



Arnout Koornneef*



The processing signature of anticipatory reading: an eye-tracking study on lexical predictions

<https://doi.org/10.1515/ling-2021-0014>

Received March 9, 2018; accepted October 4, 2019; published online February 17, 2021

Abstract: Current approaches to the human language faculty emphasize that during real-time processing anticipatory mechanisms play a vital role for people to parse and comprehend linguistic input at a sufficient pace. Consistent with this view, several Event-Related Potential (ERP) and behavioral self-paced reading (SPR) studies revealed a processing disadvantage for pre-nominal linguistic elements that (grammatically) mismatched with an expected upcoming noun. More recently, however, these findings have been challenged because the results are difficult to replicate. In the current study, I continue this line of replication research with a complementary method: eye tracking. I conducted two experiments aimed at reproducing prior findings of a SPR study of van Berkum, Jos J. A., Colin M. Brown, Pienie Zwitserlood, Valesca Kooijman & Hagoort Peter. 2005. Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31(3). 443–467. The participants read two-sentence stories constructed to elicit a strong lexical prediction about an upcoming noun. To assess whether readers were activating the lexical prediction, the noun was preceded by two gender-inflected adjectives carrying an inflectional suffix that either matched or mismatched with the syntactic gender of the predicted noun. Overall, I did not obtain evidence for strong lexical prediction as the eye-tracking metrics revealed no processing disadvantage for mismatching adjectives (i.e., contrary to the findings of van Berkum et al.). In fact, in some cases readers allocated more processing resources to pre-nominal adjectives that morphologically *matched* with the gender of the predicted noun. These intriguing findings will be discussed in the context of the time course, the processing costs, and the validation processes of lexical predictions.

*Corresponding author: Arnout Koornneef, Department of Education and Child Studies, Leiden University, Pieter de la Court Building, Wassenaarseweg 52, 2333 AK Leiden, The Netherlands, E-mail: a.w.koornneef@fsw.leidenuniv.nl

 Open Access. © 2021 Arnout Koornneef, published by De Gruyter.  This work is licensed under the Creative Commons Attribution 4.0 International License.

Keywords: computational processing costs; eye tracking; language architecture; lexical prediction; morphological inflection; reading

1 Introduction

The notion that people can and routinely will predict upcoming information while processing linguistic input, played only a minor role in the early frameworks on the architecture of the language system. In fact, in classical frameworks – in particular those stemming from the generative grammar tradition – linguists and psycholinguists explicitly argued against the feasibility of predictive processing strategies because after each word of an unfolding sentence, infinite options are available as a plausible continuation (see e.g., Kutas et al. 2011). In contrast, more recent approaches to language processing emphasize the relevance and possibly the inevitability of recruiting anticipatory (language) mechanisms to predict upcoming linguistic material. In these frameworks, it is reasoned that the only way the human language encoder can keep up with a continuous stream of noisy and informationally dense input, is to predict what will come next (for more extensive discussion and other reasons for why predictive processing is useful see e.g., Clark 2013; DeLong et al. 2014; Friston 2010; Huettig 2015; Jackendoff 2002; Kutas et al. 2011; Levinson 2000; Morris 2006; Pickering and Garrod 2007).

In line with more recent accounts on language processing, there is now an accumulating body of evidence indicating that readers and listeners anticipate upcoming linguistic information. Moreover, people seem to do so at many levels of representation, ranging from abstract syntactic structures at the sentence level, to enriched conceptual structures at the discourse level (e.g., Dikker and Pylkkanen 2013; Estevez and Calvo 2000; Federmeier 2007; Kamide et al. 2003; Lau et al. 2006; van Berkum et al. 2005). The mechanisms that give rise to these predictions, however, are not fully understood (e.g., Dikker and Pylkkanen 2013; Bott and Solstad, this issue). This is aptly illustrated by the ongoing debates on the cognitive resources that are required to elicit a linguistic prediction. Roughly two opposing viewpoints can be distinguished. On one end of the spectrum there are frameworks that assume linguistic predictions come more or less for free – e.g., because “this is simply how the human mind works” (Huettig 2015). In these accounts, it is suggested that a predictive (and action-oriented) processing mode is deeply rooted in the neural function and organization of the human mind (cf. Clark 2013; Friston 2010). In contrast, in the frameworks on the other end of the spectrum it is argued that for most linguistic predictions to emerge, an (elaborative) inference is warranted (cf. Calvo 2001; Estevez and Calvo 2000; George et al. 1997; Long and De Ley 2000; Smith and Levy 2008). Although these inferences do not require deliberate

(conscious) processing per se, they are thought to pose a strain on the cognitive resources of readers and listeners nonetheless. In other words, in these latter accounts the preparation of a linguistic prediction should come at a measurable processing cost.

In the context of these extant accounts of (linguistic) predictions the aim of the current study was to increase our understanding of the initial processing phases of anticipation, when the linguistic prediction is activated and “pre-integrated” into the mental representation of a reader. I did so by recording the eye movements of proficient adult readers while they read short stories to assess the processing costs that are associated with the *lexical prediction* of a specific upcoming noun.

1.1 Lexical predictions

Perhaps it seems that it goes without saying that people predict specific upcoming words while processing linguistic input. For one thing, in natural conversations people are capable of finishing each other’s sentences (for detailed discussions on the influence of prediction in dialogs in this issue, see Cummins and Tian; Ouyang and Kaiser; Van Bergen and Hogeweg). In addition, in more controlled settings where people are asked to complete a fragmented story, a similar pattern is observed. People tend to propose the same word to complete biased truncated texts (e.g., van Berkum et al. 2005). In spite of these observations, which intuitively seem closely related to prediction, the past few decades of research have shown that it is notoriously difficult to study whether lexical predictions are genuinely part of “normal” language comprehension processes. This is primarily due to the methodological complexities that arise when studying prediction. A proper investigation of the phenomenon requires the identification of a process that is related to information that has not been encountered yet, which has been proven difficult in many studies. Often it is impossible to disentangle truly predictive processes from its integrative consequences (e.g., Kutas et al. 2011).

There are, however, some notable exceptions to this general rule. Wicha and colleagues (Wicha et al. 2003a, 2003b, 2004) presented compelling evidence for the idea that people pre-activate a lexical item before it is encountered in a discourse. They did so by making use of the Spanish grammar system to study nominal predictions. In Spanish, nouns are preceded by articles inflected for the syntactic gender of the noun. Utilizing this specific grammatical feature in a series of Event-Related Potential (ERP) studies, Wicha and colleagues observed a different ERP waveform for articles that syntactically matched with the gender of a highly anticipated noun, relative to the ERP waveform for articles that mismatched with the gender of that noun. These findings indicated that the noun became

active, fully specified for its grammatical features, before it was encountered in the text. Adopting a similar logic, DeLong et al. (2005) and van Berkum et al. (2005) obtained equivalent results with English and Dutch materials respectively. Whereas DeLong et al. (2005) made use of a phonotactic aspect of English – mandating different indefinite articles (*a* and *an*) depending on the initial phoneme of the immediately following word – the methodology of van Berkum et al. (2005) more closely resembled the design of Wicha and colleagues. Since in the present study I made use of the stimuli of van Berkum et al. (2005) to investigate the processing costs of lexical predictions, a detailed discussion of their materials is provided below (see Table 1).

Dutch nouns carry a fixed syntactic gender feature, *common* or *neuter*, and the adjectives that modify a noun are obligatorily inflected for this feature. Whereas adjectives that modify a singular common-gender noun in indefinite noun phrases carry the inflectional suffix *-e*, adjectives that modify a neuter-gender noun carry no overtly realized inflectional suffix, also known as zero inflection (\emptyset). Van Berkum et al. (2005) took advantage of this feature of Dutch grammar by

Table 1: Examples of the two-sentence Dutch stimuli used in Experiments 1 and 2 (and approximate English translations).

Gender of predicted noun: Common	
Match (<i>e</i> -inflection)	Mismatch (\emptyset -inflection)
<p><i>Na een aantal uren onafgebroken typewerk verloor Maartje haar concentratie. Het was dus hoog tijd voor een <u> korte </u> maar hoogst <u> verfrissende </u> pauze.</i></p> <p>‘After typing for several hours straight, Maartje lost her focus. It was time for a <u>short_e</u> but very <u>refreshing_e</u> break_{common}.’</p>	<p><i>Na een aantal uren onafgebroken typewerk verloor Maartje haar concentratie. Het was dus hoog tijd voor een <u> kort </u> maar hoogst <u> verfrissend </u> dutje.</i></p> <p>‘After typing for several hours straight, Maartje lost her focus. It was time for a <u>short\emptyset</u> but very <u>refreshing\emptyset</u> nap_{neuter}.’</p>
Gender of predicted noun: Neuter	
Match (\emptyset -inflection)	Mismatch (<i>e</i> -inflection)
<p><i>De inbreker had geen enkele moeite de geheime familiekluis te vinden. Deze bevond zich natuurlijk achter een <u> groot </u> maar toch <u> onopvallend </u> schilderij.</i></p> <p>‘The burglar had no trouble locating the secret family safe. Of course, it was situated behind a <u>big\emptyset</u> but also <u>unobtrusive\emptyset</u> painting_{neuter}.’</p>	<p><i>De inbreker had geen enkele moeite de geheime familiekluis te vinden. Deze bevond zich natuurlijk achter een <u> grote </u> maar toch <u> onopvallende </u> boekenkast.</i></p> <p>‘The burglar had no trouble locating the secret family safe. Of course, it was situated behind a <u>big_e</u> but also <u>unobtrusive_e</u> bookcase_{common}.’</p>

The critical adjectives are underlined and in the English translations suffixes are added for convenience.

