GLOSSA psycholinguistics

Haeuser, K. I., & Borovsky, A. (2025). Got it right up front? Further evidence for parallel graded prediction during prenominal article processing in a self-paced reading study. *Glossa Psycholinguistics*, 4(1): 7, pp. 1–40. DOI: https://doi.org/10.5070/G6011.1636

eScholarship University of California

Got it right up front? Further evidence for parallel graded prediction during prenominal article processing in a self-paced reading study

Katja I. Haeuser, Department of Psychology, Saarland University, Germany; Collaborative Research Center Information Density and Linguistic Encoding (SFB 1102), Saarland University, Germany, khaeuser@coli.uni-saarland.de

Arielle Borovsky, College of Health and Human Sciences, Purdue University, Lafayette, Indiana, USA, aborovsky@purdue.edu

Recent studies suggest that language users generate and maintain multiple predictions in parallel, especially in tasks that explicitly instruct participants to generate predictions. Here, we investigated the possibility of parallel gradedness of linguistic predictions in a simple reading task, using a new measure that captures the probabilistic difference between multiple sentence completions (*imbalance*). We focus on prenominal gender-marked articles from German in order to obtain prediction-specific effects. Native speakers of German read predictable or unpredictable gender-marked nouns that were preceded by prediction-consistent or -inconsistent prenominal articles. Sentence frames either biased expectations more strongly towards the most likely continuation of the sentence, or they balanced expectations between the first and second most likely continuation. The results showed reading facilitation for gender-marked articles when sentences were more biased but slowing when sentences were more balanced, irrespective of article predictability. We conclude that readers issue multiple prenominal predictions and weigh those according to their likelihood, providing evidence for parallel gradedness of prenominal predictions. The results are discussed in the light of theoretical models on prediction and rational sentence processing.

1. Introduction

Language processing involves a predictive component, such that comprehenders are able to anticipate lexical (e.g., DeLong et al., 2019; 2021; Haeuser & Borovsky, 2024; Ito et al., 2016; Laszlo & Federmeier, 2009; Mantegna et al., 2019; Staub, 2015; Van Wonderen & Nieuwland, 2023), semantic (e.g., Altmann & Kamide, 1999; Borovsky et al., 2012; Federmeier & Kutas, 1999; Federmeier et al., 2002; Kuperberg et al., 2020; Mani & Huettig, 2012), or syntactic features (e.g., Dikker et al., 2010; Staub & Clifton, 2006; for review, see Ferreira & Qiu, 2021) of upcoming sentence constituents. For example, many people complete the sentence *Tim threw a rock and broke the …* with the word *window*, and not with the words *camera* or *dustbin*, even though these other completions may also be plausible. While ample evidence points to the plausibility of predictive processing, many open questions remain regarding the specificity and detail of linguistic predictions. Do predictors pre-activate broad semantic features of upcoming material, or do they predict specific word forms (lexical prediction vs. semantic prediction; e.g., Luke & Christianson, 2016)? Do people only pre-activate one word form or semantic feature, or do they predict multiple ones (all-or-nothing vs. parallel prediction; Kuperberg & Jaeger, 2016)? Finally, if multiple predictions are plausible, are they issued in a parallel or serial fashion (graded vs. serial prediction; Kuperberg & Jaeger, 2016)? We set out to explore these questions in a self-paced reading study of German gender-marked articles using a novel variable (*imbalance*) that measures the probabilistic difference between simultaneously activated sentence completions. We begin by reviewing the literature on prenominal prediction and gradedness of predictions.

1.1 Prenominal prediction

Prediction effects are often measured prenominally, for example on indefinite or definite articles that precede a noun (e.g., *You never forget how to ride a/an …*). Processing facilitation for prediction-consistent articles or adjective inflections is then taken as evidence that the noun, along with its phonological and/or morpho-syntactic features, has been predicted (e.g., DeLong et al., 2005; Fleur et al., 2020; Husband, 2022; Nicenboim et al., 2020; Otten & Van Berkum, 2008; Szewczyk & Schriefers, 2013; Szewczyk & Wodniecka, 2020; Urbach et al., 2020; Van Berkum et al., 2005; Wicha et al., 2003; Wicha et al., 2004; but see e.g., Ito et al., 2017; Kochari & Flecken, 2019; Nieuwland et al., 2018; Nieuwland et al., 2020). Since articles bear little or no semantic content (Urbach et al., 2020), prediction effects measured in this fashion are unlikely to be driven by late-stage integration (Baggio and Hagoort, 2011; Huettig, 2015; Huettig & Mani, 2016; Ito et al., 2017; Lau et al., 2008; Mategna et al., 2019; Otten & Van Berkum, 2008; Van Berkum, 2010).

Indeed, one line of research on prenominal prediction has shown that people use information provided by gender-marked articles or adjective inflections to infer the identity of upcoming referents (e.g., Cholewa et al., 2019; Dahan et al., 2000; Hopp, 2013; Huettig & Guerra, 2019;

Huettig & Janse, 2016; Lew-Williams & Fernald, 2010). For example, using the visual world eye-tracking technique, Huettig and Janse (2016) showed that people fixate on referents that are uniquely identified by preceding Dutch articles before these referents become acoustically available.

Another line of research has shown that comprehenders experience processing difficulty when prenominally presented material mismatches with their expectations. In one of the earliest studies that demonstrated such a prenominal prediction mismatch effect, Wicha and colleagues (2003) showed that prediction-inconsistent Spanish gender-marked articles (e.g. *el* in a sentence context that biases expectations towards a feminine noun requiring *la*) elicited an N400-like effect in the ERP record (the N400 is often argued to reflect lexical retrieval or semantic processing; for review see Kutas & Federmeier, 2011; Van Petten & Luka, 2012). This was interpreted as evidence that comprehenders had, indeed, anticipated predictable nouns, as well as features of their word forms (i.e., grammatical gender), during sentence processing. The fact that processing differences emerged pre-nominally, i.e. before the critical predicted nouns, may show that comprehenders indeed used the sentence context to actively predict upcoming information, instead of passively waiting for new information to become available (e.g., Baggio and Hagoort, 2011; Van Berkum, 2010).

Prenominal prediction has been addressed in many subsequent studies (for a recent meta-analysis, see Nicenboim et al., 2020; for review, see Fleur et al., 2020), and these studies have identified at least three distinct cognitive processes that may be captured in prenominal prediction effects. One hypothesis is that prenominal prediction effects may indicate that language users notice a mismatch between their noun predictions and the actual input, which is inconsistent with these predictions (*error detection*; e.g., Szewczyk & Wodniecka, 2020). According to another hypothesis, prenominal mismatch effects could reflect an updating process of some sorts, in which language users update or adjust their noun predictions according to the informativity of prenominal information and begin entertaining other likely continuations of the sentence or start generating entirely new ones (*updating of the noun*; see Chow & Chen, 2020; Fleur et al., 2020; Szewczyk & Wodniecka, 2020; also see Ness & Meltzer-Asscher, 2018). Finally, prenominal prediction effects could result from comprehenders activating prenominal article forms themselves, not (only) nouns (*article form prediction*; Fleur et al., 2020).

Possibly in line with their diverse cognitive nature, prenominal prediction effects in ERP studies vary greatly regarding their timing or polarity, and recently, some prenominal prediction effects have also failed to replicate (for Dutch gender marking, see Kochari & Flecken, 2019; Nieuwland et al., 2020; for the English *a/an* contrast, see Ito et al., 2017; Nieuwland et al., 2018; also see Ito et al., 2016; DeLong et al., 2019; Urbach et al., 2020; Yan et al., 2017). This variability illustrates that more research is needed to fully understand prenominal prediction, both research using ERPs and also research using other experimental paradigms, which are currently scarce.

1.2 Parallel gradedness of predictions

There are several definitions of *gradedness* in the extant literature, but few studies have addressed parallel gradedness of predictions. Some studies have defined gradedness as the linear relationship between predictability and measure of integration, such as N400 amplitude. For example, DeLong and colleagues (2005) found evidence of what they referred to as *gradedness of prenominal predictions*, in that the N400 amplitude elicited by phonologically aligned articles varied inversely with the articles' cloze probability: N400 amplitudes were small for more predictable articles, and became progressively larger the more unpredictable an article was (also see Kutas & Hillyard, 1984; Nieuwland et al., 2018; Urbach et al., 2020; see Brothers & Kuperberg, 2021; Levy, 2008; Reichle et al., 2003, for studies addressing the relationship between cloze probability and reading times). However, in that and many subsequent studies, N400 facilitation was only measured for one single continuation of a single sentence, without showing that a single sentence context may have cued other likely responses. Hence, these studies make it difficult to assess parallel gradedness of predictions.

Other studies have defined gradedness as the attenuated N400 response elicited by an unpredictable word when that word shares semantic features with the most likely continuation (Federmeier & Kutas, 1999; Federmeier et al., 2002), or when that unpredictable word is rendered more predictable by the presentation of a prenominal adjective (e.g., Boudewyn et al., 2015; also see Frisson et al., 2017; Heilbron et al., 2022; Szewczyk et al., 2022). These studies are important as they show that processing facilitation is dependent on degree rather than absolute semantic overlap, and that language users quickly make use of new linguistic material to update their currently held predictions. Again, though, not many of these studies measured the likelihood with which single sentences may have facilitated multiple likely responses.

More relevant to the present study are two recent studies demonstrating that language users activate multiple continuations for a given sentence context in parallel. Staub and colleagues (2015) measured participants' naming times for sentence continuations in high and low constraint sentences in a speeded cloze task to investigate parallel predictions during sentence processing. One result of that study was that participants provided faster responses in highly constraining (vs. more weakly constrained) sentence contexts across the board, not only for the modal response in the cloze task (i.e., first-best completion), but also for other, less predictable competitors. The authors also found that participants produced low-cloze competitors more quickly in sentences that had a highly probable modal response, compared to sentences that did not. These results were interpreted to mean that multiple possible continuations are activated simultaneously, racing towards a retrieval threshold, and that in some proportion of trials and due to random noise in activation levels, a low-cloze probability word can actually "win" the race (also see Ness & Meltzer-Asscher, 2021a). Crucially, the fact that latencies for competitor responses were influenced by the likelihood of the modal response means that comprehenders,

even those who did not predict the modal response, issue multiple predictions from a shared probability space, rather than sampling from their idiosyncratic probability distributions (also see Ness & Meltzer-Asscher, 2021a).

In a follow-up study to this paper, Ness and Meltzer-Asscher (2021b) found that parallel predictions in a speeded cloze test were modulated by the degree of semantic overlap between the modal response and the strength (i.e., likelihood) of its competitor: When the modal and its competitor were related, stronger competitors facilitated the modal, but when they were unrelated, stronger competitors inhibited the modal. Hence, these important studies illustrate that language users are capable of generating parallel predictions about upcoming linguistic content, and they do so in a graded manner.

However, a question these two studies do not answer is whether parallel predictions affect processing times outside of a cloze test, in a paradigm that does not ask participants to actively generate predictions about upcoming words, which can change the way people process sentences (e.g., Brothers et al., 2015; Brothers et al., 2017; Chung & Federmeier, 2023; Dave et al., 2018; Wlotko & Federmeier, 2015). A recent study by Brothers and colleagues (2023) explored parallel gradedness of predictions by using the N400 ERP component. One of the questions in that article was whether parallel predictions compete with each other by means of mutual inhibition, or if they facilitate one another as a function of their semantic relatedness (akin to Ness & Meltzer-Asscher, 2021b). The results of that study showed no support of an inhibition account. Specifically, N400 amplitudes patterned with cloze probability, but they were not additionally modulated by whether or not a word had a competitor. In contrast, the authors did find evidence for modulation of N400 amplitudes depending on semantic relationship: N400 amplitudes to both modal and competitor responses were reduced the more semantically related the two were. One of the conclusions of that article was that pre-activating multiple alternatives incurs cognitive benefits, rather than costs, as mutually co-activated representations facilitate, rather than inhibit, one another – to the extent that they are semantically related (note that we return to this study in Section 4). However, that article did not address prenominal prediction, which may be a more direct measure of linguistic prediction (as opposed to semantically-based association or integration; e.g., Ferreira & Chantavarin, 2018; Ferreira & Qiu, 2021; Kukona, 2020; Van Berkum, 2010).

