COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 46 (2022) e13201

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13201

Is the Mind Inherently Predicting? Exploring Forward and Backward Looking in Language Processing

Luca Onnis,[a,b] Alfred Lim,[c] Shirley Cheung,[d] Falk Huettig[e]

[a]*Centre for Multilingualism in Society across the Lifespan, University of Oslo*
[b]*Department of Linguistics and Scandinavian Studies, University of Oslo*
[c]*School of Psychology, University of Nottingham Malaysia Campus*
[d]*Washington University School of Medicine*
[e]*Max Planck Institute for Psycholinguistics*

Received 8 December 2020; received in revised form 23 August 2022; accepted 29 August 2022

Abstract

Prediction is one characteristic of the human mind. But what does it mean to say the mind is a "prediction machine" and *inherently forward looking* as is frequently claimed? In natural languages, many contexts are not easily predictable in a forward fashion. In English, for example, many frequent verbs do not carry unique meaning on their own but instead, rely on another word or words that follow them to become meaningful. Upon reading *take a* the processor often cannot easily predict *walk* as the next word. But the system can "look back" and integrate *walk* more easily when it follows *take a* (e.g., as opposed to **make|get|have a walk*). In the present paper, we provide further evidence for the importance of both forward and backward-looking in language processing. In two self-paced reading tasks and an eye-tracking reading task, we found evidence that adult English native speakers' sensitivity to word forward and backward conditional probability significantly predicted reading times over and above psycholinguistic predictors of reading latencies. We conclude that both forward and backward-looking (prediction and integration) appear to be important characteristics of language processing. Our results thus suggest that it makes just as much sense to call the mind an "integration machine" which is inherently backward 'looking.'

Keywords: Integration; Language model; Prediction; Reading; Sentence processing; Surprisal

Correspondence should be sent to Luca Onnis, Centre for Multilingualism in Society across the Lifespan, University of Oslo, 0317, Oslo, Norway. E-mail: lucao@uio.no

1. Introduction

In human communication, the speech signal dissipates as soon as it is produced or heard. And while printed words sit as static percepts on a page, fluent reading is a fast incremental process, with the average silent reading rate for adults in English fiction prose being estimated at 260 words per minute (Brysbaert, 2019). The inherent fleeting nature of language processing suggests that the human brain recruits mechanisms efficiently tuned to processing sequential information. One way that such mechanisms may work is to predict language, such that words are processed faster when they are more predictable in a given context. Accordingly, considerable evidence has accumulated over the last few decades that people often use context to implicitly predict how an utterance might continue, thus making comprehension a fluent process. Beyond language, the prediction has been proposed as a general mechanism of human information processing (e.g., Clark, 2013; Friston, 2005). Many current cognitive theories of language reflect this trend by placing an important role on prediction for language learning and comprehension (Altmann & Mirković, 2009; Dell & Chang, 2014; Federmeier, 2007; Ferreira & Chantavarin, 2018; Hale, 2001; Hickok, 2012; Huettig, 2015; Kuperberg & Jaeger, 2016; Levy, 2008; Norris, McQueen, & Cutler, 2016; Pickering & Gambi, 2018; Pickering & Garrod, 2013; Van Petten & Luka, 2012). It is important to note that the term prediction (or anticipation, expectation, context effects, top-down processing) has been used in different ways by different researchers and fields. Here, we refer to prediction as the preactivation of (linguistic) representations before bottom-up input has had a chance to activate them. It avoids arbitrary decisions about what constitutes prediction and what does not (cf. the distinction between expectation and prediction, e.g., Van Petten & Luka, 2012) and reflects the common language sense that prediction is about what may happen in the future (Huettig, Audring, & Jackendoff, 2022).

A considerable amount of experimental evidence consistent with the notion that preactivation is an important aspect of language processing comes from electrophysiological studies measuring the N400 ERP component, a negative-going and centro-parietally distributed component occurring approximately 400 ms after the predicted target word onset (e.g., Brothers, Swaab, & Traxler, 2015; Camblin, Gordon, & Swaab, 2007; Federmeier & Kutas, 1999; Federmeier, McLennan, De Ochoa, & Kutas, 2002; Hintz, Meyer, & Huettig, 2020; Laszlo, Stites, & Federmeier, 2012; Metusalem et al., 2012; Otten & Van Berkum, 2007, 2008; Rommers, Meyer, Praamstra, & Huettig, 2013; Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005). However, an interpretation of N400 modulation on the target word as a "direct" measure of prediction (or anticipation) is problematic, because it is difficult to establish whether N400 deflections index prediction of upcoming words (or representations), ease of integration of incoming words with preceding context, or both (see Baggio & Hagoort, 2011; DeLong, Urbach, & Kutas, 2005; Huettig, 2015; Lau, Phillips, & Poeppel, 2008; Mantegna, Hintz, Ostarek, Alday, & Huettig, 2019; Nieuwland et al., 2020; Nieuwland et al., 2018; Rabovsky, Hansen, & McClelland, 2018; Van Berkum et al., 2005; Wicha, Moreno, & Kutas, 2004 for further discussion). Indeed, integration is increasingly considered in addition to prediction in theories and electrophysiological models of language processing (see Bornkessel-Schlesewsky & Schlesewsky, 2008; Brouwer & Crocker, 2017; Brouwer, Delogu, Venhuizen, & Crocker, 2021; Brouwer, Fitz, & Hoeks, 2012; Kuperberg, 2007 for review).

The distinction between (lexical) prediction (i.e., "top-down" activation) and (lexical) integration (i.e., "bottom-up" activation) is not only important for the interpretation of ERP N400 modulations in language processing, but also for general theoretical importance. Despite the contemporary trend to interpret experimental findings in language processing within a predictive framework, recent work (Frisson, Harvey, & Staub, 2017; Huettig, 2015; Huettig & Guerra, 2019; Huettig & Mani, 2016; Luke & Christianson, 2016; Nieuwland et al., 2018; Staub, 2015) has proposed that much of language may actually proceed on integrative processes that link linguistic material just heard or read to what is already known or established (Ferreira & Chantavarin, 2018). Moreover, it has been noted that key evidence for prediction in language processing is often limited to circumscribed stimuli and task settings. Huettig and Mani (2016), for example, pointed out that the visual stimuli presented in visual-world eye-tracking experiments (another frequently used method in prediction research) provide critical scaffolding for the finding of such effects, because the visual referents of predicted words, which show anticipatory eye gazes, have been primed during the visual preview (see McQueen & Huettig, 2014, for experimental evidence supporting this claim). Similarly, Brothers, Swaab, and Traxler (2017) provide evidence that the task in ERP experiments influences the extent to which readers engage in lexical prediction.

Taken together, this casts considerable doubts on claims that humans are "prediction machines" and that forward looking is the default characteristic of human information processing (e.g., Clark, 2013; Friston, 2005). Backward looking and integration may be (at least) as important for language comprehension as forward-looking and prediction. As contemporary empirical investigations of language processing tend to predominately focus on prediction, there is a need to further explore empirically the respective contributions of forward looking and backward looking in language comprehension.

1.1. The current study

In this article, we aim to further contribute to this endeavor by assessing both forward-looking and backward-looking processes during a word-to-word reading of sentences in real-time. As a proxy for backward looking, we use a simple measure of lexical integration that can be easily derived from large language corpora and is a direct counterpart of a common measure of probabilistic prediction (namely forward surprisal). Then, we assess how these proxy measures of forward looking (prediction) and backward looking (integration) contribute to explaining sentence reading times. It is important to stress here again that our measures of forward looking (prediction) and backward looking (integration) are proxies only. They are not meant to be taken as "perfect" indexes of prediction and integration processes. Forward surprisal is typically considered a reasonable proxy for forward looking and prediction because it scales linearly with word reading times (Goodkind & Bicknell, 2018; Smith & Levy, 2013), and the size of N400 effects (Yan & Jaeger, 2020).

To understand our lexical integration proxy and framing of backward-looking as a form of integration, consider the case of the many lexical restrictions (or collocations) that English speakers implicitly master and which involve determiners and modifiers. For example, modifying verbs like *do* and *make* are near synonymous and can precede several nouns, so their predictive power is limited. But looking back, these modifiers appear to be selected largely based on the following words, as in *make some noise*, but not **do some noise*, and *do away with* but not **make away with*. Likewise, many adjective–noun combinations appear to be more backward-looking (i.e., being more informative when integrating preceding context) than predictive (i.e., involving preactivation of upcoming words), as in *strong tea* but not **powerful tea*, or *powerful resources* but not **strong resources*. In the above cases, what distinguishes the acceptability (and processability, see Onnis & Huettig, 2021) of a multiword sequence may not be so much the forward conditional probability, which is sensibly low for *noise* given both *do some* or *make some*. This happens because numerous words in the language can follow verbs like *do* and *make*, and thus the forward probability values can be empirically estimated in a corpus of language to be low. Rather, it is the backward conditional probability (e.g., *P(make some* | *noise)* that is higher in the nonstarred examples, compared to the starred examples (e.g., *P(do some* | *noise)*). Such backward-looking may not be limited to a handful of examples, or to certain classes of lexical items or verb phrases. In noun phrases, an adjective like *green* can be followed by many nouns, making the future unpredictable, but *green* often precedes the noun *papaya*, which could thus receive a processing boost when it is integrated. Likewise, in prepositional phrases, one is *under attack* but not ∗*below attack*, and something is *at work*, or *in place*, but not ∗*in work* or ∗*at place*. Here again, the first words in the phrases above appear to be selected largely based on the following words, and not vice versa.

