Great expectations:
Evidence for graded prediction of grammatical gender

Katja Haeuser (katja.haeuser@mail.mcgill.ca)
Saarland University
Department of Psychology
66123 Saarbrücken
Germany

Jutta Kray (j.kray@mx.uni-saarland.de)
Saarland University
Department of Psychology
66123 Saarbrücken
Germany

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

Abstract

Language processing is predictive in nature. But how do people balance multiple competing options as they predict upcoming meanings? Here, we investigated whether readers generate graded predictions about grammatical gender of nouns. Sentence contexts were manipulated so that they strongly biased people's expectations towards two or more nouns that had the same grammatical gender (single bias condition), or they biased multiple genders from different grammatical classes (multiple bias condition). Our expectation was that unexpected articles should lead to elevated reading times (RTs) in the single-bias condition when probabilistic expectations towards a particular gender are violated. Indeed, the results showed greater sensitivity among language users towards unexpected articles in the single-bias condition, however, RTs on unexpected gender-marked articles were facilitated, and not slowed. Our data confirm that difficulty in sentence processing is modulated by uncertainty about predicted information, and suggest that readers make graded predictions about grammatical gender.

Keywords: language processing, prediction, gender, reading, sentence comprehension

Introduction

A wealth of psycholinguistic evidence suggests that prediction, the pre-activation of semantic and form-level information, is a core component of language comprehension (Federmeier, 2007; Huettig & Mani, 2016; Kuperberg & Jaeger, 2016; Pickering & Gambi, 2018). In this paper, we investigate the pre-activation of grammatical gender of nouns, and whether this pre-activation is all-or-none or graded in nature.

In a 1999 seminal paper, Altmann and Kamide showed that, when people are presented with an auditory sentence such as "The boy will eat the cake", they start fixating edible target objects in an array of pictures well before the noun "cake" is acoustically realized. Hence, language users can leverage their linguistic and world knowledge to actively anticipate upcoming words and their meanings, rather than passively taking in new information as it incrementally occurs (as traditional linguistics frequently posits; Jackendoff, 2002).

Graded Predictions in Language Processing

In addition, research has shown that the semantic predictions people generate during language processing are graded and probabilistic, rather than all-or-none. For example, Federmeier and Kutas (1999) used ERPs to investigate whether readers not only show facilitation for highly expected nouns, but also for nouns that are unexpected but share semantic features with the most expected completion (e.g., “palms” – “pines”). In that study, sentence frames ("They wanted to make the hotel look more like a tropical resort, so along the driveway they planted rows of ...") were presented either with the most highly expected ending (palms), an unexpected noun from the same semantic category (pines), or an unexpected noun from a different semantic category (tulips). Crucially, the N400, an ERP component that indexes ease of semantic integration (Kutas & Hillyard, 1984), showed a graded pattern: It was largest for different-category nouns (tulips), somewhat smaller for same-category nouns (pines), and smallest for expected nouns (palms). Hence, there was facilitation for nouns that were unexpected but shared semantic features with expected nouns. The authors argued that readers generate strong predictions about upcoming words,

including their semantic features. When a word is encountered that is unexpected, but nevertheless has semantic overlap with the most highly expected noun, this word will be facilitated. In sum, it is the degree of semantic feature match between a word and the prediction derived from context that determine how difficult it will be to process new input: Predictions about upcoming words are not all-or-none but graded.

Subsequent studies have shown that graded predictions are not limited to semantic features of words. DeLong, Urbach, and Kutas (2005), for example, demonstrated that expectancy violations are measurable on indefinite articles that precede a critical noun and bear little or no semantic content. The authors presented sentences on a screen word by word that either ended with the most expected noun, e.g., "It was a windy day so the boy went outside to fly A kite", or with an unexpected noun that required a different indefinite article, e.g., "AN airplane". Crucially, at the time of reading the indefinite article participants could already infer whether the predictable word would complete the sentence or not. The results indeed showed that N400 amplitudes at the level of the indefinite article were correlated with its degree of predictability (its cloze probability, based on prior ratings): The higher the cloze probability of the indefinite article, the smaller the N400. The authors argued that readers not only predict semantic features based on prior context, but also phonological features of words. Crucially, this pattern of expectations is graded, and not all-or-none.

