On the time it takes to judge grammaticality

Jonathan Mirault, Jonathan Grainger

Original Article

QJEP

Quarterly Journal of Experimental Psychology
1–6

DOI: 10.1177/1747021820913296
qjep.sagepub.com
SAGE

On the time it takes to judge grammaticality

Jonathan Mirault and Jonathan Grainger

Abstract

The presentation duration of five-word sequences was varied and participants were asked to judge their grammaticality. The five-word sequences were presented for a variable duration randomly selected between 50 and 500 ms with 50-ms steps and were immediately followed by a masking stimulus. Half of the sequences were correct sentences which were randomly intermixed with ungrammatical sequences formed of the same words in scrambled order. We measured the proportion of correct responses for each presentation duration in the grammatical and ungrammatical conditions, and calculated sensitivity and bias from these measures. Both the sensitivity measure (*d′*) and the probability correct responses to grammatical and ungrammatical sequences increased as the stimulus duration increased, with a *d′* of 2 and an average percent correct close to 87% for the grammatical sequences already attained at 300 ms. The rate of increase in performance diminished beyond 300 ms. Grammatical decision times were faster and more accurate for the grammatically correct sequences, thus indicating that participants were not responding by detecting illegal word combinations in the ungrammatical sequences. On the basis of these findings, we provide an upper estimate of 300 ms as the time it takes to access reliable syntactic information from five-word sequences in French, and we discuss the implications of this constraint for models of reading.

Keywords

Reading; grammaticality judgements; grammatical decision task; good enough syntax; signal detection theory

Received: 11 July 2019; revised: 4 February 2020; accepted: 23 February 2020

Introduction

Traditionally, linguists have used a form of grammaticality judgement task to investigate the nature of syntactic structures. Often referred to as "acceptability" or "well-formedness" judgements, this paper and pencil task simply requires participants to judge how well a given sequence of words conforms to the grammar of a given language. This technique provided a window on syntax while minimising the contribution of semantics, such as when judging that the sequence of words "colorless green ideas sleep furiously" is a well-formed sentence in English (Chomsky, 1957).

A number of recent studies have used speeded grammaticality judgements as a tool to investigate the mechanisms involved in sentence reading. Simply putting time pressure on participants' responses opens up a wealth of possibilities in terms of data analysis, as certified by the rich and dynamic field of response time (RT) analyses (e.g., Luce, 1986; Ratcliff & Rouder, 1998). Moreover, speeded grammaticality judgements are for word sequences what speeded lexical decisions are for letter sequences, and the lexical decision task is by far the most commonly used technique to investigate word recognition processes in both the visual and auditory modalities (see Ferrand et al., 2018, for a review). This holds the promise of a golden future for the grammatical decision task.[1]

Our prior work focused on one specific phenomenon revealed in the grammatical decision task—transposed-word effects in decisions to ungrammatical sequences (Mirault et al., 2018). We compared performance to two types of ungrammatical word sequences, which were intermixed with grammatically correct sentences for the purposes of the grammatical decision task. The first type of sequence was formed by transposing two adjacent words

Laboratoire de Psychologie Cognitive, Centre National de la Recherche Scientifique, Aix-Marseille Université, Marseille, France

Corresponding author:
Jonathan Mirault, Laboratoire de Psychologie Cognitive, Centre National de la Recherche Scientifique, Aix-Marseille Université, 3 place Victor Hugo, 13331 Marseille, France.
Email: jonathan.mirault@gmail.com

in a grammatically correct sentence (e.g., "the white was cat big"). The second type of ungrammatical sequence was formed using the same words as in the transposed-word sequences (by recombining words from matched pairs of sequences) but could not be resolved into a correct sentence by transposing any two words (e.g., "the white was cat slowly"). Participants found it harder to reject the transposed-word sequences as being ungrammatical. This transposed-word effect in the grammatical decision task is somewhat analogous to the transposed-letter effect seen in the lexical decision task (Andrews, 1996; Bruner & O'Dowd, 1958; Chambers, 1979; Frankish & Turner, 2007; O'Connor & Forster, 1981; Perea et al., 2005).