In a recent study from our lab (Haeuser et al., 2022), we used the gender marking of the German language to measure gradedness of prenominal predictions in a self-paced reading task. Native speakers of German were presented with sentence contexts such as in (1) which constrained expectations relatively strongly towards a particular noun. Half of the sentences were continued with an unpredictable (but plausible) noun from a different grammatical class that required a different prenominal gender-marked article (see (1)a vs b). Hence, unpredictable prenominal gender-marked articles should act as a salient cue to indicate a prediction mismatch, as they are incompatible with the predictable noun.

(1) a Da Anne Angst vor Spinnen hat, geht sie zuhause nur ungern nach unten in ***den*** $_{MASC\text{-}ACC}$ ***Keller***.
'Since Anne is afraid of spiders, she does not like going down into **the basement**.'

b Da Anne Angst vor Spinnen hat, geht sie zuhause nur ungern nach unten in ***das*** $_{NEUT\text{-}ACC}$ ***Schlafzimmer***.
'Since Anne is afraid of spiders, she does not like going down into **the bedroom**.'

Crucially, we took into account not only the predictability of the more and less predictable sentence completions, which we used as an index for all-or-nothing prediction. We also introduced a new variable that measures to what extent single sentence contexts may have biased expectations towards a highly dominant completion or towards multiple probable continuations (named *imbalance* of a sentence frame, see below).[1] Thus, imbalance is a measure of parallel gradedness of prediction. Unlike the two studies cited above (Ness & Meltzer-Asscher, 2021b; Staub et al., 2015), which computed the cloze probabilities of the competitor words out of those cases in which participants did not complete a given sentence fragment with the modal response, the present study used a version of cloze task in which each participant was instructed to provide two completions to any given sentence fragment (henceforth, first- and second-best response). This multiple-response cloze procedure allowed us to investigate if the sentence frames cued competitor responses, over and above the most predictable (or first-best) response in individual participants. Crucially, though, participants were not required to give either a first-best or a second-best response (i.e., they could leave both continuations blank). This procedure allowed us to identify and exclude items which did not particularly cue expectations in one or the other direction.

We defined *imbalance* as the relative probabilistic difference between the two most frequent first-best and second-best responses, by subtracting the likelihood of the second-best response from the likelihood of the first-best response. By *first-best* and *second-best* we mean the best (i.e., most probable) of the responses provided as first and second alternatives. For example, in a sentence like *Sie trinkt ihren Kaffee mit … (She drank her coffee with…)* , participants might provide a first-best response of *milk/der Milch* approximately 70% of the time, and a second-best response of *cream/der Sahne* approximately 50% of the time, yielding an imbalance of 0.2 (resulting from the subtraction of 0.7 minus 0.5). It is important to note here that, due to the multiple probe nature of the cloze task, the summed probabilities of the first- and second-best responses can range between 0 and 2 – not 0 and 1, as in typical single-response cloze tasks, and that the imbalance value can range between –1 and +1. Larger (absolute) imbalance values

[1] We acknowledge that, in a follow-up to their main analysis, Ness & Meltzer-Asscher (2021b) did consider effects of what we refer to imbalance. According to the results, a model that included this variable did not fit the data better than a model that included competitor cloze instead of imbalance; even though qualitatively, the models showed the same pattern of results.

identify sentence frames that are more strongly biased towards one single noun continuation and its gender marked article (i.e., the first-best continuation), whereas smaller imbalance values indicate items in which the first-best and the second-best response are relatively equally probable. The fact that imbalance takes into account the probabilistic difference between two possible completions for a given sentence frame makes it different from *constraint*, another common predictability measure. *Constraint* measures the probability of the first-best response, but does not take into account that there may be a possible competitor. That being said, constraint and imbalance obviously correlate because imbalance is computed as the probabilistic difference between the first-best cloze probability and a competitor probability, and constraint is equivalent to the cloze probability of that first-best response (i.e., predictability; we return to this in Section 4). Note that we also present evidence in this article that imbalance adds something to the study of predictability that constraint alone does not address.

Predictability and imbalance were then used to predict reading times of gender-marked articles in a group of German-speaking young adults (n = 84) who read sentences silently for comprehension in a cumulative moving-window self-paced reading task (note that a group of older adults was additionally tested to address questions of aging and prediction which are not explored in the current study). Critical sentences were presented with prenominal adjectives between the gender-marked article and the head noun (adverbs and adjectives; e.g., *den/das oftmals schlecht belüftete/n Keller*, *the often badly ventilated basement*), inserted to make sure that spill-over effects from the article would not be confounded with integration effects at the noun.

The results of that study showed that, irrespective of article predictability, reading times of gender-marked articles were facilitated in more biased sentences, i.e., those items in which a dominant first-best response outweighed a far less likely alternative. In contrast, reading times at the article were slowed down when a sentence frame was more balanced, i.e., in those items in which first- and second-best response were relatively equally probable. Crucially, predictability (i.e., the likelihood of the actually presented gender-marked articles and nouns) did not modulate reading times. We concluded that adult language users not only generate parallel predictions about multiple prenominal articles instead of limiting their predictions to one first-best completion (lack of an effect for predictability), but that language users are also sensitive to the relative probability difference between these varying predictions. In other words, comprehenders weigh their parallel predictions according to their likelihood – evidence for parallel gradedness of predictions.

1.3 The present study

The goal of the present study was to establish further evidence for the possibility of parallel graded predictions, by extending the effects obtained in that earlier investigation to another experimental paradigm – word-by-word self-paced reading in an online study. One of our larger

goals in other studies was to test the robustness and generalizability of parallel gradedness of predictions in broader populations across a variety of ages and backgrounds, including children. We made the change to the word-by-word self-paced reading paradigm to make the experiment "web friendly", in order to support recruitment of a broader sample. Here, a central word location was available within our chosen software solution (LabVanced; Finger et al., 2017). This yielded fewer potential problems arising from fixed line breaks in a moving-window paradigm when PC screens varied in format.

We specified two predictors of interest, predictability and imbalance. Our expectations for the effects of these two variables were as follows. First and foremost, we expected a negative-going effect of imbalance at gender-marked articles, in other words, facilitated reading for articles when a sentence frame was more biased compared to when it was more balanced. Since word-by-word reading produces stronger spill-over effects than moving-window reading (Keating & Jegerski, 2015; Witzel et al., 2012), we also anticipated that effects of imbalance may more readily spill over onto subsequent critical regions in the sentence, e.g., onto the adjectival spill-over region after the article. For predictability, we expected to find globally facilitated reading for more predictable articles and nouns, but slowing for more unpredictable items, in line with prior research (e.g., DeLong et al., 2005; Van Berkum et al., 2005).

2. Methods

2.1 Power analysis

We conducted a power analysis, based on the data from the first 16 participants who completed the study, in order to estimate the sample size needed to replicate the findings reported in our original study. We chose 16 subjects, because our idea was to base the power analysis on a full counterbalance of the four experimental lists used in the study. We ran a linear mixed-effects model on the data of those 16 subjects to estimate the effects of article imbalance on log-RTs of the gender-marked article in the replication study, while controlling for predictability, trial number and RTs of the previous word. The full model specifications of the model were

```
Log(Article_RT) ~ Imbalance + Predictability + Trial_Number +
        Previous_Word_RT + (1 | Subject) + (1 | Item)[2]
```

[2] By-item random slopes for imbalance were not added, since imbalance did not vary over the two versions of an experimental item. By-subject random slopes for imbalance were warranted by the design, but had to be dropped because of fitting problems during power simulations (maybe because of a lack of variability that could account for this random effect).

In that model, the coefficient ($\hat{\beta}$ -value) for the scaled[3] imbalance value was –1.4 ms in raw RTs (which means a 1-SD increase in imbalance facilitated reading by 1.4 ms), and –0.013 in log RTs.

We then used the package *simr* in R (Green & McLeod, 2016) to estimate how many more subjects we needed in order to find an effect that was two thirds of the effect reported in our original study. The idea to estimate power for a smaller effect size was based on the idea that the obtained effect size is likely an overestimation of the true effect (e.g., Gelman & Carlin, 2014; also see Fleur et al., 2020). That power simulation showed that we needed a total of 80 subjects in order to obtain the specified effect with a power of 90.70% [95% CI: 88.73, 92.43]. When running the experiment, the recruitment methods were more successful than planned, yielding a slightly larger sample than the original goal which we opted to include in the final analysis. These additional participants were not excluded from the final analysis reported below.

However, due to the small sample size involved in this simulation, the power estimates obtained from the first 16 subjects may be misleading. Therefore, in order to inform future studies looking into potentially replicating the results obtained here, we also ran power simulations for a prospective study.[4] These power simulations estimated the sample size that would be needed if one wanted to obtain an imbalance effect that is 65%, 75%, 85% and 95% of the size of the effect found in the present study with a power of 80%. **Table 1** shows the simulation results.

Table 1: Prospective Power Simulation Results.

Effect Size	Scaled Imbalance	95% CI	Sample Size Needed
65%	$\hat{\beta}$ = –2.99 ms	[76.24, 81.39]	350
75%	$\hat{\beta}$ = –3.45 ms	[78.11, 83.10]	240
85%	$\hat{\beta}$ = –3.91 ms	[77.49, 82.53]	140
95%	$\hat{\beta}$ = –4.37 ms	[78.64, 83.58]	110

2.1.1 Participants

Eighty-seven native German adults were initially recruited as part of the broader aims of the study, but eight were excluded up front, because they were older than 45 years. We chose to exclude these older participants because prediction effects have been reported to vary with age (e.g., DeLong et al., 2012; Federmeier et al., 2002; Haeuser et al., 2018; Haeuser et al., 2019; Payne & Federmeier, 2018; Wlotko et al., 2012; also see Dave et al., 2018; Pichora-Fuller et al.,

[3] Throughout this article, when we scale predictors, we mean that we center them around their means, and additionally divide by their standard deviation. This means that *b* values for scaled predictors are comparable (as they are on the same scale), and model intercepts reflect reading times at the midpoint of the scaled predictors.

[4] We thank an anonymous reviewer for this suggestion.

1995; Steen-Baker et al., 2017; Tun & Wingfield, 1994; Wingfield et al., 1985; Wingfield & Stine-Morrow, 2000), and because we knew from our earlier investigation (Haeuser et al., 2022) that imbalance effects are less likely to emerge in older adults. Note that we ended up excluding one further participant because of low performance in the comprehension questions (see below). The remaining sample included the 16 subjects used for the power analysis and consisted of seventy-nine native German younger adults (i.e., under the age of 45, mean age = 28 years, range = 18–41; 45 female, 33 male, 1 non-binary), who participated for financial compensation (participants recruited through Prolific, $n = 61$) or course credit (participants recruited through the university's study recruitment website, $n = 18$). All participants had normal or corrected-to-normal vision and reported no neuropsychiatric medication and/or history of language impairments at the time of testing. The study was run online, using the stimulus presentation software *LabVanced* (Finger et al., 2017).

To ensure that our participants were performing norm-typically, all participants completed a battery of cognitive tests. These included the Digit Symbol Substitution Test (DSST; Salthouse, 1992) to measure processing speed, a measure of receptive vocabulary (i.e. the Peabody Picture Vocabulary Test–Version 4, German translation; Dunn & Dunn, 1997) and the verbal fluency test (category fluency and letter-S fluency). The results of these cognitive tests are presented in **Table 2**, and align with results obtained in previous studies on younger adults (Borovsky et al., 2012; Rommers et al., 2015;).

Table 2: Group Performance on the Cognitive Tests.