Language users do appear to make use of backward looking in many experimental settings, for example, to fill in words they have missed in comprehension (Gwilliams, Linzen, Poeppel, & Marantz, 2018; Lieberman, 1963), to repeat words that help to access upcoming words (Harmon & Kapatsinski, 2021), or to fill in acoustically ambiguous words. For example, in the classic experiment of phoneme restoration by Warren (1970), subjects heard spoken sentences in which a word mid-sentence had been partly covered by a cough (signaled here by a *), and thus was made acoustically ambiguous. The lexical context following the ambiguous word was manipulated as follows:

1. It was found that the *eel was on the axle
2. It was found that the *eel was on the shoe
3. It was found that the *eel was on the orange
4. It was found that the *eel was on the table

As an effect of this manipulation, subjects restored the ambiguous word **eel* with ease contextually based on the following final word context, often not even realizing the presence of the cough (they had the perception that they had heard the correct word whole). An integration explanation of this effect is that participants implicitly assessed the backward probability of the ambiguous word, activating the one that yielded the highest backward probability. For instance, *P* (*heel was on the* | *shoe*) is more likely than *P* (*peel/wheel/eel was on the* | *shoe*).

We conjecture that backward looking and integration is not limited to compensatory processes that aid to disambiguate noisy signals in the preceding context but is part and parcel of linguistic inference more broadly during natural language processing, even when the signal is clear. Likewise, we conjecture that backward-looking processes are crucial beyond specifically crafted examples of psycholinguistic experiments, and thus its effect should extend to naturalistic samples of natural language sentences more broadly. Specifically, we expect that sentence reading times should be impacted by both forward and backward types of contextual information. While prediction-based models of language—that is, those explicitly based on estimating forward conditional probabilities, from n-gram to the popular class of recurrent neural networks (RNNs) (Goldstein et al., 2022)—engage exclusively on predictive inference, here we assessed the extent to which human reading times can also be explained by a model that explicitly tracks backward conditional probabilities, which we take as a proxy measure of backward looking and integration. We also test two-way interactions between prediction, integration, and position of a word in the sentence to investigate possible temporal dynamics emerging from the interaction of these variables.

To establish robust results from our analyses, we analyzed three separate datasets of human sentence reading tasks, involving three subject samples reading the same set of sentences. Common to the three datasets is the fact that they involve reading whole natural language sentences, which, unlike many psycholinguistic experiments, do not present a well-crafted manipulation at a specific point in the sentence. Thus, the first advantage of the sentence set we used is that the obtained results can be said to have broad coverage, being applicable to whole sentences in natural language (Frank, Monsalve, Thompson, & Vigliocco, 2013). A second advantage of using the same sentence pool with different subjects is that it allows replicability of results, an aspect that has become critical to validate psychological science (Anderson et al., 2015; Munafò et al., 2017; Simons, 2014). In particular, the first and second datasets use the same self-paced reading task in which native speakers of English read sentences one word at a time, with words appearing centrally and sequentially on a screen, and word-level reading times were collected. The third dataset is an eye-tracking study in which the third sample of English native speakers read sentences naturally on a computer display and their reading patterns were recorded via an eye-tracker. While the third dataset is not a direct replication of the first two, it can be used to assess whether the findings from a self-paced reading task extend to a form of sentence processing that is closest to naturalistic human reading behavior. For clarity, the first and third datasets come from Frank et al. (2013), while the second dataset was collected afresh by us.

Across the three datasets, for each word read in a sentence, we measured predictive and integrative relations in corpus-derived forward and backward probabilities from n-gram models. We hypothesized that higher sensitivity to backward probability will significantly reduce reading latencies, such that words are read faster when the immediately preceding words can be more easily integrated with the current target word.

2. Method

2.1. Stimuli

For all three datasets in this study, English sentences came from the 361-sentence UCL corpus Frank et al. (2013) explicitly created to evaluate language models on word reading times. These sentences were drawn from original English narratives. For 166 of the sentences, yes/no comprehension questions were constructed and used to keep participants on task.

2.2. *Dataset 1: Frank et al. (2013)'s self-paced reading study*

The original dataset from Frank et al. (2013) contains data from all 175 first-year students (92 females, 70 native English speakers, mean age = 18.9 years) of psychology at University College London. In the dataset, sentence stimuli had been selected at random without replacement for each subject from the 361 experimental sentences until 40 min had elapsed. Subjects read on average 212 sentences, ranging from a minimum of 65 to a maximum of 349.

2.3. *Dataset 2: Replication of Frank et al. (2013)'s self-paced reading task*

2.3.1. *Participants*

As a near replication of the Frank et al. dataset, we recruited 48 novel young adult native speakers of English (35 women: age $M = 22.6$, $SD = 6.0$) at a large university in the United Kingdom. They were tested individually in a quiet room at their university and were compensated a small monetary token for their participation.

2.3.2. *Procedure*

Each participant was randomly assigned to one of 10 groups, each containing 36 unique test sentences in English from the University College London UCL corpus and five practice sentences. Test sentences were presented in random order. As in the original Frank et al. data, the words were displayed one at a time, progressing across the screen in their natural position with successive presses of the spacebar. Approximately half of the sentences were followed by a yes-no question regarding the content of what was just read to maintain the attention of the participants. In terms of comparison with Dataset 1, the difference between our data collection and Frank et al. is that in collecting Dataset 2, we exposed participants to a fixed number of sentences ($n = 36$), while in Dataset 1, participants were exposed to a variable number of sentences until 40 min had elapsed. Other than that, both studies sampled sentences within-subject from the same pool of sentence stimuli, both shuffled the order of sentences, both asked the same comprehension questions, and both presented the stimuli on screen in the same way.

2.4. *Dataset 3: Frank et al. (2013)'s eye-tracking study*

2.4.1. *Materials and subjects*

The original eye-tracking data from Frank et al. (2013) contain 43 subjects (27 females, 37 native English speakers, mean age = 25.8 years). Of the original 361 sentence stimuli, the subjects had read the 205 that fit on a single line of the display to be used in the eye-tracking study. The dataset contains word-level information about the first fixation time on the current word (or 0 if word not fixated), first-pass reading time (or 0 if word not fixated), right-bounded reading time (or 0 if word not fixated), and go-past reading time (or 0 if word not fixated).

Following Frank et al. (2013), words attached to punctuation (including all sentence-final words) and nonfixated words were not included in the analysis.

2.5. *Variables predicting reading times*

Processes of word reading in a sentence, as measured by reading times and eye movements, are known to be affected by basic word properties, among which are base frequency, word length in the number of characters, and position in the sentence. We considered these variables to model our three datasets. In addition, a more efficient reader may adapt her reading times to the words' probability of occurrence given its context, such that more probable words are read more quickly (Levy, 2008; Smith & Levy, 2013). To assess whether processing involves both prediction and integration, we considered two measures of sensitivity to a word probability given its immediate context—forward and backward probabilities, as estimated using the 2019 version of the Google Books Ngram corpus, 200 billion words of data in both the American and British English datasets (Michel et al., 2011), between years 1900 and 2009. We present analyses based on trigram statistics, for the simplicity and reliability of their empirical measurement given available mega corpora, and because a vast literature points to them as robust and ubiquitous predictors of psycholinguistics processes, ranging from learning to memory and retrieval, across the lifespan. After splitting the sentences from the dataset into 1- to 3-grams, from the corpora we obtained three lexical statistics for each n-gram in the dataset: (1) the frequency of each n-gram; from which we derived (2) the forward conditional probability related to the last word of each n-gram, and (3) the backward conditional probability related to the last word of each n-gram.

In order to assign an empirically grounded probability to word sequences, we used the equation for forward conditional probability of any word w_t given its previous two-word context w_{t-2}, w_{t-1} based on n-gram frequencies, as follows:

$$P\left(w_t|w_{t-2}, w_{t-1}\right) = \frac{Freq\left(w_{t-2}, w_{t-1}, w_t\right)}{Freq\left(w_{t-2}, w_{t-1}\right)} \tag{1}$$

Likewise, we assessed the probability of prior context w_{t-2}, w_{t-1} given the current word w_t using the equation for backward conditional probabilities as follows:

$$P(w_{t-2}, w_{t-1}|w_t) = \frac{Freq(w_{t-2}, w_{t-1}, w_t)}{Freq(w_t)} \tag{2}$$

In line with recent psycholinguistic work, we converted probabilities into bits of information, through the information-theoretic function of surprisal (Hale, 2001; Levy, 2008). Conceptually, surprisal estimates how unexpected a given event is. Improbable events carry more information than expected ones and should be perceived as more "surprising," so that surprisal is inversely related to probability, through a logarithmic function. In the statistical analyses that follow, we refer to the amount of surprisal derived from forward conditional probabilities as prediction surprisal:

$$effort\,(t) \propto prediction\,surprisal(w_t) = -log(P(w_t|w_{t-2}, w_{t-1})) \tag{3}$$

Consistently, we use log-transformed backward conditional probabilities to refer to lexical integration:

$$\mathit{effort}\,(t) \propto \mathit{lexical\,integration}(w_t) = -log(P(w_{t-2}, w_{t-1}|w_t)) \tag{4}$$

In psycholinguistics, higher prediction surprisal values have been shown to be associated with longer reading times (Boston, Hale, Kliegl, Patil, & Vasishth, 2008; Demberg & Keller, 2008; Smith & Levy, 2013). However, it has been less investigated to what degree lexical integration is also predictive of reading times. We investigate this question in a series of statistical regression models where forward surprisal and lexical integration are entered as independent variables of sentence reading times.

