of reading times across languages and fixation measurements, and that its effects in L2 are weaker with respect to L1.

1 Introduction

Context-dependent predictive processes have been proposed as a core component of the human cognitive system (Bar, 2007; Clark, 2013). In the language processing literature, a clear picture that is progressively emerging is that speakers spontaneously pre-activate the upcoming lexical material before they encounter it (Huettig, 2015; Schlenter, 2019; Staub, 2015). This pre-allocation of resources to predictable material is evidenced by the fact that unpredictable words are a major cause of processing costs, as measured through self-paced reading times (Frank and Hoeks, 2019; Fernandez Monsalve et al., 2012), eye movements (Ehrlich and Rayner, 1981) and pupil size (Frank and Thompson, 2012) in reading, and EEG responses (Kutas and Hillyard, 1984; Frank et al., 2015). The role of prediction in language processing was, in particular, characterized via computational modeling, with the information-theoretic notion of *surprisal* being extended to psycholinguistics (Hale, 2001; Levy, 2008). Surprisal quantitatively captures how unpredictable a word is in terms of the negative logarithm of the probability of a word conditioned by the preceding sentence context (1).

$$\text{surprisal}(w_i) = -\log_2 P(w_i|w_1, w_2 \dots w_{i-1}) \quad (1)$$

In this perspective, surprisal acts as a linking function between cognitive effort and predictability (Fernandez Monsalve et al. 2012, but see Brothers and Kuperberg 2020), where the former is measured empirically, and the latter is estimated probabilistically. Levy (2008) demonstrated that the surprisal of a word given the previous context is mathematically equivalent to the Kullback-Leibler divergence (i.e. relative entropy) between probability distributions[1]. Under this view, surprisal effects can therefore be interpreted as the cognitive costs associated to a shift between probability distributions.

Computational linguistics has proven itself very useful to derive word probability estimates (Frank et al., 2013; Demberg and Keller, 2008; Levy, 2008), and the psychometric predictive power of a language model – i.e., how well it can account for human processing times – is a linear function of that model's quality, measured as its perplexity (Goodkind and Bicknell, 2018; Wilcox et al., 2020). Computational studies on prediction in sentence processing have the indisputable merit of testing the effects of predictability at large scale and in the context of naturalistic reading. However, if compared to psycholinguistic studies on prediction, they generally focus on:

[1] In its original formulation, surprisal theory was employed to account for syntactic processing. Probability shifts were thus defined over syntactic parses.

Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022, pages 138–144
November 20–23, 2022.

i **Gaze duration.** Differently from psycholinguistic research (Frisson et al., 2005; Rayner et al., 2011), computational studies tend to consider only the eye-tracking measure that is typically best fitted by surprisal estimates, namely gaze duration (Aurnhammer and Frank, 2019; Goodkind and Bicknell, 2018; Smith and Levy, 2013; Wilcox et al., 2020), ignoring other cognitively relevant eye-tracking metrics.

ii **Germanic languages.** A vast body of findings corroborates the effects of lexical prediction in English (Aurnhammer and Frank, 2019; Frank and Bod, 2011; Frank et al., 2015; Fernandez Monsalve et al., 2012; Wilcox et al., 2020; Goodkind and Bicknell, 2018; Smith and Levy, 2013), Dutch (Frank and Hoeks, 2019; Brouwer et al., 2010) and German (Boston et al., 2008; Brouwer et al., 2021); however, evidence from other language families is far more limited (although see Fan and Reilly, 2020; Kuribayashi et al., 2021).

iii **L1.** Within the computational framework, most of the studies reported insofar targeted sentence processing in the dominant languages (but see Berzak and Levy, 2022; Frank, 2014, 2021), while the psycholinguistic community is witnessing an increasing interest in predictive processing in L2 (Cop et al., 2015; Grüter et al., 2014, 2017; Kaan et al., 2010; Martin et al., 2013).

We argue that these three limitations might undermine both the internal and the external validity of the results.