In this work, we used a data-limited version of the grammatical decision task, as opposed to the response-limited version used in our prior work. Thus, rather than examining RTs to word sequences that remain on screen until response, here we examined accuracy in responding to word sequences that remained on screen for a variable amount of time and were immediately followed by a backward mask. Directly relevant for this study is another data-limited paradigm that has been used in several recent studies. This is the Rapid Parallel Visual Presentation (RPVP) technique whereby a sequence of words is presented for a brief duration (200 ms) and followed by a pattern mask and a positional cue for identification of one word in the sequence (Snell & Grainger, 2017; Wen et al., 2019). In these studies, we once again compared grammatically correct sequences (e.g., *the boy can run*) with ungrammatical scrambled sequences of the same words (e.g., *run boy the can*), and the target was the same word at the same position in both sequences ("boy" in these examples). We found that word identification was more accurate in the grammatical sequences than the ungrammatical sequences—a sentence superiority effect.

The results obtained with the post-cued RPVP paradigm suggest that some form of elementary syntactic representation can be extracted from very briefly presented sequences of words. This suggests that participants in these experiments were processing several words in parallel and associating parts-of-speech to these words to construct an initial "good enough" syntactic structure that could then constrain ongoing word identification processes (Declerck et al., 2019). However, participants in those studies only had to identify one word, and it is possible that benefits in word identification could have been driven by the syntactic compatibility of one or two adjacent words (e.g., determiner-noun vs. noun-determiner) rather than the grammaticality of the entire sequence. In this study, we asked our participants to judge the grammaticality of a complete sequence of words while varying the amount of time that the sequence was displayed for prior to backward masking. We expect that the timing estimates obtained from this manipulation will provide strong constraints on models of sentence reading and syntactic processing, and particularly with respect to the serial versus parallel word processing debate (e.g., Reichle et al., 2009; Snell & Grainger, 2019).

Method

Participants

One hundred participants (75 female) were recruited at Aix-Marseille University, France. They were all native French speakers and received either course credit or monetary compensation (€10/hr). The participants reported normal or correct-to-normal vision and ranged in age from 18 to 26 years ($M=22.07$, $SD=2.77$). They were naïve as to the purpose of the experiment and signed an informed consent form before starting the experiment. Ethics approval was obtained from the *Comité de Protection des Personnes* SUD-EST IV (No. 17/051).

Design and stimuli

We constructed 240 five-word sequences in French (120 grammatically correct and 120 grammatically incorrect) with an average length of 24 letters by sequence. The ungrammatical sequences were formed in a number of different ways such that they could differ from a correct sentence just by one word (e.g., a noun replaced by a verb), by incorrect verb tense, or by more than one word creating the ungrammaticality. The words had an average length of 3.47 letters and an average frequency of 4,317 occurrences per million which is equivalent to 3.63 Zipf (van Heuven et al., 2014). Word frequencies were the film subtitle frequencies (New et al., 2007) in the Lexique2 database (New et al., 2004). The complete set of stimuli is provided in Supplemental Appendix.

Apparatus

Stimuli were displayed using OpenSesame (Mathôt et al., 2012) with each sentence occupying a single line. The sentences were displayed on a 24.5-in. LCD monitor with a refresh rate of 60 Hz and a screen resolution of $1,024 \times 768$ pixels ($54.5 \times 31\ cm^2$). Stimuli were presented in lower case 24-point monospaced font (droid sans mono; the default monospaced font in OpenSesame) and the text was presented in black ($31.80\ cd/m^2$) on a light grey background ($40.32\ cd/m^2$). Participants were seated 86 cm from the monitor, such that every 2.35 characters equalled approximately 1° of visual angle.