constructing short stories strongly biased towards a specific noun. For example, in a two-sentence discourse such as, *After typing for several hours straight, Maartje lost her focus. It was time for a short but very refreshing...*, people very strongly anticipate that the noun *break* will follow, before they process the adjectives *short* and *refreshing*. The critical manipulation was that when the participants encountered the adjectives, the suffix of the adjectives either agreed with the grammatical gender of the predicted noun (see top-left story in Table 1) or that it did not (see top-right story in Table 1). Van Berkum et al. (2005) observed that adjectives whose inflectional morphology did not agree with the features of the predicted noun elicited a differential ERP effect. In addition, they conducted a self-paced moving-window reading experiment (i.e., participants repeatedly pressed a button to read a text in a word-by-word fashion, hereafter referred to as SPR) and observed increased reading times for the second prediction-inconsistent adjective (i.e., *refreshing* in the example above). Hence, just like the findings of Wicha et al. (2003a, 2003b, 2004) these electrophysiological and behavioral results strongly suggested that the participants must have predicted the specific noun that is bound to follow (see also Otten and van Berkum 2008, 2009; Otten et al. 2007).¹

1.2 The processing costs of lexical predictions

The studies discussed above suggest that people anticipate lexical elements before these elements are encountered in an unfolding discourse. In addition, the findings also present some insight into the processes and the associated computational costs that precede the moment at which the lexical prediction actually occurs in the input. To fully appreciate the implications for these early stages of anticipation, I will (informally) distinguish several prediction phases. From a functional perspective, a full processing cycle of a lexical prediction (or any prediction for that matter) consists of three main phases. First, the prediction must be activated. In the case of a nominal prediction this entails that the noun becomes pre-activated in the mental lexicon due to the constraining properties of the discourse. Second, the lexical prediction must be updated or even pre-integrated (see below) into the developing mental representation. Third, when people encounter the sentence position where strongly anticipated lexical items (should) occur, a final phase evaluates the lexical prediction against all the available evidence.

¹ In the current contribution I will not provide a detailed discussion of the polarity, latency, and scalp topography of the ERPs reported for the different studies in Dutch, Spanish, and English. As it turns out, these studies revealed a very mixed picture for these aspects of the ERPs, making it difficult to reflect on the type of processes that may underlie the patterns obtained (for discussion and an overview, see Kochari and Flecken 2019).

Based on this simplified framework of activating, updating (pre-integrating), and validating a prediction, three sources of potential processing costs can be distinguished. In the final phase, processing costs may arise when strongly anticipated input is not received and, hence, a prediction turns out to be wrong in the context at hand. There is a large body of evidence, both from electrophysiological and behavioral studies, in support of this hypothesis (e.g., Ehrlich and Rayner 1981; Kutas and Hillyard 1980, 1984; Morris 1994, 2006; for an overview see; Kutas et al. 2011). In most accounts, these costs are thought to resemble some sort of additional processing because the initial prediction must be revised, overridden, re-analyzed, or inhibited – or all of the above. In a way, very similar processing costs may arise while updating the lexical prediction, before the predicted item is actually encountered. That is, the observed differential ERP waveforms to prediction-consistent and prediction-inconsistent determiners and adjectives (DeLong et al. 2005; Otten and van Berkum 2008, 2009; Otten et al. 2007; Wicha et al. 2003a, 2003b, 2004) are often interpreted as reflecting increased syntactic integration efforts, or as the processing consequences of adjusting the nominal prediction (e.g., van Berkum et al. 2005).

The potential processing costs of the first phase, when the lexical prediction is activated, are less well documented. As mentioned earlier, whether this phase demands additional cognitive resources may not be a particularly relevant question in frameworks where a predictive processing mode is deeply grounded in the default functioning of the human mind (cf. Clark 2013; Friston 2010; Huettig 2015). There are, however, some good reasons to assume that the preparation of a prediction comes at a processing cost. This is perhaps most obvious in the case of an elaborate predictive inference (e.g., Calvo 2001; Estevez and Calvo 2000; George et al. 1997; Long and De Ley 2000). In addition, there are more subtle implementations of this hypothesis. For instance, Smith and Levy (2008) put forward a formal model to describe and explain predictability effects on reading times at arbitrary points in written texts. Their model is based on the general idea of *optimal preparation*. The language processor predicts what lies ahead, but at the same time attempts to minimize the trade-off between the processing benefits of a prediction and the resources spent on preparing that prediction. In other words, since preparing to process a word quickly comes at a cost, people only devote their resources to linguistic prediction if it is worth the effort (Kutas et al. 2011; Wlotko and Federmeier 2015).

1.3 All-or-none prediction

Debates on the processing costs of lexical prediction are closely tied to discussions on whether lexical prediction should be interpreted as an *all-or-none* or *graded*

phenomenon. Whereas all-or-none prediction is considered as an active and potentially resource-consuming affair, graded prediction is conceived of as passive, diffuse, global, and cost-free (Luke and Christianson 2016). The ERP studies discussed above seem to provide evidence in favor of the hypothesis that strong, all-or-none lexical prediction is a genuine aspect of language comprehension processes. Otherwise it would be difficult to explain why abstract (semantically arbitrarily) morphosyntactic features of a noun play a role *before* that noun is being processed. Some recent studies and insights, however, cast doubt on this idea. For example, in a comprehensive eye-tracking study, Luke and Christianson (2016) concluded that strong lexical predictions occur in highly constraining contexts only and that continuous, graded prediction would be a better characterization of linguistic pre-activation processes. It should be noted that Luke and Christianson do not dismiss the idea of all-or-none prediction. Instead, they emphasize that the way ERP studies are traditionally designed and conducted may encourage more detailed predictions. They emphasize that ERP studies with written materials often employ a word-by-word presentation mode in which each word is presented for 350–500 ms. These moderate real-time constraints of the methodology offer the participants significantly more time to read, thereby inviting all-or-none prediction. Consistent with this idea, Wlotko and Federmeier (2015) observed in an ERP study that a speeded presentation rate of written stimuli decreases the likelihood that predictive processing will affect ongoing comprehension.

Other developments in the field also call into question whether the design principles of prior studies provide a proper assessment of strong (all-or-none) prediction. As pointed out by Nieuwland et al. (2017), the *a/an* manipulation in DeLong et al.'s study (2005) with English materials may not be a good test case. The manipulation is based on the phonological form of the next word and, hence, is independent of the upcoming noun (e.g., an adjective may intervene between the article and the noun: *an ENORMOUS kite*). Furthermore, the ERP studies conducted on Dutch materials also suffer from a complicating factor. As pointed out by Kochari and Flecken (2019), articles and adjectival forms in Dutch are not exclusively indicative of the syntactic gender of an upcoming (singular) noun. The definite article marking common singular gender (*de* 'the_{common}') is used to mark plural nouns as well, and the definite article marking singular neuter gender (*het* 'the_{neuter}') is used to mark all diminutive derivations (*de taart* – *het taartje* 'the_{common} cake' – 'the_{neuter} tiny cake'). Likewise, in contexts with indefinite determiners (*een* 'a') adjectives modifying a singular diminutive always carry \emptyset -inflection, even if the original noun is of the common-gender type. Perhaps as the result of these complicating factors, some of the effects as reported in previous Dutch and English ERP studies do not appear to be robust and replicable: a multi-lab study by Nieuwland et al. (2018) failed to replicate the results of DeLong et al.

(2005) and a large-sample replication study by Kochari and Flecken (2019) failed to fully reproduce the results of Otten and van Berkum (2009).

1.4 The present study

The discussion above revealed a complicated picture. On the one hand, many ERP studies on lexical prediction showed that an upcoming noun must be activated before that noun is actually encountered in the input. On the other hand, two recent, large-scaled ERP studies failed to replicate these findings. Furthermore, the studies that provided evidence for the hypothesis that lexical predictions are routinely being made employed a design in which the participants either listened to the critical stories (Otten et al. 2007; van Berkum et al. 2005; Wicha et al. 2003a), or alternatively, read the stories in relatively slow, non-self-paced word-by-word manner (DeLong et al. 2005; Otten and van Berkum 2008, 2009; Wicha et al. 2003b, 2004). Hence, these studies cannot provide an answer to the question of whether lexical predictions are an intrinsic aspect of normal *reading* comprehension, when readers move their eyes freely (and rapidly) over a text. Moreover, many other open issues remain, relating to the nature of lexical prediction (graded vs. all-or-none), the associated processing costs, and whether lexical prediction occurs regularly or only in highly constrained contexts. Consequently, novel avenues of experimentation are required to move forward (Nieuwland et al. 2018).

In the current study I try to contribute to this endeavor and at the same time my study resembles recent attempts to reproduce seemingly well-established findings. As discussed in Section 1.1, van Berkum et al. (2005) observed in their behavioral SPR study that readers showed a processing advantage at a second pre-nominal prediction-consistent adjective (i.e., the adjectives *refreshing* and *unobtrusive* in the examples presented in Table 1). However, just like the presentation mode of the visual stimuli in the ERP studies was somewhat artificial (DeLong et al. 2005; Otten and van Berkum 2008, 2009; Wicha et al. 2003b, 2004), a similar objection holds for the word-by-word moving-window SPR paradigm as employed in the study of van Berkum et al. (2005). For example, the somewhat moderate real-time constraints of the methodology offer the participants significantly more time to read (i.e., during *first-pass* reading), relative to the natural reading pace of most individuals. In addition, it has been argued that readers adapt to the word-by-word presentation mode by resorting to a more incremental processing strategy, in which they more rapidly use the information afforded by each word – i.e., to generate, pre-integrate, and validate a lexical prediction – than they would do in unconstrained reading (cf. Koornneef et al. 2019). These concerns related to the ecological validity of word-by-word SPR do not invalidate the results obtained with

the methodology, but they do point to the need for additional, less obtrusive measures (cf. Mitchell 2004).

