HeCz: A large scale self-paced reading corpus of newspaper headlines

Jan Chromý (jan.chromy@ff.cuni.cz)
Institute of Czech Language and Theory of Communication, Charles University
nám. Jana Palacha, Prague 1, 116 38 Czechia

Markéta Cehákova (marketa.cehakova@ff.cuni.cz)
Institute of Czech Language and Theory of Communication, Charles University
nám. Jana Palacha, Prague 1, 116 38 Czechia

James Brand (james.brand@ff.cuni.cz)
Institute of Czech Language and Theory of Communication, Charles University
nám. Jana Palacha, Prague 1, 116 38 Czechia

Abstract

Linguistic corpora have been a vital resource for understanding not only how we use language, but also how we process words and sentences. In order to better understand language processing, researchers have recently been creating corpora that integrate both traditional text annotations as well as behavioural measurements collected from human participants. In this paper we introduce the HeCz Corpus, which to our knowledge is the largest such example of a behavioural corpus, containing 1,919 newspaper headlines taken from a Czech language news website. The sample consisted of 1,872 participants, each reading approximately 120 headlines. Each headline was read using a self-paced reading, meaning that every word in the corpus can be analyzed for reading time. After reading each headline, each participant answered a question relating to a specific information contained within the headline, providing a measurement of comprehension. To facilitate better understanding of participant level variation in how the headlines are processed, we collected data on the participant's mood state immediately prior to their participation, along with other basic demographic information. We also collected data from a subset of participants who read the stimuli in the initial testing round, but also completed the same experiment in a second round after a one-month gap, which can provide new insights into how texts are processed and understood when being re-read. In order to highlight the practical uses of the corpus, our analyses focus on how reading times are modulated by i) headline length in words, ii) trial order, and iii) testing round, in addition to examining the role of targeted information location in comprehension accuracy. HeCz thus provides a unique and novel resource that can be used by psycholinguists and cognitive scientists more generally, in order to gain new insights into how real-world language is processed and understood.

Keywords: psycholinguistics; reading; comprehension; reaction times; self-paced reading; sentence processing

Introduction

Linguistic corpora have been crucial to the study of language for nearly six decades (Kučera & Francis, 1967), offering important insights into language use and serving as a tool not only for linguistics, but also for numerous disciplines within cognitive science. Corpora may differ in various aspects, such as their size, represented languages, types of texts, annotation, and modality (cf. McEnery & Hardie, 2011), highlighting their broad and diverse range of uses.

More recently, researchers have begun to create corpora that not only include annotated texts, but also include behavioral measures of how these texts are processed by individuals while reading. Two techniques have been used for these purposes, namely eye-tracking and self-paced reading. Eye-tracking corpora (e.g., Kennedy, Pynte, Murray, & Paul, 2013; Kliegl, Grabner, Rolfs, & Engbert, 2004, etc.) contain data on the eye-movements recorded during reading, whereas self-paced reading corpora (e.g., Futrell et al., 2021) are based on reaction times (RTs) to individual parts of the stimulus (typically words or phrases within a sentence or paragraph). Participants read presented stimuli by button pressing: each button press reveals a certain part of the stimulus (typically a word or a phrase) and simultaneously hides the previous one (see more below). RTs for each button press are recorded and the general idea is that the longer it takes the participant to react, the more demanding is to process the given part of the stimulus (Chromý & Dotlačil, 2022; Jegerski, 2013; Just & Carpenter, 1980).

Importantly, various eye-tracking and self-paced reading corpora have been collected for different languages (cf. Siegelman et al., 2022), types of stimuli (natural or modified texts of various genres, isolated sentences), and additional information regarding both participants and stimuli (such as predictability norms, cf. Luke & Christianson, 2018). Such corpora have the potential to shed light not only on how language is being used in the broader society, but importantly, also on how language is processed and comprehended by its users.