First (**i**), only considering the best-performing eye-tracking measure does not provide any insight as to *when* such predictability effects take place during natural reading. An analysis of the time range where predictability effects can be detected is however crucial to disentangle between predictive and integrative processes (Cevoli et al., 2022; Staub, 2015). Indeed, a higher processing cost induced by an unpredictable word might not be due to anticipatory processes, but also to a difficulty in integrating the unpredictable word in the phrasal context. While early measurements such as first fixation duration are thought to reflect lexical or pre-lexical processes (and thus a genuine effect of predictability; Staub, 2015), gaze duration can be considered as a "midmeasure" (Roberts and Siyanova-Chanturia, 2013), and thus it is not sufficient to disentangle between integrative and predictive processing.

Second (**ii**), some of the results that were obtained in English within the framework of surprisal theory were not replicated in other languages. For instance, Kuribayashi et al. (2021) have shown that the negative relationship between a language model's perplexity and its psychometric accuracy does not hold for the Japanese language. Hence, the rather limited typological variability in the language samples considered leaves an open question as to whether prediction itself should be considered as a core processing mechanism that generalizes across languages.

Third (**iii**) , the study of predictive processing in non-native reading is of crucial relevance since more than half of the global population is bilingual (Ansaldo et al., 2008). The role of anticipation in bilingual individuals is attracting growing interest in second language acquisition studies, and large-scale data-driven approaches might shed light on a complex picture currently characterized by little consensus. The Reduced Ability to Generate Expectations hypothesis (RAGE, Grüter et al., 2014, 2017) proposes that even highly proficient L2 speakers differ from native speakers in their abilities to anticipate the upcoming linguistic material. However, the results supporting this theory have been questioned (Hartsuiker et al., 2016; Leal et al., 2017); they are generally derived from offline tasks in small-scale studies (Grüter et al., 2014), and restricted to circumscribed linguistic phenomena (such as gender information in determiners, see Grüter et al., 2012; Lew-Williams and Fernald, 2010). Instead, it would be desirable to test the effects of word prediction in L2 when reading naturalistic, contextualized texts (see for instance Berzak and Levy, 2022; Cop et al., 2015), as opposed to artificially constructed experimental materials, presented out of context and repeated many times. Berzak and Levy (2022) have overcome these limitations by testing the effects of predictability in L2 at scale. They reported a *larger* effect of surprisal in non-native reading, which is at odds with the psycholinguistic evidence reported before, and difficult to explain. As mentioned by the authors, context-contingent expectations are statistically demanding to compute, and it is not clear why the effects of such a complex processing mechanism should be stronger in L2 than in L1.

In the present study we address these limita-

tions in the literature by considering different eye-tracking measurements, including early fixation measurements that are expected to reflect predictive processes (**i**); extending our sample to 12 diverse languages, belonging to five language families and written in five different scripts (**ii**); and comparing the effect of prediction in L1 and L2 (**iii**).

2 Methods

2.1 Eye-tracking data

The MECO-L1 corpus (Siegelman et al., 2022) is a large-scale collection of high-quality eye movement records in 13 languages[2] collected in a naturalistic reading task. Participants were presented with 12 texts composed by multiple sentences, consisting in encyclopedic entries on a variety of topics. The MECO-L2 corpus provides eye movement data on English texts read by non-native speakers (Kuperman et al., 2020). In our study, we analyze three eye-tracking measurements, that are considered an early, an intermediate, and a late processing measure, respectively (Demberg and Keller, 2008; Roberts and Siyanova-Chanturia, 2013):

- *First fixation (FF):* the duration of the first fixation landing on the target word. This measure is often assumed to reflect lexical access and low-level oculomotor processes.
- *Gaze duration (GD):* the sum of the duration of the fixations on the target word before the gaze leaves it for the first time. This measure is thought to be indicative of semantic and early syntactic processing.
- *Total reading time (TT):* the sum of the duration of all the fixations on the target word. This measure is thought to be indicative of integrative processes.

The fixations considered by different eye-tracking measures are organized in a relationship of inclusion (FF $\subseteq$ GD $\subseteq$ TT); hence, intermediate and late processing measures inevitably incorporate information about early processing. However, since the inclusion relationship is asymmetrical, early measures do not include information about late processing. Hence, predictability effects that can be detected in early eye-tracking measures can be ascribed to predictive processing (Staub, 2015).