To address this issue, I repeated the word-by-word SPR experiment of van Berkum et al. (2005) in two *eye-tracking* experiments where Dutch university students freely read through the same materials (i.e., in contrast to the SPR experiment, the texts were presented in their entirety). My main objective was straightforward. I intended to reproduce the grammatical gender effect as reported in ERP and SPR studies. That is, if readers generate strong (all-or-none) lexical predictions during unconstrained reading, a gender-mismatching adjective should come as a surprise and, hence, should induce longer reading times and more regressive eye movements relative to its gender-matching counterpart. I should emphasize, however, that the current study is not merely a replication study as it complements prior (replication) studies in two important ways. First, a research methodology (i.e., eye tracking) was employed that, to my knowledge, has not been used before to study pre-nominal (grammatical gender) effects. A second novel aspect of the current study is that the Dutch common-neuter gender dichotomy will be addressed in more detail. Whereas prior studies used both common and neuter gender nouns (and the corresponding articles or inflected adjectives) to control for potentially confounding factors, I will follow the recommendation of Kochari and Flecken (2019) and explore how syntactic gender modulates the time course of lexical predictions.

2 Experiment 1

2.1 Method

2.1.1 Participants

Participants were 24 undergraduate students from the Utrecht University community (23 female, mean age 21, range 18–34 years) who received money for their participation. In this and the following experiment participants were native speakers of Dutch, without a diagnosed reading or learning disability, and normal or corrected-to-normal vision.

2.1.2 Materials

The stimulus set of the SPR experiment of van Berkum et al. (2005, Experiment 3; see Table 1 for examples) was used in the current eye-tracking study. This set consisted of 40 two-sentence items containing a context sentence followed by a critical target sentence. For each item, there were two versions of the target

sentence. In one version the final noun was a highly expected noun, in the other version it was a much less expected noun (van Berkum et al. assessed the strength of this manipulation in two paper-and-pencil cloze tasks, see their paper for details). The structure of the critical region in the target sentence was held constant across items and conditions, and adhered to the following template: [indefinite article] [adjective-1] [connector] [adverb] [adjective-2] [noun]. The critical manipulation was that at the moment the readers encountered the adjectives during first pass reading, the suffix of the two adjectives either agreed with the grammatical gender of the discourse-predictable noun or that it did not. In half of the items the predictable noun was a neuter-gender noun, in the other half the predictable noun was a common-gender noun. The final noun in prediction-consistent story versions was the discourse-predictable noun (e.g., *painting*). The final noun in prediction-inconsistent story versions was a much less predictable noun of alternative gender (e.g., *bookcase*). Note, however, that the prediction-consistent and prediction-inconsistent story versions were both fully grammatical and semantically coherent.

The stimuli were divided into two counterbalanced lists, with each list containing 20 prediction-consistent story versions (10 with a common gender noun, and 10 with a neuter gender noun) and 20 prediction-inconsistent story versions (10 with a common gender noun, and 10 with a neuter gender noun). Forty stories of an unrelated experiment, examining how the meaning of verbs influences the interpretation of pronominals, were included as fillers (an example of a typical filler item is: *David and Linda were both driving pretty fast. At a busy intersection they crashed hard into each other. David apologized to Linda because he was the one to blame.*). One pseudo-randomization was used for both lists. The original randomization order was used for one half of the participants, the reversed order for the other half. Half of the experimental and filler trials were followed by a statement about the story to encourage discourse comprehension. Participants had to indicate whether the statement about the story was correct or false (half were correct and half were false). On average, participants provided the correct answer to these statements in 94% of the cases (range: 85–100%).

2.1.3 Procedure

Eye movements were recorded with a head-mounted SMI eye tracker that monitored the gaze location of the right eye at a sampling rate of 250 Hz. All participants were individually tested in a sound-treated booth at Utrecht University. The stories were presented in their entirety on a CRT-screen at a viewing distance of approximately 60 cm. Before presentation, a fixation mark appeared on screen at the position of the first word of the first sentence. Participants were instructed to fixate this mark before they made a story visible by pressing a button. After reading a

story the participants again pressed this button to progress. The comprehension questions were answered using two buttons on the same response box. Each session started with written instructions, after which the eye-tracker was mounted and calibrated. Upon successful calibration the experiment started with five practice trials, two followed by a question. Before the experimental trials were presented the eye tracker was re-calibrated. This procedure was repeated three times throughout the experiment. A session was completed within 50 min.

2.1.4 Dependent variables

In eye-movement studies researchers typically report several different, yet inter-related measures (Clifton et al. 2007). In the current study, four commonly reported (first-pass) reading time measures were computed: First-Fixation durations (the duration of the very first fixation on a word), First-Gaze durations (the sum of all fixations on a word before the reader either moves on, or looks back into the text), Right-Bounded durations (the sum of all fixations on a word before moving on progressively) and Regression-Path durations (the sum of all fixation durations from the time when the reader fixates a word, to the time when the reader moves on progressively). In addition to these continuous reading time measures, I will report the categorical measures Fixation Probability (the likelihood that a region receives at least one fixation during first-pass reading) and Regression Probability (the likelihood of a regressive eye-movement after a word is fixated during first-pass).

2.2 Results

For each reading time measure, separate analyses were conducted for three regions of interest: the first adjective, the second adjective, and the final noun. Prior to all analyses, 5.6% of the trials was removed because major tracker losses and eye blinks made it impossible to determine the course of fixations in these critical regions. Furthermore, words that were skipped during first pass reading were treated as missing data. Table 2 reports the average values of the remaining data of the dependent variables as a function of Match (two levels: *match* with highly predictable noun or *mismatch* with highly predictable noun), Predicted Gender (two levels: the highly predictable noun is of the *common* or *neuter* gender type²) and sentence region.

² Note that the labels *common* and *neuter* for the two levels of the factor Predicted Gender refer to the gender of the *predicted* noun, and not to the actual inflection on the adjective, nor to the final noun itself.

Table 2: Mean reading times (in ms) and the fixations and regressions probabilities in Experiment 1 as a function of Predicted Gender, Match, and sentence region.

Measure	Predicted Gender	Match	Sentence region		
			First adj. Mean (SE)	Second adj. Mean (SE)	Noun Mean (SE)
First-fixation	Common	Match	195 (4)	193 (6)	224 (7)
		Mismatch	197 (4)	208 (8)	229 (7)
	Neuter	Match	197 (4)	183 (5)	236 (9)
		Mismatch	196 (4)	187 (6)	247 (8)
First-gaze	Common	Match	213 (6)	235 (10)	294 (14)
		Mismatch	215 (6)	245 (12)	332 (12)
	Neuter	Match	211 (6)	235 (13)	313 (15)
		Mismatch	227 (9)	221 (11)	362 (15)
Right-bounded	Common	Match	224 (7)	251 (11)	294 (15)
		Mismatch	215 (6)	249 (12)	356 (14)
	Neuter	Match	216 (7)	239 (14)	326 (16)
		Mismatch	228 (8)	230 (12)	372 (16)
Regression-path	Common	Match	282 (17)	354 (20)	674 (55)
		Mismatch	248 (11)	351 (29)	778 (56)
	Neuter	Match	279 (25)	354 (37)	689 (58)
		Mismatch	273 (14)	299 (21)	689 (47)
Fixation prob.	Common	Match	0.81 (0.03)	0.80 (0.03)	0.67 (0.03)
		Mismatch	0.78 (0.03)	0.69 (0.03)	0.85 (0.03)
	Neuter	Match	0.76 (0.03)	0.69 (0.03)	0.81 (0.03)
		Mismatch	0.85 (0.02)	0.76 (0.03)	0.85 (0.02)
Regression prob.	Common	Match	0.13 (0.02)	0.36 (0.04)	0.99 (0.01)
		Mismatch	0.10 (0.02)	0.19 (0.03)	0.96 (0.02)
	Neuter	Match	0.11 (0.02)	0.23 (0.03)	0.97 (0.01)
		Mismatch	0.11 (0.02)	0.20 (0.03)	0.93 (0.02)

SE, standard error of mean. Adj., adjective. Prob., probability.

Linear mixed-effects regression models were fitted for the continuous reading time measures (with the response variable log-transformed to correct for right skewness) and generalized mixed-effects regression models were fitted for the categorical dependent measures. I estimated the models with the R package LME4 (version 1.1–20). All models that are reported in this study included the fixed factors Match (match vs. mismatch) and Predicted Gender (common vs. neuter), and the interaction of these factors. Participants and items were included as crossed random effects (Baayen et al. 2008). Sum coding was applied in the main analyses (*match* was coded as -0.5 and *mismatch* as 0.5 ; *common* was coded as -0.5 and *neuter* as 0.5). I will report and discuss effects of Match, and if present, the interactions of Match and Predicted Gender. In the case of a significant

interaction, dummy-coded follow-up analyses were conducted (i.e., I fitted identical models, yet dummy-coded the independent variables and adjusted the reference category to examine the relevant simple main effects). Fixed-effects estimates, t -values (for the continuous dependent variables), z -values (for the categorical dependent variables), and the associated p -values of the main analyses will be reported in tables (see Tables 3 and 5). The results of the follow-up (dummy-coded) analyses will be provided in the text. Note that, since it is not clear how to determine the degrees of freedom for the t -values of the models fitted for the continuous dependent measures (Baayen et al. 2008), the associated p -values are based on z -statistics as well (Barr et al. 2013).

2.2.1 First and second adjectives

Significant interactions of Match and Predicted Gender were observed for the dependent variable Fixation Probability at both the first and second adjective (see Table 3). Follow-up analyses showed that in the neuter-gender conditions participants were more likely to fixate adjectives carrying an inflection that mismatched with the gender of the predicted noun (first adjective: $\beta = 0.67$, $SE = 0.25$, $z = 2.7$, $p < 0.01$; second adjective: $\beta = 0.54$, $SE = 0.24$, $z = 2.3$, $p = 0.02$). A very different pattern was observed for the common-gender conditions. That is, no effect was observed at the first adjective ($\beta = -0.25$, $SE = 0.23$, $z = -1.1$, $p = 0.29$) and at the second adjective an increased fixation probability was observed for adjectives that morphologically *matched* with the predictable noun ($\beta = -0.79$, $SE = 0.25$, $z = -3.2$, $p < 0.01$). The Regression Probability measure revealed a main effect of Match at the second adjective: participants were more likely to regress to earlier sections of the mini story when the adjective *matched* the gender of the predictable noun.