Task	Measure	Mean	SD	Range
DSST	Reaction times (ms)	1371	179	908–1770
PPVT	Raw Score	210	19	118–222[5]
VF Category Fluency	Correct Responses	25.55	6.97	5–40
VF Letter Fluency	Correct Responses	18.42	5.48	9–34

Note. DSST: Digit Symbol Substitution Test, processing speed. PPVT: Peabody Picture Vocabulary Test, receptive vocabulary. VF: Verbal fluency test.

[5] The average raw score in the Peabody Picture Vocabulary Test corresponds to a *T* score of 58 (*CI* [56; 60]) in the test manual for 17-year olds (which the oldest age group tested in the German PPVT norms). The *T* scale has a mean of 50 and a standard deviation of 10; hence a *T* score of 50 indicates a raw score equal to the mean, whereas a *T* score of 40 indicates a raw score one standard deviation below the mean. **Table 2** shows that the receptive vocabulary size of our sample was relatively norm-typical.

2.2 Materials

All materials are presented on the OSF site of this article (https://osf.io/8xubf/?view_only=50e2a7216b2e4d89bfef071f34f33d3e). Materials consisted of 48 sentence frames (e.g., *Nachdem Paul seinen Führerschein erhalten hatte, fuhr er ständig mit …*, English translation: *When Paul finally got his driver's license, he was always driving (around) with …*), presented in a predictable and unpredictable condition (Predictable: *dem[dative neuter] Auto*, English: *the car*; vs Unpredictable: *der[dative feminine] Gruppe*, English: *the group*), such that the pre-nominal gender-marked article (*dem* vs *der* in that item) could be used as an early cue that indicated whether or not the predictable noun would follow (the items were identical to the ones used in our earlier study). Each sentence was concluded by a sentence continuation (identical for predictable and unpredictable versions), which was inserted to allow for spill-over effects after the noun (e.g., *von Freunden auf den Landstraßen herum*, English: *of friends on the country roads*).

To allow for spill-over effects from prenominal gender-marked articles, adjectives were inserted before the phrase-final noun (e.g., for this item, *old but reliable*; note that the third word in the adjective region was gender-marked in half of the items, such that for those items, predictable and unpredictable versions of an item differed from one another.[6] These 48 items were used to analyze effects of predictability.

2.2.1 Predictability

We defined predictability as the cloze probability of the actually presented predictable and unpredictable gender-marked articles and nouns (see **Figure 1** for illustration of predictability and imbalance). Predictability was assessed by means of a cloze test in which 40 Psychology students were asked to complete each sentence with the first word that came to mind. For each item, participants were also asked to provide a second-best completion that could alternatively complete the sentence (see below). The predictability of gender-marked articles and nouns was then calculated by means of the first-best response that participants gave in the cloze test. Some items yielded responses in which the most frequent prenominal completions were indefinite articles or possessive pronouns (e.g., *ein Bier*, *ihre Mutter*; *a beer*, *her mother*). We conjectured that the presentation of definite articles in sentence contexts which do not specifically license definiteness (or, respectively, invite possessive markers rather than definite articles) may lead to a prediction violation based on pragmatic rather than lexical grounds, so we chose to exclude these items (see Fleur et al., 2020, for supporting empirical evidence). According to the results of the cloze test, predictable articles and nouns in the first responses were relatively highly predictable (M = 0.81, SD = 0.14, Range = 0.43–1.00; and M = 0.78, SD = 0.16, Range =

[6] Note that gender marking of the adjective region did not change our interpretation of the results (see exploratory analyses presented on the OSF site of this article; also see the discussion in Section 4).

0.30–1.00, respectively). We then chose unpredictable nouns for each sentence context (see **Figure 1**, bottom panel) making sure that the gender of the unpredictable noun was different from the gender of the noun provided as the first-best and second-best continuation. In other words, unpredictable gender-marked articles were truly unexpected, because they did not get named in the cloze test as potential first- or second-best continuations. Consequently, unpredictable gender-marked articles yielded near-zero probabilities ($M = 0.04$, $SD = 0.05$) and unpredictable nouns yielded cloze probabilities of zero. Note that for nineteen out of forty-eight items, we selected an unpredictable article form (German *die*) that is ambiguous in that it can be interpreted to foreshadow an unpredictable noun from a different grammatical gender, or the plural form of a noun. We chose to not remove these items from the item pool as we conjectured that both readings of the unpredictable article (e.g., different-gender noun vs plural noun) would likely elicit a lexical-semantic mismatch effect with the predictable noun. However, we ran follow-up models that estimated the effects of predictability at the gender-marked article when excluding those 19 items. The results remained the same as in our main analysis.

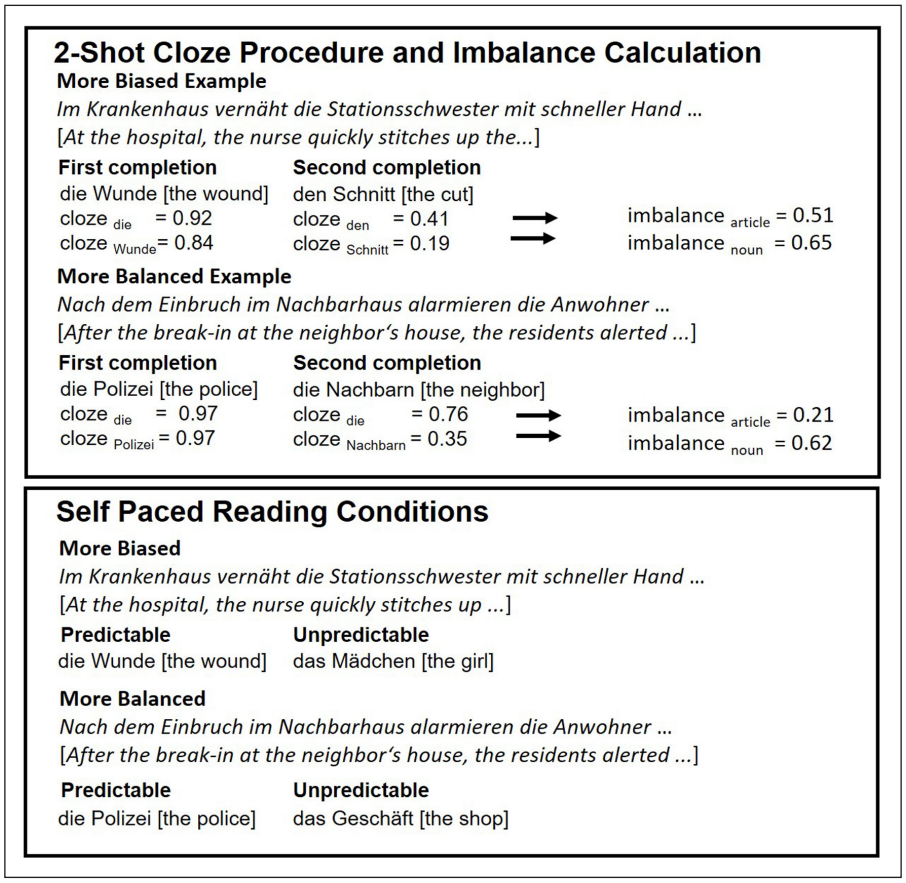

Figure 1: Stimulus Design. The graph illustrates the calculation of imbalance and the experimental items used in the self-paced reading task.

2.2.2 Imbalance

For the imbalance analysis, we compared both responses from the two-shot cloze procedure which included the first-best response in the cloze test as well as the second alternative ending that participants had provided for each sentence frame. Only a subset ($n = 32$) of the original 48 sentences could be used for this analysis, for several reasons. For example, some items yielded lexically identical, near-synonymous or semantically overlapping first and second completions (e.g., bus-bus; stove-stove top, goal-goal line; ship-boat; grass-lawn; trash-trash can); such items were excluded. Other items yielded no particular gender-marked article in the second completion, i.e. participants left the second completion blank (this happened frequently when first-response articles and nouns had cloze probabilities larger than 0.8, which suggests that it was difficult for participants to come up with a plausible second completion when the first completion was highly dominant). For these reasons, 16 from the initial 48 items had to be excluded. The resulting item pool, which allowed for a systematic investigation of imbalance consisted of 32 items.

Once the 32 items were identified, we calculated imbalance scores for each item by subtracting the by-sentence cloze probability of the second-best completion of the by-sentence cloze probability of the first-best completion (see **Figure 1**, top panel). This imbalance measure was calculated separately for gender-marked articles and nouns.[7] Hence, imbalance was a continuous variable. To illustrate this, when a sentence had yielded a cloze probability of 0.8 for the first-best article, and a cloze probability of 0.3 for the second-best article, the resulting imbalance value was 0.5. Alternatively, when a sentence had yielded a cloze probability of 0.8 in the first-best response, and a cloze probability of 0.5 in the second-best response, the resulting imbalance value was 0.3. Larger imbalance values indicate items which more strongly bias responses towards one dominant article or noun completion, because in such items, the top two responses are further away from another in the probability space. Smaller imbalance values indicate that the sentence context supports multiple equi-probable completions, because in balanced items, the top two completions are relatively equally likely. Because imbalance is derived from the difference from a two-shot cloze procedure, this measure is distinct from traditional measures of constraint / cloze, which are calculated from the top completion given a sentence context. However, since imbalance, constraint and cloze probability are all derived from the top-best completion of a sentence frame, these measures obviously correlate. The correlation between imbalance and constraint was .67 [95% CI: .51, .79]. The correlation between imbalance and the cloze probability of the predictable word was .67 [95% CI: .42, .82]. The correlation between constraint and the cloze probability of the predictable word was 1.

[7] We also calculated imbalance for items in which the gender-marked article was identical for first- and second-best responses, despite the fact that such items may, at first sight, appear biased towards a single continuation. However, in those items, the cloze probability of the second-best article was markedly lower than the cloze probability of the first best article, thereby warranting that we take into account the probabilistic difference between items rather than their gender alone.

Article imbalance varied between –0.09 and 0.62 ($M = 0.26$, $SD = 0.18$).[8] Imbalance values for nouns varied between 0.06 and 0.81 ($M = 0.47$, $SD = 0.22$). Note that noun imbalance values were never below zero, but article imbalance values were. Below-zero imbalance values indicate that the likelihood of the second-best response was higher than the likelihood of the first-best response. This illustrates that the sentences used in the experiment constrained expectations relatively strongly towards a particular noun, whereas expectations for articles were not nearly as strongly constrained. **Table 3** shows the average cloze probability values of the first- and second-best responses in the subset of items used to investigate imbalance, split out by articles and nouns. Histograms of article and noun imbalance values are shown in **Figure 2**.

Table 3: Mean Cloze Probabilities (and Ranges) of First and Second Completions in the Subset of 32 Items Used to Investigate Imbalance.

	First completion	**Second completion**
Article	0.79 (Range = 0.43–1.00)	0.53 (Range = 0.32–0.81)
Noun	0.76 (Range = 0.30–1.00)	0.29 (Range = 0.14–0.78)

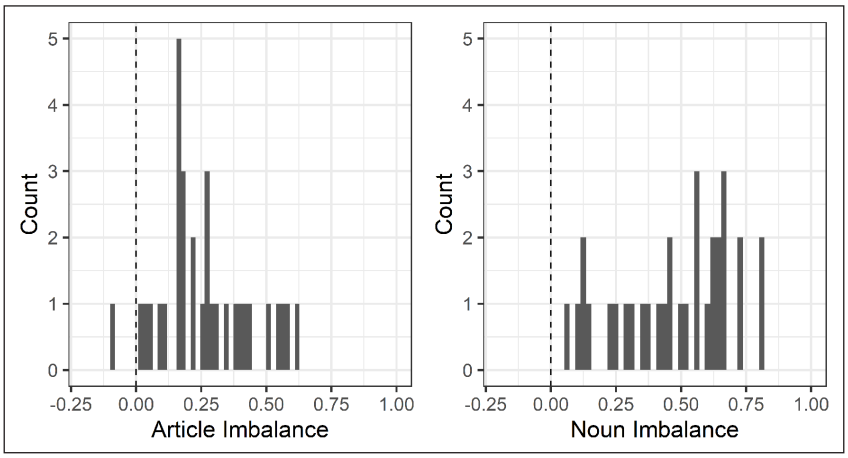

Figure 2: Histograms of Article and Noun Imbalance Values. Imbalance values are cloze probability difference scores, reflecting the probabilistic difference between the first-best and second-best completion in a cloze task in which participants were asked to provide two possible completions for each and every sentence frame. The dashed vertical line indicates the zero point of the cloze probability distribution. Negative values mean that the cloze probability of the second-best completion was higher than the cloze probability of the first-best completion.