2.2 Model and metrics

Our probability estimates are derived with $\text{mBERT}_{\text{BASE}}$'s native masked language modelling component (Devlin et al., 2019), which has been shown to generate probability estimates that are good predictors of eye movement data (Hollenstein et al., 2021). To derive word-level probability estimates, we freeze the model weights and mask all the sentence tokens iteratively. Except for the first and the last token of each sequence, where the model predictions are conditioned only by the right and the left context, mBERT predicts the token in the masked position relying upon the bidirectional context. Note that the formula in (1) implicitly refers to auto-regressive, left-to-right models. Dealing with a bidirectional encoder, we calculate the bidirectional $\text{surprisal}_{\text{B}}$ of a word w_i in a sentence of N tokens as the negative logarithm of the word probability conditioned by both the left ($w_1 \dots w_{i-1}$) and the right context ($w_{i+1} \dots w_N$, see 2).

$$\text{surprisal}_{\text{B}}(w_i) = -\log_2 P(w_i | w_1 \dots w_{i-1}, w_{i+1} \dots w_N) \quad (2)$$

2.3 Analyses

In our analyses, we discard all the surprisal estimates of multi-token words[3]. We fit all our models as linear mixed-effects models, with random intercepts for participants and items. As a baseline, we include word frequency (derived from multilingual large-scale frequency estimates, Speer et al., 2018), length, and their interaction; additionally, we include as covariates the same indexes relative to the previous w_{i-1} word, to account for spillover effects. Then, we include the effect of surprisal relative to both w_i and w_{i-1}. We first fit 36 separate models (12 languages $\times$ 3 fixation measurements) to assess the effects of surprisal for each individual language at different processing stages; then, we fit an overall model for each fixation measurement including languages as random slopes and intercepts.

In a second part of the study, we compare predictability effects across L1 and L2; to do so, we merge the two MECO datasets, and dummy-code whether each trial is recorded in an individual's

[2] We excluded the Estonian data in our study since we could not find frequency estimates comparable with the other languages.

[3] Indeed, while with standard auto-regressive models multi-token probabilities can be computed via the application of the chain rule, the same cannot be done with masked language models. See Table 1, column "%" for the percentage of the original items that were included in the analyses.

			First fixation duration				Gaze duration				Total reading time			
Language	N	%	Estimate	SE	t	p	Estimate	SE	t	p	Estimate	SE	t	p
Dutch	44,843	66%	0.0222	0.0085	2.6226	0.0088	0.0233	0.0087	2.6718	0.0076	0.0456	0.0100	4.5477	$\ll .0001$
English	65,421	77%	0.0156	0.0084	1.8574	0.0634	0.0112	0.0082	1.3612	0.1736	0.0145	0.0087	1.6619	0.0967
Finnish	20,277	31%	0.0464	0.0175	2.6515	0.0083	0.0393	0.0173	2.2789	0.0230	0.0372	0.0182	2.0387	0.0419
German	49,608	66%	0.0267	0.0112	2.3800	0.0175	0.0314	0.0117	2.6822	0.0074	0.0522	0.0125	4.1661	$\ll .0001$
Greek	56,738	51%	0.0111	0.0150	0.7363	0.4617	0.0331	0.0143	2.3064	0.0212	0.0565	0.0148	3.8106	0.0001
Hebrew	22,718	34%	0.0110	0.0128	0.8549	0.3929	0.0313	0.0124	2.5262	0.0118	0.0233	0.0144	1.6184	0.1060
Italian	56,738	65%	0.0361	0.0087	4.1286	$\ll .0001$	0.0400	0.0084	4.7448	$\ll .0001$	0.0279	0.0087	3.2228	0.0013
Korean	8,283	23%	0.0182	0.0132	1.3836	0.1667	0.0365	0.0132	2.7624	0.0058	0.0095	0.0132	0.7232	0.4696
Norwegian	33,930	54%	0.0190	0.0079	2.4048	0.0162	0.0240	0.0077	3.1272	0.0018	0.0354	0.0077	4.5788	$\ll .0001$
Russian	33,109	48%	0.0062	0.0118	0.5290	0.5969	0.0174	0.0111	1.5691	0.1169	0.0108	0.0116	0.9307	0.3522
Spanish	66,097	76%	0.0105	0.0063	1.6646	0.0960	0.0075	0.0061	1.2283	0.2194	-0.0022	0.0062	-0.3604	0.7186
Turkish	11,546	36%	0.0133	0.0114	1.1654	0.2440	0.0211	0.0113	1.8749	0.0610	0.0501	0.0116	4.3164	$\ll .0001$