Procedure

Each sequence was presented at each of the 10 possible durations from 50 to 500 ms with steps of 50 ms. Presentation duration was randomised with a different

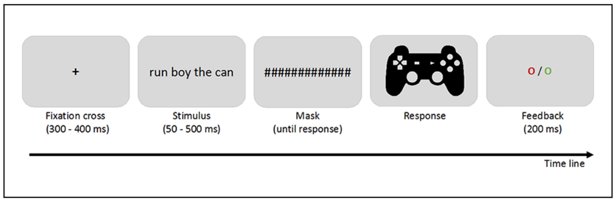

Figure 1. Procedure for one trial.

random order for each participant. On each trial, a fixation cross was presented in the centre of the screen for a random duration between 300 and 400 ms. Then, the word sequence was presented for one of the 10 possible durations and immediately followed by a pattern mask. Participants had to decide if the sequence was grammatically correct or not using an ergonomic gamepad. They had to press with their index fingers the right button if the sequence was judged to be grammatically correct and the left button otherwise. Feedback was then provided during 200 ms in the form of a dot in green for correct responses and red for incorrect responses (see Figure 1). Participants were requested to respond spontaneously and not to dwell upon their response. As is typical with data-limited techniques, the main dependent variable was accuracy, but we also recorded RTs. The accuracy data were used to calculate the sensitivity (d') and bias (c) measures of signal detection theory (Swets et al., 1961) on the basis of the percentage of hits (correct responses to grammatical sequences) and false alarms (incorrect responses to ungrammatical sequences).

Results

The raw data were merged using R (R Core Team, 2017) and we rejected 14 participants due to low average performance (i.e., <50%). All the other participants ($N=86$) performed with average accuracy equal to $M=73.32\%$ ($SD=44.22$). The final dataset consisted of 10,320 data points per condition, which is largely superior to the recommendation of Brysbaert and Stevens (2018). We used *psignifit* (Schütt et al., 2015), a toolbox for Bayesian psychometric estimation for the curve adjustments in MATLAB (version 2018b).

Proportion correct

We first calculated the probability of correct responses for the grammatical and ungrammatical sequences as a function of stimulus duration. The results are shown in Figure 2. We observed a clear increase in performance from 50 to 300 ms, with a clear plateau reached at that point for the grammatical sequences. The average accuracy was 81.00% ($SD=39.22$) for grammatical sequences and 69.32% ($SD=46.11$) for ungrammatical sequences. The complete set of condition means (hits, correct rejections, false alarms, misses) is provided in Table 1.

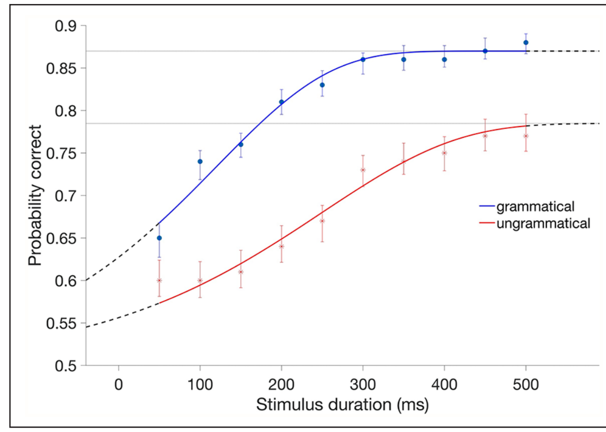

Figure 2. Evolution of the probability of correct responses for the grammatical (in blue) and ungrammatical (in red) conditions at each stimulus duration. Error bars represent standard errors.

Sensitivity and bias

Next, we computed sensitivity (d') and bias (c) from the percentage of hits and false alarms in the grammatical decision task, with grammatically correct sentences being the signal and the incorrect sequences the noise. Figure 3 plots the values of d' and c for each duration. Just like the percent correct to grammatical sequences (Figure 2), there is a steady increase in d' from 50 to 300 ms, and very little increase thereafter. It can also be revealed from Figure 3 that there was a small response bias[2] in favour of grammatical responses, which remained stable across stimulus durations after 100 ms. This can partly explain why response accuracy was greater for grammatical sequences than ungrammatical sequences.