[8] For one item, the cloze probability of the second-best article was higher than the cloze probability of the first-best article. This happened because several subjects had completed the corresponding sentence frame with the indefinite, rather than the definite, article, even though the cloze probability of the associated first-best noun was uniformly high (cloze = 0.80). This item was not removed from the final analysis. Follow-up analysis showed that removing this item resulted in the same pattern of findings as reported below.

2.3 Procedure

All 48 items were presented to participants (corresponding to 96 unique items). The 96 predictable and unpredictable items were arranged on two experimental lists, using a Latin Square design. A total of 30 moderately predictable filler sentences (taken from the Potsdam sentence corpus) was added to each list to make sure that participants continued to generate predictions in the course of the experiment, despite having predictions disconfirmed multiple times. The proportion of unpredictable sentences per each list was, therefore, 44%. Simple yes/no comprehension questions were added for 30% of all sentences on each list. As a final step, and to prevent trial order effects, we created two reversed-order lists from Lists 1 and 2, resulting in a total of four experimental lists. During data acquisition, each subject was assigned to one of the four lists.

The experiment consisted of two major sections. The first section was the self-paced reading task, in which participants read the 48 experimental and 30 filler items. This task was set up as a word-by-word reading task (i.e., no moving-window set up; no mask); participants saw one word appear in the center of the screen at a time and pressed the space bar to reveal the next word. Participants were instructed to read all sentences as fast as possible, and to answer all true/false comprehension questions as accurately as possible by pushing the "J" (Yes, correct) and "N" (No, incorrect) keys on the keyboard. Sentences were separated by a 500 ms fixation cross.

The second part of the experiment consisted of the three individual difference tasks, which were administered after the SPR task in the same order. Average completion time of the experiment was 35 minutes (SD = 11 minutes).

3. Results

We report our findings in two sections. The first section reports the effect of predictability on reading times; the second section reports the effects of imbalance. In each section, results are reported for the maximal number of items that allowed for the respective analysis. Thus, effects of predictability are reported for the full set of 48 items in the experiment. Effects of imbalance are reported for the subset of 32 items which allowed for this analysis (see 2.2). The dependent variable for all analyses was the reading times in the four critical regions of interest, log-transformed to avoid skewness. The critical regions consisted of the gender-marked article (e.g., *the*), the three-word spill-over region after the article (*often badly ventilated*)[9], the noun (e.g., *basement/bedroom*), and the two-word spill-over region after the noun (e.g., *by her parents*;

[9] For the sake of brevity, we chose to run a single model on the article spill-over region. Note that models run on single words in the spill-over region yielded the same qualitative results (see additional analysis presented on the article's OSF site, https://osf.io/8xubf/?view_only=50e2a7216b2e4d89bfef071f34f33d3e).

see **Table 4**, for schematic display of the four critical regions). All analyses are presented on this paper's OSF link under https://osf.io/8xubf/?view_only=50e2a7216b2e4d89bfef071f34f33d3e.

Table 4: Critical Regions.

	Article	**Article Spill Over**			**Noun**	**Noun Spill Over**	
		A+1	**A+2**	**A+3**		**N+1**	**N+2**
Predictable	den	oft	schlecht	belüfteten	Keller	ihrer	Eltern
	the $_{ACC\text{-}MASC}$	*often*	*badly*	*ventilated*	*basement*	*by her*	*parents*
Unpredictable	das	oft	schlecht	belüftete	Schlafzimmer	ihrer	Eltern
	the $_{ACC\text{-}NEUT}$	*often*	*badly*	*ventilated*	*basement*	*by her*	*parents*

Prior to analysis, and based on visual inspection of the data, RT data per word were trimmed minimally by excluding reading times faster than 100 ms and slower than 1500 ms (for all target words before the noun), or 2000 ms after the noun, which affected less than 1% of all data points (for similar outlier criteria, see e.g., Linzen & Jaeger, 2016).

To analyze the data statistically, we used linear mixed-effects models (Baayen et al., 2008), as implemented in the lme4 library (Bates et al., 2015) in R (R Core Team, 2021). All models were initially fit using the maximal random slope structure warranted by the design (Barr et al., 2013), while suppressing the correlations between random slopes and random intercepts, to facilitate convergence (Winter, 2019). P-values in model outputs were estimated using the Satterthwaite degrees of freedom method, as implemented in the *lmerTest* package (Kuznetsova et al., 2017). The effects of predictability and imbalance on de-logged reading times are shown in **Figure 3**.

3.1 Comprehension question accuracy

One subject was excluded from further analysis because their accuracy rate on the comprehension questions (0.54 correct) was markedly below the average. Average accuracy rates of the remaining subjects were high (0.97, range: 0.81–1), and did not differ between predictable (M = 0.966) and unpredictable (M = 0.967) items, $t(154) = -0.13$, $p = .89$. Thus, participants were attentive during the reading task and understood the sentences they were reading.

3.2 Effects of predictability

Fixed effects in models that estimated effects of predictability included the cloze probability of the first-best gender-marked article (for all prenominal words; i.e., for the article and the three-word spill-over region) or noun (for the noun and the spill-over region after the noun) as a scaled continuous variable. We also added a control predictor for frequency. (This predictor

was only added when a critical region consisted of a single word; frequency estimates were added in log per million, and were obtained from the movie subtitle norms from Brysbaert et al., 2011). Each model also contained control predictors for region length, reading times of the previous word, trial number, and word position in the sentence.[10] Initial models also included a control variable indicating where participants were recruited from (Prolific vs. university), but since that variable showed no effects, it was dropped from further analysis. **Table 5** shows the $\hat{\beta}$ coefficients, standard errors and *t*-values for all predictors added in models of predictability.

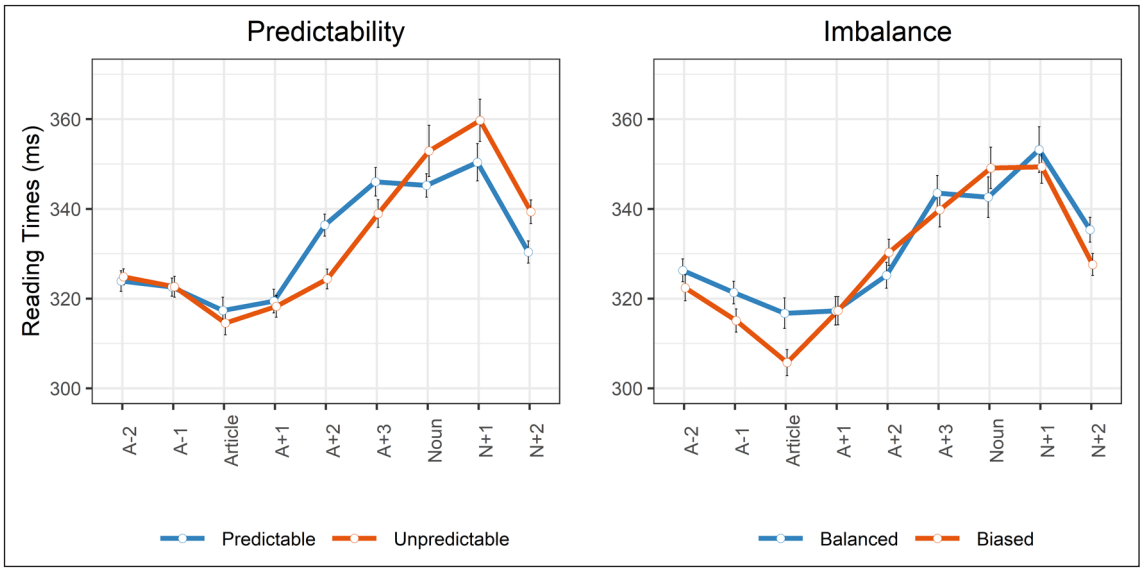

Figure 3: Self-Paced Reading Times on Words in the Critical Region Illustrating the Effects of Predictability and Imbalance. The graph shows effects of predictability and imbalance based on a median split of both continuous predictors. All statistical models were run using the scaled continuous variable. Predictability refers to the cloze probability of the actually presented gender-marked article. Imbalance refers to the cloze probability difference between the first- and second-best articles and nouns named for each item in the cloze task. Predictability effects are shown for the full set of 48 items; imbalance effects are shown for the 32 items that allowed for this analysis (see 2.2). Error bars represent SE and are adjusted for within-subject designs (Cousineau, 2005). A-2: Second word before the article; A-1: First word before the article. A+1 ... A+3: First and third word after the article. N+1, N+2: First (second) word after the noun.

To paraphrase the results, in line with our previous study, article predictability did not affect reading rates of pre-nominal gender-marked articles ($p = .26$). However, there was a positive-going effect of article predictability on the three-word spill-over region after the article ($p = .01$),

[10] Word length was not added to models of article RTs because the article was always three characters long. Imbalance was not added as control predictor to the predictability analysis, because imbalance values are not specified for all 48 items used for this analysis (see 2.2).

Table 5: Effect sizes (b), Standard Errors (SE), and T-Values for Models Estimating the Effects of Predictability on Log-Transformed RTs of the Critical Regions.

	Article			***Article Spill-Over***			***Noun***			***Noun Spill-Over***		
	$\hat{\beta}$	***SE***	***t***	$\hat{\beta}$	***SE***	***t***	$\hat{\beta}$	***SE***	***t***	$\hat{\beta}$	***SE***	***t***
Fixed Effects												
Intercept	5.57	0.10	53.49	6.82	0.03	232.98	5.74	0.04	151.17	6.46	0.03	201.89
Predictability	0.005	0.004	1.13	0.01	0.003	2.59	0.001	0.004	0.32	–0.01	0.005	–2.07
Trial	–0.07	0.003	–20.45	–0.08	0.003	–27.85	–0.07	0.004	–18.94	–0.07	0.003	–23.60
Frequency	0.02	0.02	1.11	–			0.006	0.008	0.80	–		
Length	–			0.01	0.005	1.58	0.02	0.01	3.26	0.01	0.005	1.11
Word Position	–0.01	0.01	–1.08	0.01	0.005	2.30	0.004	0.005	0.71	0.01	0.005	2.67
Previous Word RT	0.10	0.01	19.46	0.10	0.004	23.35	0.10	0.01	17.32	0.06	0.005	13.43
Random Effects												
Subjects	0.06			0.07			0.07			0.08		
Predictability	Ø			Ø			Ø			0.0003		
Items	0.001			0.0008			0.05			0.0007		
Predictability	0.0003			0.0002			Ø			0.0003		

Note. N-dash indicates predictors that were not applicable (e.g., word frequency for multi-word regions, word length for a fixed-length predictor). Predictors that had to be removed due to convergence issues are indicated by Ø. T-v alues larger than 2.0 are usually considered significant at the .05 level. The intercept reflects the log-RT values when all predictors in the model are zero.

which suggested that reading times on pre-nominal adjectives were *slowed* when these adjectives followed more predictable articles. Predictability did not significantly affect reading rates of critical nouns ($p = .75$), but there was a significant, negative-going effect of predictability on the spill-over region after the noun ($p = .04$), showing that participants read more quickly when the previously encountered noun was highly predictable.

In sum, the predictability models showed two effects of interest. First, pre-nominal adjectives were read more slowly when they followed highly predictable gender-marked articles. Second, the spill-over region after the noun was read more quickly when nouns were more highly predictable.