2.2.2 Final noun

The analyses for the final noun revealed main effects for the factor Match in several reading time measures. First-Gaze, Right-Bounded, and Regression-Path durations all displayed shorter reading times for the highly predictable noun than for the less predictable noun. In addition, a main effect of Match and a Match x Predicted Gender interaction were observed for Fixation Probability. Follow-up analyses showed that participants were more likely to fixate the less predictable noun than the highly predictable noun, but only when the predicted noun was of the common-gender type ($\beta = 1.3$, $SE = 0.26$, $z = 5.2$, $p < 0.01$). When the predicted noun was of the neuter-gender type, fixation probabilities for the anticipated and unanticipated nouns did not differ ($\beta = 0.34$, $SE = 0.27$, $z = 1.3$, $p = 0.20$).

Table 3: Fixed-effects estimates and the associated statistics of the sum-coded models fitted for the dependent variables in Experiment 1.

Measure	Fixed effect	First adj.					Second adj.					Sentence Region				
		β	SE	t/z	p		β	SE	t/z	p		β	SE	t/z	p	
First-fixation	Match	0.005	0.02	0.30	0.76	0.019	0.03	0.75	0.46	0.041	0.03	1.56	0.12			
	Gender	0.001	0.02	0.07	0.95	-0.061	0.05	-1.15	0.25	0.034	0.03	1.31	0.19			
First-gaze	Match x gender	-0.010	0.04	-0.27	0.79	0.018	0.05	0.34	0.74	0.037	0.05	0.72	0.47			
	Match	0.031	0.02	1.29	0.20	-0.022	0.03	-0.66	0.51	0.151	0.03	4.69	<0.01			
Right-bounded	Gender	0.004	0.02	0.17	0.87	-0.070	0.08	-0.83	0.41	0.049	0.03	1.40	0.16			
	Match x gender	0.039	0.05	0.82	0.41	0.009	0.07	0.14	0.89	0.021	0.06	0.32	0.75			
Regression-path	Match	0.004	0.02	0.15	0.88	-0.028	0.03	-0.84	0.40	0.181	0.03	5.60	<0.01			
	Gender	-0.003	0.03	-0.11	0.91	-0.084	0.09	-0.91	0.36	0.056	0.04	1.31	0.19			
Fixation prob.	Match x gender	0.074	0.05	1.51	0.13	0.064	0.07	0.95	0.34	-0.055	0.06	-0.85	0.39			
	Match	-0.019	0.03	-0.56	0.57	-0.069	0.05	-1.42	0.16	0.142	0.06	2.46	0.01			
Regression prob.	Gender	-0.006	0.04	-0.18	0.86	-0.125	0.09	-1.36	0.18	-0.047	0.08	-0.59	0.55			
	Match x gender	0.072	0.07	1.09	0.28	0.037	0.10	0.38	0.70	-0.118	0.12	-1.02	0.31			
Noun	Match	0.212	0.17	1.25	0.21	-0.128	0.17	-0.74	0.46	0.832	0.18	4.51	<0.01			
	Gender	0.159	0.26	0.61	0.54	-0.253	0.43	-0.59	0.55	0.459	0.28	1.63	0.10			
Noun	Match x gender	0.925	0.34	2.71	0.01	1.332	0.34	3.86	<0.01	-0.983	0.37	-2.68	0.01			
	Match	-0.208	0.26	-0.80	0.42	-0.595	0.20	-3.01	<0.01	-1.024	0.57	-1.81	0.07			
Noun	Gender	-0.068	0.33	-0.21	0.84	-0.259	0.29	-0.9	0.37	-0.969	0.72	-1.34	0.18			
	Match x gender	-0.046	0.52	-0.09	0.93	0.732	0.39	1.86	0.06	0.084	1.12	0.08	0.94			

Adj., adjective. Prob., probability. Estimates of Fixation and Regression Probabilities reflect logit scores.

2.3 Discussion

Consistent with the findings of many studies, the results revealed that highly predictable nouns were processed more quickly than less predictable nouns – and are fixated less often when the highly predictable noun is of the common-gender type. Our main interest, however, lies in how the two critical adjectives are processed before readers encounter the actual noun. Previous ERP and behavioral experiments revealed a processing disadvantage for pre-nominal linguistic elements that grammatically mismatched with an expected upcoming noun (e.g., Otten and van Berkum 2008, 2009; van Berkum et al. 2005; Wicha et al. 2003a, 2003b, 2004). The results of Experiment 1 do not replicate these findings.

First of all, none of the continuous measures revealed a reading time delay for mismatching adjectives. Furthermore, although participants were less likely to fixate an adjective that morphologically matched with a highly predictable neuter-gender noun, it is unclear whether these inflated skipping rates should be attributed to lexical prediction. That is, the common-gender conditions revealed a very different, arguably opposite pattern, with increased skipping rates for mismatching adjectives (note that this effect was significant at the second adjective only). Hence, perhaps a more parsimonious explanation for the observed “cross-over” interactions is to interpret them as main effects of inflection instead: adjectives with e-inflection are simply fixated more often than adjectives with \emptyset -inflection. On this view, the overall pattern of fixation probabilities at the critical adjectives should be attributed to features of e- and \emptyset -inflection that are orthogonal to the influence of lexical prediction. For example, word length effects (number of letters, number of syllables, spatial extent; see Barton et al. [2014] for a review) should be taken into account as e-inflected adjectives tend to be longer – and longer words tend to be skipped less often (e.g., Rayner et al. 2011). But many other features of e- and \emptyset -inflection could be of relevance here, such as their morphological complexity (the surface structure – not the deep structure – of e-inflected adjectives is morphologically more complex) and how frequently they occur in day-to-day life (the distribution of e-inflected adjectives and \emptyset -inflected adjectives is skewed with the former outnumbering the latter, see Blom et al. [2008]).

Not only did Experiment 1 reveal no clear evidence of a processing disadvantage for mismatching adjectives, but also it provided some results that suggested the exact opposite. More specifically, participants were more likely to make a regressive eye movement out of the second adjective region if that adjective morphologically *matched* with the gender of the predicted noun. On the assumption that increased regression rates are indicative of increased cognitive effort, this would provide evidence against the idea that mismatching adjectives

are more difficult to process, and more speculatively, against the idea that (all-or-none) lexical predictions are generated by readers.

Obviously, the evidence for this elaborate interpretation of the results is weak, particularly since the influence of parafoveal preview of the critical noun may have contaminated the results for the second adjective: it is difficult to disentangle whether the increased rate of regressions is due to the second adjective itself or arises as a consequence of a preview effect of the final noun instead. In all, to avoid reporting and interpreting spurious results, a replication experiment was conducted in which the same eye-tracking methodology was used, and the exact same critical stories were presented to a new – and larger – sample of university students.

3 Experiment 2

3.1 Method

3.1.1 Participants

Participants were 59 undergraduate students from the Utrecht University community (49 female, mean age 23, range 19–31 years) who received money for their participation. None of them participated in Experiment 1.

3.1.2 Materials

The critical stimuli were identical to the stimuli of Experiment 1. In addition, the filler items were held constant across experiments, with one small exception: the total set of fillers was increased from 40 to 48 items (the eight additional filler items were of the same type as the items in the original filler set).

3.1.3 Procedure

The experimental procedure was kept constant across experiments, with some minor exceptions. First, in Experiment 2 the eye movements were recorded with a desktop-mounted EyeLink 1000 eye tracker, sampling at a rate of 500 Hz. Second, the stories were presented on a LCD screen. Third, Experiment 2 was part of a larger reading study consisting of two 90-min sessions (the two sessions never took place on the same day). The eye-tracking experiment was the first experiment in the second session.

3.2 Results

The procedure for the analyses was identical to the procedure in Experiment 1. Trials with major tracker losses and too many eye blinks in the critical regions were removed from the analyses (< 1%). Furthermore, words that were skipped during first-pass reading were treated as missing data in the reading duration variables. Table 4 reports the average values of the remaining data of the dependent variables as a function of Match, Predicted Gender, and sentence region. Table 5 reports the results of the mixed-effects analyses.

Table 4: Mean reading times (in ms) and the fixations and regressions probabilities in Experiment 2 as a function of Predicted Gender, Match, and sentence region.

Measure	Predicted Gender	Match	Sentence region		
			First adj. Mean (SE)	Second adj. Mean (SE)	Noun Mean (SE)
First-fixation	Common	Match	231 (4)	260 (5)	217 (6)
		Mismatch	220 (4)	246 (5)	253 (8)
	Neuter	Match	217 (4)	255 (6)	241 (8)
		Mismatch	221 (4)	251 (5)	241 (6)
First-gaze	Common	Match	242 (5)	338 (8)	231 (6)
		Mismatch	229 (5)	290 (6)	297 (10)
	Neuter	Match	226 (4)	295 (7)	276 (10)
		Mismatch	231 (5)	323 (9)	267 (7)
Right-bounded	Common	Match	257 (6)	381 (9)	256 (9)
		Mismatch	239 (5)	331 (8)	365 (14)
	Neuter	Match	235 (5)	360 (10)	306 (11)
		Mismatch	243 (5)	397 (12)	325 (11)
Regression-path	Common	Match	314 (16)	567 (28)	556 (55)
		Mismatch	284 (9)	448 (21)	882 (58)
	Neuter	Match	290 (13)	632 (30)	816 (50)
		Mismatch	292 (11)	669 (31)	934 (50)
Fixation prob.	Common	Match	0.73 (0.02)	0.94 (0.01)	0.59 (0.02)
		Mismatch	0.68 (0.02)	0.90 (0.01)	0.86 (0.01)
	Neuter	Match	0.65 (0.02)	0.88 (0.01)	0.69 (0.02)
		Mismatch	0.72 (0.02)	0.91 (0.01)	0.75 (0.02)
Regression prob.	Common	Match	0.12 (0.02)	0.28 (0.02)	0.34 (0.03)
		Mismatch	0.15 (0.02)	0.19 (0.02)	0.44 (0.02)
	Neuter	Match	0.13 (0.02)	0.38 (0.02)	0.59 (0.02)
		Mismatch	0.13 (0.02)	0.34 (0.02)	0.61 (0.02)

SE, standard error of mean. Adj., adjective. Prob., probability.