Table 1: Effects of surprisal across languages on the three fixation measurements considered. The first two columns indicate the language from which the reading data were obtained, the number of data points on which the regression coefficients were computed, and the percentage of items that were not discarded in the analyses (see §2.3). The following columns indicate the regression coefficients of surprisal, their standard error (SE), the t statistic and the respective p-value for FF, GD and TT.

dominant or non-dominant language. Then, we test the interaction between language dominance (L1-L2) and surprisal. Once again, we fit our models with random intercepts for participants and items; the former random effects are particularly relevant in this analysis in order to account for differences in proficiency levels across participants. Note that since frequency and surprisal are naturally correlated, we also include in our models an interaction between surprisal and lexical frequency, as well as a main effect of language dominance. Lexical frequency is a non-contextual measure; hence, the interaction between frequency and language dominance can also be informative in studying the role of context-independent prediction in L1 and L2 (see Berzak and Levy, 2022, for similar considerations).

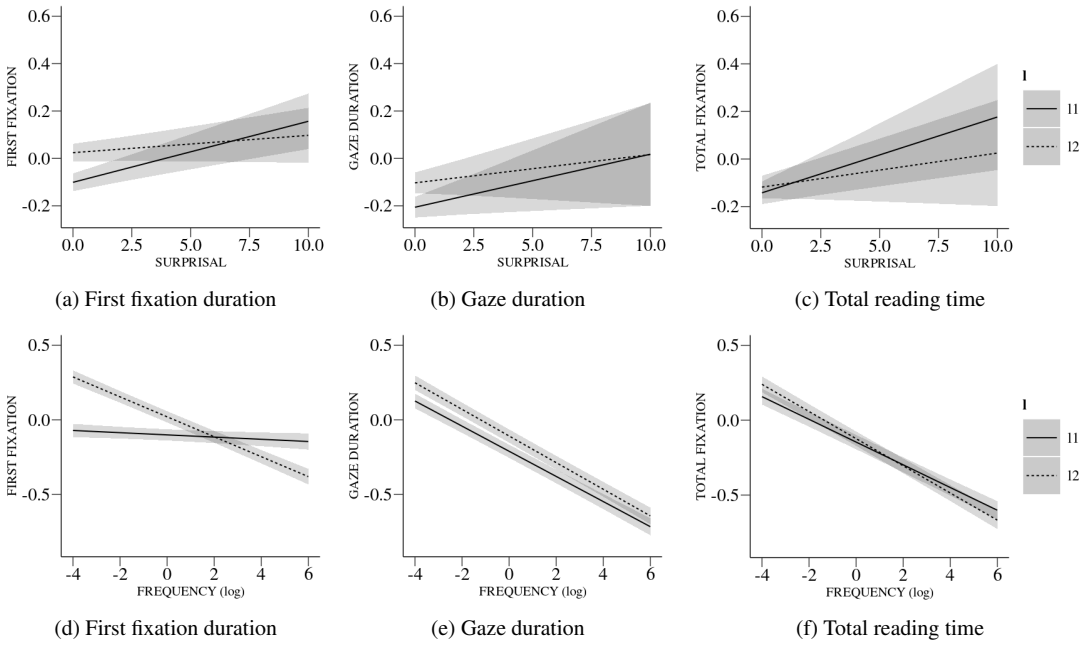
3 Results

Our language-wise results in L1 reading are summarized in Table 1; analyzing the effects separately for each language, surprisal is a significant predictor of FF in five languages; this number raises up to eight when considering GD, and seven with TT. However, a joint model with language-wise random slopes and intercepts shows a significant effect of surprisal in all the fixation measurements considered (FF: $\hat{B} = 0.0203$, $t = 5.6659$, $p < 0.001$; GD: $\hat{B} = 0.0239$, $t = 6.1418$, $p < 0.0001$; TT: $\hat{B} = 0.0258$, $t = 5.8616$, $p < 0.0001$). The presence of an effect in FF is particularly indicative, since it can be considered as a sign of predictive processing.