3.3 Effects of imbalance

Models estimating the effects of imbalance maintained scaled cloze probability as a control predictor, and specified scaled imbalance as an additional fixed effect (such that each model only took into account the subset of 32 items that were specified for this analysis).[11] All models also included the interaction between imbalance and predictability, in order to test whether more balanced or biased items potentially elicited diverging effects of predictability (e.g., more biased items might elicit larger effects of unpredictability). We did not include constraint in this analysis, as it led to multicollinearity with imbalance, but we present an analysis below that examines the separate contributions of constraint and imbalance on model fit. The control predictors for each region of interest were the same as specified in 3.2. **Table 6** shows $\hat{\beta}$ coefficients, standard errors and *t*-values for all model predictors in the four regions of interest.

Our expectation was that there should be reading facilitation in more biased items, but reading slowdown in more balanced items. In line with these expectations, there was a negative-going effect of imbalance on reading times of the gender-marked article ($p = .04$). Thus, when items were more biased towards a single gender-marked article, there was reading facilitation. In contrast, when items were balanced, there was slowing. Neither the model of article RTs, nor any other model in this section, showed interactions between predictability and imbalance (all p's $> .24$).

Imbalance did not significantly modulate reading times of the three-word spill-over region after the article ($p = .99$), the noun ($p = .37$), or the spill-over region after the noun ($p = .35$).[12]

---

[11] We replicated all findings reported below when using absolute, not scaled, values for imbalance. Since predictability and imbalance are correlated ($r = .67$), we checked whether the simultaneous inclusion of predictability and imbalance led to multicollinearity. We found no evidence attesting to this.

[12] Of note, we ran a follow-up model that not only specified imbalance, but also its interaction with a (sum-coded) binary variable that indicated whether the gender of the first- and the second-best completions were identical. The rationale for this model was that same-gender conditions might lead to pooled, rather than differential, activation, which would render our current definition of imbalance inappropriate for such items. We found no evidence for such an interaction at the article or any other word in the critical region.

Table 6: Effect sizes (b), Standard Errors (SE), and T-Values for Models Estimating the Effects of Imbalance on Log-Transformed RTs of the Critical Regions.

	Article			***Article Spill-Over***			***Noun***			***Noun Spill-Over***		
	$\hat{\beta}$	***SE***	***t***	$\hat{\beta}$	***SE***	***t***	$\hat{\beta}$	***SE***	***t***	$\hat{\beta}$	***SE***	***t***
Fixed Effects												
Intercept	5.51	0.11	51.20	6.82	0.03	240.54	5.73	0.04	143.72	6.45	0.03	206.51
Imbalance	–0.01	0.01	–2.11	0.001	0.01	0.02	0.003	0.01	0.46	–0.01	0.01	–0.95
Predictability	0.01	0.004	1.87	0.01	0.001	2.16	0.01	0.01	0.90	–0.01	0.01	–1.86
Trial	–0.05	0.004	–14.32	–0.07	0.003	–21.31	–0.07	0.001	–16.31	–0.06	0.004	–18.16
Frequency	0.03	0.02	1.54	–			0.01	0.01	1.10	–		
Length	–			0.001	0.01	0.42	0.02	0.01	2.35	0.003	0.005	0.61
Word Position	–0.01	0.01	–1.51	0.02	0.01	3.10	0.01	0.01	1.63	0.01	0.01	2.80
Previous Word RT	0.12	0.01	19.46	0.11	0.01	20.07	0.10	0.01	14.54	0.07	0.01	12.31
Imbalance * Predictability	0.002	0.01	0.33	–0.003	0.004	–0.73	0.001	0.01	0.24	0.002	0.01	0.36
Random Effects												
Subjects	0.21						0.08			0.07		
Imbalance	0.02						Ø			Ø		
Predictability	0.002						0.001			Ø		
Items	0.02						0.001			0.0004		
Predictability	0.02						0.001			0.0003		

Note. N-dash indicates predictors that were not applicable (e.g., word frequency for multi-word regions, word length for a fixed-length predictor). Predictors that had to be removed due to convergence issues are indicated by Ø. T-values larger than 2.0 are usually considered significant at the .05 level.

3.4 Exploratory analysis: Interactions with trial number

Prior studies have suggested that participants adapt their predictions when taking part in experiments in which predictions are frequently disconfirmed (e.g., Delaney-Busch et al., 2019; Ness & Meltzer-Asscher, 2021b; but see Nieuwland, 2021; Van Wonderen & Nieuwland, 2023). This has been interpreted as evidence for *rational* sentence processing, i.e., comprehenders adapting the degree to which they engage in predictive processing by tuning in to the contingencies of the current input. During peer review of this article, we were asked to examine if any of our predictability and imbalance effects were additionally modulated by trial number. We found no interactions, with one exception: The predictability effect on the spill-over region after the article became progressively smaller in the course of the experiment, as suggested by a trending interaction between cloze probability of the article and trial number for RTs of the article spill-over region, $\hat{\beta} = -0.004$, SE = 0.002, $t = -1.78$, $p = .08$). In other words, in early trials, participants showed slowing when reading the three-word adjective region following predictable gender-marked articles, but this effect gradually washed out in the course of the experiment, likely reflecting adaptation. We return to this point in section 4.4.

3.5 Comparing imbalance and constraint

Whereas imbalance takes into account the cloze probability difference between the first and second most likely completion of a sentence, another common psycholinguistic measure, constraint, refers to the average cloze probability of the most likely completion of a sentence (we return to this in section 4). During peer review of this article, we were asked to demonstrate that the imbalance measure adds something to the study of predictive processing that is not already incorporated in constraint alone. To address this, we present two analyses using model comparisons below.

The first analysis investigated whether a model that includes imbalance (but not constraint) fits the data as well, if not better, than a model that includes constraint (but not imbalance). The second analysis compared, for both models, the improvement in model fit from a base model that did not include either predictor. We take into account log likelihood, a measure for the goodness of fit of a model, and the Akaike Information Criterion (AIC), a measure of fit that penalizes a model for having more variables (such that larger values indicate worse fit, corrected for the number of variables; Winter, 2019). Better model fit is indicated by a larger log likelihood and lower AIC.

Analysis 1. We fit two models that included the exact same number of parameters. The imbalance model was identical to the model presented in the main analysis. Its syntax was: *lmer(RT) ~ scale(imbalance) * scale(cloze) + scale(trial) + log-frequency + scale(word_position) + scale(previous_word_RT) + (1 + scale(imbalance)||subject) + (1 + scale(cloze)||item)*. The constraint model was identical to the imbalance model except that it specified constraint

instead of imbalance. The syntax of the constraint model was, *lmer(RT) ~ scale(constraint) * scale(cloze) + scale(trial) + log-frequency + scale(word_position) + scale(previous_word_RT) + (1 + scale(constraint)||subject) + (1 + scale(cloze)||item)*. **Table 6** shows the results. According to the values in **Table 7**, the imbalance model fit the data slightly better and resulted in lower AIC than the constraint model. Notably, in the constraint model, the main effect of constraint was not statistically significant ($\hat{\beta} = -0.01$, SE = 0.01, $t = -1.80$, $p = .07$).

Analysis 2. We compared the improvement in model fit for the imbalance and constraint models when each was compared to a base model that did not include imbalance or constraint. **Table 6** summarizes the results. The imbalance model resulted in larger log likelihood and AIC increment than the constraint model. Both models significantly improved model fit compared to the base model (p's $< .01$), but imbalance improved model fit more (χ^2 constraint model vs. base: 12.52; χ^2 imbalance model vs. base: 13.21).

Table 7: Parameters of Model Fit for Base, Imbalance and Constraint Models.

	AIC	**BIC**	**Log Likelihood**
Base	–1264.80	–1206.70	642.40
Base w. Imbalance	–1272.00	–1196.40	649.00
Base w. Constraint	–1271.30	–1195.70	648.66
Increment (Imbalance-Base)	–7.20	10.30	6.60
Increment (Constraint-Base)	–6.50	11.00	6.26

4. Discussion

Do language users predict multiple sentence continuations during language processing or do they predict predominantly one continuation (parallel vs. all-or-nothing prediction)? If multiple predictions are generated, are they issued in a serial or graded fashion? Two recent studies had obtained evidence supporting parallel gradedness of prediction during sentence processing using tasks that explicitly asked participants to predict upcoming nouns (e.g., Ness & Meltzer-Asscher, 2021b; Staub et al., 2015). In present study, we aimed to extend these findings to a simple sentence reading paradigm that merely asked participants to read for comprehension. To measure parallel gradedness of prenominal predictions, we introduced a novel measure, imbalance, which indicates the probabilistic difference between multiple continuations of a sentence. We also chose to investigate parallel gradedness of *prenominal* predictions, because our goal was to obtain prediction-specific effects, which may be different from semantic association or late-stage integration effects (e.g., Van Berkum et al., 2005; Van Berkum, 2010; but see Ferreira & Chantavarin, 2018, and Ferreira & Qiu, 2021, for more recent accounts).

We used a word-by-word self-paced reading task, in which native speakers of German were asked to read predictable and unpredictable sentences of German for comprehension. All experimental materials constrained expectations relatively strongly towards a particular noun. Unpredictable sentence continuations were nouns and articles from a different grammatical class than the predictable noun. Hence, the prenominal article in unpredictable sentences was a salient cue to foreshadow a prediction violation at the noun.

In order to gain insight into the possibility of single vs. parallel gradedness of prenominal prediction, we defined two measures of interest, predictability and imbalance. Predictability referred to the cloze probability of the actually presented predictable and unpredictable articles and nouns. We argued that predictability effects at the article would demonstrate that language users generate expectations about primarily one article form, as predictability effects indicate processing differences between one highly predictable and a generally far less predictable word. Hence, predictability served as a measure of one-shot prediction.

Imbalance, in turn, captures the extent to which sentence contexts cued parallel graded predictions for multiple prenominal articles and nouns. More specifically, we defined imbalance as the probabilistic difference between the first-best and second-best continuations (article and noun) of a sentence, as normed by a cloze task in which sentence frames were truncated before the gender-marked article and participants were instructed to provide, for each single sentence frame, a first-best and a second-best continuation. We argued that balanced sentence contexts indicate a smaller probabilistic difference between the first-best and the second-best completion, and therefore cue expectations towards multiple continuations that are roughly equally probable and balanced. More biased contexts, in turn, have a larger probabilistic difference between the first- and second-best completion, and, therefore, might cue a stronger expectation towards one particular continuation over others. We used imbalance as a measure for parallel gradedness of predictions, because effects of this variable would not only constitute evidence that language users generate multiple predictions about upcoming content when reading a sentence (gradedness of predictions, as reflected by the inclusion of two likely continuations of a sentence), but that language users actually weigh these simultaneously activated predictions according to their likelihood (parallel prediction, as reflected by the fact that imbalance is a difference score between two probabilities).

We used predictability and imbalance to predict the reading times of seventy-seven German speaking adults participating in an online study that asked participants to read sentences silently for comprehension. Overall, our findings are most in line with parallel gradedness of predictions, in line with prior research (e.g., DeLong et al., 2005; Ness & Meltzer-Asscher, 2021b; Staub et al., 2015; Szewczyk & Schriefers, 2013). We unpack our findings for predictability and imbalance, as well as their implications, in greater detail below.

4.1 Findings for predictability

In the present study, predictability effects emerged in two critical regions: the spill-over region after the phrase-final predictable or unpredictable noun, and the article spill-over region after the critical gender-marked article.

Turning first to predictability effects associated with unpredictable nouns, we found that participants slowed their reading when processing the spill-over region after an unpredictable noun. In other words, reading of the noun spill-over region was facilitated following predictable nouns. This result is in line with a long line of previous research suggesting that, in strongly constraining sentence contexts, language users are able to predict a single lexical item (e.g., DeLong et al., 2019; Frisson et al., 2017; Ito et al., 2016; Laszlo & Federmeier, 2009; Kukona, 2020; Luke & Christianson, 2016; among others). Therefore, this aspect of our findings is in line with one-shot accounts of predictions.