RTs

We computed RTs as the elapsed time between the onset of the stimulus and the response of the participant. These were calculated separately for hits, misses, false alarms, and correct rejections. There were no significant correlations between stimulus duration and RTs for any of these response categories (all $rs < .14$), which was to be expected given the data-limited procedure that was used, and that accuracy of responding was stressed more than speed. Averaged across stimulus duration, the mean RT was 1,256 ms for correct grammatical responses (hits), 1,868 ms for correct ungrammatical responses (correct rejections), 1,375 ms for incorrect grammatical responses (false alarms), and 2,355 ms for incorrect ungrammatical responses (misses).

Table 1. Average percentage responses for each response category (hits, correct rejections, false alarms, misses) at each stimulus duration and the corresponding standard errors.

Durations (ms)	Hits		Correct rejections		False alarms		Misses	
	Percentage	*SE*	Percentage	*SE*	Percentage	*SE*	Percentage	*SE*
50	64.77	0.02	60.25	0.02	39.77	0.02	35.21	0.02
100	73.57	0.01	60.10	0.02	39.91	0.02	26.33	0.01
150	75.90	0.01	61.34	0.02	38.71	0.02	24.02	0.01
200	80.09	0.01	64.29	0.02	35.75	0.02	18.87	0.01
250	83.18	0.01	66.69	0.02	33.37	0.02	16.59	0.01
300	85.55	0.01	72.87	0.01	27.21	0.01	14.26	0.01
350	86.18	0.01	74.32	0.01	25.79	0.01	13.52	0.01
400	86.37	0.01	74.91	0.02	25.22	0.01	13.40	0.01
450	87.29	0.01	77.11	0.01	23.01	0.01	12.45	0.01
500	87.84	0.01	77.38	0.02	22.87	0.02	11.86	0.01

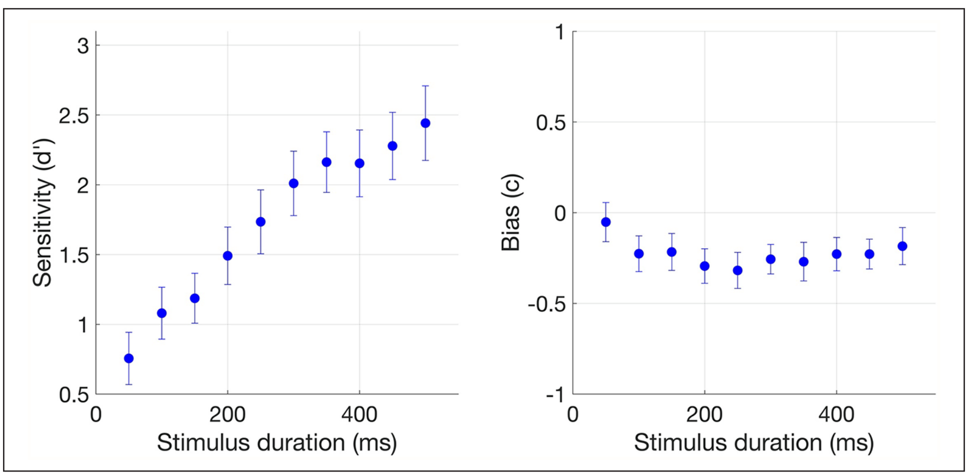

Figure 3. Evolution of sensitivity (d') and bias (c) as a function of stimulus duration. Error bars represent 95% confidence intervals.

Discussion

In this work, we displayed sequences of five words for a variable duration randomly selected between 50 and 500 ms with 50-ms steps and asked participants to judge whether the word sequence formed a grammatically correct sentence or not (grammatical decision task). We calculated percentage correct grammatical and ungrammatical decisions at each duration and also the sensitivity (d') and bias (c) statistics from the percentage hits (correct classification of a sentence) and false alarms (incorrect classification of an ungrammatical sequence as a sentence). We found a steady increase in average percent correct to grammatical sequences and d' from 50 to 300 ms, followed by a reduced increase from 300 to 500 ms. Furthermore, grammatical decision times were faster and more accurate for the grammatically correct sequences, thus indicating that participants were not performing the task by trying to detect illegal word combinations, as that would have led to faster and more accurate responding to ungrammatical sequences. In other words, participants were likely processing the complete sequence of words before responding. There was evidence for a small response bias in favour of grammatical decisions that, crucially, remained relatively stable from the 100-ms duration on.