3. Analytical pipeline

Our data, scripts for statistical analyses, and results are digitally available and reproducible on the Open Science Framework (OSF) repository: https://osf.io/q6zep/wiki/home/?view_only = 9f2f22dffcaf45ef86608e635b24f19f (blinded temporarily for peer review). The original experimental materials from Frank et al. (2013) are available as supplementary material in the online version of their article (https://doi.org/10.3758/s13428-012-0313-y).

3.1. Exclusion criteria

Following Frank et al. (2013), words attached to punctuation (including all sentence-final words) and nonfixated words (in the eye-tracking dataset) were not included in the analysis (the percentage of nonfixated words was 34.0% overall but varied widely among subjects, from 5.5% to 60.4%). Also, following Frank et al. (2013), we discarded data from subjects who had an error rate above 25% on the yes/no comprehension questions. We also did not consider self-paced reading times that were extreme (below 80 ms or above 3000 ms). In Dataset 3, we excluded eye tracking first pass times above 1000 ms. Finally, because we report results based on 3-grams as the basis for context effects, we considered reading times starting from the third word in any sentence.

For the three datasets, to investigate how readers' sensitivity to various language properties, including context surprisal, is related to their online reading, the collected data were analyzed by generalized additive mixed models (GAMMs) performed and computed with R version 4.1.0 (R Core Team, 2021). The fitting engine used for all models was the *mgcv* package v1.8-35 (Wood, 2011). GAMMs are an extension of the linear mixed model that make it possible to model a response variable as a nonlinear function of one or more predictor variables, using, for example, thin plate regression splines. GAMMs have recently been applied to various linguistic and psycholinguistic data (R. H. Baayen & Linke, 2019; Coupé, 2018; Murakami, 2016; Sóskuthy, 2017).

The software available in the *mgcv* package for R by Wood (2011) provides statistical tools for the modeling of both fixed-effect factors, random effects, covariates, and their interactions. For consistency, all three datasets were analyzed using the same statistical modeling workflow described below.

3.1.1. Dependent variables

The dependent variable was the Word reading times log transformed to improve normality of distribution, according to standard practice in reading tasks (though the figures display them in the original metric). Upon suggestion by a Reviewer, we also report separate regression analyses on the word following the critical word (w_{t+1}), to account for possible spillover effects. This relates to the ongoing processing of the previous word, for example, when a reader has not yet fully processed a word but he/she already reads the next word. In other words, the processing of one word spills over to the next word. Finally, all continuous predictors were centered and scaled to improve model convergence, and to allow for comparison of variable estimators.

3.2. Inferential statistics

To predict word reading times in each dataset, the degree of autocorrelation (AR) in the data (H. Baayen, Vasishth, Kliegl, & Bates, 2017; R. H. Baayen, van Rij, de Cat, & Wood, 2018) was first established empirically by fitting each data to a null model that contained only the intercept. The value of AR Rho was established for each dataset and included in more complex models. This AR model did not contain random and fixed effects. Visual inspection of residuals' distribution confirmed that including an AR for the previous word effectively eliminated AR from the models in Datasets 1 and 2.

We then fitted a full model that considered AR and added the following random and fixed effects.

3.2.1. Random effects structure

We included random effects (intercepts) for Subject, Item, and Sentence Position in the sequence of trials. The latter was expected to capture variance due to fatigue or changes in attention to the stimuli as the task progressed. We also included random slopes for subjects for the terms of theoretical importance, namely single word surprisal, forward surprisal, and lexical integration. These terms were specified as nonlinear terms.

3.2.2. Fixed effects structure

Fixed effects included the following item-specific effects: a fixed smooth predictor for Word Position in the sentence and for Word Length in characters; fixed tensor products (and by-subject random slopes) for unigram surprisal, forward surprisal, and lexical integration. Finally, we included linear interactions of forward surprisal by lexical integration, the interaction of word position by forward surprisal, and the interaction of word position by lexical integration. Theoretically, adding unigram surprisal to the model estimated the effect on reading times based on the surprisal of the word being read (i.e., the negative log of its probability in a corpus), as if each word was read independenty of its context of occurrence. Adding forward surprisal tested whether reading times are affected by the probability of the immediately preceding context. Adding lexical integration tested whether the probability of the immediately preceding context given the target word influences reading times on the target word.

Finally, interaction terms tested whether forward surprisal and lexical integration interacted with each other, and independently with the position of the word in a sentence.

3.2.3. Model selection

One important advantage of GAMMs fitted with the *mgcv* function is that there is in principle no need to conduct separate step-wise variable selection (Marra & Wood, 2011). Model estimation and model selection are integrated into *mgcv*. The penalization procedure available for regression/smoothing splines leaves a simple linear term or reduces it to zero if not justified (this is achieved practically by specifying *select* = *TRUE* in the model formula in R code).

To assess the model fit for each dataset, we compared deviance explained for the full model with deviance explained for a null model (intercept only) in the following way: Overall deviance explained for the full model was estimated by:

$$Overall\,deviance = \frac{deviance(null\,model) - deviance(full\,model)}{deviance(null\,model)} \tag{5}$$

Likewise, partial deviance explained by each variable in the full model was estimated by:

$$Partial\,deviance = \frac{deviance\,(model\,without\,target\,variable) - deviance\,(full\,model)}{deviance(null\,model)} \tag{6}$$

4. Results from Dataset 1: Frank et al. (2013) self-paced reading task

As a first indication that forward and backward probabilities should be considered as independent predictors of language processing, the correlation between forward and backward log-transformed probabilities for the trigrams composing the sentences in the self-paced reading task was small: $r(3818) = -.11, p < .001$.

Table 1 reports the fixed effects from the GAM models when fitted to predict the current word as well as the subsequent word (for spillover effects). As the shape of the regression line is not interpretable from the summary table, visualization is an important tool for interpreting nonlinear regression models. Fig. 1 reports partial effects, that is, the isolated effects of one predictor, or interaction (Fig. 2).

The colored area around the regression curves indicates 95% confidence intervals. The model predicting current word latencies yielded significant nonlinear effects such that words were read slower when they appeared in mid-sentence. Importantly, lexical integration measured as sensitivity to backward trigram probabilities also decreased reading times, while neither single word suprisal nor prediction surprisal was significant. Note that when interpreting surprisal as a value, higher values represent more surprising (i.e., unpredictable) n-grams. The model also indicated a Word position by Prediction surprisal interaction, and a Word position by Word surprisal interaction, such that these effects were stronger later rather than earlier in the sentence.

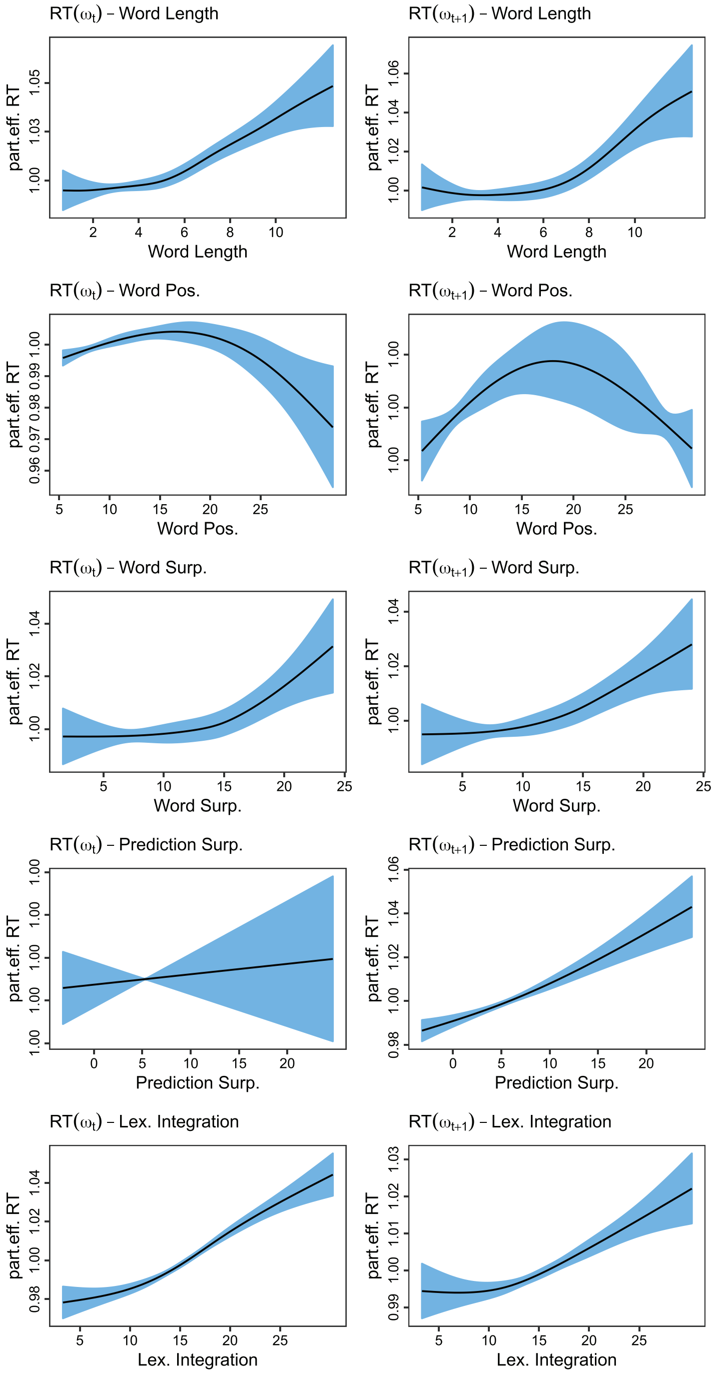

Fig. 1. Plots of fixed effects for the GAMM terms predicting word reading times (RT) in Dataset 1: Data from the original Frank et al. (2013) study. The left-hand column reports predictors of current word RT, while the right-hand column reports predictors of spillover effects on the next word.