Turning now to the predictability effect that surfaced on the prenominal adjective region, our results showed that, following more predictable gender-marked articles, participants were slowed in reading the adjective region. There was also some suggestion from an exploratory analysis that this effect became smaller in the course of the experiment, in other words, even though participants' initial reading at the series of adjectives was slowed, this reading slow-down gradually washed out in the course of the experiment. Even though we found no predictability effects for the article itself, it is remarkable that the reading slow-down on the adjective region emerged primarily after participants had read more predictable gender-marked articles, not unpredictable ones. This aspect of our findings could suggest that the slowdown on the adjective region was, in fact, driven by the gender-marked article – even though the direction of the effect is in clear contrast to what would be expected, based on prior research. Despite the counter-intuitive nature of this finding, one possible interpretation could be that reading a more predictable article may have "re-assured" participants that the predictable noun must follow immediately – an expectation that was then disconfirmed by the series of prenominal adjectives, which resulted in slowing. This interpretation of our results is somewhat in line with the *updating of the noun* hypothesis (Fleur et al., 2020), i.e., the hypothesis that people may use prenominal information in the article to update, or potentially revise, their noun predictions. Hence, participants may have used the gender information conveyed by the predictable article to more strongly anticipate, or cue a state of readiness (Ferreira & Chantavarin, 2018) for, the highly predictable noun. A second potential explanation of this effect is that predictable articles not only cued participants' "readiness" specifically for the predictable noun, but for any kind of noun, in line with the statistical regularities of languages, which make it more likely for people to encounter an article-noun sequence rather than an article-adjective-noun sequence (see, e.g., Luke & Christianson, 2016). In line with this account of the data, the reading times at the adjective region seemed to increase incrementally at the spill-over region, both for predictable

and unpredictable sentences (see **Figure 3**, left panel). The reading slowdown may constitute evidence for morpho-syntactic prediction: Participants may have been unconsciously aware of the statistical contingencies of German (where gender marked articles are normally followed by a noun), and they may have used this kind of knowledge during sentence reading to anticipate that a noun, not an adjective, will follow an article (Dikker et al., 2010; Lau et al., 2006; Matchin et al., 2017; Staub & Clifton, 2006; for a recent review of the literature on syntactic prediction, see Ferreira & Qiu, 2021). Hence, under this account, the reading slow-down would reflect the cost of sustaining a prediction and integrating the adjectival material (e.g., Gibson, 1998). The gradual decrease of this effect (see section 4.4) throughout the experiment would, then, constitute an adaptation process in which participants progressively tuned in to (and became less surprised by) the relatively high proportion of article-adjective-noun phrases in the experiment, which accounted for 62% of all experimental sentences in a given list.

A limitation of the present study is that it cannot address how, and to what extent, the prenominal adjectives themselves may have impacted the processing of the noun. We can probably rule out the possibility that the prenominal adjectives elicited their own gender mismatch effects, since, across all experimental items, only the third word in the adjective region was consistently gender-marked, and for that third word, there were no gender mismatch effects.[13] However, other potential effects elicited by the adjectives are more difficult to come by. One problem is that the prenominal adjectives were obviously not semantically vacant, and, in being so, may have changed participants' expectations of the phrase-final noun (e.g., Boudewyn et al, 2015; Szewczyk et al., 2022). We know from another cloze test ($n = 30$ participants) in which sentence frames were truncated just before the noun that the series of adjectives did not lower participants' expectations of the predictable noun (in fact, noun cloze probabilities in that second cloze test were slightly higher than the ones from the original cloze test in which sentence frames were truncated before the article: $M = 0.87$, Range = 0.50–1 vs $M = 0.78$, Range = 0.30–1). Even though this makes us somewhat confident that, as a whole, the article spill-over region did not dramatically change noun predictability, it is difficult to estimate how single words in the spill over region may have pushed around noun predictability effects. For example, a participant may strongly anticipate *basement* after having read *Since Anne is afraid of spiders, she does not like going down into the …*, but they may be less likely to anticipate that same noun after reading *badly ventilated*, if bad ventilation is not a core feature that they associate with basements. To shed more light on this issue, a series of cloze tests would have to be conducted that estimate noun

[13] We ran follow-up models that specified, as an outcome variable, RTs on single words in the spill-over region, and, as predictor variables, adjective gender-marking (a binary variable, yes or no) in interaction with predictability. The models on the first and second word in the adjective region showed singular fit warnings, probably resulting from the fact that, in only three out of 48 items, these two words were gender-marked (most of them were adverbs or adpositions, which are not gender-marked in German).

predictability after every single word in the adjective spill-over region. While we consider this a limitation of the present study, we note that the adjective region could not have modulated reading times of the critical gender-marked articles in any way, as the articles preceded the adjective region.

4.2 Findings for imbalance

Imbalance is a measure of parallel gradedness of predictions, as it indicates whether a sentence context biases expectations towards one dominant continuation (when imbalance values are large), or towards multiple ones that are equally likely (when imbalance values are small). More biased contexts have larger imbalance values, as they bias expectations more strongly towards the first-best continuation of the sentence. Balanced contexts have smaller imbalance values, suggesting that the two most likely continuations are similarly predictable. Imbalance effects during reading indicate that readers generate multiple expectations in parallel, and that they are sensitive to graded probabilistic differences between these multiple expectations. Therefore, we used imbalance to infer *gradedness* of *parallel* linguistic predictions.

We find an effect of imbalance on reading times of gender-marked articles, in line with our earlier study on this topic (Haeuser et al., 2022). Specifically, when participants were reading gender-marked articles in items which were balanced, allowing for the use of multiple gender-marked articles, there was reading slow-down. In contrast, articles in more biased items, that is, those which pointed towards one highly likely gender-marked article, were read more quickly. In a nutshell, this means that readers experienced processing difficulty at the level of the gender-marked article when the preceding sentence context licensed other gender-marked articles as likely continuations of the sentence. In turn, processing was facilitated when contexts were more strongly biased towards one dominant continuation.

Taken together, this aspect of our findings constitutes evidence for parallel gradedness of lexical predictions. Specifically, these results illustrate that, at least in cases where there are multiple potential options for continuing a sentence, comprehenders generate multiple predictions about those possible sentence continuations in parallel, and that the strength of competition between these options can influence the time course of comprehension during reading. Therefore, this aspect of our data is most consistent with a parallel account of prediction, rather than serial or all-or-nothing accounts. However, future work would be needed to explore to what extent a parallel mechanism might best explain predictive processing for other aspects of linguistic processing. For instance, prior work on anticipation of syntactic structure suggests that listeners may generate potential expectations in a serial fashion, rather than entertaining multiple competing structural options simultaneously (e.g., Traxler et al., 1998; Van Gompel et al., 2005). Future work would be needed to try to identify how, when, and to what extent linguistic prediction effects may happen in a parallel or serial fashion.

An open question in this context is why imbalance effects emerged on the article, not on the spill-over region. Since word-by-word reading produces stronger spill-over effects than moving-window reading, our expectation had been that imbalance effects would emerge more readily at the spill-over region after the article, and not directly on the article, as in our initial study. We can only speculate as to why this was not the case. However, we note that, both in the present study and in an earlier investigation using moving-window reading, there was a visible trend for facilitated reading in more biased items, even before the critical gender-marked article (see **Figure 3**), i.e., on the word preceding the gender-marked article. Notably, though, the effect was smaller in the present study, where word-by-word, not moving-window reading, was used. Hence, there was a small indication that imbalance effects can vary in timing, depending on the stimulus presentation paradigm. This suggests that a converging methods approach may be necessary to fully understand parallel activation in predictive processing using an imbalance measure approach.

4.3 Imbalance and other expectancy measures

Imbalance measures the probabilistic difference between the first two prominent completions of a sentence, as inferred by a cloze test in which participants were tasked with completing each sentence fragment with two possible continuations. To what extent is imbalance different from two other psycholinguistic measures that are sometimes used to index expectancy effects, for example constraint (e.g., Federmeier et al., 2007; Schwanenflugel & Shoben, 1985) and entropy (e.g., Frank, 2013; Hale, 2006; Linzen & Jaeger, 2016; Lowder et al., 2018)? We discuss this question here.

Constraint normally measures the likelihood of a sentence ending with the most probable word, taking into account only the first-best completion of a sentence in a one-shot cloze test. This means that constraint does not take into account the probabilistic difference between multiple continuations of a sentence, as it only reflects one single value. To illustrate this, assume that there are two sentence frames in which the first-best completion has a cloze probability of 0.80 (for both sentences) and the second-best completions have cloze probabilities of 0.5, and 0.2, respectively. The constraint of these two sentence frames would be 0.80, for both sentence frames, whereas imbalance values for these same two sentence frames would be 0.3 (i.e., a relatively more balanced sentence frame) and 0.6 (i.e., a relatively more biased sentence frame), respectively, since imbalance additionally takes into account the likelihood of the second-best completion. However, since imbalance is computed from the cloze probability of the first-best response (i.e., constraint), constraint and imbalance necessarily correlate with one another. In the present study, the correlation was: $r = .67$ [*95% CI*: .42, .82]. Despite this high correlation, we found that a model that included imbalance (but not constraint) had a better fit to the data than a model that included constraint (but not imbalance). In addition, when compared to a

base model that did not include either predictor, a model that additionally included imbalance improved model fit more than a model that additionally included constraint. This shows that imbalance adds something to the study of predictive processing that is not already captured by constraint. However, we would like to emphasize that our goal here is not to argue that the imbalance measure is superior to constraint. Instead, our goal is to report an empirical result that we hope informs research on gradedness of predictions. Both measures (constraint and imbalance) likely have advantages and disadvantages in their conception (e.g., constraint does not explicitly consider the presence or absence of a highly-likely second option or competitor, but, in contrast to imbalance, has a wealth of research on how it can explain variance in a range of psycholinguistic phenomena). Future research is needed to compare these measures across a range of phenomena in which constraint and imbalance could be applied, to fully understand their impact.

Entropy, in turn, is thought to reflect a person's uncertainty about possible sentence completions: Entropy is higher the more possible completions there are and the more equally likely those completions are. It is calculated as the sum of the log base cloze probabilities of all possible continuations of a sentence, and it can range from zero to infinity. Even though entropy takes into account a wider range of cloze probabilities than constraint does, entropy values are also normally computed from a one-shot cloze test. This means that, for entropy, the likelihood of the modal response constrains the likelihood of all the other sentence continuations. For example, a sentence frame in which 80% of the cloze probability mass is centered in one value could maximally yield 20% as the cloze probability of the next best competitor (under the assumption that all of the remaining cloze probability mass is taken up by one single value), since all cloze values need to sum up to zero. For imbalance, which is computed by taking into account two distinct responses for a given sentence frame, each with their own probability distribution, no such constraints apply. One additional difference between imbalance and entropy is that entropy is computed over the distribution of between-subjects one-shot responses, whereas imbalance captures the possible within-subjects distribution of word activations (or has the potential to capture it).[14]

Hence, imbalance is different from constraint, and it is also different from entropy. By explicitly asking the same participant to provide multiple sentence completions for each item, this measure closely aligns with the potential construct of interest – that of parallel graded predictions. It would be illuminating in future work to generate entropy, constraint and imbalance measures on the same set of items, by collecting cloze data from single and multiple completion measures, to systematically explore how these measures are related, and which best explains their impact on processing. In fact, it would be interesting to compare responses in a cloze task that asks

[14] We thank an anonymous reviewer for suggesting this.

participants to provide even more responses to a single item (i.e., more than only one or two). This type of work would support clearer and more context-sensitive theoretical specifications of the breadth and specificity of activation that accompanies real-time language processing.

4.4 Open questions and avenues for future research

Our key findings for imbalance show that language users slow down their reading when their expectations about upcoming words are more balanced, which could be a result of co-activating multiple likely representations. One aspect that needs to be fleshed out in future studies is what cognitive mechanism accounts for this slowing. One possibility is that co-activated representations compete with one another and that readers need to resolve the ambiguity. Alternatively, having multiple words with the same probability co-activated in memory may tax working memory resources more than when a single more prominent word is pre-activated. In their current form, our data cannot speak to this question, and it will take future studies to uncover the cognitive mechanics behind the imbalance effect.