However, a large-scale replication attempt that spanned nine labs in the UK (Nieuwland *et al.*, 2018) failed to reproduce the findings of DeLong *et al.* (2005). In that study, there was no critical correlation between N400 amplitudes and the cloze probability of the indefinite article. One of the conclusions of that study was that the a/an manipulation might not be well suited to measure prediction, because the indefinite article only needs to align with the phonological characteristics of the upcoming word – which does not necessarily have to be a noun (e.g., "an ENORMOUS kite").

Prediction of Grammatical Gender

Another way to measure predictability effects before the critical noun is to take into account gender-marked definite articles. Whereas English does not have a gender-marking system, nouns in many other languages such as German, Dutch, Italian, or Spanish require gender-marking in their preceding definite articles (e.g. in German ***die*** $_{fem}$ ***Sonne***, ***der*** $_{masc}$ ***Mond*** [the sun, the moon]). Indeed, gender manipulations might be better suited to measure prediction, because a gender-marked article always needs to align with its head noun, irrespective of whether the noun is preceded by modifiers (e.g., ***die*** *hell scheinende* ***Sonne***, ***der*** *weiße* ***Mond*** [the bright shining sun, the white moon]).

Many studies have observed that people use article gender information to infer whether or not a sentence will continue with the anticipated noun (for review, see Kochari & Flecken, 2019). For example, an ERP study by Wicha and colleagues identified processing difficulties when readers encountered Spanish definite articles whose gender did not agree with the gender of the expected noun (Wicha, Moreno, & Kutas, 2003). In that study, people were presented with sentences such as "El sabía que cuando su padre muriera podría al fin ponerse" ("He knew that when his father died he would finally be able to wear ..."), that strongly biased their expectation towards a particular gender-marked noun (in this case, "corona", which requires the feminine article "la"). The authors found an enhanced N400-like response for unexpected articles (in this case, "el"), indicating that readers predicted the gender of the upcoming noun and noticed the mismatch between the expected noun and preceding article. Hence, language users make use of context to actively predict word-level information of nouns, including their gender.

But are prediction-inconsistent effects of gender also probabilistic/graded? Prior studies do not speak to this issue, because they normally presented gender-marked articles that were either globally predictable or globally unpredictable. This question is crucial because, in the light of the failed replication attempt discussed earlier, graded effects of predictability that are not semantic in nature have not been previously demonstrated.

The Present Study

The question we ask in this paper is whether prediction-inconsistent effects at the article are modulated by the strength of the gender bias that the previous sentence context creates. For example, in some cases a sentence context might create a strong bias towards a particular gender-marked article because it biases nouns that have the same grammatical gender (e.g., because both nouns require a masculine article; henceforth "single bias condition"). In other cases, when a sentence context biases nouns that do not have the same gender, there might be a less consistent, weaker bias towards multiple gender-marked articles ("multiple bias condition"). Here, we hypothesized that a prediction-inconsistent article should lead to greater processing disruptions in the single-bias condition than in the multiple bias condition. This should be the case because the single-bias condition elicits a strong probabilistic expectation towards one particular gender-marked article, which is violated when a prediction-inconsistent article is presented. Here, we use a self-paced reading task to investigate this hypothesis with definite articles in German, a language that has three gender classes (masculine, feminine, and neuter), which all require different definite articles.

Method

Participants

Eighty-four psychology students (55 female, 29 male) between the ages of 18 and 35 (mean age=21 years, SD=3) participated for course credit. All participants were native

speakers of German, had normal or corrected-to-normal vision, and reported no neuropsychological disorders.

Materials

Cloze ratings for an initial set of 73 items were obtained by 40 Saarland University students who did not participate in the main experiment. Participants were presented with sentence frames that were truncated before the definite article and asked to generate a definite article and noun that best completed the sentences. Participants were additionally asked to generate a second-best sentence completion (article and noun) that indicated how the sentence could be completed otherwise.