Table 1
Summary of nonlinear trigram models with parametric interactions of Dataset 1: Self-paced reading times collected by Frank et al. (2013)

	RT(w_t)	RT(w_{t+1})
(Intercept)	5.57**	5.57**
	(258.85)	(260.52)
Prediction Surp. X Lex. Integration	−0.00	0.00
	(−0.78)	(0.28)
Word Pos. X Prediction Surp.	0.00**	−0.00
	(3.26)	(−0.59)
Word Pos. X Lex. Integration	−0.00	−0.00*
	(−1.33)	(−2.56)
Word Pos. X Word Surp.	−0.00*	−0.00
	(11.74)	(3.93)
Word Length	3.05	3.18
	(34.41)	(28.30)
Word Position	2.73**	1.29**
	(10.23)	(5.58)
Word Surprisal	2.44	2.06
	(24.64)	(28.97)
Prediction Surprisal	0.19	1.99**
	(0.12)	(118.04)
Lexical Integration	3.18**	2.60**
	(503.02)	(60.77)
AIC	−13,152.33	−17,110.44
BIC	−2278.33	−5986.19
Log Likelihood	7630.80	9644.39
Deviance	13,709.50	12,131.38
Deviance explained	0.53	0.53
Dispersion	0.06	0.06
*R*2	.53	.53
GCV score	−5529.57	−7522.91
Num. obs.	222,063	201,507
Num. smooth terms	11	11

Note: The table presents linear interactions first, with beta indicating the standardized regression weights, and *t*-values in brackets. For nonlinear main effects, the table reports the estimated degrees of freedom (EDF), which is an estimate of the "wigglyness" of the relationship (EDF = 1 corresponds to a straight line), and *F*-values in brackets. Asterisks indicate the significance level: $^*p < .025$, $^{**}p < .005$.

The model testing spillover effects revealed similar effects, with the exception that prediction surprisal also reduced time latencies, and the effect of lexical integration was stronger later in the sentence. Table 2 reports the partial deviance explained by each term in the model.

5. Results from Dataset 2: New data from the self-paced reading task

Table 3 reports fixed effects, while Figs. 3 and 4 show visualizations of significant partial smooth effects and interactions, respectively. The model predicting current word latencies yielded a significant effect on word length, such that longer words were read more slowly.

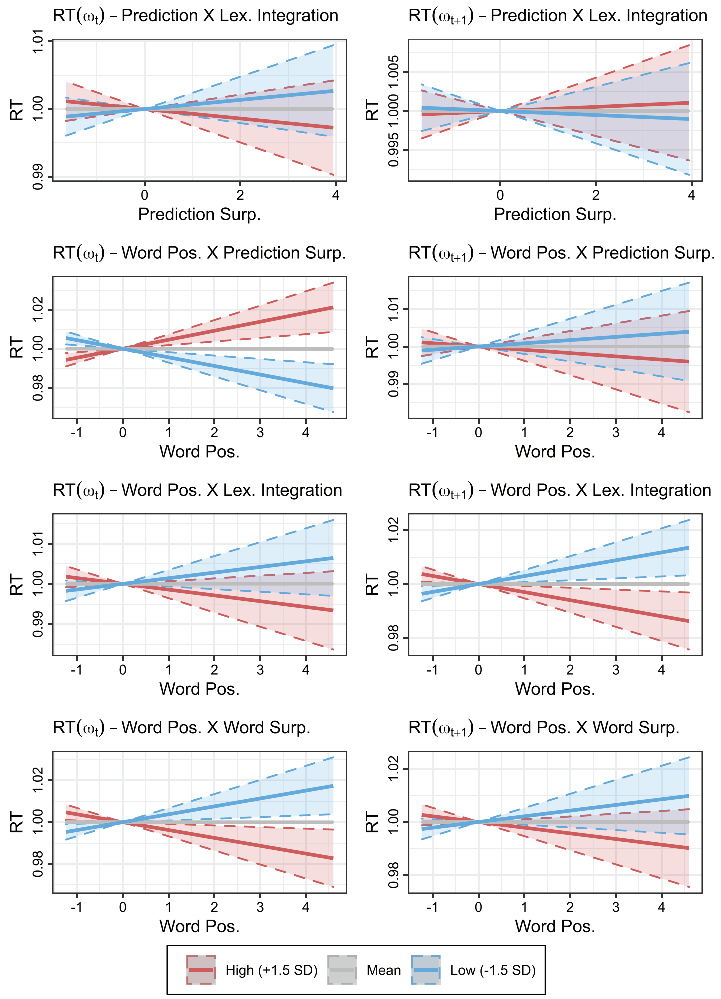

Fig. 2. Dataset 1: Plots of two-way interaction effects for current word RT and spillover effects.

Nonlinear effects were obtained such that words were read slower when they appeared in mid-sentence. Importantly, lexical integration measured as sensitivity to backward trigram probabilities also decreased reading times, while prediction surprisal was not significant.

The model testing spillover effects revealed similar effects, but with a significant linear effect of word surprisal and a nonsignificant effect of integration surprisal. Table 4 reports the partial deviance explained by each term in the model.

6. Results from Dataset 3: Frank et al. (2013) eye-tracked reading task

We report results based on two separate dependent variables: (1) First-fixation time (the duration of only the first fixation on the current word), and (2) Go-past time (the summed duration of all fixations from the first fixation on the current word up to (but not including)

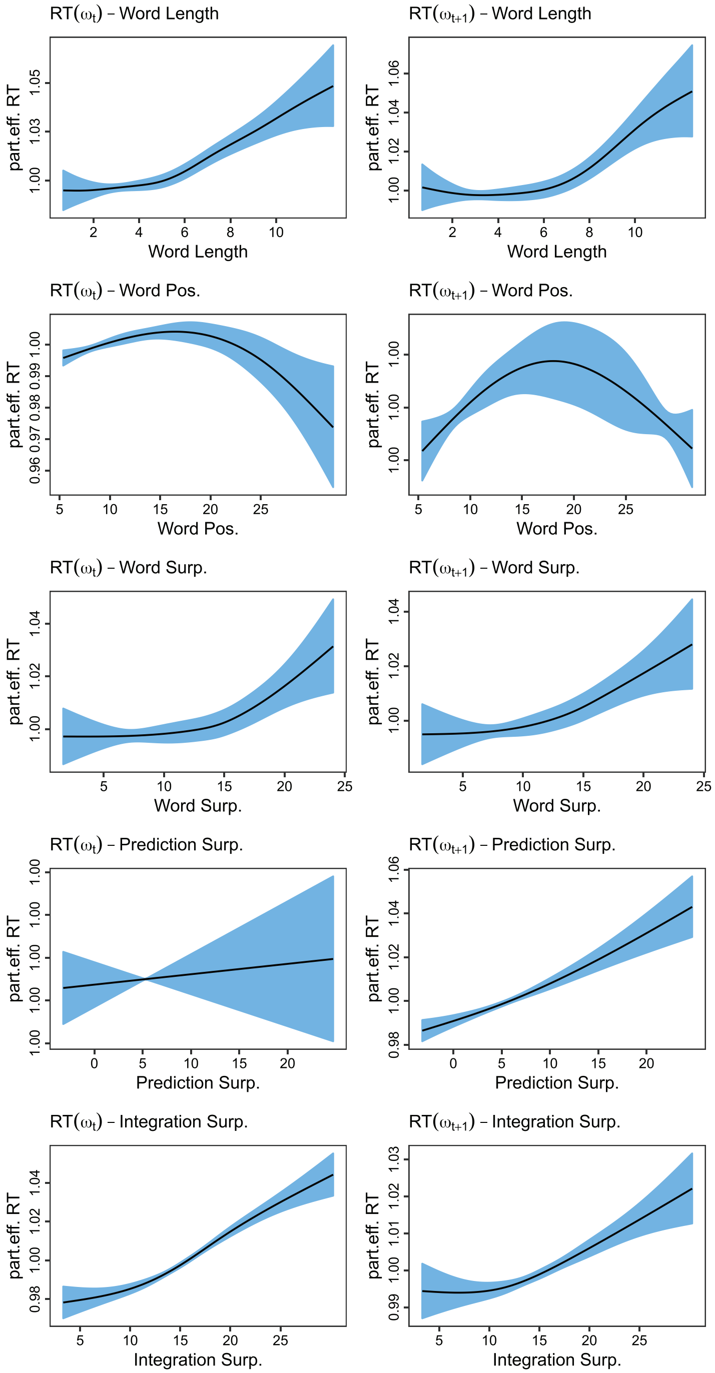

Fig. 3. Plots of fixed effects on reading times (RT) for current word (left-hand column) and following word (right-hand column), for Dataset 2: Newly collected self-paced reading times.