Table 5: Fixed-effects estimates and the associated statistics of the sum-coded models fitted for the dependent variables in Experiment 2.

Measure	Fixed effect	First adj.						Second adj.						Sentence region											
		Noun			Verb			Noun			Verb			Noun			Verb								
		β	SE	t/z	p	β	SE	t/z	p	β	SE	t/z	p	β	SE	t/z	p	β	SE	t/z	p				
First-fixation	Match	-0.014	0.01	-0.94	0.35	-0.020	0.02	-1.30	0.19	0.064	0.02	2.90	<0.01	-0.021	0.02	-1.11	0.27	-0.012	0.03	-0.38	0.70	0.018	0.04	0.42	0.67
	Gender	0.076	0.03	2.52	0.01	0.048	0.03	1.52	0.13	-0.107	0.04	-2.42	0.02	-0.018	0.02	-1.09	0.28	-0.027	0.02	-1.43	0.15	0.088	0.02	3.70	<0.01
First-gaze	Match	-0.023	0.02	-0.94	0.35	-0.043	0.05	-0.84	0.40	0.011	0.06	0.20	0.84	-0.023	0.02	-0.94	0.35	-0.043	0.05	-0.84	0.40	0.011	0.06	0.20	0.84
	Gender	0.086	0.03	2.64	0.01	0.180	0.04	4.65	<0.01	-0.184	0.05	-3.85	<0.01	0.086	0.03	2.64	0.01	0.180	0.04	4.65	<0.01	-0.184	0.05	-3.85	<0.01
Right-bounded	Match	-0.019	0.02	-1.09	0.27	-0.023	0.02	-1.11	0.27	0.167	0.03	6.40	<0.01	-0.019	0.02	-1.09	0.27	-0.023	0.02	-1.11	0.27	0.167	0.03	6.40	<0.01
	Gender	-0.030	0.03	-1.18	0.24	0.010	0.06	0.16	0.87	0.014	0.07	0.18	0.85	-0.030	0.03	-1.18	0.24	0.010	0.06	0.16	0.87	0.014	0.07	0.18	0.85
Regression-path	Match x gender	0.115	0.03	3.34	<0.01	0.203	0.04	4.83	<0.01	-0.211	0.05	-4.03	<0.01	0.115	0.03	3.34	<0.01	0.203	0.04	4.83	<0.01	-0.211	0.05	-4.03	<0.01
	Match	-0.009	0.02	-0.36	0.72	-0.088	0.03	-2.90	<0.01	0.216	0.04	5.44	<0.01	-0.009	0.02	-0.36	0.72	-0.088	0.03	-2.90	<0.01	0.216	0.04	5.44	<0.01
Fixation prob.	Match	-0.025	0.03	-0.77	0.44	0.150	0.09	1.61	0.11	0.298	0.18	1.66	0.10	-0.025	0.03	-0.77	0.44	0.150	0.09	1.61	0.11	0.298	0.18	1.66	0.10
	Gender	0.090	0.05	1.88	0.06	0.249	0.06	4.06	<0.01	-0.217	0.08	-2.71	0.01	0.090	0.05	1.88	0.06	0.249	0.06	4.06	<0.01	-0.217	0.08	-2.71	0.01
Regression prob.	Match	-0.108	0.20	-0.54	0.59	-0.359	0.47	-0.76	0.45	-0.197	0.33	-0.60	0.55	-0.108	0.20	-0.54	0.59	-0.359	0.47	-0.76	0.45	-0.197	0.33	-0.60	0.55
	Gender	0.569	0.20	2.85	<0.01	1.004	0.31	3.22	<0.01	-1.767	0.24	-7.39	<0.01	0.569	0.20	2.85	<0.01	1.004	0.31	3.22	<0.01	-1.767	0.24	-7.39	<0.01
Regression prob.	Match	0.185	0.15	1.20	0.23	-0.440	0.11	-4.00	<0.01	0.324	0.14	2.39	0.02	0.185	0.15	1.20	0.23	-0.440	0.11	-4.00	<0.01	0.324	0.14	2.39	0.02
	Gender	-0.036	0.23	-0.16	0.88	0.765	0.28	2.68	0.01	1.283	0.53	2.42	0.02	-0.036	0.23	-0.16	0.88	0.765	0.28	2.68	0.01	1.283	0.53	2.42	0.02
Regression prob.	Match x gender	-0.210	0.31	-0.67	0.50	0.519	0.22	2.32	0.02	-0.350	0.27	-1.28	0.20	-0.210	0.31	-0.67	0.50	0.519	0.22	2.32	0.02	-0.350	0.27	-1.28	0.20
	Match	0.185	0.15	1.20	0.23	-0.440	0.11	-4.00	<0.01	0.324	0.14	2.39	0.02	0.185	0.15	1.20	0.23	-0.440	0.11	-4.00	<0.01	0.324	0.14	2.39	0.02

Adj., adjective. Prob., probability. Estimates of Fixation and Regression Probabilities reflect logit scores.

3.2.1 First and second adjectives

At the first adjective, I observed significant interactions between the factors Match and Predicted Gender for First-Fixation durations, First-Gaze durations, Right-Bounded durations, and Fixation Probability. Follow-up analyses revealed that in the common-gender conditions the reading times were longer for *matching* adjectives than for *mismatching* adjectives (First-Fixation: $\beta = -0.052$, $SE = 0.021$, $t = -2.4$, $p = 0.01$; First-Gaze: $\beta = -0.061$, $SE = 0.023$, $t = -2.6$, $p < 0.01$; Right-Bounded: $\beta = -0.076$, $SE = 0.024$, $t = -3.1$, $p < 0.01$). These differences in reading times, however, were absent in the neuter-gender conditions (First-Fixation: $\beta = 0.024$, $SE = 0.021$, $t = 1.1$, $p = 0.26$; First-Gaze: $\beta = 0.025$, $SE = 0.023$, $t = 1.1$, $p = 0.27$; Right-Bounded: $\beta = 0.039$, $SE = 0.024$, $t = 1.6$, $p = 0.11$). The follow-up analyses for Fixation Probability showed that there was no effect of Match in the common-gender conditions ($\beta = -0.27$, $SE = 0.14$, $z = -1.8$, $p = 0.07$), yet in the neuter-gender conditions matching adjectives were skipped more often than mismatching adjectives ($\beta = 0.30$, $SE = 0.14$, $z = 2.2$, $p = 0.03$).

At the second adjective, I observed significant interactions for First-Gaze durations, Right-Bounded durations, Regression-Path durations, Fixation Probability, and Regression Probability. Follow-up analyses showed that in the common-gender conditions adjectives that matched with the gender of the highly predictable noun were fixated more often, induced longer reading times, and triggered more regressions than the adjectives that did not match (First-Gaze: $\beta = -0.12$, $SE = 0.028$, $t = -4.3$, $p < 0.01$; Right-Bounded: $\beta = -0.12$, $SE = 0.030$, $t = -4.2$, $p < 0.01$; Regression-Path: $\beta = -0.21$, $SE = 0.044$, $t = -4.9$, $p < 0.01$; Regression Probability: $\beta = -0.70$, $SE = 0.17$, $z = -4.1$, $p < 0.01$; Fixation Probability: $\beta = -0.65$, $SE = 0.23$, $z = -2.8$, $p < 0.01$). For the neuter-gender conditions a different pattern was observed. There was no difference between matching and mismatching adjectives in Regression-Path durations ($\beta = 0.036$, $SE = 0.043$, $t = 0.9$, $p = 0.39$), Fixation Probability ($\beta = 0.36$, $SE = 0.20$, $z = 1.8$, $p = 0.08$), and Regression Probability ($\beta = -0.18$, $SE = 0.14$, $z = -1.3$, $p = 0.21$). However, in First-Gaze and Right-Bounded durations mismatching adjectives induced longer reading times than matching adjectives did (First-Gaze: $\beta = 0.063$, $SE = 0.027$, $t = 2.3$, $p = 0.02$; Right-Bounded: $\beta = 0.078$, $SE = 0.029$, $t = 2.7$, $p < 0.01$).

3.2.2 Final noun

The analyses for the final noun revealed a main effect of Match for Regression Probability and Match x Predicted Gender interactions for First-Fixation durations, First-Gaze durations, Right-Bounded durations, Regression-Path durations and Fixation Probability. The overall pattern was that highly-predictable nouns were

skipped less often, induced shorter reading times, and triggered fewer regressions than did less predictable nouns. However, these mismatching effects were more pronounced in the common gender-conditions (First-Fixation: $\beta = 0.12$, $SE = 0.032$, $t = 3.7$, $p < 0.01$; First-Gaze: $\beta = 0.18$, $SE = 0.035$, $t = 5.2$, $p < 0.01$; Right-Bounded: $\beta = 0.27$, $SE = 0.038$, $t = 7.2$, $p < 0.01$; Regression-Path: $\beta = 0.32$, $SE = 0.058$, $t = 5.6$, $p < 0.01$; Fixation Probability: $\beta = 2.1$, $SE = 0.18$, $z = 11.4$, $p < 0.01$) than in the neuter-gender conditions (First-Fixation: $\beta = 0.010$, $SE = 0.030$, $t = 0.33$, $p = 0.74$; First-Gaze: $\beta = -0.0042$, $SE = 0.033$, $t = -0.1$, $p = 0.90$; Right-Bounded: $\beta = 0.061$, $SE = 0.036$, $t = 1.7$, $p = 0.09$; Regression-Path: $\beta = 0.11$, $SE = 0.055$, $t = 2.0$, $p = 0.05$; Fixation Probability: $\beta = 0.33$, $SE = 0.15$, $z = 2.2$, $p = 0.03$).