Based on the ratings, we computed the cloze probabilities for completions that were produced with the highest frequency in the first- and second-best guess ratings (for articles and nouns, separately).

In a series of successive steps, we then excluded items that did not fit. In a first step, we excluded items that did not elicit definite articles (which limited the item set down to 60). We then excluded items whose highest-cloze article and highest-cloze noun did not agree in gender, whose first- and second-best guess noun were identical (e.g., bus-bus), or whose first- and second-best guess nouns were synonyms or near-synonyms (e.g., stove-stove top, goal-goal line, alps-mountains). This selection procedure reduced item set to 35. Finally, three more items were excluded (two items whose highest-cloze article was indefinite; another item whose noun cloze probability was below 0.25). The final set of experimental stimuli consisted of 32 items. First-guess articles and nouns had an average cloze probability of 0.79 and 0.76, respectively (range: 0.43-1.0 and 0.30-1.0). Second-guess articles and nouns had an average cloze probability of 0.53 and 0.29, respectively (range: 0.32-0.81 and 0.14-0.78). Items were then sorted into two categories: items whose first- and second-best guess articles were identical in gender ("single bias"; e.g., "In der Nachmittagshitze war der Wein warm geworden, also stellte Johanna ihn in" [In the heat of the afternoon the wine had become warm, and so Johanna put it in] ***den*** masc ***Kühlschrank*** (*the fridge*) / ***den*** masc ***Schatten*** (*the shade*)), and items whose first- and second-best guess articles were NOT identical ("multiple bias"; e.g., "Nach dem Überfall auf das Schiff vergruben die Piraten" [After the robbery of the ship, the pirates buried ...] ***den*** masc ***Schatz*** (the treasure) / ***die*** fem ***Beute*** (the booty)). Hence, items with gender-identical ratings created a strong bias towards one particular gender-marked article, whereas items with non-identical ratings created a weaker bias towards multiple gender-marked articles.

In a final step, the experimenters chose low-cloze, unexpected article-noun continuations for each one of the 32 sentence stems, making sure that a) the low-cloze nouns had a different grammatical gender than the first-guess highly expected noun, b) the low-cloze nouns were never produced as first- or second-guess completions in the cloze ratings, and c) the unexpected completions matched with the expected ones in frequency (based on the Zipf scale from the SUBTLEX DE data base, Brysbaert *et al.*, 2011). For example, the low-cloze, unexpected completions for the above two items were **die** fem **Badewanne** (the bathtub) and ***das*** neut ***Gold*** (the gold), respectively. The average cloze probability of unexpected definite articles was 0.02 (range: 0.00 – 0.1). The average cloze probability of the expected nouns was < 0.001.

Finally, to account for spill-over from the definite article onto subsequent words of the sentence during reading, the experimenters inserted adjectives and adverbs between the definite article and the noun (e.g., **die kalte, gut gefüllte Badewanne**; *the cold, well-filled fridge/bathtub*). In order to make sure that the critical nouns were never encountered in sentence-final position, the sentence stems were padded with identical words that continued both expected and unexpected endings plausibly. In the final item set, 30 items were presented in the accusative case, requiring the definite articles *den masc*, *die fem* and *das neut*, two items were presented in the dative case, requiring the definite articles *dem masc/neut* and *der fem*. Table 1 shows that the final distribution of article types over high-cloze and low-cloze sentences was roughly balanced.

Table 1: Definite articles in the experiment.

	accusative case			dative case	
	das	den	die	der	dem
High cloze	10	11	9	0	2
Low cloze	6	11	13	2	0

Finally, the 64 sentences (32 high cloze, 32 low cloze) were distributed on two experimental lists so that each subject viewed only one experimental version of each item during testing. We included 38 sentences from the Potsdam sentence corpus as fillers (yielding 70 total sentences per list) to ensure that participants continued to make predictions during reading despite the large number of unexpected sentence continuations, (Fine, Jaeger, Farmer, & Quian, 2013; Jaeger & Snider, 2013). Comprehension questions (simple yes/no questions) were created for 25% of all sentences to make sure that participants read for content. Experimental items and fillers were randomly distributed on each list, with two constraints: 1) No more than four unexpected items in a row, and 2) no more than three items with comprehension questions in a row.