Table 2
Partial deviance explained by each term entered in the generalized additive model predicting word reading times for Dataset 1: the original data from Frank et al. (2013)

Term	RT(w_t)	RT(w_{t+1})
Random effects		
Sentence Position	11.45	11.614
Subject	26.79	26.901
Word	0.241	0.334
Word Surprisal by Subject	0.049	0.021
Prediction Surprisal by Subject	0.04	0.033
Lexical Integration by Subject	0.092	0.053
Fixed effects		
Word Length	0.006	0
Word Position	0.017	0.001
Word Surprisal	0.011	0
Prediction Surprisal	0.012	0.012
Lexical Integration	0.013	0.007
Prediction Surp. X Integration	0.012	0
Word Pos. X Prediction Surp.	0.013	0
Word Pos. X Lex. Integration	0.011	0.002
Word Pos. X Word Surp.	0.014	0.002

Note: Partial deviance is the deviance explained by all terms minus the deviance in the submodel in which the term of interest is removed.

the first fixation on a word further to the right. Go-past time is also known as regression-path time, and often includes fixations on words to the left of the current word. For this reason, it may be a particularly indicative index of integration processes. The outcomes of the analyses for the two dependent variables are reported side-by-side for easier comparison in Table 5. The outcome of the analyses involving first-fixation time and go-past time was very similar and thus we report it jointly. Longer words and words at the end of a sentence were read slower. Both prediction surprisal and lexical integration were significant, such that more informative word relations forward and backward lead to faster reading times (see the visualization of main partial smooth effects in Fig. 5). In addition, significant interactions (Fig. 6) indicated that the effect of lexical integration on reading times was larger later in the sentence, and when Prediction surprisal was highest. This last interaction can be interpreted as readers relying more on integration when predicting is unreliable. Table 6 reports the partial deviance explained by each term in the model.

7. Discussion

Mainstream views of prediction posit that the mind is in stable "predictive mode," attempting to constantly anticipate sensory input. The question that we have been addressing in this study is whether forward-looking processes are the one central characteristic of real-time language processing as is frequently claimed (e.g., Clark, 2013; Friston, 2005) or just one,

Table 3
Summary of nonlinear trigram models with parametric interactions of Dataset 2: Newly collected self-paced reading times in replication of Dataset 1

	RT(w_t)	RT(w_{t+1})
(Intercept)	5.78***	5.80**
	(162.31)	(162.72)
Prediction Surp. X Lex. Integration	0.01	−0.00
	(2.07)	(−0.41)
Word Pos. X Prediction Surp.	0.01	−0.01
	(1.34)	(−2.11)
Word Pos. X Lex. Integration	0.00	−0.00
	(0.72)	(−0.56)
Word Pos. X Word Surp.	−0.00	0.01
	(0.64)	(0.53)
Word Length	0.97**	1.94**
	(6.18)	(2.03)
Word Position	2.80**	2.62**
	(137.59)	(82.86)
Word Surprisal	1.53	1.51**
	(2.69)	(8.07)
Prediction Surprisal	0.00	0.29
	(0.00)	(0.12)
Lexical Integration	1.54**	1.24
	(12.87)	(0.78)
AIC	12,804.13	11,873.81
BIC	14,261.77	13,282.14
Log Likelihood	−6215.83	−5754.68
Deviance	2244.82	2049.11
Deviance explained	0.35	0.34
Dispersion	0.12	0.12
*R*2	.34	.33
GCV score	6557.59	6083.64
Num. obs.	18,523	16,792
Num. smooth terms	11	11

Note: The table presents linear interactions first, with beta indicating the standardized regression weights, and *t*-values in brackets. For nonlinear main effects, the table reports the EDF and *F*-values in brackets. Asterisks indicate the significance level: *$p < .025$, **$p < .005$, ***$p < .0005$.

albeit important, processing principle among others. We have provided additional experimental evidence that points to an important role for backward-looking and integration, which is understudied (or at least currently underappreciated, cf. Ferreira & Chantavarin, 2018; Ferreira & Qiu, 2021) in the psycholinguistic literature in favor of purely forward models. Indeed, given the prevalent view that the brain is a "prediction machine" (Clark, 2013), we conjecture that our findings are of general theoretical importance for reevaluating the

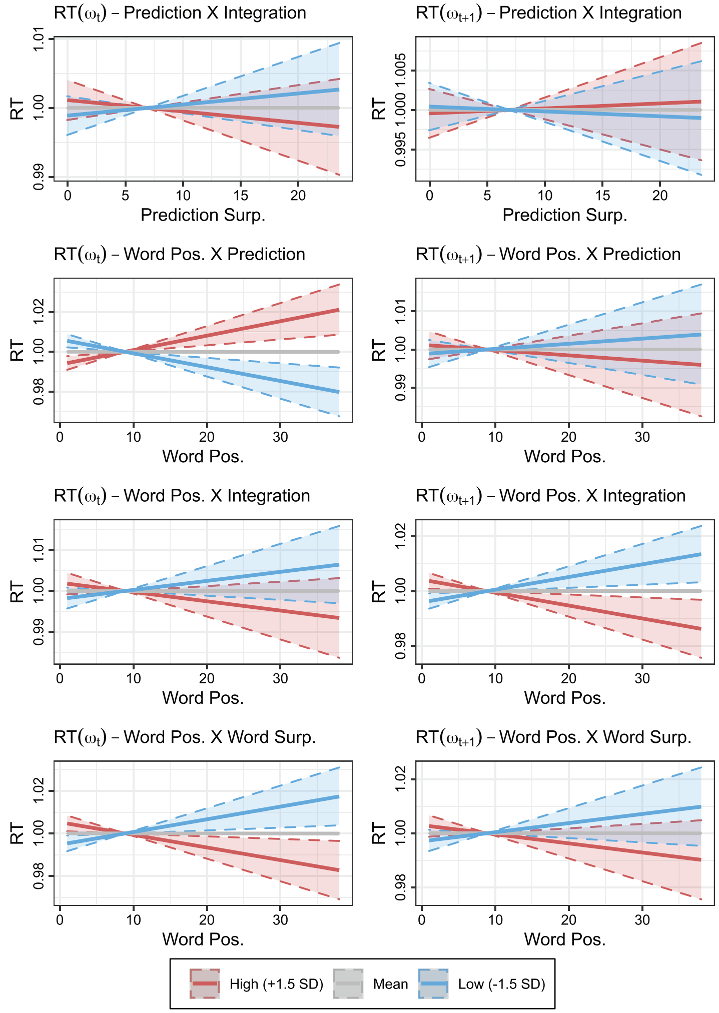

Fig. 4. Dataset 2: Plots of two-way interaction effects for current word RT and spillover effects.

relative contribution of prediction and integration to processing, or at the very least, call for considering integration more explicitly in contemporary models of language processing (cf. Aurnhammer & Frank, 2019; Bornkessel-Schlesewsky & Schlesewsky, 2008; Brouwer & Crocker, 2017; Brouwer et al., 2021; Kuperberg, 2007). To put this differently, our findings suggest that the brain is also an “integration machine” and inherently backward looking.

One strength of our study is that we assessed language processing not just on specific target words, but over entire sentences sampled from authentic sources. In addition, we carried out similar statistical analyses independently on three different linguistic datasets and involving two distinct dependent measures (reading times and eye-tracked fixations). We believe that in using this approach, the observed effects of forward-looking and backward-looking are