3.3 Discussion

The results of Experiment 2 partly confirmed but in addition clearly extended the findings of Experiment 1. In both experiments the analyses at the noun revealed a processing advantage for highly predictable nouns, relative to their less predictable alternatives. However, in Experiment 2 these processing advantages for anticipated nouns emerged reliably for all dependent measures in the common-gender conditions, which was, somewhat surprisingly, not the case in the neuter-gender conditions – only Regression-Path duration, Fixation Probability, and Regression Probability revealed a relatively weak processing advantage for anticipated nouns. This could be taken to suggest that lexical predictions were less prominent (or attenuated) for neuter gender nouns – however, note that the critical nouns were not matched across conditions, because van Berkum et al. (2005) optimized their design to study *pre*-nominal incongruency effects.

The analyses of Experiment 1 produced some isolated, yet intriguing results in the adjectival regions. First, matching adjectives induced *more* regressive eye-movements than mismatching adjectives. Second, mismatching adjectives were more likely to be fixated than matching adjectives in the neuter-gender conditions, yet the opposite pattern was observed for the common-gender conditions where mismatching adjectives were skipped more often than matching adjectives. This latter pattern (i.e., a cross-over interaction of Match and Predicted Gender, with a match effect for common-gender conditions and a mismatch effect for neuter-gender conditions) emerged more consistently in Experiment 2: interactions were observed in most (if not all) dependent variables at both the first and second adjective. As mentioned in Section 2.3, a relatively straightforward interpretation for these results is to attribute them to features of *e*- and \emptyset -inflection that are independent of the influence of lexical prediction. That is, *e*-inflected adjectives may require more processing resources than \emptyset -inflected adjectives due to, for

example, word length, morphological complexity, and frequency effects. There is one caveat, however: the matching effect for common-gender nouns is more pronounced than the mismatching effect for neuter-gender nouns. This is most apparent in the first adjective region where the common-gender conditions induced match effects in numerous eye-tracking measures, yet the mismatch effect in the neuter-gender condition was only reliable in the Fixation Probability metric – i.e., reading time metrics revealed no difference between matching and mismatching adjectives in the neuter-gender conditions.

4 General discussion

Many studies suggested that strong (all-or-none) lexical predictions are routinely being made in both listening (Otten et al. 2007; van Berkum et al. 2005; Wicha et al. 2003a) and reading paradigms (DeLong et al. 2005; Otten and van Berkum 2008, 2009; Wicha et al. 2003b, 2004). More recently, however, these findings have been challenged for several reasons. First, two recent, large-scaled studies failed to replicate crucial findings (Kochari and Flecken 2019; Nieuwland et al. 2018). Second, concerns have been raised about the materials that were presented to the participants; they may be too constraining and not the best test case to examine lexical prediction (Luke and Christianson 2016; Nieuwland et al. 2018). Third, in reading paradigms a relatively slow presentation mode may have invited readers to engage in all-or-none lexical prediction (Luke and Christianson 2016; Wlotko and Federmeier 2015). Hence, even if evidence in favor of lexical prediction was obtained in reading studies, it is unclear whether lexical prediction will occur in more natural reading settings (note that this does not apply to the ERP studies that use a listening paradigm). In the context of these open issues and concerns, my main research objective was straightforward. In two eye-tracking experiments, I examined whether readers generate strong lexical predictions in a relatively naturalistic reading setting and I evaluated whether these predictions are activated or updated in the same vein as shown in previous studies. In addition to this main objective I explored how syntactic gender features (i.e., common vs. neuter) modulate the time course of nominal predictions in Dutch.

A synthesis of the results of the two experiments reveals a somewhat puzzling pattern that can be summarized as follows. First, highly-anticipated nouns are processed more rapidly than less anticipated nouns. Second, this predictability advantage appears to be more prominent for common-gender nouns than for neuter-gender nouns. Third, pre-nominal adjectives that morphologically match with an anticipated common-gender noun require prolonged processing (a *match* effect). Fourth, pre-nominal adjectives that morphologically match with an

anticipated neuter-gender noun require less processing (a *mismatch* effect). Fifth, the match effect for adjectives in the common-gender conditions is more pronounced than the mismatch effect for the adjectives in the neuter-gender conditions.

Based on previous findings, evidence consistent with strong prediction would have been obtained if mismatching adjectives in both the common-gender and the neuter-gender conditions induced longer reading times and/or more regressive eye-movements than matching adjectives. In that sense I fail to replicate the findings of prior studies; most notably the behavioral SPR study of van Berkum et al. (2005) in which identical critical stimuli (and nearly-identical filler stimuli) were presented to the readers. Hence, in the current study lexical predictions are clearly *not* activated and/or updated in the same vein as shown in previous studies. On a general level this shows that the usage of complementary research methods is vital, even when studying ostensibly well-established phenomena (cf. Nieuwland et al. 2018). Furthermore, because the results of Experiments 1 and 2 were similar, yet not identical, my study also highlights the importance of repeating an experiment several times. Finally, a more speculative conclusion that can be drawn is that, in line with a proposal of Luke and Christianson (2016), only highly-constrained contexts in which readers process the incoming information at a relatively slow pace, will induce strong, all-or-none lexical predictions.

4.1 Do lexical predictions play no role in the current study?

On the one hand, the results of the current study do not present evidence for all-or-none prediction – at least not at first glance. On the other hand, it cannot be ruled out that nominal predictions were generated by the readers. After all, anticipated nouns were processed more quickly than unanticipated nouns and, in both experiments, readers were sensitive to the manipulation at the gender-inflected adjectives – albeit in a puzzling way. I will therefore explore several alternative, prediction-oriented explanations that could also account for the intriguing data of the current study. The time course, processing costs, and validation processes of prediction will be accentuated in this discussion.

In Sections 2.3 and 3.3, I raised the possibility that the interaction effects at the adjective regions emerged for reasons that are unrelated to lexical prediction: e-inflected adjectives simply require more processing resources than \emptyset -inflected adjectives. This, however, does not rule out that all-or-none nominal predictions are activated during reading. The main difference between prior studies and the current study would then be that in prior studies the morphosyntactic properties of the adjectives are used to validate predictions, whereas in the current study they

are not. On that view, the reading-time constraints that are enforced by a research method do not determine whether all-or-none lexical predictions are activated (cf. Luke and Christianson 2016), but they *do* affect whether prediction incongruencies are detected and repaired on the fly. This would be in line with frameworks on sentence and text validation mechanisms in which certain stages of validation are a resource-consuming affair and under strategic control of the reader (Isberner and Richter 2014) (see also Section 4.2).

This explanation of the data disregards any influence of lexical prediction on how the critical adjectives are processed by the participants. Although this makes sense in the current situation, as it reflects a plausible and perhaps the most parsimonious interpretation of the data, it does not seem to tell the whole story. Recall that the match effect (matching adjectives induce more processing costs than mismatching adjectives) in the common-gender conditions was far more pronounced than the mismatch effect (mismatching adjectives induce more processing costs than matching adjectives) in the neuter-gender conditions. In fact, the only reliable mismatch effect observed at the first adjective in the neuter-gender conditions was that mismatching adjectives were skipped more often. If we assume that the findings for the adjective regions *do* reflect lexical prediction processes and that the first adjective presents a more reliable region of interest than the second adjective (i.e., the results for the latter region are potentially contaminated by a parafoveal preview of the critical noun) two interesting issues arise. Namely, (1) why did readers slow down while they were processing adjectives that morphologically matched with an anticipated noun and (2) why did this match effect surface if the predicted noun carried a common-gender feature, but no effect was observed when the predicted noun carried a neuter-gender feature?

4.2 Why do prediction-*consistent* adjectives induce a processing delay?

At the outset of this contribution, I distinguished two opposing viewpoints on the processing costs of lexical prediction. Whereas some frameworks assume that linguistic predictions come for free, other frameworks state that the preparation of a linguistic prediction should come at a measurable processing cost (cf. Calvo 2001; Clark 2013; Estevez and Calvo 2000; Friston 2010; George et al. 1997; Huettig 2015; Long and De Ley 2000; Luke and Christianson 2016). In the context of these extant accounts of linguistic prediction, the results appear to be more consistent

with the latter type of frameworks: a processing advantage of highly-anticipated (common-gender) nouns comes at the expense of increased processing costs in preceding sentence regions (i.e., in this case the adjectival regions). Depending on the exact time course of lexical prediction in the current study, these increased processing costs may reflect (all-or-none) *activation* processes or, alternatively, they may reflect processes of *updating* or *pre-integration*.

If the match effect reflects the processing costs of activating a prediction, one must assume that due to the real-time constraints of the reading task, it became less feasible for the readers to generate a lexical prediction in the current study than in prior (ERP) studies. Consequently, their lexical predictions were delayed, or at least not fully active, when they encountered the critical adjectives: only at the moment readers encounter the first inflected adjective, the continuation of the sentence becomes constrained to such an extent that it becomes worthwhile to generate an all-or-none lexical prediction. This approach presupposes a hybrid prediction mechanism in which non-taxing, graded prediction can evolve into more resource-consuming, all-or-none prediction. Hence, on this view an important issue for future research is to examine when and how this transition in the linguistic prediction system takes place and what kind of linguistic information would be sufficient to unleash all-or-none prediction.