An open question in this context is why multiple pre-activated representations would be more difficult to process at all, in the light of conflicting findings from the divided attention literature. Specifically, previous studies have sometimes found that total activation builds faster, not slower, when there are two targets present, as opposed to when there is only one. This phenomenon is called *statistical facilitation*: When an observer must make the same speeded response to either a visual or auditory signal, reaction times are sometimes facilitated when both signals are presented, rather than when only one is presented (i.e., pooled activation; Miller & Ulrich, 2003; Mordkoff & Yantis, 1991). Such an account predicts that pre-activated representations during predictive processing facilitate one another, rather than slowing each other down, which conflicts with the present findings. Indeed, two previous studies on predictive processing have found evidence for neighbor facilitation rather than slowing, provided that the pre-activated representations are semantically related. For example, Brothers and colleagues (2023) examined the N400 ERP amplitude to second-best continuations for three-sentence mini stories (e.g., *Stephen wanted to do something special for his girlfriend. He decided to make her a hand-made card. On it, he drew some … flowers.*). According to the results, N400 amplitudes were reduced (i.e. more facilitation occurred) for those second-best continuations that were more strongly related to the first-best continuation. Converging results emerged in the article by Ness & Meltzer-Asscher (2021b), who found that modal responses in a speeded cloze task were issued faster when the second-best response had a high cloze probability and was semantically similar to the modal. Hence, in these studies, pre-activated representations facilitated, rather than slowed, each other, as a function of their semantic relatedness. Even though the findings obtained here may seem to conflict with such results at first sight, we do not believe that there is a radical difference between our findings and the ones in these two studies. Specifically, the

design of the present study intentionally excluded semantically similar or overlapping responses (see 2.2), meaning that our materials included items whose first- and second-best responses were semantically dissimilar. When this element of our experimental design is taken into account, our findings seem to align, rather than conflict, with findings obtained by, e.g., Ness and Meltzer-Asscher (2021b): When first- and second-best responses were semantically less related in that study (as they likely were in the present study), response times for modals were slowed, as long as the competitor was more likely (see **Figure 2** of that paper). This pattern mirrors our findings for imbalance, where reading times were slowed when the likelihood of the second-best response approached the likelihood of the modal. Together, these findings indicate that, when language users generate parallel predictions, behavioral responses can be either facilitated or slowed, depending on the semantic similarity of the pre-activated representations. Future studies could examine whether this process is additionally modulated by word frequency, as has been suggested for competitor activation in word recognition research (e.g., Luce & Pisoni, 1998).

5. Conclusion

Our current study adds to a growing body of evidence that linguistic prediction is not only psycholinguistically plausible – readers entertain multiple competing continuations in parallel. Future work is needed to explore how these effects extend in more real-life texts that more closely mimic everyday reading and conversation, and also to extend this work to linguistic phenomena beyond that of lexical prediction for nouns that are preceded by gender-marked articles. Doing so would help to continue to flesh out when and how one type of domain general cognitive process – prediction – is recruited in the service of language comprehension.

Data Accessibility Statement

All data files, analysis scripts and materials can be found on this paper's project page using the link https://osf.io/8xubf/?view_only=50e2a7216b2e4d89bfef071f34f33d3e.

Ethics and consent

Ethics approval was granted by the *Deutsche Forschungsgesellschaft* DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

Acknowledgments

The authors would like to thank the three anonymous reviewers, who provided helpful feedback on this manuscript.

Funding Statement

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

Competing Interests

The authors have no competing interests to declare.

Author contributions (CRediT)

Conceptualization and Data curation: KH and AB. Formal Analysis: KH. Funding acquisition: KH. Investigation and Methodology: KH and AB. Project administration, Resources, Software, Supervision, Validation, Visualization: KH. Writing – original draft: KH, Writing – review & editing: KH and AB.

ORCiD IDs

Katja Haeuser: https://orcid.org/0000-0001-6553-3551

Arielle Borovsky: https://orcid.org/0000-0001-5869-0241

References

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. https://doi.org/10.1016/S0010-0277(99)00059-1

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

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. https://doi.org/10.1080/01690965.2010.542671

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Borovsky, A., Elman, J. L., & Fernald, A. (2012). Knowing a lot for one's age: Vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. *Journal of Experimental Child Psychology*, *112*(4), 417–436. https://doi.org/10.1016/j.jecp.2012.01.005

Boudewyn, M. A., Long, D. L., & Swaab, T. Y. (2015). Graded expectations: Predictive processing and the adjustment of expectations during spoken language comprehension. *Cognitive, Affective, & Behavioral Neuroscience*, *15*, 607–624. https://doi.org/10.3758/s13415-015-0340-0

Brothers, T., & Kuperberg, G. R. (2021). Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language*, *116*, 104174. https://doi.org/10.1016/j.jml.2020.104174

Brothers, T., Morgan, E., Yacovone, A., & Kuperberg, G. (2023). Multiple predictions during language comprehension: Friends, foes, or indifferent companions? *Cognition*, *241*, 105602. https://doi.org/10.1016/j.cognition.2023.105602

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. https://doi.org/10.1016/j.cognition.2014.10.017

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. https://doi.org/10.1016/j.jml.2016.10.002

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, *58*, 412–424. https://doi.org/10.1027/1618-3169/a000123

Cholewa, J., Neitzel, I., Bürsgens, A., & Günther, T. (2019). Online-processing of grammatical gender in noun-phrase decoding: An eye-tracking study with monolingual German 3rd and 4th graders. *Frontiers in Psychology*, *10*, 2586. https://doi.org/10.3389/fpsyg.2019.02586

Chow, W. Y., & Chen, D. (2020). Predicting (in) correctly: listeners rapidly use unexpected information to revise their predictions. *Language, Cognition and Neuroscience*, *35*(9), 1149–1161. https://doi.org/10.1080/23273798.2020.1733627

Chung, Y. M. W., & Federmeier, K. D. (2023). Read carefully, because this is important! How value-driven strategies impact sentence memory. *Memory & Cognition*, 1–16. https://doi.org/10.3758/s13421-023-01409-3

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology*, *1*(1), 42–45. https://doi.org/10.20982/tqmp.01.1.p042

Dahan, D., Swingley, D., Tanenhaus, M. K., & Magnuson, J. S. (2000). Linguistic gender and spoken-word recognition in French. *Journal of Memory and Language*, *42*(4), 465–480. https://doi.org/10.1006/jmla.1999.2688

Dave, S., Brothers, T. A., & Swaab, T. Y. (2018). 1/f neural noise and electrophysiological indices of contextual prediction in aging. *Brain Research*, *1691*, 34–43. https://doi.org/10.1016/j.brainres.2018.04.007

Delaney-Busch, N., Morgan, E., Lau, E., & Kuperberg, G. R. (2019). Neural evidence for Bayesian trial-by-trial adaptation on the N400 during semantic priming. *Cognition*, *187*, 10–20. https://doi.org/10.1016/j.cognition.2019.01.001

DeLong, K. A., Chan, W. H., & Kutas, M. (2019). Similar time courses for word form and meaning preactivation during sentence comprehension. *Psychophysiology*, *56*(4), e13312. https://doi.org/10.1111/psyp.13312

DeLong, K. A., Chan, W. H., & Kutas, M. (2021). Testing limits: ERP evidence for word form preactivation during speeded sentence reading. *Psychophysiology*, *58*(2), e13720. https://doi.org/10.1111/psyp.13720

DeLong, K. A., Groppe, D. M., Urbach, T. P., & Kutas, M. (2012). Thinking ahead or not? Natural aging and anticipation during reading. *Brain and Language*, *121*(3), 226–239. https://doi.org/10.1016/j.bandl.2012.02.006

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. https://doi.org/10.1038/nn1504

Dikker, S., Rabagliati, H., Farmer, T. A., & Pylkkänen, L. (2010). Early occipital sensitivity to syntactic category is based on form typicality. *Psychological Science*, *21*(5), 629–634. https://doi.org/10.1177/0956797610367751

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test (PPVT)*. Circle Pines, MN: American Guidance Service. https://doi.org/10.1037/t15145-000

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. https://doi.org/10.1006/jmla.1999.2660

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. https://doi.org/10.1111/1469-8986.3920133

Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research*, *1146*, 75–84. https://doi.org/10.1016/j.brainres.2006.06.101

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. https://doi.org/10.1177/0963721418794491

Ferreira, F., & Qiu, Z. (2021). Predicting syntactic structure. *Brain Research*, *1770*, 147632. https://doi.org/10.1016/j.brainres.2021.147632

Finger, H., Goeke, C., Diekamp, D., Standvoß, K., & König, P. (2017). LabVanced: A unified JavaScript framework for online studies. In *International Conference on Computational Social Science (Cologne)* (pp. 1–3). Cologne: University of Osnabrück.

Fleur, D. S., Flecken, M., Rommers, J., & Nieuwland, M. S. (2020). Definitely saw it coming? The dual nature of the pre-nominal prediction effect. *Cognition*, *204*, 104335. https://doi.org/10.1016/j.cognition.2020.104335

Frank, S. L. (2013). Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science*, *5*(3), 475–494. https://doi.org/10.1111/tops.12025

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. https://doi.org/10.1016/j.jml.2017.04.007

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Green, P., & MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Haeuser, K., & Borovsky, A. (2024). Predictive processing suppresses form-related words with overlapping onsets. In L. K. Samuelson, S. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds), *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 46). Rotterdam, The Netherlands: Cognitive Science Society.

Haeuser, K. I., Demberg, V., & Kray, J. (2018). Surprisal modulates dual-task performance in older adults: Pupillometry shows age-related trade-offs in task performance and time-course of language processing. *Psychology and Aging*, *33*(8), 1168. https://doi.org/10.1037/pag0000316

Haeuser, K. I., Demberg, V., & Kray, J. (2019). Effects of aging and dual-task demands on the comprehension of less expected sentence continuations: Evidence from pupillometry. *Frontiers in Psychology*, *10*, 709. https://doi.org/10.3389/fpsyg.2019.00709

Haeuser, K. I., Kray, J., & Borovsky, A. (2022). Hedging Bets in Linguistic Prediction: Younger and Older Adults Vary in the Breadth of Predictive Processing. *Collabra: Psychology*, *8*(1), 36945. https://doi.org/10.1525/collabra.36945

Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, *30*(4), 643–672. https://doi.org/10.1207/s15516709cog0000_64

Heilbron, M., Armeni, K., Schoffelen, J. M., Hagoort, P., & De Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, *119*(32), e2201968119. https://doi.org/10.1073/pnas.2201968119

Hopp, H. (2013). Grammatical gender in adult L2 acquisition: Relations between lexical and syntactic variability. *Second Language Research*, *29*(1), 33–56. https://doi.org/10.1177/0267658312461803

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research*, *1626*, 118–135. https://doi.org/10.1016/j.brainres.2015.02.014

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. https://doi.org/10.1016/j.brainres.2018.11.013

Huettig, F., & Janse, E. (2016). Individual differences in working memory and processing speed predict anticipatory spoken language processing in the visual world. *Language, Cognition and Neuroscience*, *31*(1), 80–93. https://doi.org/10.1080/23273798.2015.1047459

Huettig, F., & Mani, N. (2016). Is prediction necessary to understand language? Probably not. *Language, Cognition and Neuroscience*, *31*(1), 19–31. https://doi.org/10.1080/23273798.2015.1072223

Husband, E. M. (2022). Prediction in the maze: Evidence for probabilistic pre-activation from the English a/an contrast. *Glossa Psycholinguistics*, *1*(1). https://doi.org/10.5070/G601153

Ito, A., Corley, M., Pickering, M. J., Martin, A. E., & Nieuwland, M. S. (2016). Predicting form and meaning: Evidence from brain potentials. *Journal of Memory and Language*, *86*, 157–171. https://doi.org/10.1016/j.jml.2015.10.007