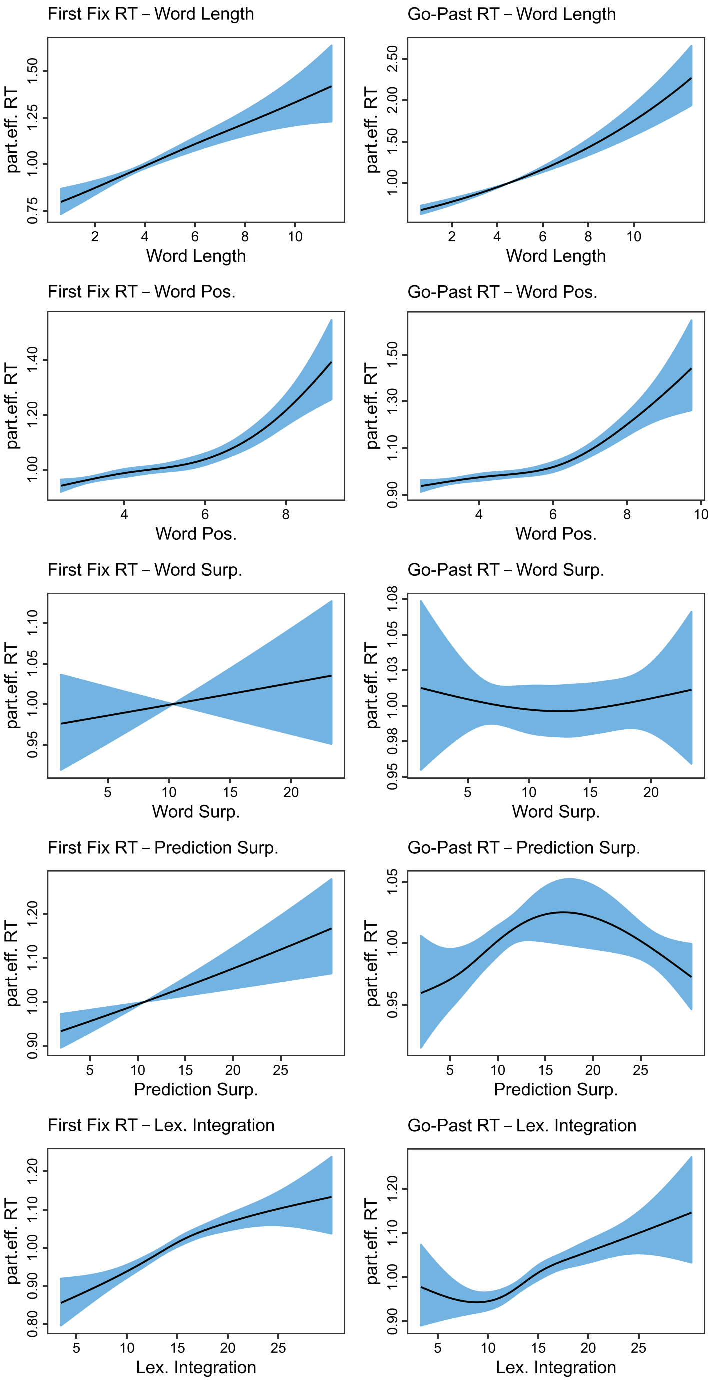

Fig. 5. Dataset 3: Plots of effects first fixation and go-past reading times (RT) for current word. Error bands represent 95% confidence intervals.

Table 4
Partial deviance explained by each term in the nonlinear model used for Dataset 2

Term	RT(w_t)	RT(w_{t+1})
Random effects		
Sentence Pos.	5.873	5.648
Subject	24.709	24.883
Word	0.337	0.549
Word Surprisal by Subject	0.004	0
Prediction Surprisal by Subject	0.082	0.029
Lexical Integration by Subject	0.101	0.157
Fixed effects		
Word Length	0	0.069
Word Position	2.057	1.264
Word Surprisal	0.008	0.051
Prediction Surprisal	0	0
Lexical Integration	0.014	0.021
Prediction Surp. X Lex. Integration	0.031	0
Word Pos. X Prediction Surp.	0.004	0
Word Pos. X Lex. Integration	0.005	0
Word Pos. X Word Surp.	0.003	0

Note: Calculated as the deviance explained by all terms minus the deviance in the submodel in which the term of interest is removed.

robust and generalizable. The pattern of results obtained supports the notion that backward information that estimates how much preceding context is likely given the current word being read (or to put it differently how likely it is that I read the preceding context given the current word)–may effectively reduce reading effort on the current word, in fact, at least in the present investigation, more so than forward information that estimates how much the current word is likely given preceding context. One interpretation of the results and arguably an intriguing (or to some provocative) possibility is that if readers rely more on backward than forward distributional cues, this suggests that it is more efficient to be engaged in integration than in prediction. Be that as it may, and we explicitly encourage further work on this, the important point is not to deny that prediction is one important part of language processing but rather that currently too little empirical emphasis is paid to the investigation of backward-looking in language processing. Such evidence is important to evaluate computational models of language electrophysiology such as the one by Brouwer and colleagues (2017, 2021) which explicitly includes an integrative component. It is also important for detailed mechanistic accounts about the interplay of prediction and integration and the notion that integration is a prerequisite for prediction (cf. Hale, 2001; Levy, 2008; Venhuizen, Crocker, & Brouwer, 2019a, 2019b).

What then is the relationship between context and backward looking? One (psychological level) candidate mechanism for backward looking is a form of integration whereby the processing system "waits" for a given perceptual input (a word in this case) and then processes it faster if the preceding context is a good fit (or, to put it probabilistically, it is more likely to

Table 5
Summary of fixed effects in generalized additive models for Dataset 3: First fixation and Go-past reading times

	First Fix RT	Go-Past RT
(Intercept)	−0.11*	0.03
	(−2.37)	(0.83)
Prediction Surp. X Lex. Integration	0.02*	0.02*
	(2.35)	(2.60)
Word Pos. X Prediction Surp.	−0.01	0.01
	(−1.03)	(0.53)
Word Pos. X Lex. Integration	−0.03**	−0.03**
	(−3.23)	(−3.44)
Word Pos. X Word Surp.	−0.00	0.02
	(0.98)	(0.95)
Word Length	1.60**	1.00**
	(316.59)	(2641.84)
Word Position	3.01**	2.94**
	(144.76)	(519.38)
Word Surprisal	0.39	0.17
	(4.93)	(3.34)
Prediction Surprisal	0.92**	1.34*
	(49.23)	(37.98)
Lexical Integration	2.32**	3.07**
	(127.41)	(176.77)
AIC	87,938.61	100,028.71
BIC	92,489.70	106,580.75
Log Likelihood	−43,430.41	−49,248.77
Deviance	25,183.12	29,120.43
Deviance explained	0.19	0.26
Dispersion	0.74	0.77
*R*2	.18	.25
GCV score	44,305.13	50,655.53
Num. obs.	34,380	38,494
Num. smooth terms	11	11

Note: The table presents linear interactions first, with beta indicating the standardized regression weights, and *t*-values in brackets. For nonlinear main effects, the table reports the EDF and *F*-values in brackets. Asterisks indicate the significance level: $*p < .025$, $** \ p < .005$.

precede it). Prediction and integration are of course related but, we believe, can be dissociated in useful ways. We consider prediction to be akin to the preactivation of upcoming words (or representations, e.g., semantic, phonological, though in the present paper, we focus on the "word level") ahead of time. Integration, in contrast, we define as the combination of incoming words into a higher order (e.g., sentential) representation in absence of such preactivation. According to our view then, context can modulate both prediction and integration. More precisely, context and prediction are straightforwardly related because context can preactivate upcoming words. Importantly, context can also affect integration because even

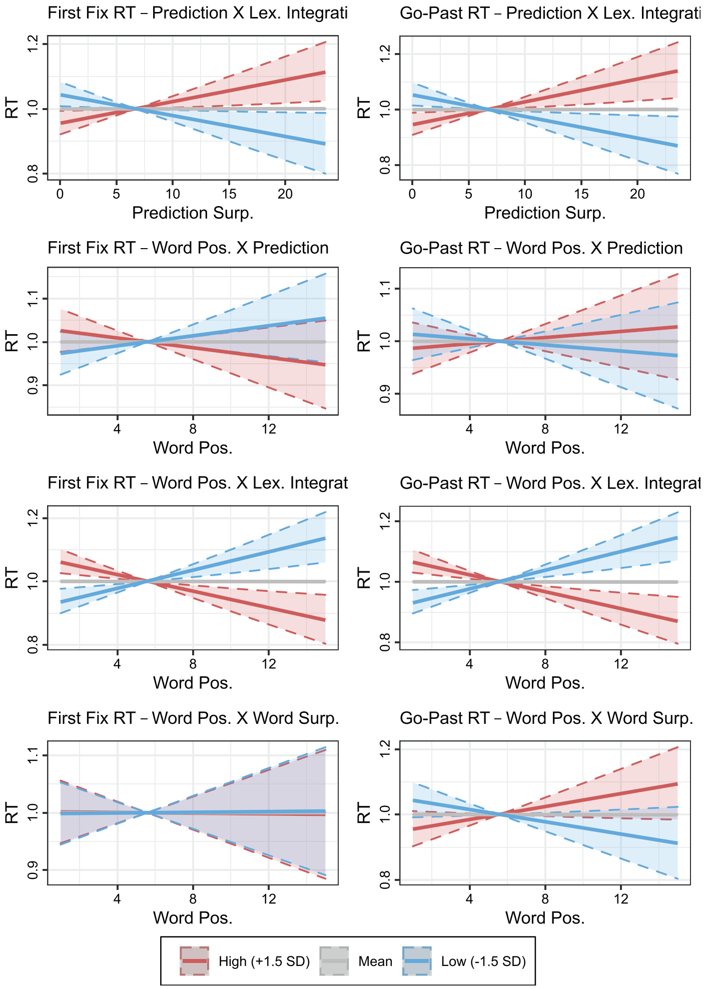

Fig. 6. Dataset 3: Plots of two-way interaction effects for first fixation and go-past reading times (RT).

without preactivation a word may be easier or more difficult to integrate with the preceding context after it has become activated (e.g., on reading the word) in a “bottom-up” fashion. This is in line with evidence that language input is often fast and suboptimal and may in some situations “afford” a little forward-looking (cf. Huettig & Mani, 2016). It is important to stress here again that prediction and integration are necessarily related, for example, because predicted words can be more easily integrated than nonpredicted words.