For example, based on the Dundee Corpus, Kennedy et al. (2013) examined parafoveal processing in relation to predictability. Their findings reveal that more predictable parafoveal words produce longer foveal fixations. Luke and Christianson (2016) analyzed the Provo Corpus and found that while reading, predictability of the full word form plays a much smaller role in reading facilitation than general semantic and morphosyntactic predictability. The analysis of the multilingual MECO corpus (Siegelman et al., 2022) showed that readers of different languages vary considerably in their skipping rate (i.e. their tendency not to fixate certain words), but at the same time do not differ in their fixation times. Moreover, processing corpus data may serve as a useful benchmark for the comparison between behavioral and neurological measures, such as in Wehbe et al. (2021) who used the Natural Stories Corpus (Futrell et al., 2021) to compare

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

self-paced reading and fMRI data.

However, these processing-based corpora tend to be relatively small. For instance, the Natural Stories Corpus (Futrell et al., 2021), which is based on self-paced reading, consists of 10,245 word tokens across 485 sentences, read by 181 native speakers of English, resulting in a total of 848,768 observations. Similarly, the self-paced reading part of the UCL Corpus (Frank, Fernandez Monsalve, Thompson, & Vigliocco, 2013) consists of 361 English sentences drawn from novels, read by 117 participants, which provides 274,893 observations. Eye-tracking corpora are even smaller in terms of number of participants/items.

The inherently limited sample size imposes constraints on the breadth of research that can be conducted using these corpora and the range of research questions that can be reasonably investigated. Certain linguistic elements may not even be present in such datasets, including specific words, phrases, or grammatical structures. Despite these inherent limitations, these corpora provide us data to test various hypotheses in relation to general questions about language processing.

In this paper, we present the HeCz corpus which is a large scale self-paced reading corpus of newspaper headlines in Czech. We will first present the design of the corpus and its data structure and then, we will present an initial analysis of the data.

The HeCz corpus

Unlike other existing behavioural corpora, HeCz uses headlines taken from a popular online Czech news website. This was motivated by the fact that such headlines are often read in isolation, meaning that they do not need extra context to be comprehended in detail and should be easily processed in isolation. Moreover, the tokens included within headlines should be varied and have a diverse content, meaning the types of words will range in terms of their frequency and there will be a wide range of grammatical constructions.

The corpus consists of reading times and response accuracy data for each headline. Our primary aim was to collect this data in order to better understand the ways that headlines are read and processed. Additionally, we also aimed to test intra-participant variability, we conducted two rounds of data collection with a one-month interval in between. All participants from the first round were asked to participate again in the second round and they were presented with the identical stimuli presented in the first round (only the trial order was randomized again and therefore different). Additionally, the corpus provides comprehensive demographic data about the participants, including their age, L2, self-declared L2 proficiency, and foreign language exposure. Moreover, participants also completed the Czech adaptation of the Profile of Mood States questionnaire (McNair, Lorr, & Droppleman, 1971; Stuchlíková, Man, & Hagtvet, 2005), assessing the following moods ANGER, CONFUSION, DEPRESSION, FATIGUE, and TENSION. The corpus thus offers a possibility to relate participants' mood with their behavioral measures (RTs and comprehension accuracy).

Materials

The corpus is composed of headlines obtained from Seznamzpravy.cz, a widely read online news portal in the Czech Republic. Initially, we randomly selected 2,500 headlines from the year 2019 to deliberately avoid the pervasive theme of COVID-19, ensuring a more diverse range of topics while also working with the latest available headlines. Subsequently, we manually checked these headlines for their suitability to be included in the corpus. We excluded headlines which did not form a sentence or were generally incoherent (e.g., *Police dog training, tips for the weekend and the mood in society after the revolution*). In the next phase, yes–no comprehension questions were created for each headline. These questions were designed to target various types of information (e.g., subject, object, locative adjunct, etc.) preventing participants from focusing on specific segments of the sentence. However, for certain headlines, formulating a meaningful question proved challenging, leading to the exclusion of those specific headlines from the final stimuli set.

After filtering was completed, our final set contained 1,919 headlines, with 23,634 tokens. The mean length of each headline was 12.3 words (sd = 2.71) and the mean word length was 5.53 characters (sd = 2.8). The headlines were randomly distributed into 16 lists, each comprising either 119 or 120 headlines. A participant was randomly assigned an individual list of headlines to read during the experiment.