To test whether the effects of surprisal are similar in their extent across L1 and L2, we concatenate the MECO-L1 and MECO-L2 dataframes, dummy-code whether each trial is recorded in L1 or L2, and test for an interaction between language dominance and surprisal. The surprisal $\times$ language interaction is a significant predictor of reading times across all the fixation measurements we analyzed (FF: $\hat{B} = -0.0184$, $t = -5.626$, $p < 0.0001$, see Figure 1a,; GD: $\hat{B} = -0.0104$, $t = -3.4640$, $p = 0.0005$, 1b; TT: $\hat{B} = -0.01756$, $t = -5.723$, $p < 0.0001$, 1c). These results indicate that the surprisal effect in L1 is larger than in L2 across all three fixation measurements, since the slope for surprisal is consistently steeper in L1 (see Figure 1 for a graphical depiction of the interactions). Additionally, we also report the results of the interaction between frequency and language dominance. This interaction is significant when considering FF ($\hat{B} = -0.0594$, $t = -14.4350$, $p < 0.0001$, 1d) and TT ($\hat{B}$ -0.0148, $t = -3.7970$, $p < 0.001$, 1e; although from a graphical inspection it is clear that the largest effect is found in the case of FF); however, it does not reach statistical significance in the case of GD ($\hat{B} = -0.0050$, $t = -1.306$, $p = 0.1915$, 1f). Notably, in this case the direction of the interactions is reversed, with steeper slopes in L2 than L1.

4 Discussion

In this study, we show that prediction is a widespread processing mechanism that can be detected across a variety of languages and language families; while we fail to report significant effects in some of the languages taken individually, the consistent direction of the effects and the results of the large linear models including multiple lan-



(a) First fixation duration (b) Gaze duration (c) Total reading time

(d) First fixation duration (e) Gaze duration (f) Total reading time

Figure 1: Plots of the interactions between surprisal and language (upper row) and frequency and language (bottom row). Note that surprisal, frequency estimates and fixations were standardized. All the surprisal × language interactions are statistically significant with $p < 0.0001$, and across all the fixation measurements the slope for surprisal is steeper in L1. Conversely, the frequency × language interactions are significant in the cases of first fixation duration ($p < 0.0001$) and total reading time ($p = 0.0002$), with a steeper slope in L2.

guages strongly support the idea that natural reading involves the active anticipation of the following linguistic material. This finding complements previous results in computational psycholinguistics, showing that predictability effects are not confined to English and the few other Germanic languages which are usually considered in the surprisal literature. Crucially, surprisal exerts a cross-lingual effect even in FF, an eye-tracking metric that is thought to reflect the earliest stages of word processing. This supports our claim that the effects of surprisal that we report are the result of truly predictive processes, and do not reflect a difficulty in integrating unpredictable words in the phrasal context. Our results also highlight some interesting differences in the reading behaviour of native and non-native speakers: the role of predictive processing in the non-dominant language appears to be significantly reduced when compared with the dominant one. On the other hand, eye movements in L2 are more strongly impacted by context-independent expectations, as operationalized with unigram word frequencies. This is particularly evident in the earliest fixation measure considered, namely FF. The early onset of this L1-L2 dissociation – which would not have been detected if considering only GD – suggests a potential role of non-contextual prediction in L2: while L1 speakers might rely more strongly on the phrasal context to predict the next word, L2 speakers might base their expectations primarily on prior probabilities of the lexical material. Context-based predictions are harder to estimate in real-time reading than their context-independent counterparts; hence, language experience might influence the extent to which a speaker relies on simple frequency estimates or context-sensitive predictions to calibrate her/his expectations on the following word (Berzak and Levy, 2022).