To our knowledge, the way we measured lexical integration is different from studies that have incorporated backward conditional probabilities in language processing. For example, McDonald and Shillcock (2003a, 2003b) defined a backward probability over the target word

Table 6
Partial deviance explained by each term in the generalized additive model for Dataset 3

Term	First Fix RT	Go-Past RT
Random effects		
Sentence Pos.	0.166	0.236
Subject	8.803	4.895
Word	5.338	12.596
Word Surprisal by Subject	0.453	0.245
Prediction Surprisal by Subject	0.183	0.108
Lexical Integration by Subject	0.068	0.053
Fixed effects		
Word Length	0	0
Word Position	0.07	0
Word Surprisal	0	0.011
Prediction Surprisal	0	0.001
Lexical Integration	0	0.016
Prediction Surp. X Integration	0.028	0.029
Word Pos. X Prediction Surp.	0.002	0
Word Pos. X Lex. Integration	0.028	0.017
Word Pos. X Word Surp.	0	0

Note: Calculated as the deviance explained by all terms minus the deviance in the submodel in which the term of interest is removed.

and the next one:

$$Backward\ TP(w_t, w_{t+1}) = P(w_t|w_{t+1}) = \frac{Freq\,(w_t, w_{t+1})}{Freq\,(w_{t+1})} \tag{7}$$

The fact that readers are sensitive to this latter type of backward probability has been interpreted as evidence that they can extract lexical information parafoveally when reading text. This type of "successor surprisal" has also been found in settings where preview is unavailable (Angele et al., 2015; Van Schijndel & Linzen, 2018). However, it is important to note that successor surprisal is a distinct effect from lexical integration as we intend it here, as it is assessed using a different metric. Thus, to our knowledge, none of the previous studies that attempt to capture comprehension effort in sentence processing have measured explicitly the contribution of backward-looking in the way we do here.

Note that in the present study, we used forward and backward probability as correlates of forward and backward looking, respectively, we did not consider these measures to provide any "pure reflection" of prediction and integration. We explicitly acknowledge the proxy aspect of both forward surprisal and lexical integration measures. It is also conceivable that both measures index to varying degrees prediction and integration processes. We do conjecture, however, that forward surprisal is a better proxy for forward-looking and prediction than our lexical integration measure. Similarly, we believe that lexical integration as measured in the present study is a better proxy for backward-looking and integration than forward surprisal.

In the current study, we have provided a crucial piece of evidence that backward looking and integrating linguistic context is an important part of language processing in real-time, at least in English. This in turn raises the possibility, to be investigated in future studies, that the prevalent word order of a given language fosters different habits of processing that maximize the informativeness of information. As such, our results raise intriguing questions for more cross-linguistic research (cf. Evans & Levinson, 2009; Henrich, Heine, & Norenzayan, 2010). Currently, the majority of psycholinguistic research is still carried out on English or closely related Indo-European languages. One promising avenue is then to test the hypothesis that the relative importance of forward versus backward-looking processing is partly language-specific, specifically by directly comparing the role of prediction and integration in speakers of left- and right-branching languages. For example, a straightforward prediction is that real-time sentence comprehension in speakers of left-branching languages, such as Korean or Hindi, should rely more heavily on predictive processes than in right-branching languages.

There are several other ways in which future work could build on the current results. First, in this study, we have modeled prediction and integration on language processing using a proxy for context based on trigrams. There exist arguably more powerful language models, for example, those based on contemporary RNNs, and which predict self-paced reading times better than n-gram models (Goodkind & Bicknell, 2018; Wilcox et al., 2020). Indeed, a trigram model cannot capture several important dependencies, including the example of ambiguity resolution in noise offered by Warren (1970) (which would require a 5-gram model). One of the reasons our trigram-based measures contributed little variance in the regression models may be attributable to the fact that readers make use of longer stretches of contextual information to reduce reading times. This could be tested using RNN models. One of the reasons why we did not rely on such models here is that they are intrinsically predictive, and thus cannot explicitly model integrative processes explicitly. And even if they did, their architecture may not be able to separate the contribution of prediction and integration. In other words, RNNs focus on the accumulation of a memory context in the forward direction only. There exists a class of bi-directional RNNs used in artificial intelligence (known as bi-directional long-short-term memory, LSTM) which improve target word prediction based on left and right contexts, however, this procedure can be applied successfully to words for which the entire sentence is available at once, and not incrementally as in a self-paced reading task where the right side of the context relative to any target work is still unavailable:

$$[\text{left\_context}]\left[\text{target\_word}\right]\left[\text{right\_context}\right]$$

When left and right contexts are available, for example, to a computer that can access an entire sentence at once, bi-directional LSTMs can integrate their forward estimate of $P\,(\mathit{Target}\mid\mathit{left\_context}) = x$ with information of $P\,(\mathit{Target}\mid\mathit{right\_context}) = y$. For instance, in a sentence such as “Today several … came out of school early,” it is possible for bi-directional RNNs to guess the word “pupils” with more certainty using (un)certainty from both right context and left context concurrently.

Arguably, a bi-directional language model may be a good candidate to assess successor effects, but not the case of the partial unfolding of language in real-time where the system can only rely on the preceding context for any given target word. Our measure of lexical integration is thus different from the measure that current bi-directional RNNs can offer. Rather, it corresponds to assessing P (*left_context* | *Target*) $= z$, and to our knowledge, it has no equivalent implementation in current neural language models.

Thus, so far, n-gram models represent an accessible source of estimation of probabilistic information in a forward and backward way, and in the context of the partial unfolding of the previous context alone. Furthermore, if the effects of prediction and integration on human reading times can be estimated even with simple n-gram models, this finding can be taken as a lower bound that more sophisticated and realistic language models are likely to improve upon, capturing language statistics even more accurately (e.g., Frank et al., 2013). However, we do not expect such language models to yield results that would obliterate or reverse our findings.

In future studies, it may also be possible to assess the independent contribution of forward surprisal and lexical integration on multiple real-time processing tasks, such as self-paced reading, repetition, and phrase recognition, by selectively manipulating the informativeness of each cue. For example, it is possible to sample from a large representative corpus multiword sequences that are matched in forward surprisal but differ in lexical integration, and vice versa. Based on our findings, we predict facilitatory effects of processing (e.g., faster reading times, more accurate repetitions, and faster recognition) for both types of stimuli.

Future work could also integrate and extend the metric of lexical integration to quantify the expectancy of a word with other information-theoretic metrics proposed to explain word informativity. For example, entropy refers to the uncertainty of a particular outcome, the greater the number of possible outcomes, the greater the entropy value. In the context of incremental sentence processing, the more possible continuations for a given part of the sentence at any given time, the more effort is expected to process the sentence continuation. In line with this prediction, experimental evidence suggests that the degree of uncertainty about an upcoming structure (Linzen & Jaeger, 2016) or the next word (Lowder, Choi, Ferreira, & Henderson, 2018) correlated with longer reading times. Relatedly, the entropy reduction induced by a word, measured as the difference between entropy and surprisal, quantifies the extent to which a word decreases the amount of uncertainty about what is being communicated (Frank, 2013).

Finally, it is important to note that the measure of entropy discussed above is a type forward, that is, prediction entropy, as it measures the number of possible word continuations given an initial sentence fragment. Just as suprisals based on forward and backward conditional probabilities can be distinct predictors, a measure of *integration entropy* (how many different words can precede a target word being read) might turn out to explain reading times variance independent of prediction entropy. An open empirical question is thus how several information-theoretic measures calculated in the backward direction looking at previous context will be relevant predictors of processing difficulty along known forward measures.

Acknowledgments

We thank Pia Knoeferle, Roger Levy, Vsevolod Kapatsinski, Ben Falandays, Michael Spivey, Layla Unger, and four anonymous reviewers for useful comments that led to an improved manuscript. This work was partly supported by the Research Council of Norway through its Centres of the Excellence funding scheme, project number 223265.

Open Research Badges

This article has earned Open Data and Open Materials badges. Data and materials are available at https://osf.io/7vkdt/.

References

Altmann, G. T., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, *33*(4), 583–609.

Anderson, J. E., Aarts, A. A., Anderson, C. J., Attridge, P. R., Attwood, A., Axt, J., Babel, M., Bahník, Š., Baranski, E., Barnett-Cowan, M., et al. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251).

Angele, B., Schotter, E. R., Slattery, T. J., Tenenbaum, T. L., Bicknell, K., & Rayner, K. (2015). Do successor effects in reading reflect lexical parafoveal processing? evidence from corpus-based and experimental eye movement data. *Journal of Memory and Language*, *79*, 76–96.

Aurnhammer, C., & Frank, S. L. (2019). Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia*, *134*, 107198.

Baayen, H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, *94*, 206–234.

Baayen, R. H., & Linke, M. (2019). An introduction to the generalized additive model. *A practical handbook of corpus linguistics*. Springer, Berlin. (forthcoming).

Baayen, R. H., van Rij, J., de Cat, C., & Wood, S. (2018). Autocorrelated errors in experimental data in the language sciences: Some solutions offered by generalized additive mixed models. *Mixed-effects regression models in linguistics* (pp. 49–69). Springer.