The results of this study suggest that sufficient syntactic information has been extracted from a five-word sequence presented for 300 ms (and followed by a pattern mask) to quite accurately judge the grammaticality of that sequence. The only means for a strictly serial model to achieve such performance would be to assume that processing of the word sequence continues after removal of the stimulus and presentation of the pattern mask, and the relatively long RTs suggest that this was indeed what participants were doing. In fact, the timing estimate for five-word sequences derived from this work fits remarkably well with the estimated 50–60 ms per word necessary for fluent reading in the disappearing text paradigm (e.g., Liversedge et al., 2004; Rayner et al., 1981, 2003). Moreover, Rayner et al. (1981) also reported that uptake of information from the parafovea benefitted from longer stimulus presentation times (prior to masking), with performance continuing to improve up to the maximum 150 ms that was tested in that study. We would therefore tentatively suggest that information concerning the entire sequence is stored in some form of short-term memory buffer such as visual short-term memory (VSTM), and 300 ms would reflect the time it takes to enter sufficient information from the entire sequence into this buffer to continue processing the individual words and extract syntactic information.

Under the assumption of serial lexical processing, the information stored in this short-term buffer would have to be sublexical in nature—either visual features, letters, or letter combinations. This is because a rapid scan of the word sequence with 50-ms glimpses per word would not appear to be a practical solution. The problem then is that current estimates of the capacity of VSTM are in the order of four different items (e.g., Cowan, 2001), thus suggesting that it is word identities that are stored in this short-term visual buffer rather than sublexical information (Snell et al., 2018). We are currently testing this possibility with a VSTM version of the same–different matching task with word sequences that we have used in prior research (Pegado & Grainger, 2020a, 2020b). In line with our work on grammaticality judgements (Mirault et al., 2018), same–different judgements to sequences of words (400-ms presentation of the reference sequence followed by presentation of the test sequence until response) were harder to make when the sequences differed by a word transposition compared with word substitutions. Crucially, we also found quite large transposed-word effects when the reference was an ungrammatical sequence of words, and a greatly reduced effect when the reference and target sequence were formed of pseudowords. Our prediction is that when the delay between reference and target sequences is increased to 1 s and accompanied with articulatory suppression (to avoid rehearsal of verbal materials), this should not affect performance for word sequences but should greatly impact on performance to pseudoword sequences. This would be in line with our hypothesis that it is sequences of word identities that are stored in VSTM not sequences of letters.

In conclusion, we have shown how a data-limited version of the grammatical decision task can provide important constraints relative to the computations involved in extracting syntactic structures from print. Future research could manipulate the nature of the ungrammatical sequences tested in this task, forcing participants to pay more or less attention to word order, for example, by creating ungrammaticality by word transpositions or by word substitutions.

Acknowledgements

The authors thank Stéphane Dufau for technical and methodological advice and Agnès Guerre-Genton, Fanny Broqua, Loreley Fargère, Cécile Mahnich, and Charlotte Leflaëc for their help in running the experiment.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by grants from the French Agence Nationale de la Recherche (ANR; Nos ANR-11-LABX-0036 and ANR-15-CE33-0002-01) and by a grant from the European Research Council (ERC; No. 742141).

ORCID iD

Jonathan Mirault https://orcid.org/0000-0003-1327-7861

Open practices

The data and materials from the present experiment are publicly available at the Open Science Framework website: https://osf.io/kwu8b/

Supplementary material

The Supplementary Material is available at qjep.sagepub.com

Notes

1. We introduce the term "grammatical decision task" here for the first time, having used "grammaticality judgments" in our previous work. We use this term given the obvious analogy with the highly popular lexical decision task used in word recognition research.
2. A value of –1 indicates a shift of the response criterion of one standard deviation in favour of "signal" (i.e., grammatical) responses.

References

Andrews, S. (1996). Lexical retrieval and selection processes: Effects of transposed-letter confusability. *Journal of Memory and Language*, *35*, 775–800.