Participants

In the first round of data collection, data was obtained from 1,872 native Czech speakers (undergraduates of Charles University who participated for course credit). Out of these, 1,162 participants were tested again in the second round (after approximately one month).

Each participant read the headlines from one list in a randomized order, with an average of 117 participants per list in the first round (range = 108–141 participants) and 72.6 in the second round (range = 61–86 participants).

Procedure

The data collection was web-based using JATOS (Lange, Kühn, & Filevich, 2015) and programmed using JsPsych v 7.0.0 (De Leeuw, 2015). First, participants received a demographic questionnaire, they were then presented with the Czech adaptation of the Profile of Mood States questionnaire (Stuchlíková et al., 2005) with each question being presented on a separate screen. Participants were then presented headlines in a self-paced reading moving-window presentation, i.e. sentences were presented as a series of underscores and each button press uncovered one word of the headline and hid the previous one. After reading each headline they received a comprehension question which they responded to using a yes/no button. See (1) for an example of translated stimulus.

(1) *Headline:* Řidič vyrazil s formulí na dálnici D4. Policie pátrá po něm i po svědcích.

'The driver set off with the formula on the D4 highway. The police are looking for him and for witnesses.'
Question: Pátrá stále policie po řidiči? / 'Is the police still looking for the driver?'

Data overview and current state

The self-paced reading data contain 4,840,075 observations (i.e. RTs) with 2,986,615 observations from the first round of data collection and 1,853,460 from the second round. Thus, the data set is by far the most extensive self-paced reading corpus to date.

The data collection was finished in the summer of 2023. It was then processed and the corpus was automatically lemmatized using MorphoDiTa (Straková, Straka, & Hajič, 2014). The lemmatization is currently being manually controlled and corrected. The first version of the corpus is planned to be made openly available in the first half of 2024.

Analysis

To highlight the practical utility of the HeCz Corpus, we conducted two general analyses of the data. We focus on RTs and comprehension accuracy. In both analyses, we use the data from participants who completed both testing rounds. Three main effects are analyzed:

i The effect of the data collection round (i.e. the intra-participant reliability between the rounds)

ii The effect of stimulus order (i.e. its relative position in the experiment for the given participant)

iii The effect of headline length

In the comprehension accuracy analysis, we also examined the differences between question types and the word order position of the information targeted by the question.

All analyses were exploratory, we did not have any hypotheses which were being tested directly.

Reaction times

For this analysis, we used data only from the participants who took part in both data collection rounds. The RTs were trimmed for the analysis so that the results would not be influenced by outliers. First, RTs below 100ms and above 10,000ms were removed on the basis that these would be too quick or slow to be considered reliable measurements of attentive reading. Then, RTs were log transformed and those that were more than 3 standard deviations from the mean were also removed (the cut-off point was 7.24 log(ms), i.e. 1395.65ms). Altogether, 1.4% of all values were thus removed from the data. As the dependent variable in all models, log transformed RTs were used.

The data were then aligned for each participant and item so we compared directly the RTs for each particular word and participant in the two rounds. The general correlation

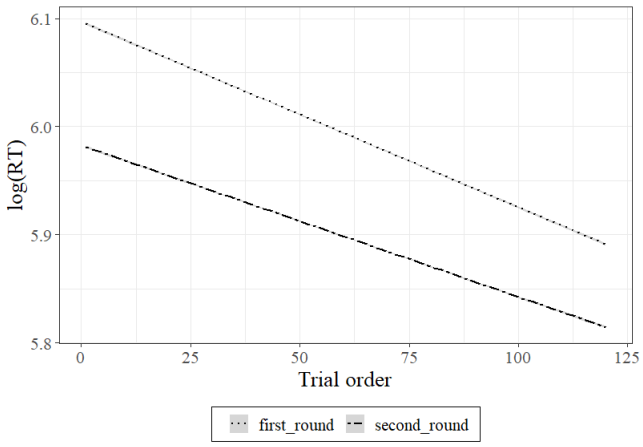

Figure 1: Relationship between trial order and log RTs for words.