Baggio, G., & Hagoort, P. (2011). The balance between memory and unification in semantics: A dynamic account of the N400. *Language and Cognitive Processes*, *26*(9), 1338–1367.

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2008). An alternative perspective on 'semantic P600' effects in language comprehension. *Brain Research Reviews*, *59*(1), 55–73.

Boston, M. F., Hale, J., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the potsdam sentence corpus. *Journal of Eye Movement Research*, *2*(1).

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2015). Effects of prediction and contextual support on lexical processing: Prediction takes precedence. *Cognition*, *136*, 135–149.

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2017). Goals and strategies influence lexical prediction during sentence comprehension. *Journal of Memory and Language*, *93*, 203–216.

Brouwer, H., & Crocker, M. W. (2017). On the proper treatment of the N400 and P600 in language comprehension. *Frontiers in Psychology*, *8*, 1327.

Brouwer, H., Delogu, F., Venhuizen, N. J., & Crocker, M. W. (2021). Neurobehavioral correlates of surprisal in language comprehension: A neurocomputational model. *Frontiers in Psychology*, *12*, 615538.

Brouwer, H., Fitz, H., & Hoeks, J. (2012). Getting real about semantic illusions: rethinking the functional role of the P600 in language comprehension. *Brain Research*, *1446*, 127–143.

Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language*, *109*, 104047.

Camblin, C. C., Gordon, P. C., & Swaab, T. Y. (2007). The interplay of discourse congruence and lexical association during sentence processing: Evidence from ERPs and eye tracking. *Journal of Memory and Language*, *56*(1), 103–128.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204.

Coupé, C. (2018). Modeling linguistic variables with regression models: Addressing non-gaussian distributions, non-independent observations, and non-linear predictors with random effects and generalized additive models for location, scale, and shape. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00513

Dell, G. S., & Chang, F. (2014). The p-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634), 20120394.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8* (8), 1117–1121.

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

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, *32* (5), 429–448.

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44* (4), 491–505.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, *41* (4), 469–495.

Federmeier, K. D., McLennan, D. B., De Ochoa, E., & Kutas, M. (2002). The impact of semantic memory organization and sentence context information on spoken language processing by younger and older adults: An ERP study. *Psychophysiology*, *39* (2), 133–146.

Ferreira, F., & Chantavarin, S. (2018). Integration and prediction in language processing: A synthesis of old and new. *Current Directions in Psychological Science*, *27* (6), 443–448.

Ferreira, F., & Qiu, Z. (2021). Predicting syntactic structure. *Brain Research*, *in press*.

Frank, S. L. (2013). Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science*, *5* (3), 475–494.

Frank, S. L., Monsalve, I. F., Thompson, R. L., & Vigliocco, G. (2013). Reading time data for evaluating broad-coverage models of english sentence processing. *Behavior Research Methods*, *45* (4), 1182–1190.

Frisson, S., Harvey, D. R., & Staub, A. (2017). No prediction error cost in reading: Evidence from eye movements. *Journal of Memory and Language*, *95*, 200–214.

Friston, K. (2005). A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological Sciences*, *360* (1456), 815–836.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., … & Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, *25* (3), 369–380.

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, 10–18.

Gwilliams, L., Linzen, T., Poeppel, D., & Marantz, A. (2018). In spoken word recognition, the future predicts the past. *Journal of Neuroscience*, *38* (35), 7585–7599.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, 1–8.

Harmon, Z., & Kapatsinski, V. (2021). A theory of repetition and retrieval in language production. *Psychological review*, *128* (6), 1112.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33* (2-3), 61–83.

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience*, *13* (2), 135–145.

Hintz, F., Meyer, A. S., & Huettig, F. (2020). Activating words beyond the unfolding sentence: Contributions of event simulation and word associations to discourse reading. *Neuropsychologia*, 107409.

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research*, *1626*, 118–135.

Huettig, F., Audring, J., & Jackendoff, R. (2022). A parallel architecture perspective on pre-activation and prediction in language processing. *Cognition*, *224*, 105050.

Huettig, F., & Guerra, E. (2019). Effects of speech rate, preview time of visual context, and participant instructions reveal strong limits on prediction in language processing. *Brain Research*, *1706*, 196–208.

Huettig, F., & Mani, N. (2016). Is prediction necessary to understand language? Probably not. *Language, Cognition and Neuroscience*, *31*(1), 19–31.

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, *1146*, 23–49.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59.

Laszlo, S., Stites, M., & Federmeier, K. D. (2012). Won't get fooled again: An event-related potential study of task and repetition effects on the semantic processing of items without semantics. *Language and Cognitive Processes*, *27*(2), 257–274.

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics:(De) constructing the N400. *Nature Reviews Neuroscience*, *9*(12), 920–933.

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

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, *6*(3), 172–187.

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

Lowder, M. W., Choi, W., Ferreira, F., & Henderson, J. M. (2018). Lexical predictability during natural reading: Effects of surprisal and entropy reduction. *Cognitive Science*, *42*, 1166–1183.

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, *88*, 22–60.

Mantegna, F., Hintz, F., Ostarek, M., Alday, P. M., & Huettig, F. (2019). Distinguishing integration and prediction accounts of ERP N400 modulations in language processing through experimental design. *Neuropsychologia*, *134*, 107199.

Marra, G., & Wood, S. N. (2011). Practical variable selection for generalized additive models. *Computational Statistics and Data Analysis*, *55*(7), 2372–2387. https://doi.org/10.1016/j.csda.2011.02.004

McDonald, S. A., & Shillcock, R. C. (2003a). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science*, *14*(6), 648–652.

McDonald, S. A., & Shillcock, R. C. (2003b). Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research*, *43*(16), 1735–1751.

McQueen, J. M., & Huettig, F. (2014). Interference of spoken word recognition through phonological priming from visual objects and printed words. *Attention, Perception, & Psychophysics*, *76*(1), 190–200.

Metusalem, R., Kutas, M., Urbach, T. P., Hare, M., McRae, K., & Elman, J. L. (2012). Generalized event knowledge activation during online sentence comprehension. *Journal of Memory and Language*, *66*(4), 545–567.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., … Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Du Sert, N. P., Simonsohn, U., & Wagenmakers, E.-J. Ware, J. J., & Ioannidis, J. P., (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 1–9.

Murakami, A. (2016). Modeling systematicity and individuality in nonlinear second language development: The case of English grammatical morphemes. *Language Learning*, *66*(4), 834–871.

Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., … Von Grebmer Zu Wolfsthurn, S. (2020). Dissociable effects of prediction and integration during language comprehension: Evidence from a large-scale study using brain potentials. *Philosophical Transactions of the Royal Society B*, *375*(1791), 20180522.

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., … Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife*, *7*, e33468.

Norris, D., McQueen, J. M., & Cutler, A. (2016). Prediction, Bayesian inference and feedback in speech recognition. *Language, Cognition and Neuroscience*, *31*(1), 4–18.

Onnis, L., & Huettig, F. (2021). Can prediction and retrodiction explain whether frequent multi-word phrases are accessed 'precompiled' from memory or compositionally constructed on the fly? *Brain Research*, *1772*(147674), 1–6. https://doi.org/10.1016/j.brainres.2021.147674

Otten, M., & Van Berkum, J. J. (2007). What makes a discourse constraining? Comparing the effects of discourse message and scenario fit on the discourse-dependent N400 effect. *Brain Research*, *1153*, 166–177.

Otten, M., & Van Berkum, J. J. (2008). Discourse-based word anticipation during language processing: Prediction or priming? *Discourse Processes*, *45*(6), 464–496.

Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin*, *144*(10), 1002.

Pickering, M. J., & Garrod, S. (2013). Forward models and their implications for production, comprehension, and dialogue. *Behavioral and Brain Sciences*, *36*(4), 377–392.

R Core Team. (2021). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, *2*(9), 693–705.

Rommers, J., Meyer, A. S., Praamstra, P., & Huettig, F. (2013). The contents of predictions in sentence comprehension: Activation of the shape of objects before they are referred to. *Neuropsychologia*, *51*(3), 437–447.

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*(1), 76–80.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Sóskuthy, M. (2017). Generalised additive mixed models for dynamic analysis in linguistics: A practical introduction. *arXiv preprint arXiv:1703.05339*.

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

Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443.

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

Van Schijndel, M., & Linzen, T. (2018). Can entropy explain successor surprisal effects in reading? *Proceedings of the Society for Computation in Linguistics (SCiL)*, 1–7.

Venhuizen, N. J., Crocker, M. W., & Brouwer, H. (2019a). Expectation-based comprehension: Modeling the interaction of world knowledge and linguistic experience. *Discourse Processes*, *56*(3), 229–255.

Venhuizen, N. J., Crocker, M. W., & Brouwer, H. (2019b). Semantic entropy in language comprehension. *Entropy*, *21*(12), 1159.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, *167*(3917), 392–393.

Wicha, N. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *Journal of Cognitive Neuroscience*, *16*(7), 1272–1288.

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. P. (2020). On the predictive power of neural language models for human real-time comprehension behavior. 1707–1713.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(1), 3–36.

Yan, S., & Jaeger, T. F. (2020). (Early) context effects on event-related potentials over natural inputs. *Language, Cognition and Neuroscience*, *35*(5), 658–679.