It is also possible – perhaps even more plausible – that the main issue is not so much when the nominal prediction becomes fully active, but whether and how the prediction is pre-integrated into the developing mental model of the reader. Earlier the conjecture was made that in order to explain *any* processing differences between matching and mismatching adjectives, one must assume that at least a rudimentary form of syntactic pre-integration takes place in which the parser checks the syntactic features of the adjective to those of the anticipated noun (van Berkum et al. 2005). There is no a priori reason, however, to assume that processes of pre-integration should be limited to *syntactic* pre-integration, i.e., the adjective may also be pre-integrated *semantically* with the anticipated noun. Then, based on the assumptions (1) that semantic pre-integration of the adjective and the noun requires some cognitive effort and (2) that semantic integration is only initiated if the inflection on the adjective corresponds with the gender of the anticipated noun, the match effect at the critical adjectives can be accounted for: matching adjectives (temporarily) induce a higher cognitive load than mismatching adjectives because the former are semantically pre-integrated right away, whereas the latter are not.

This explanation of the match effect presupposes a cognitive language architecture in which (morpho)syntactic processing precedes – and in case of an

ungrammatical dependency even blocks – subsequent semantic processing.³ In addition to these claims about the sequential architecture of the human language system, we also have to assume that in the case of a morphological mismatch the reader does not initiate an attempt to adjust or repair the lexical prediction right away – after all, this should incur measurable costs at the mismatching adjectives. At first glance this seems incompatible with the widely held belief that language comprehension is a highly incremental affair (i.e., a reader continuously updates and meticulously checks his or her mental representation of an unfolding text; e.g., Kutas et al. 2011; van Berkum et al. 2005). However, there is also an accumulating body of evidence suggesting that language comprehension does not proceed fully incrementally in all circumstances. Parsing and integration decisions are often postponed by language comprehenders. This “wait-and-see” approach has been reported, for example, in studies examining the resolution of ambiguous pronouns (MacDonald and MacWhinney 1990; Stewart et al. 2007). Similarly, readers often construct an underspecified syntactic representation of a sentence, in particular in the case of garden-path sentences (e.g., von der Malsburg and Vasishth 2013). As a final example, recently O’Brien and Cook (2016) presented a model on text comprehension that assumes that connections formed in the integration stage of comprehension are subsequently checked against information in memory in a validation stage (cf. Isberner and Richter 2014). They explicitly mention that in particular the validation stage – which may trigger processes of re-analyses and repair – has the potential to have a delayed influence on comprehension. Putting aside the discrepancies between these studies and frameworks, they point in the same direction. Although readers often use the information in a sentence or discourse right at the moment it becomes available, there are also circumstances in which the available information is used only partly, in a delayed manner, or not at all. Extrapolated to the current study this means that even in the face of morphological evidence against a specific prediction, readers may “decide” to postpone processes of re-analyses and repair.

³ Although a sequential architecture is upheld in many ‘single-stream’ theories of language comprehension (e.g., Frazier and Rayner 1982; Friederici and Kotz 2003; Koornneef 2008; Reuland 2001, 2011) there are also many “multi-stream” theories (e.g., Ferreira and Patson 2007; Karimi and Ferreira 2016; Kuperberg 2007; van Herten et al. 2006) claiming that a separate semantic (or heuristic) representation can be constructed independently of the surface structure of a sentence or text (for a discussion of single- vs. multi-stream frameworks and a defense of single-stream frameworks cf. Brouwer et al. [2012]; Koornneef [2008]). Furthermore, note that multi-stream models would not readily predict a match effect in the current explanation.

4.3 Why is the influence of prediction primarily observed for common-gender nouns?

If we assume that the match effect in the common-gender conditions is directly related to the activation or pre-integration of lexical predictions, then a puzzling finding is that the experimental manipulation did not result in a match effect for the neuter-gender conditions. This could indicate that readers were not generating a specific nominal prediction when the stories were biased towards a neuter-gender noun, which would be consistent with the observation that in the neuter-gender conditions of Experiment 2 attenuated reading time differences emerged between the highly and less predictable nouns. However, this attenuated effect at the final noun was not observed in Experiment 1. Moreover, the idea that only common-gender nouns can be predicted by a reader seems somewhat peculiar and would clearly deviate from the conclusions of previous studies (Otten and van Berkum 2008, 2009; Otten et al. 2007; van Berkum et al. 2005).

Although it is difficult to provide a straightforward solution to this final puzzle, I would like to point out that so-called *deflection* phenomena could play a role here. Deflection is the tendency of a language “to get rid of” its inflectional morphology (Bennis 2010). This phenomenon is observed in Dutch and holds for many Germanic languages. Although the influence of deflection is most clearly visible for verbal inflection, adjectival inflection in Dutch seems to be under pressure as well. That is, e-inflection (the default in Dutch) tends to become more dominant over time, which induces overgeneralization (i.e., e-inflection is used for neuter-gender nouns after an indefinite determiner) and may even result in a gradual disappearance of the usage of \emptyset -inflected adjectives (Bennis 2010; Bennis and Hinskens 2014). On the assumption that this gradual disappearance of \emptyset -inflection is real – and is already affecting the syntactic features of the entries of neuter nouns in the lexicon of Dutch readers – one could make the following conjecture: inflected adjectives are informative if the predicted noun is of the common-gender type (i.e., e-inflection is compatible with a common noun, yet \emptyset -inflection is incompatible with a common-gender noun), whereas inflected adjectives are not (or less informative) if the predicted noun is of the neuter-gender type (i.e., e-inflection and \emptyset -inflection are, to some extent, both compatible with a neuter-gender noun). Although this would provide an elegant explanation for why nominal predictions affect the processing signature of adjectives in the common-gender conditions but not (or differently) in the neuter-gender conditions, it does require devious argumentation – and only holds when deflection is ongoing in Dutch but not completed. Moreover, there are other complicating factors that could play a role here. For example, as pointed out by Kochari and Flecken (2019), in contexts with

indefinite determiners, adjectives modifying a singular diminutive always carry \emptyset -inflection, even if the original noun is of the common-gender type. So, in all, although my data and the discussion above reveal that specific features and tendencies of Dutch morphology may have had a profound influence on the processing signature of lexical predictions in the current and prior studies, it is not possible to provide a detailed picture of how they exerted their influence exactly.

5 Conclusion

The two eye-tracking experiments presented in this contribution revealed a complex pattern of results. Several explanations of the data were considered and connected to different viewpoints on the architecture of the language faculty, as well as to ongoing debates on the time course and processing costs of (linguistic) predictions. Perhaps the most parsimonious explanation of the data is that the participants did not engage in all-or-none lexical prediction due to the nature of the reading task, allowing a higher reading pace than the tasks of prior studies. However, it also became clear that we cannot rule out all-or-none lexical prediction as a genuine phenomenon in reading, even in the current study. Given the speculative nature of some of the explanations discussed throughout this contribution, I refrain from committing to one interpretation and merely state that the data is inconclusive. Hence, the results speak neither against nor in favor of (resource-consuming) all-or-none prediction and, similarly, neither against nor in favor of (non-taxing) graded prediction. Having said that, the results of the current eye-tracking study are highly relevant because, above all, they clearly diverge from the results obtained in prior ERP and behavioral studies – using identical or very similar manipulations and items (Otten and van Berkum 2008, 2009; Otten et al. 2007; van Berkum et al. 2005; see Koehne et al. in this issue for another example of conflicting evidence between eye-tracking and ERP experiments). Moreover, the results illustrate that syntactic gender may modulate the time course of lexical predictions, an issue that has not been examined in great detail before (Kochari and Flecken 2019).

Obviously, I do not wish to claim that the current findings completely undermine prior conclusions. The point being made here is that the present findings force us to see the results of previous studies with different eyes and novel research is required to fully explain the discrepancies between studies. In my opinion, a reasonable next step for future research on lexical prediction is to co-register the eye-movements and ERPs of readers in a single set-up. Although such co-registration research will come with challenges of its own, it also presents clear opportunities for a more profound understanding of how we should synthesize the

results of eye-tracking and ERP studies (Dimigen et al. 2011; Kliegl et al. 2012). As such, it will deepen our understanding of the early stages of linguistic prediction and, in addition, it may solve the puzzles raised by the current study on the side.

Acknowledgments: Research for this article was in part supported by an NWO (Netherlands Organization for Scientific Research) VENI grant (grant number 275-89-012) awarded to the author. I am grateful to the participants, my colleagues in Utrecht and Leiden, and two anonymous reviewers for their involvement in this research. Special thanks go to Jos van Berkum, who inspired me to conduct the experiments presented here.

References

- Baayen, R. Harald, J. Davidson Douglas & M. Bates Douglas. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59(4). 390–412.
- Barr, Dale J., Roger Levy, Christoph Scheepers & Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68(3). 255–278.
- Barton, Jason J. S., Hashim M. Hanif, Laura Eklinder Björnström & Charlotte Hills. 2014. The word-length effect in reading: A review. *Cognitive Neuropsychology* 31(5/6). 378–412.
- Bennis, Hans. 2010. A dynamic perspective on inflection. In C. Jan-Wouter Zwart & Mark de Vries (eds.), *Structure preserved—studies in syntax for Jan Koster*, 49–56. Amsterdam & Philadelphia: John Benjamins.
- Bennis, Hans & Frans Hinskens. 2014. Goed of fout. *Nederlandse Taalkunde* 19(2). 131–184.
- Blom, Elma, Daniela Polišenská & Fred Weerman. 2008. Articles, adjectives and age of onset: The acquisition of Dutch grammatical gender. *Second Language Research* 24(3). 297–331.
- Brouwer, Harm, Hartmut Fitz & John Hoeks. 2012. Getting real about semantic illusions: Rethinking the functional role of the p600 in language comprehension. *Brain Research* 1446. 127–143.
- Calvo, Manuel G. 2001. Working memory and inferences: Evidence from eye fixations during reading. *Memory* 9(4–6). 365–381.
- Clark, Andy. 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* 36(3). 181–204.
- Clifton, Charles, Adrian Staub, Keith Rayner, Roger P. G. van Gompel, H. Martin, Wayne S. Fischer & Murray. 2007. Eye movements in reading words and sentences. In Robin L. Hill (ed.), *Eye movements: A window on mind and brain*, 341–372. Amsterdam: Elsevier.
- DeLong, Katherine A., Melissa Troyer & Marta Kutas. 2014. Pre-processing in sentence comprehension: Sensitivity to likely upcoming meaning and structure. *Language and Linguistics Compass* 8(12). 631–645.
- DeLong, Katherine A., Thomas P. Urbach & Marta Kutas. 2005. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience* 8(8). 1117–1121.