5 Limitations and further directions

In this study, we considered L2 processing as a homogeneous cognitive phenomenon. However, it has been suggested that L2 proficiency might modulate some differences between native and non-native reading, including predictive processing (Berzak and Levy, 2022; Bovolenta and Marsden, 2021; Ito et al., 2018). We leave for future research an assessment of whether the difference in contextual and non-contextual prediction is better explained by a categorical distinction between L1 and L2, or rather a graded account of language proficiency.

References

Ana Inés Ansaldo, Karine Marcotte, Lilian Scherer, and Gaelle Raboyeau. 2008. Language therapy and bilingual aphasia: Clinical implications of psycholinguistic and neuroimaging research. *Journal of Neurolinguistics*, 21(6):539–557.

Christoph Aurnhammer and Stefan L Frank. 2019. Comparing gated and simple recurrent neural network architectures as models of human sentence processing.

Moshe Bar. 2007. The proactive brain: using analogies and associations to generate predictions. *Trends in cognitive sciences*, 11(7):280–289.

Yevgeni Berzak and Roger Philip Levy. 2022. Eye movement traces of linguistic knowledge.

Marisa Ferrara Boston, John Hale, Reinhold Kliegl, Umesh Patil, and Shravan Vasishth. 2008. Parsing costs as predictors of reading difficulty: An evaluation using the potsdam sentence corpus. *Journal of Eye Movement Research*, 2(1).

Giulia Bovolenta and Emma Marsden. 2021. Prediction and error-based learning in l2 processing and acquisition: a conceptual review. *Studies in Second Language Acquisition*, pages 1–26.

Trevor Brothers and Gina Kuperberg. 2020. Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language*, 116.

Harm Brouwer, Francesca Delogu, Noortje J Venhuizen, and Matthew W Crocker. 2021. Neurobehavioral correlates of surprisal in language comprehension: A neurocomputational model. *Frontiers in Psychology*, 12:615538.

Harm Brouwer, Hartmut Fitz, and John Hoeks. 2010. Modeling the noun phrase versus sentence coordination ambiguity in dutch: Evidence from surprisal theory. In *Proceedings of the 2010 workshop on cognitive modeling and computational linguistics*, pages 72–80.

Benedetta Cevoli, Chris Watkins, and Kathleen Rastle. 2022. Prediction as a basis for skilled reading: insights from modern language models. *Royal Society Open Science*, 9(6):211837.

Andy Clark. 2013. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204.

Uschi Cop, Denis Drieghe, and Wouter Duyck. 2015. Eye movement patterns in natural reading: A comparison of monolingual and bilingual reading of a novel. *PloS one*, 10(8):e0134008.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Susan F Ehrlich and Keith Rayner. 1981. Contextual effects on word perception and eye movements during reading. *Journal of verbal learning and verbal behavior*, 20(6):641–655.

Xi Fan and Ronan Reilly. 2020. Reading development at the text level: an investigation of surprisal and embeddingbased text similarity effects on eyemovements in chinese early readers. *Journal of Eye Movement Research*, 13(6).

Irene Fernandez Monsalve, Stefan Frank, and Gabriella Vigliocco. 2012. Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 398–408, Avignon, France. Association for Computational Linguistics.

Stefan Frank. 2014. Modelling reading times in bilingual sentence comprehension.

Stefan Frank. 2021. Toward computational models of multilingual sentence processing. *Language Learning*, 71(S1):193–218.

Stefan Frank and Rens Bod. 2011. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological science*, 22(6):829–834.

Stefan Frank, Irene Fernandez Monsalve, Robin L Thompson, and Gabriella Vigliocco. 2013. Reading time data for evaluating broad-coverage models of english sentence processing. *Behavior research methods*, 45(4):1182–1190.

Stefan Frank and John CJ Hoeks. 2019. The interaction between structure and meaning in sentence comprehension. recurrent neural networks and reading times.

Stefan Frank, Leun J Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The erp response to the amount of information conveyed by words in sentences. *Brain and language*, 140:1–11.