Bruner, J. S., & O'Dowd, D. (1958). A note on the informativeness of parts of words. *Language and Speech*, *1*, 98–101.

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, 1, Article 9.

Chambers, S. M. (1979). Letter and order information in lexical access. *Journal of Verbal Learning and Verbal Behavior*, *18*, 225–241.

Chomsky, N. (1957). *Syntactic structures*. De Gruyter Mouton.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87–185.

Declerck, M., Wen, Y., Snell, J., Meade, G., & Grainger, J. (2019). Unified syntax in the bilingual mind. *Psychonomic Bulletin & Review*, *27*, 149–154.

Ferrand, L., Méot, A., Spinelli, E., New, B., Pallier, C., Bonin, P., Dufau, S., Mathôt, S., & Grainger, J. (2018). Megalex: A megastudy of visual and auditory word recognition. *Behavior Research Methods*, *50*, 1285–1307.

Frankish, C., & Turner, E. (2007). SIHGT and SUNOD: The role of orthography and phonology in the perception of transposed letter anagrams. *Journal of Memory and Language*, *56*, 189–211.

Liversedge, S. P., Rayner, K., White, S. J., Vergilino-Perez, D., Findlay, J. M., & Kentridge, R. W. (2004). Eye movements when reading disappearing text: Is there a gap effect in reading? *Vision Research*, *44*, 1013–1024.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford University Press.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*, 314–324.

Mirault, J., Snell, J., & Grainger, J. (2018). You that read wrong again! A transposed-word effect in grammaticality judgments. *Psychological Science*, *29*, 1922–1929.

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, *28*, 661–677.

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, *36*, 516–524.

O'Connor, R. E., & Forster, K. I. (1981). Criterion bias and search sequence bias in word recognition. *Memory & Cognition*, *9*, 78–92.

Pegado, F., & Grainger, J. (2020a). A transposed-word effect in same-different judgments to sequences of words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. https://doi.org/10.1037/xlm0000776

Pegado, F., & Grainger, J. (2020b). Dissociating lexical and sublexical contributions to transposed-word effects. *Acta Psychologica*, *201*, Article 102943.

Perea, M., Rosa, E., & Gómez, C. (2005). The frequency effect for pseudowords in the lexical decision task. *Perception & Psychophysics*, *67*, 301–314.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*, 347–356.

Rayner, K., Inhoff, A. W., Morrison, R. E., Slowiaczek, M. L., & Bertera, J. H. (1981). Masking of foveal and parafoveal vision during eye fixations in reading. *Journal of Experimental Psychology: Human Perception and Performance*, *7*, 167.

Rayner, K., Liversedge, S. P., White, S. J., & Vergilino-Perez, D. (2003). Reading disappearing text: Cognitive control of eye movements. *Psychological Science*, *14*, 385–388.

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Reichle, E. D., Warren, T., & McConnell, K. (2009). Using EZ Reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic Bulletin & Review*, *16*(1), 1–21.

Schütt, H., Harmeling, S., Macke, J., & Wichmann, F. (2015, September). Psignifit 4: Pain-free Bayesian inference for psychometric functions. Paper presented at the 15th Annual Meeting of the Vision Sciences Society (VSS 2015), St. Pete Beach, FL, United States.

Snell, J., & Grainger, J. (2017). The sentence superiority effect revisited. *Cognition*, *168*, 217–221.

Snell, J., & Grainger, J. (2019). Readers are parallel processors. *Trends in Cognitive Sciences*, *23*, 537–546.

Snell, J., van Leipsig, S., Grainger, J., & Meeter, M. (2018). OB1-reader: A model of word recognition and eye movements in text reading. *Psychological Review*, *125*, 969–984.

Swets, J. A., Tanner, W. P., Jr., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review*, *68*, 301–340.

Van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, *67*, 1176–1190.

Wen, Y., Snell, J., & Grainger, J. (2019). Parallel, cascaded, interactive processing of words during sentence reading. *Cognition*, *189*, 221–226.