- Dikker, Suzanne & Liina Pylkkanen. 2013. Predicting language: MEG Evidence for lexical preactivation. *Brain and Language* 12(1). 55–64.
- Dimigen, Olaf, Werner Sommer, Annette Hohlfeld, Arthur M. Jacobs & Reinhold Kliegl. 2011. Coregistration of eye movements and EEG in natural reading: Analyses and review. *Journal of Experimental Psychology: General* 140(4). 552–572.
- Ehrlich, Susan F. & Rayner Keith. 1981. Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior* 20(6). 641–655.
- Estevez, Adelina & Manuel G. Calvo. 2000. Working memory capacity and time course of predictive inferences. *Memory* 8(1). 51–61.
- Federmeier, Kara D. 2007. Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology* 44(4). 491–505.
- Ferreira, Fernanda & Nikole D. Patson. 2007. The good enough approach to language comprehension. *Language and Linguistics Compass* 1(1/2). 71–83.
- Frazier, Lyn & Keith Rayner. 1982. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology* 14(2). 178–210.
- Friederici, Angela D. & Sonja A. Kotz. 2003. The brain basis of syntactic processes: Functional imaging and lesion studies. *Neuroimage* 20. s8–s17.
- Friston, Karl. 2010. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience* 11(2). 127–138.
- George, Marie St, Suzanne Mannes, James E. Hoffman. 1997. Individual differences in inference generation: An ERP analysis. *Journal of Cognitive Neuroscience* 9(6). 776–787.
- Huettig, Falk. 2015. Four central questions about prediction in language processing. *Brain Research* 1626. 118–135.
- Isberner, Maj-Britt & Tobias Richter. 2014. Comprehension and validation: separable stages of information processing? A case for epistemic monitoring in language comprehension. In David N. Rapp & Jason L. G. Braasch (eds.), *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences*, 245–276. Cambridge, MA: MIT Press.
- Jackendoff, Ray. 2002. *Foundations of language: Brain, meaning, grammar, evolution*. Oxford: Oxford University Press.
- Kamide, Yuki, Gerry T. M. Altmann & Sarah L. Haywood. 2003. The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language* 49(1). 133–156.
- Karimi, Hossein & Fernanda Ferreira. 2016. Good-enough linguistic representations and online cognitive equilibrium in language processing. *The Quarterly Journal of Experimental Psychology* 69(5). 1013–1040.
- Kliegl, Reinhold, Michael Dambacher, Olaf Dimigen, Arthur M. Jacobs & Werner Sommer. 2012. Eye movements and brain electric potentials during reading. *Psychological Research* 76(2). 145–158.
- Kochari, Arnold R. & Monique Flecken. 2019. Lexical prediction in language comprehension: A replication study of grammatical gender effects in Dutch. *Language, Cognition and Neuroscience* 34(2). 239–253.
- Koornneef, Arnout. 2008. *Eye-catching anaphora*. Utrecht: Netherlands Graduate School of Linguistics.

- Koornneef, Arnout, Astrid Kraal & Marleen Danel. 2019. Beginning readers might benefit from digital texts presented in a sentence-by-sentence fashion. But why? *Computers in Human Behavior* 92. 328–343.
- Kuperberg, Gina R. 2007. Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research* 1146. 23–49.
- Kutas, Marta, Katherine A. DeLong & Nathaniel J. Smith. 2011. A look around at what lies ahead: Prediction and predictability in language processing. In Moshe Bar (ed.), *Predictions in the brain: Using our past to generate a future*, 190–207. Oxford: Oxford University Press.
- Kutas, Marta & Steven A. Hillyard. 1980. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science* 207. 203–205.
- Kutas, Marta & Steven A. Hillyard. 1984. Brain potentials during reading reflect word expectancy and semantic association. *Nature* 307. 161–163.
- Lau, Ellen, Stroud Clare, Silke Plesch & Colin Phillips. 2006. The role of structural prediction in rapid syntactic analysis. *Brain and Language* 98(1). 74–88.
- Levinson, Stephen C. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT Press.
- Long, Debra L. & Logan De Ley. 2000. Implicit causality and discourse focus: The interaction of text and reader characteristics in pronoun resolution. *Journal of Memory and Language* 42(4). 545–570.
- Luke, Steven G. & Kiel Christianson. 2016. Limits on lexical prediction during reading. *Cognitive Psychology* 88. 22–60.
- MacDonald, Maryellen C. & Brian MacWhinney. 1990. Measuring inhibition and facilitation from pronouns. *Journal of Memory and Language* 29(4). 469–492.
- Malsburg, Titus von der & Shravan Vasishth. 2013. Scanpaths reveal syntactic underspecification and reanalysis strategies. *Language & Cognitive Processes* 28(10). 1545–1578.
- Mitchell, Don C. 2004. On-Line methods in language processing: Introduction and historical review. In Manuel Carreiras & Charles Clifton (eds.), *The on-line study of sentence comprehension: Eyetracking, ERPs and beyond*, 15–32. New York & Hove: Psychology Press.
- Morris, Robin K. 1994. Lexical and message-level sentence context effects on fixation times in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20(1). 92–103.
- Morris, Robin K. 2006. Lexical processing and sentence context effects. In Matthew Traxler & Gernsbacher Morton (eds.), *Handbook of psycholinguistics*, 2nd edn., 377–401. Oxford: Academic Press.
- Nieuwland, Mante, Stephen Politzer-Ahles, Evelien Heyselaar, Katrien Segaert, Emily Darley, Nina Kazanina, Sarah von Grebmer zu Wolfsturn, Federica Bartolozzi, Vita Kogan, Aine Ito, Diane Mézière, Dale J. Barr, Guillaume Rousselet, Heather J. Ferguson, Simon Busch-Moreno, Xiao Fu, Jyrki Tuomainen, Eugenia Kulakova, E. Matthew Husband, David I. Donaldson, Zdenko Kohút, Shirley-Ann Rueschemeyer & Huettig Falk. 2018. Limits on prediction in language comprehension: A multi-lab failure to replicate evidence for probabilistic pre-activation of phonology. *eLife* 7. e33468.
- O'Brien, Edward J. & Anne E. Cook. 2016. Coherence threshold and the continuity of processing: The RI-Val model of comprehension. *Discourse Processes* 53(5/6). 326–338.
- Otten, Marte & Jos J. A. van Berkum. 2008. Discourse-based word anticipation during language processing: Prediction or priming? *Discourse Processes* 45(6). 464–496.
- Otten, Marte & Jos J. A. van Berkum. 2009. Does working memory capacity affect the ability to predict upcoming words in discourse? *Brain Research* 1291. 92–101.

- Otten, Marte, Mante S. Nieuwland & Jos J. A. van Berkum. 2007. Great expectations: Specific lexical anticipation influences the processing of spoken language. *BMC Neuroscience* 8. <https://doi.org/10.1186/1471-2202-8-89>.
- Pickering, Martin J. & Simon Garrod. 2007. Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences* 11(3). 105–110.
- Rayner, Keith, Timothy J. Slattery, Denis Drieghe & Simon P. Liversedge. 2011. Eye movements and word skipping during reading: effects of word length and predictability. *Journal of Experimental Psychology: Human Perception and Performance* 37(2). 514–528.
- Reuland, Eric. 2001. Primitives of binding. *Linguistic Inquiry* 32(3). 439–392.
- Reuland, Eric. 2011. *Anaphora and language design*. Cambridge, MA: MIT Press.
- Smith, Nathaniel J. & Roger Levy. 2008. Optimal processing times in reading: A formal model and empirical investigation. *Proceedings of the Annual Meeting of the Cognitive Science Society* 30. 595–600.
- Stewart, Andrew J., Judith Holler & Evan Kidd. 2007. Shallow processing of ambiguous pronouns: Evidence for delay. *The Quarterly Journal of Experimental Psychology* 60(12). 1680–1696.
- van Berkum, Jos J. A., Colin M. Brown, Pienie Zwitserlood, Valesca Kooijman & Hagoort Peter. 2005. Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31(3). 443–467.
- van Herten, Marieke, Dorothee J. Chwilla & Herman H. J. Kolk. 2006. When heuristics clash with parsing routines: ERP evidence for conflict monitoring in sentence perception. *Journal of Cognitive Neuroscience* 18(7). 1181–1197.
- Wicha, Nicole Y. Y., Elizabeth A. Bates, Eva M. Moreno & Marta Kutas. 2003a. Potato not pope: Human brain potentials to gender expectation and agreement in Spanish spoken sentences. *Neuroscience Letters* 346(3). 165–168.
- Wicha, Nicole Y. Y., Eva M. Moreno & Marta Kutas. 2003b. 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.
- Wicha, Nicole Y. Y., Eva M. Moreno & Marta Kutas. 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.
- Wlotko, Edward W. & Kara D. Federmeier. 2015. Time for prediction? The effect of presentation rate on predictive sentence comprehension during word-by-word reading. *Cortex* 68. 20–32.