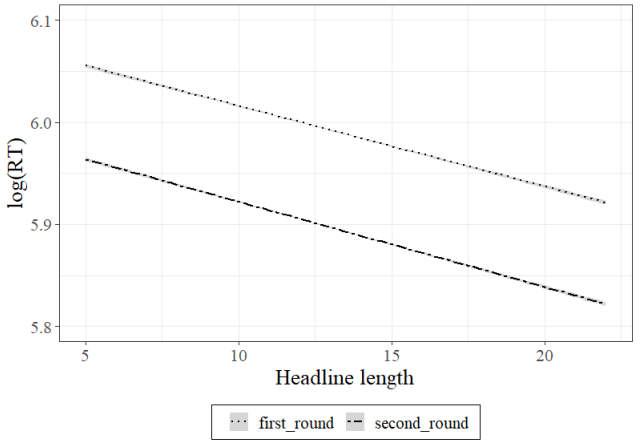

Figure 2: Relationship between headline length in words and log RTs for words.

was moderate ($r = 0.42, p < 0.001$). We then used a linear mixed-effects model with testing round (treatment coded with first round as a baseline), scaled trial order, and scaled sentence length in words as fixed effects and participant and item as random effects (without random slopes due to convergence problems). This model yielded significant effects for all three variables. Participants had a clear tendency to be faster in the second round in comparison to the first round ($\beta = -0.097, SE = 0.0003, t = -309.93, p < 0.001$), they read the headlines progressively faster in the second round ($\beta = -0.055, SE = 0.0001, t = -349.13, p < 0.001$), and the longer the sentence was, the faster the RTs were for individual words ($\beta = -0.013, SE = 0.0002, t = -60.02, p < 0.001$). Figure 1 shows the regression line for the relationship between trial order and log transformed RTs and Figure 2 shows the relationship between RTs and headline length.

Table 1: Comprehension accuracy for six question types. Values are reported as mean percentages of correct answers for the whole sample. Values in square brackets represent 95% confidence intervals of the mean.

Question target	Round 1	Round 2
attribute	88.6 [88.2–89.0]	89.4 [89.1–89.8]
locative adjunct	90.7 [90.3–91.0]	91.0 [90.6–91.4]
object	90.5 [90.1–90.9]	91.0 [90.6–91.3]
subject	88.7 [88.3–89.0]	89.4 [89.0–89.8]
temporal adjunct	85.8 [85.1–86.5]	86.7 [86.1–87.4]
verb	90.6 [90.2–90.9]	90.9 [90.6–91.2]

This points out that task adaptation effects (cf., Prasad & Linzen, 2021) are not only present in each of the testing rounds, but may persist for a month, suggesting that they could remain for even longer periods. Moreover, it shows that the length of a sentence in words is related to processing time spent with reading of each of the words – the longer the sentence is, the faster each of the words tends to be read.

Comprehension accuracy

As was the case for the analysis of RTs, we will also focus only on those participants who completed both data collection rounds for the analysis of comprehension accuracy. Participants' average response accuracy in the first round was 89.47% (with sd=5.79%) and in the second round, it was 90.05% (sd=6.04%). The general correlation between the two testing rounds was rather weak ($r = 0.27, p < 0.001$). We examined the same three fixed effects in the model as we did for the RTs analysis. Figure 3 shows the relationship between accuracy and the trial order, whilst Figure 4 presents accuracy and headline length. Moreover, we also analyzed the position of the information targeted by the question, based on the word order location within the headline (Figure 5).

Additionally, we analyzed the differences between the types of comprehension questions. We focused on six types of comprehension questions, namely on questions targetting (i) attributes (85,975 data points), (ii) subjects (77,129 data points), (iii) verbs (80,544 data points), (iv) temporal adjuncts (27,632 data points), (v) locative adjuncts (63,533 data points), and (vi) objects (71,044 data points). The general accuracy rate for each question type under analysis is presented in Table 1. We can see that the overall response accuracy is very high for all question types and that the differences between question types are minimal.