Procedure

We used the "moving-window" format for self-paced reading. Participants read sentences on a screen word-by-word. Each trial started with the presentation of the first word of the sentence, next to a number of underscores, separated by spaces, indicating the number of words to follow. By pushing the space bar with their dominant hand, participants proceeded to the next word, and the letters of the previous word were replaced with underscores. Participants were instructed to read the sentences as quickly

and accurately as possible, and to answer all true/false comprehension questions by pushing the "J" (Yes, correct) and "N" (No, incorrect) bars on the keyboard. Trials were separated by a 500 ms fixation cross.

Results

Accuracy on the comprehension questions was near ceiling for both high- and low-cloze sentences (M = 0.94 for both, range: 0.81-1.00), suggesting that participants were attentive during the experiment and understood the sentences they were reading.

We identified three dependent variables. These included residual (i.e. length-corrected) reading times (RTs) of the article (Gibson & Levy, 2016), the spill-over region (consisting of the three words after the article before the noun), and the noun. Independent variables were bias (a two-level factor: multiple bias vs single bias), as well as cloze probability of the article (for article and spill-over RTs) and cloze probability of the noun (for noun RTs).

Prior to analysis, raw RTs for all DVs were trimmed for outliers by identifying RTs that were more than 3sd deviations away from the participant/condition mean, and replacing those RTs with the respective cut-off value of the participant/condition mean. This affected less than 2% of all data points.[1]

We constructed separate linear mixed effects models for each of the three dependent variables as implemented in the lme4 library (Bates, Maechler, Bolker, & Walker, 2013; version 1.1-19) in R (R Development Core Team, 2016; version 3.5.2). Independent variables were entered into the models including their interaction terms; the factor bias was dummy-coded (0, 1), with "multiple bias" as the reference category. All models contained random intercepts for subjects and items and were initially fit with the maximal random slope structure warranted by the design (i.e. by-subject random slopes for cloze and bias, including their interaction). In the case of non-converging models, the random structure of each model was simplified using the least-variance approach (see Barr, Levy, Scheepers & Tily, 2013, for guidelines). Trial number was added as a scaled control predictor to all models. P-values were estimated using the Satterthwaite degrees of freedom method. Means and standard deviations of raw reading times for articles, spill over regions, and nouns are displayed in Table 2.

Article RTs The model for residual RTs of the definite article was fit with an additional control variable for case marking. We found a significant interaction between bias and article cloze probability (b = -10.79, SE = 4.93, t = -2.19, p = .029). Figure 1 shows that, in contrast to what we predicted, RTs for relatively less expected, low-cloze articles were faster when there was a strong bias towards a single gender-marked article. Vice versa, RTs were slower when there was a weak bias towards multiple gender-marked articles.

Table 2: Raw reading times (in ms) in the experiment, split by bias and cloze probability (high vs low cloze, HC and LC). Standard deviations in parentheses.

	article		spill-over		noun	
	HC	LC	HC	LC	HC	LC
Multiple Bias	396 (153)	404 (189)	1275 (472)	1260 (474)	465 (249)	514 (348)
Single Bias	383 (163)	382 (153)	1256 (482)	1271 (459)	453 (244)	560 (441)

Spill-over RTs Except for a main effect of trial number, which suggested faster reading times for later trials (b = -160.45, SE = 5.264, t = -30.49, p < .001), no critical main effects or interactions were significant in the model for residual RTs of the spill-over region.