Stefan Frank and Robin Thompson. 2012. Early effects of word surprisal on pupil size during reading. In *Proceedings of the annual meeting of the cognitive science society*, volume 34.

Steven Frisson, Keith Rayner, and Martin J Pickering. 2005. Effects of contextual predictability and transitional probability on eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(5):862.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)*, pages 10–18.

Theres Grüter, Casey Lew-Williams, and Anne Fernald. 2012. Grammatical gender in l2: A production or a real-time processing problem? *Second Language Research*, 28(2):191–215.

Theres Grüter, Hannah Rohde, and Amy Schafer. 2014. The role of discourse-level expectations in non-native speakers' referential choices. In *Proceedings of the annual Boston university conference on Language Development*.

Theres Grüter, Hannah Rohde, and Amy J Schafer. 2017. Coreference and discourse coherence in l2: The roles of grammatical aspect and referential form. *Linguistic Approaches to Bilingualism*, 7(2):199–229.

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

Robert Hartsuiker, Aster Dijkgraaf, and Wouter Duyck. 2016. Predicting upcoming information in native-language and non-native-language auditory word recognition. *Bilingualism*, -1.

Nora Hollenstein, Federico Pirovano, Ce Zhang, Lena Jäger, and Lisa Beinborn. 2021. Multilingual language models predict human reading behavior. *arXiv preprint arXiv:2104.05433*.

Falk Huettig. 2015. Four central questions about prediction in language processing. *Brain research*, 1626:118–135.

Aine Ito, Martin Corley, and Martin J Pickering. 2018. A cognitive load delays predictive eye movements similarly during l1 and l2 comprehension. *Bilingualism: Language and Cognition*, 21(2):251–264.

Edith Kaan, Andrea Dallas, Frank Wijnen, JW Zwart, and M de Vries. 2010. Syntactic predictions in second-language sentence processing. *Structure preserved*, pages 207–214.

Victor Kuperman, Noam Siegelman, Sascha Schroeder, A Alexeeva, C Acartürk, Simona Amenta, S Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, et al. 2020. Text reading in english as a second language: Evidence from the multilingual eye-movements corpus (meco). *Studies in Second Language Acquisition*.

Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 2021. Lower perplexity is not always human-like. *arXiv preprint arXiv:2106.01229*.

Marta Kutas and Steven A Hillyard. 1984. Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307(5947):161–163.

Tania Leal, Roumyana Slabakova, and Thomas A Farmer. 2017. The fine-tuning of linguistic expectations over the course of l2 learning. *Studies in Second Language Acquisition*, 39(3):493–525.

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

Casey Lew-Williams and Anne Fernald. 2010. Real-time processing of gender-marked articles by native and non-native spanish speakers. *Journal of memory and language*, 63(4):447–464.

Clara D Martin, Guillaume Thierry, Jan-Rouke Kuipers, Bastien Boutonnet, Alice Foucart, and Albert Costa. 2013. Bilinguals reading in their second language do not predict upcoming words as native readers do. *Journal of memory and language*, 69(4):574–588.

Keith Rayner, Timothy J Slattery, Denis Drieghe, and Simon P Liversedge. 2011. Eye movements and word skipping during reading: effects of word length and predictability. *Journal of Experimental Psychology: Human Perception and Performance*, 37(2):514.

Leah Roberts and Anna Siyanova-Chanturia. 2013. Using eye-tracking to investigate topics in l2 acquisition and l2 processing. *Studies in Second Language Acquisition*, 35(2):213–235.

Judith Schlenter. 2019. *Predictive language processing in late bilinguals: Evidence from visual-world eye-tracking*. Ph.D. thesis, Universität Potsdam.

Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, et al. 2022. Expanding horizons of cross-linguistic research on reading: The multilingual eye-movement corpus (meco). *Behavior research methods*, pages 1–21.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Robyn Speer, Joshua Chin, Andrew Lin, Sara Jewett, and Lance Nathan. 2018. Luminosoinsight/wordfreq: v2.2.

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

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. *arXiv preprint arXiv:2006.01912*.