We analyzed the comprehension accuracy with a logit mixed-effects model. Four fixed effects were included in the analysis, namely data collection round, scaled headline length in words, the scaled trial order (position of the headline during the experiment for the participant), and scaled word order position of the targeted information. As random effects, participant and item were included (with trial order and headline length as random slopes for items and no random slopes for participants).

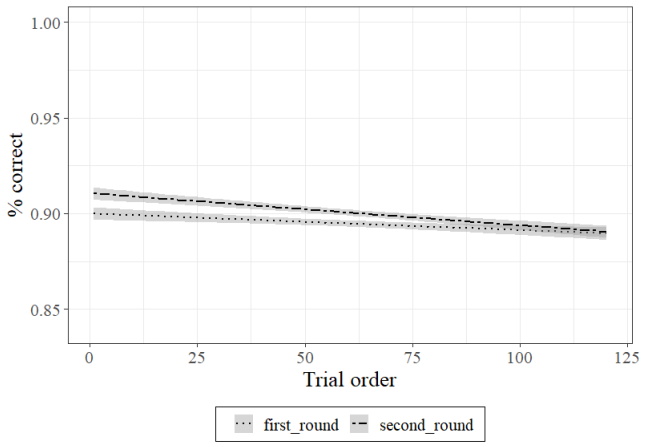

Figure 3: Relationship between trial order and comprehension accuracy.

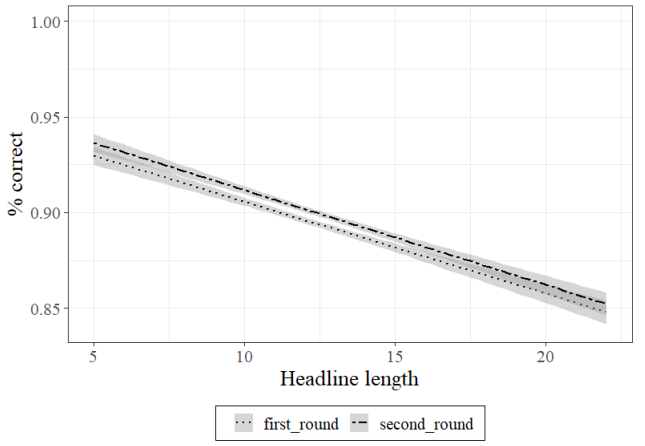

Figure 4: Relationship between headline length in words and comprehension accuracy.

The model yielded the following fixed effects: (i) data collection round ($\beta = 0.074, SE = 0.014, z = 5.468, p < 0.001$), (ii) trial order ($\beta = -0.065, SE = 0.008, z = -7.442, p < 0.001$), (iii) headline length ($\beta = -0.291, SE = 0.027, z = -10.624, p < 0.001$), and (iv) targeted information position ($\beta = 0.431, SE = 0.027, z = 15.81, p < 0.001$).

Similarly to the analysis of RTs, the results show participants had significantly higher comprehension accuracy in the second round of testing. We also documented a detrimental effect of trial order: the later the headline appeared in the experiment, the lower the comprehension accuracy. Furthermore, the results demonstrate that comprehension accuracy decreases as sentence length increases. In other words, longer headlines were harder to answer than shorter ones. And finally, the position of the information was also significant: if the information being targeted by the comprehension ques-

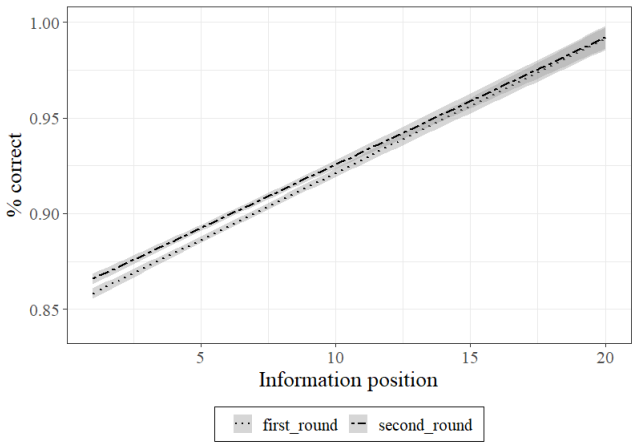

Figure 5: Relationship between position of the targeted information and comprehension accuracy.

tion was closer to the end of the headline, the higher was the comprehension accuracy.