Noun RTs The model for noun residual RTs was run with a scaled control variable for noun frequency. This model showed a simple effect of cloze probability (b = -26.29, SE = 6.90, t = -3.80, p < .001), suggesting that, irrespective of bias, relatively expected, high-cloze, nouns were read more quickly than relatively unexpected, low-cloze, nouns (see Figure 2).[2]

Discussion

Nearly half of the languages in the world use gender marking to separate nouns into grammatical classes. Whenever a noun is preceded by a definite article, the article must agree with the gender of the head noun. Prior studies have shown that people use gender-marking on articles as cues that can foreshadow whether or not a noun that is predictable based on context will continue the sentence or not. However, up to date it has been unknown whether people's predictions go beyond noticing that a gender-marked article is globally predictable or unpredictable, or whether language users show graded sensitivity to the statistics of the expected input.

[1] Note that we replicated all effects reported below when using untrimmed, raw RT values.

[2] Note that we addressed the possibility of a change in predictive processing over time (e.g., a de-sensitization to violated predictions as the experiment progressed) by running additional models for RTs on the article, the spill-over region, and the noun that included the three-way interaction between cloze, bias, and trial number. In no model, the three-way interaction reached statistical significance (all p's > .1).

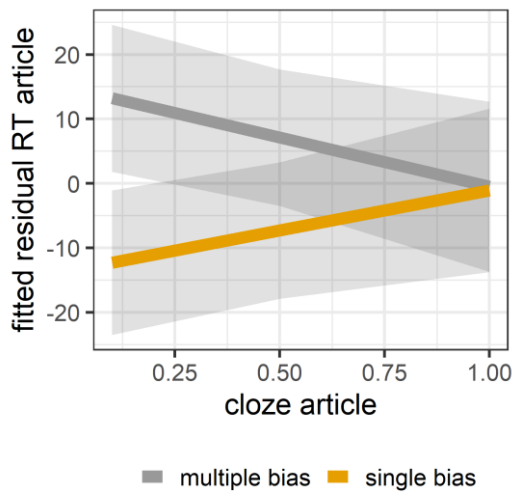

Figure 1: Partial effects plot of the estimated effects of bias and article cloze probability on residual reading times of the article. Error bands represent 95% confidence intervals.

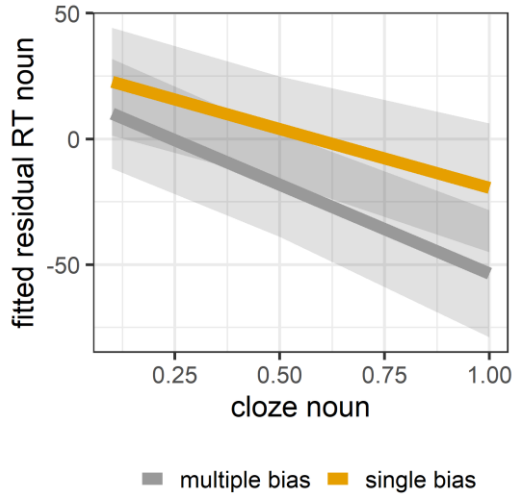

Figure 2: Partial effects plot of the estimated effects of bias and noun cloze probability on residual reading times of the noun. Error bands represent 95% confidence intervals.

Here we used self-paced reading to investigate whether pre-activation of gender information in German is probabilistic, by manipulating the gender bias of preceding context. For example, sometimes a sentence context might cue a strong expectation towards a particular gender-marked article by biasing two or more noun completions that agree in gender class. In other cases, there may be no such strong expectation towards a particular gender-marked article, because the nouns that are predictable based on context have a different grammatical gender.

Our goal here was to investigate whether prediction-inconsistent gender-marked articles lead to greater processing disruptions in such "single-bias" conditions, in other words, in situations when there is a strong probabilistic bias towards a particular gender-marked article. We used single-bias sentence contexts (*Nachdem Paul seinen Führerschein erhalten hatte, fuhr er ständig mit* ***dem*** $_{masc}$ ***Auto*** */* ***dem*** $_{masc}$ ***Motorrad*** [When Paul got his drivers‘ license, he was constantly driving around with the car / the motorbike]) and multiple-bias sentence contexts (*Da Anne Angst vor Spinnen hat, geht sie bei sich zuhause nur ungern nach unten in* ***den*** $_{masc}$ ***Keller / die*** $_{fem}$ ***Garage*** [Anne is scared of spiders, so when she’s at home, she doesn’t like going down into the basement / the garage]), and presented unexpected article-noun combinations (e.g., ***der*** $_{fem}$ ***Gruppe*** [the group (of friends)], ***das*** $_{neut}$ ***Schlafzimmer*** [the bedroom], respectively). We measured reading times at the level of the prediction-inconsistent definite article, the spill-over region (adjectives and adverbs between article and noun (e.g., *dem alten aber zuverlässigen Auto* [the old but reliable car]), and the critical noun.