Ito, A., Martin, A. E., & Nieuwland, M. S. (2017). How robust are prediction effects in language comprehension? Failure to replicate article-elicited N400 effects. *Language, Cognition and Neuroscience*, *32*(8), 954–965. https://doi.org/10.1080/23273798.2016.1242761

Keating, G. D., & Jegerski, J. (2015). Experimental designs in sentence processing research: A methodological review and user's guide. *Studies in Second Language Acquisition*, *37*(1), 1–32. https://doi.org/10.1017/S0272263114000187

Kochari, A. R., & Flecken, M. (2019). Lexical prediction in language comprehension: a replication study of grammatical gender effects in Dutch. *Language, Cognition and Neuroscience*, *34*(2), 239–253. https://doi.org/10.1080/23273798.2018.1524500

Kukona, A. (2020). Lexical constraints on the prediction of form: Insights from the visual world paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(11), 2153. https://doi.org/10.1037/xlm0000935

Kuperberg, G. R., Brothers, T., & Wlotko, E. W. (2020). A tale of two positivities and the N400: Distinct neural signatures are evoked by confirmed and violated predictions at different levels of representation. *Journal of Cognitive Neuroscience*, *32*(1), 12–35. https://doi.org/10.1162/jocn_a_01465

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*, 32–59. https://doi.org/10.1080/23273798.2015.1102299

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, *62*, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Kutas, M., & Hillyard, S. A. (1984). Event-Related Brain Potentials (ERPs) Elicited by Novel Stimuli during Sentence Processing a. *Annals of the New York Academy of Sciences*, *425*(1), 236–241. https://doi.org/10.1038/307161a0

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *Journal of statistical software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Laszlo, S., & Federmeier, K. D. (2009). A beautiful day in the neighborhood: An event-related potential study of lexical relationships and prediction in context. *Journal of Memory and Language*, *61*(3), 326–338. https://doi.org/10.1016/j.jml.2009.06.004

Lau, E., Stroud, C., Plesch, S., & Phillips, C. (2006). The role of structural prediction in rapid syntactic analysis. *Brain and Language*, *98*(1), 74–88. https://doi.org/10.1016/j.bandl.2006.02.003

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics:(de) constructing the N400. *Nature reviews neuroscience*, *9*(12), 920–933. https://doi.org/10.1038/nrn2532

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Lew-Williams, C., & Fernald, A. (2010). Real-time processing of gender-marked articles by native and non-native Spanish speakers. *Journal of Memory and Language*, *63*(4), 447–464. https://doi.org/10.1016/j.jml.2010.07.003

Linzen, T., & Jaeger, T. F. (2016). Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive Science*, *40*(6), 1382–1411. https://doi.org/10.1111/cogs.12274

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. https://doi.org/10.1111/cogs.12597

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*(1), 1–36. DOI: https://doi.org/10.1097/00003446-199802000-00001

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology, 88*, 22–60. https://doi.org/10.1016/j.cogpsych.2016.06.002

Mani, N., & Huettig, F. (2012). Prediction during language processing is a piece of cake—But only for skilled producers. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(4), 843. https://doi.org/10.1037/a0029284

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. https://doi.org/10.1016/j.neuropsychologia.2019.107199

Matchin, W., Hammerly, C., & Lau, E. (2017). The role of the IFG and pSTS in syntactic prediction: Evidence from a parametric study of hierarchical structure in fMRI. *Cortex*, *88*, 106–123. https://doi.org/10.1016/j.cortex.2016.12.010

Miller, J., & Ulrich, R. (2003). Simple reaction time and statistical facilitation: A parallel grains model. *Cognitive Psychology*, *46*(2), 101–151. https://doi.org/10.1016/S0010-0285(02)00517-0

Mordkoff, J. T., & Yantis, S. (1991). An interactive race model of divided attention. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(2), 520–538. https://doi.org/10.1037/0096-1523.17.2.520

Ness, T., & Meltzer-Asscher, A. (2018). Predictive pre-updating and working memory capacity: evidence from event-related potentials. *Journal of Cognitive Neuroscience*, *30*(12), 1916–1938. https://doi.org/10.1162/jocn_a_01322

Ness, T., & Meltzer-Asscher, A. (2021a). From pre-activation to pre-updating: A threshold mechanism for commitment to strong predictions. *Psychophysiology*, *58*(5), e13797. https://doi.org/10.1111/psyp.13797

Ness, T., & Meltzer-Asscher, A. (2021b). Love thy neighbor: Facilitation and inhibition in the competition between parallel predictions. *Cognition*, *207*, 104509. https://doi.org/10.1016/j.cognition.2020.104509

Nicenboim, B., Vasishth, S., & Rösler, F. (2020). Are words pre-activated probabilistically during sentence comprehension? Evidence from new data and a Bayesian random-effects meta-analysis using publicly available data. *Neuropsychologia*, *142*, 107427. https://doi.org/10.1016/j.neuropsychologia.2020.107427

Nieuwland, M. S. (2021). How 'rational'is semantic prediction? A critique and re-analysis of. *Cognition*, *215*, 104848. https://doi.org/10.1016/j.cognition.2021.104848

Nieuwland, M. S., Arkhipova, Y., & Rodríguez-Gómez, P. (2020). Anticipating words during spoken discourse comprehension: A large-scale, pre-registered replication study using brain potentials. *Cortex*, *133*, 1–36. https://doi.org/10.1016/j.cortex.2020.09.007

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., et al., (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife*, *7*, e33468. https://doi.org/10.7554/eLife.33468

Otten, M., & Van Berkum, J. J. (2008). Discourse-based word anticipation during language processing: Prediction or priming?. *Discourse Processes*, *45*(6), 464–496. https://doi.org/10.1080/01638530802356463

Payne, B. R., & Federmeier, K. D. (2018). Contextual constraints on lexico-semantic processing in aging: Evidence from single-word event-related brain potentials. *Brain Research*, *1687*, 117–128. https://doi.org/10.1016/j.brainres.2018.02.021

Pichora-Fuller, M. K., Schneider, B. A., & Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *The Journal of the Acoustical Society of America*, *97*(1), 593–608. https://doi.org/10.1121/1.412282

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

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The EZ Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences*, *26*(4), 445–476. https://doi.org/10.1017/S0140525X03000104

Rommers, J., Meyer, A. S., & Huettig, F. (2015). Verbal and nonverbal predictors of language-mediated anticipatory eye movements. *Attention, Perception, & Psychophysics*, *77*, 720–730. https://doi.org/10.3758/s13414-015-0873-x

Salthouse, T. A. (1992). Influence of processing speed on adult age differences in working memory. *Acta Psychologica*, *79*(2), 155–170. https://doi.org/10.1016/0001-6918(92)90030-H

Schwanenflugel, P. J., & Shoben, E. J. (1985). The influence of sentence constraint on the scope of facilitation for upcoming words. *Journal of Memory and Language*, *24*(2), 232–252. https://doi.org/10.1016/0749-596X(85)90026-9

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. https://doi.org/10.1111/lnc3.12151

Staub, A., & Clifton Jr, C. (2006). Syntactic prediction in language comprehension: evidence from either... or. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(2), 425–436. https://doi.org/10.1037/0278-7393.32.2.425

Staub, A., Grant, M., Astheimer, L., & Cohen, A. (2015). The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language*, *82*, 1–17. https://doi.org/10.1016/j.jml.2015.02.004

Steen-Baker, A. A., Ng, S., Payne, B. R., Anderson, C. J., Federmeier, K. D., & Stine-Morrow, E. A. L. (2017). The effects of context on processing words during sentence reading among adults varying in age and literacy skill. *Psychology and Aging, 32*(5), 460–472. https://doi.org/10.1037/pag0000184

Szewczyk, J. M., Mech, E. N., & Federmeier, K. D. (2022). The power of "good": Can adjectives rapidly decrease as well as increase the availability of the upcoming noun? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 48*(6), 856–875. https://doi.org/10.1037/xlm0001091

Szewczyk, J. M., & Schriefers, H. (2013). Prediction in language comprehension beyond specific words: An ERP study on sentence comprehension in Polish. *Journal of Memory and Language*, *68*(4), 297–314. https://doi.org/10.1016/j.jml.2012.12.002

Szewczyk, J. M., & Wodniecka, Z. (2020). The mechanisms of prediction updating that impact the processing of upcoming word: An event-related potential study on sentence comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(9), 1714–1734. https://doi.org/10.1037/xlm0000835

Traxler, M. J., Pickering, M. J., & Clifton Jr, C. (1998). Adjunct attachment is not a form of lexical ambiguity resolution. *Journal of Memory and Language*, *39*(4), 558–592. https://doi.org/10.1006/jmla.1998.2600

Tun, P. A., & Wingfield, A. (1994). Speech recall under heavy load conditions: age, predictability, and limits on dual-task interference. *Aging and Cognition, 1*(1), 29–44. https://doi.org/10.1080/09289919408251448

Urbach, T. P., DeLong, K. A., Chan, W. H., & Kutas, M. (2020). An exploratory data analysis of word form prediction during word-by-word reading. *Proceedings of the National Academy of Sciences*, *117*(34), 20483–20494. https://doi.org/10.1073/pnas.1922028117

Van Berkum, J. J. (2010). The brain is a prediction machine that cares about good and bad-any implications for neuropragmatics? *Italian Journal of Linguistics*, *22*, 181–208. https://hdl.handle.net/11858/00-001M-0000-0012-C6B0-9

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), 443–467. https://doi.org/10.1037/0278-7393.31.3.443

Van Gompel, R. P., Pickering, M. J., Pearson, J., & Liversedge, S. P. (2005). Evidence against competition during syntactic ambiguity resolution. *Journal of Memory and Language*, *52*(2), 284–307. https://doi.org/10.1016/j.jml.2004.11.003

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176–190. https://doi.org/10.1016/j.ijpsycho.2011.09.015

Van Wonderen, E., & Nieuwland, M. S. (2023). Lexical prediction does not rationally adapt to prediction error: ERP evidence from pre-nominal articles. *Journal of Memory and Language*, *132*, 104435. https://doi.org/10.1016/j.jml.2023.104435

Wicha, N. Y., Moreno, E. M., & Kutas, M. (2003). Expecting gender: An event related brain potential study on the role of grammatical gender in comprehending a line drawing within a written sentence in Spanish. *Cortex, 39(3)*, 483–508. https://doi.org/10.1016/S0010-9452(08)70260-0

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. https://doi.org/10.1162/0898929041920487

Wingfield, A., Poon, L. W., Lombardi, L., & Lowe, D. (1985). Speed of processing in normal aging: Effects of speech rate, linguistic structure, and processing time. *Journal of Gerontology*, *40*(5), 579–585. https://doi.org/10.1093/geronj/40.5.579

Wingfield, A., & Stine-Morrow, E. A. L. (2000). Language and speech. In F. I. M. Craik & T. A. Salthouse (Eds.), *The handbook of aging and cognition* (2nd ed., pp. 359–416). Mahwah, NJ: Erlbaum.

Winter, B. (2019). Statistics for linguists: An introduction using R. Routledge. https://doi.org/10.4324/9781315165547

Witzel, N., Witzel, J., & Forster, K. (2012). Comparisons of online reading paradigms: Eye tracking, moving-window, and maze. *Journal of Psycholinguistic Research*, *41*, 105–128. https://doi.org/10.1007/s10936-011-9179-x

Wlotko, E. W., & Federmeier, K. D. (2015). Time for prediction? The effect of presentation rate on predictive sentence comprehension during word-by-word reading. *Cortex*, *68*, 20–32. https://doi.org/10.1016/j.cortex.2015.03.014

Wlotko, E. W., Federmeier, K. D., & Kutas, M. (2012). To predict or not to predict: Age-related differences in the use of sentential context. *Psychology and Aging, 27*(4), 975–988. https://doi.org/10.1037/a0029206

Yan, S., Kuperberg, G. R., & Jaeger, T. F. (2017). Prediction (or not) during language processing. A commentary on Nieuwland et al. (2017) and DeLong et al. (2005). *BioRxiv*, 143750. https://doi.org/10.1101/143750