General Discussion

In this paper, we introduced the HeCz corpus and an exploratory analysis of the data. We focused both on RTs and comprehension accuracy. We found several effects that shed new light on some general sentence processing issues.

First, we found a significant difference between the two testing rounds. Participants were both faster and more accurate in the second round of testing. The second round took place about a month after the first one and used the same items for each participant (only in a different, randomized trial order). Also, the correlation between the two rounds was relatively weak ($r = 0.42$ for RTs and $r = 0.27$ for comprehension accuracy). This underscores the important point that task adaptation is not a fleeting phenomenon confined to the duration of an experiment but persists over more extended periods, such as a month. Even after this period, participants are generally faster in their reading and are apparently also able to comprehend the information represented within the headline more accurately.

Second, we found significant effects of trial order. Importantly, these effects were different for RTs and comprehension accuracy. For RTs, we documented a typical "task adaptation" effect (cf. Prasad & Linzen, 2021), i.e. participants were gradually reading faster during the experiment. However, the response accuracy was gradually dropping. This points out that task adaptation effects – clearly documented in the analysis of testing round – may be, at least in part, a result of gradual loss of motivation and attentiveness during the experiment, and not an adaptation per se (i.e. a result of a learning process). Interestingly, this finding is in contrast to the findings reported by Chromý and Tomaschek (submitted) who documented task adaptation effects in both RTs and comprehension accuracy in six consecutive experiments with large sample sizes. The crucial difference between their study and ours is the comprehension measure. They used open-ended questions, whereas we used yes–no questions. This inconsistency between these findings may thus lie in the general task difference – for example, answering open-ended questions is definitely more demanding and participants may be pushed to focus more attentively on providing responses by this very fact.

Third, we documented robust headline length effects on both RTs and comprehension accuracy. Perhaps unsurprisingly, we showed that the longer the headline, the lower the comprehension accuracy. But importantly, we also found that the longer the headline, the faster the RTs for each word (see Figure 2). It remains unclear to what extent this effect plays a role in reading of longer texts, where the sentence length is not as easily perceptible, which is the case for self-paced reading (moving window) paradigms.

Fourth, it has been shown that the word order position of the information targeted by the comprehension question influenced the accuracy. The later in the sentence the information was presented, the higher the comprehension accuracy. This may be interpreted as a manifestation of a recency effect well known from memory literature (Greene, 1986).

All these findings have noteworthy implications for sentence processing research in general. For RT research, sentence length seems to be a crucial factor influencing RTs for individual words. This should be taken into account during analyses of experimental data, especially when comparing results from different studies using different stimuli or testing different languages. Moreover, we may assume differences in RTs and also response accuracy between participants who previously participated in similar experiments and participants who have not. Results of studies focusing on comprehension measures are more likely to be influenced by the length of stimuli used and also by the position of the targeted information in the sentence. Finally, task adaptation effects may not only be due to learning, but also due to a loss of motivation during the experiment. The number of stimuli used in an experiment is thus an important factor influencing both the RTs and comprehension accuracy.

Future directions

The HeCz corpus is, to our knowledge, the largest example of a behavioural corpus that uses the self-paced reading paradigm, providing researchers with a range of behavioural measures at the word and sentence level. Whilst we have presented an initial analysis of data that may be of interest to psycholinguists specifically, the corpus itself has a much broader potential, which we hope will appeal to cognitive scientists from a range of backgrounds.

For example, integrating the behavioural measures and the lemmatization of the corpus will provide computational linguists with access to a unique resource that is in a morphologically rich language. The combination of having a sample size

that is substantially large and having detailed participant demographic information, including the profile of mood states, will open up new avenues of research for those interested in understanding how individual differences may play a role in the processing of texts and how we may read things differently depending on the way we feel or who we are. We are also working on collecting lexical decision time data for each of the words in the corpus. This will provide further resources to researchers interested in the differences between processing words in isolation, and how they are processed in context.