Two key findings emerged. First, in contrast to our hypothesis that prediction-inconsistent articles should be more disruptive in the single-bias condition, we found that reading times for low cloze, unexpected articles were actually faster, and not slower, in the single-bias condition than in the multiple-bias condition. Second, we found that at the level of the noun, expected high-cloze nouns were read faster than unexpected low cloze nouns, an effect that was identical in size for both gender bias conditions. No effects emerged for the spill-over region.

Pre-activation of Gender-marked Articles

Although we predicted readers would be slower when they encountered unexpected article gender in conditions where potential completions aligned in gender (i.e. in the single bias condition), we found the opposite pattern. Instead, people were faster when confronted with an unexpected gender-marked article in the single vs. multiple bias condition. Hence, people did show sensitivity towards the expectancy manipulation, albeit somewhat differently from what we expected. One possibility is that people indeed processed the article as prediction-inconsistent, but instead of slowing down, they sped up to get more quickly to the head noun of the phrase. Under conditions of normal reading when the full sentence is visible at all times, it is conceivable that participants would have regressed out of the article region towards the noun, in order to quickly make sense of the sentence that was not progressing as expected. In the self-paced reading task we used here, there was no possibility for participants to launch a regressive eye movement into a later portion of the sentence because, at the time of reading the definite article, the noun phrase was hidden by underscores. Hence, it is possible that the speed-up reflects some sort of a compensation for the inability to make a regression – people just read more quickly to get to the noun. Whatever the reason for the speed-up in reading times, our results align with prior studies by showing that people are sensitive to gender information conveyed by definite articles when comprehending sentences. However, above and beyond the findings of earlier studies, our results suggest that pre-activation of gender depends on the bias

created by the previous sentence context: People process prediction-inconsistent gender-marked articles differently when these occur in a context that strongly biases an article from a different grammatical class. Hence, pre-activation of grammatical gender is graded in nature.

Noun Reading Times

In contrast to the definite article, reading times of the noun were not modulated by gender bias. Instead, there was a strong effect of cloze probability, suggesting that unexpected, low-cloze nouns were read considerably more slowly than expected, high-cloze nouns. This slow-down in RTs amounted to 75 ms in raw RTs, and the effect also continued onto subsequent words of the sentence (as additional models showed, not reported here), where the effect gradually washed out. The lack of a gender bias effect at the level of the noun is not too surprising - it merely suggests that processes of semantic (re)analysis and integration overruled at that point of time in processing the sentence, so that gender information was less prevalent.

Conclusions and Future Research

Our data suggest that, not only do comprehenders anticipate the grammatical gender of highly predictable nouns, the nature of their expectations is graded, depending on the gender bias of the prior context. These findings are relevant in the light of previous studies that have investigated pre-activation of grammatical gender using ERPs. For prediction-inconsistent articles, these studies have yielded a puzzling mixture of ERP components, with very little overlap in either the timing, scalp distribution, or even the polarity of the effects (see Wicha *et al.*, 2003; Otten & Van Berkum, 2009; Wicha *et al.*, 2004). Potential differences between these studies that could account for the conflicting findings have been addressed only recently, among them, for example, the stimulus presentation rate or the use of predictable or unpredictable fillers (see Kochari & Flecken, 2019). The data presented here raise the possibility that yet another - previously unaccounted for - factor could help explain the conflicting findings: the fact that people's expectancies about upcoming words and their gender are not all-or-nothing but graded, and contingent on the granularity of the gender bias created by the sentence context.