Funding

This work was supported by the European Regional Development Fund project "Beyond Security: Role of Conflict in Resilience-Building" (reg. no.: CZ.02.01.01/00/22_008/0004595).

Acknowledgments

We would like to express our sincere thanks to everyone who initially helped with the initial choice of headlines and with the creation of comprehension questions, namely to Petra Čechová, Lucie Guštarová, Alžběta Králová, Radim Lacina, Eva Pospíšilová, Mikuláš Preininger, and Anna Staňková.

References

Chromý, J., & Dotlačil, J. (2022). Čtení vlastním tempem: kritické představení metody. *Časopis pro moderní filologii*, *104*(2), 177–193. doi: 10.14712/23366591.2022.2.2

Chromý, J., & Tomaschek, F. (submitted). Learning or boredom? Task adaptation effects in sentence processing experiments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior research methods*, *47*, 1–12. doi: 10.3758/s13428-014-0458-y

Frank, S. L., Fernandez Monsalve, I., Thompson, R. L., & Vigliocco, G. (2013). Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior research methods*, *45*, 1182–1190. doi: 10.3758/s13428-012-0313-y

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S. T., & Fedorenko, E. (2021). The Natural Stories corpus: a reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, *55*, 63–77. doi: 10.1007/s10579-020-09503-7

Greene, R. L. (1986). Sources of recency effects in free recall. *Psychological Bulletin*, *99*(2), 221–228. doi: 10.1037/0033-2909.99.2.221

Jegerski, J. (2013). Self-paced reading. In J. Jegerski & B. VanPatten (Eds.), *Research methods in second language psycholinguistics* (pp. 36–65). Routledge.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. *Psychological review*, *87*(4), 329–354. doi: 10.1037/0033-295X.87.4.329

Kennedy, A., Pynte, J., Murray, W. S., & Paul, S.-A. (2013). Frequency and predictability effects in the Dundee Corpus: An eye movement analysis. *Quarterly Journal of Experimental Psychology*, *66*(3), 601–618. doi: 10.1080/17470218.2012.676054

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European journal of cognitive psychology*, *16*(1-2), 262–284. doi: 10.1080/09541440340000213

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence: Brown University Press.

Lange, K., Kühn, S., & Filevich, E. (2015). "Just another tool for online studies" (JATOS): An easy solution for setup and management of web servers supporting online studies. *PloS one*, *10*(6), e0130834. doi: 10.1371/journal.pone.0130834

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

Luke, S. G., & Christianson, K. (2018). The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, *50*, 826–833. doi: 10.3758/s13428-017-0908-4

McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

McNair, D. M., Lorr, M., & Droppleman, L. F. (1971). *POMS Manual for the Profile of Mood States*. San Diego, CA: Educational and Industrial Testing Service.

Prasad, G., & Linzen, T. (2021). Rapid syntactic adaptation in self-paced reading: Detectable, but only with many participants. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(7), 1156–1172. doi: 10.1037/xlm0001046

Siegelman, N., Schroeder, S., Acartürk, C., Ahn, H.-D., Alexeeva, S., Amenta, S., ... others (2022). Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO). *Behavior research methods*, *54*(6), 2843–2863. doi: 10.3758/s13428-021-01772-6

Straková, J., Straka, M., & Hajič, J. (2014). Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 13–18). Baltimore, Maryland: Association for Computational Linguistics.

Stuchlíková, I., Man, F., & Hagtvet, K. (2005). Dotazník k měření afektivních stav\uu: konfirmační faktorová analỳza krátké české verze. *Československá psychologie*, *49*(5), 459–467.

Wehbe, L., Blank, I. A., Shain, C., Futrell, R., Levy, R., von der Malsburg, T., … Fedorenko, E. (2021). Incremental language comprehension difficulty predicts activity in the language network but not the multiple demand network. *Cerebral Cortex*, *31*(9), 4006–4023.