Let us briefly consider how the present data would look like if they were to support an all-or-nothing account of gender prediction - in other words, the idea that readers do not entertain multiple (gender) predictions during language processing, but predominantly the one with the highest cloze probability. In this case, our data would have shown a simple effect of predictability at the level of the article, with faster reading times for predictable gender-marked articles and slower reading times for unpredictable articles, with no interaction suggesting that predictability effects are contingent on the strength of the gender bias. In the light of these observations, note that the model for article RTs did not show a main effect of cloze. Thus, had we merely considered article cloze probability in the present investigation (and not the additional interaction with bias), we would have been forced to conclude that gender information has no behavioral effect before the critical noun – which is the conclusion drawn in a prior study that investigated article prediction in Spanish-speaking participants and that predominantly looked at article cloze probability (see Guerra, Nicenboim & Helom, 2018).

In sum, our findings support proposals arguing that language comprehension entails the parallel computation of multiple linguistic possibilities, each with some degree of probabilistic support, which are continuously updated once more bottom-up input is available. As such, the cloze probability of a given word not only reflects the one word that completes a sentence best. Instead, it reflects a "degree of belief" (Kuperberg & Jaeger, 2013) of parallel expectations that language users compute based on the environmental contingencies.

Under this account, the gender bias of a given context or sentence fragment could be expressed as the weighted average of all the possible nouns that require a masculine article, a feminine article, and so forth. Crucially, at any point in time the parser computes a probabilistic image of all the possible gender-marked articles rather than committing to one and only one prediction (and their gender).

These observations are in line with information-theoretic accounts arguing that comprehension difficulty in human sentence processing is not only accounted for by suprisal (i.e. the amount of new information a word conveys, see Frank, Otten, Galli, & Vigliocco, 2015), but also by the amount of uncertainty about possible alternative parses (i.e., entropy; Linzen & Jaeger, 2014): Entropy is lower in situations in which there are many possible, low-probability continuations (i.e., multiple bias condition) compared to situations in which there are few, high-probability continuations (i.e., strong bias condition).

In addition, the parser adapts flexibly whenever new input renders earlier predictions inappropriate by narrowing the scope of possible sentence continuations (Delaney-Busch, Morgan, Lau, & Kuperberg, 2019). Future research could address under which conditions such adaptations occur, and what happens when individuals (for example with learning disabilities or communicative disorders) are unable to make such adaptations.

References

Altmann, G. T., & , Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*, 247-264.

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.

Bates, D., Maechler, M., Bolker, B., &Walker, S. (2013). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1-48.

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.

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.

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

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

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PloS One*, *8*(10).

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1-11.

Gibson, E., & Levy, R. (2016). An attempted replication of Hackl, Koster-Hale, Varvoutis (2012). arXiv [Preprint]. doi: 1605.00178.

Guerra, E., Nicenboim, B., & Helo, A. (2018). A crack in the crystal ball: Evidence against pre-activation of gender features in sentence comprehension. Poster presentation, *Architectures and Mechanisms for Language Processing (AMLaP)*, Berlin, Germany, September 06-08.

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

Ito, A., Gambi, C., Pickering, M. J., Fuellenbach, K., & Husband, E. M. (2020). Prediction of phonological and gender information: An event-related potential study in Italian. *Neuropsychologia*, *136*, 107291.

Jackendoff, R. (2002). Foundations of language. New York: Oxford University Press.

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

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*, 161-163.

Linzen, T & Jaeger, F. Investigating the role of entropy in sentence processing. Proceedings of the Fifth Workshop on Cognitive Modeling and Computational Linguistics, ACL, Baltimore, MD (2014), pp. 10-18.

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

Otten, M., & Van Berkum, J. J. (2009). Does working memory capacity affect the ability to predict upcoming words in discourse? Brain Research, 1291, 92-101.

Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. Psychological Bulletin, 144, 1002-1044.

R Development Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

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*, 483-508.

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*, 1272-1288.