RUSSIAN SENTENCE CORPUS: BENCHMARK MEASURES OF EYE MOVEMENTS IN READING IN CYRILLIC

Laurinavichyute A.K.(1)*, Sekerina I.A.(2), Alexeeva S.A.(3), Bagdasaryan K.A.(1)
alaurinavichute@hse.ru
1 – National Research University Higher School of Economics, Moscow; 2 – College of Staten Island of The City University of New York, New York; 3 – St.-Petersburg State University, St.-Petersburg

Eye-movement corpora are an indispensable tool for basic research in cognitive psychology as well as in psycholinguistics and its applications in education and treatment of developmental and acquired literacy disorders. First, they serve as a repository of basic benchmarks of eye-movement characteristics for languages with typologically diverse orthographies and grammars. As a result, eye-movement corpora function as an important testing ground for models of eye-movements in reading, e.g., the *E-Z Reader* model (Reichle et al. 1998) and the *SWIFT* model (Engbert et al. 2005). Second, eye movements reflect typical linguistic behavior, i.e., silent reading process, and serve to evaluate theories of language processing: for example, the predictions of Gibson's (2000) Dependency Locality theory were tested on eye-movement data in English (Demberg & Keller 2008) and Hindi (Husain et al. 2015), and the predictions of the Surprisal account (Hale 2001) were confirmed on the Potsdam Sentence Corpus (Boston et al. 2008). Finally, corpora provide the necessary control data source to study acquisition of literacy in unskilled (Ashby et al. 2005) and bilingual adults (Cop et al. 2016), developmental reading difficulties in children with and without learning disabilities (Tiffin-Richards & Schroeder 2015), and acquired reading disorders in adults with cognitive impairments such as aphasia (Ablinger et al. 2014) and Alzheimer's disease (Crawford et al. 2015).

In this article, we introduce *the Russian Sentence Corpus (RSC)*. This is the first systematic corpus of eye movements in reading in Russian by skilled young adults that extends the existing eye-movement corpora of European Roman-based and Asian logographic languages to include Cyrillic. Russian is the most representative exemplar of the Cyrillic-based languages, with more than 160 million speakers in the Russian Federation alone. The transparency of its writing system puts it in the middle of the continuum, between shallow (Finnish) and deep (English) orthographies. Several characteristics of Russian, especially phonological (e.g., non-systematic stress patterns, conditional pronunciation in the form of vowel reduction and consonant assimilation, complex syllable structure, and long polysyllabic and polymorphemic words) as well as morphological (rich inflectional and derivational morphology), are of interest for comparative reading research.

Design and materials. The materials were designed following the protocol of the Potsdam Sentence Corpus (Kliegl et al. 2004). First, 144 words were randomly selected from the *StimulStat* database (stimul.cognitivestudies.ru, Alexeeva et al. 2015) based on the pre-defined criteria for a 3x3x2 design, i.e., part of speech (adjective/noun/verb), length (3–4, 5–7, and 8–10 characters), and frequency (> 50 ipm or <10 ipm). Using the resulting list of 144 words, we selected sentences from the Russian National Corpus that included the words in such a way that their position ranged from the third from the beginning to the third one from the end of the sentence. The selected sentences were subjected to acceptability norming. Participants (N = 215) read each sentence online and were asked to judge its acceptability on a scale ranging from 1 "*totally unacceptable*" to 5 "*perfectly acceptable*". The four sentences that received the score below 3.5 were modified by our research team. Third, the resulting 144 sentences were used in a predictability norming study: participants (N = 750) started with a blank screen and were asked to type any word. The script then would replace the word typed by the participant by the first actual word from one of the 144 sentences, and the participant had to guess the second word (and after that all the following words) in such a way that the resulting phrase was a possible word combination in Russian. We collected responses online and included data from every participant that made more than 20 guessing attempts out of the total

number of 1362 words in the corpus. The main and final step was to collect eye movements from 96 monolingual Russian-speaking participants as they read the entire RSC.

Procedure. Sentences were presented in the middle of a 24-inch ASUS VG248QE monitor (resolution: 1920 x 1080 px, response time: 1 ms, frame rate: 144 Hz, font face: 22pt Courier New) controlled by a ThinkStation computer. The presentation of the materials and recording of the eye movements were implemented by Experiment Builder (SR Research Ltd.). Participants were tested individually with the Eyelink 1000+ desktop mount eye-tracker using a chin rest. They were seated at a distance of 55 cm from the camera and 90 cm from the monitor. Calibration consisting of 9 points was performed before the beginning of the experiment and after every 15 sentences afterwards. Before each new calibration participants were asked if they wanted a short break. Eye-movements (only the right eye) were recorded at 1000 Hz rate.

Each trial began with a fixation point at the position of the first letter of the first word in the sentence. If the participant fixated it for at least 500 ms, the sentence presentation automatically commenced; otherwise, after 2 s, 9-point calibration was repeated. After finishing reading the sentence, participants were instructed to look at the red dot in the lower right corner of the screen. To ensure that participants read the sentences for comprehension, 33% of them were followed by a multiple-choice three-alternative comprehension question, and choice was recorded from a mouse click on the response. The program advanced to the next trial after a 1 s delay.

Data Analyses. Eye-movement data were split into fixations and saccades based on the algorithm from the *Data Viewer* package (SR Research Ltd). The first and last words in every sentence were excluded from the analyses. Linear mixed-effects models (R Core Team 2016) with random intercepts for participants, sentences, and words, were used to estimate the impact of the following variables on the inspection times: (a) centered and scaled word length (linear and quadratic trends), (b) logarithm (base 10) of word frequency, and (c) logit-transformed predictability [*log(p/(1-p))*]. The effects were estimated for the following dependent measures: first fixation duration (FFD), single fixation duration (SFD), gaze duration (GD), total reading time (TT).

Results. Figure 1 presents average duration times and their confidence intervals for all corpus words as a function of word's length (A), frequency (B), and predictability (C). The means (and their respective SD aggregated first by participants, and then for the whole dataset) are as follows: FFD – 217 (23) ms, SFD – 228 (26) ms, GD – 259 (42) ms, TT – 318 (79) ms.

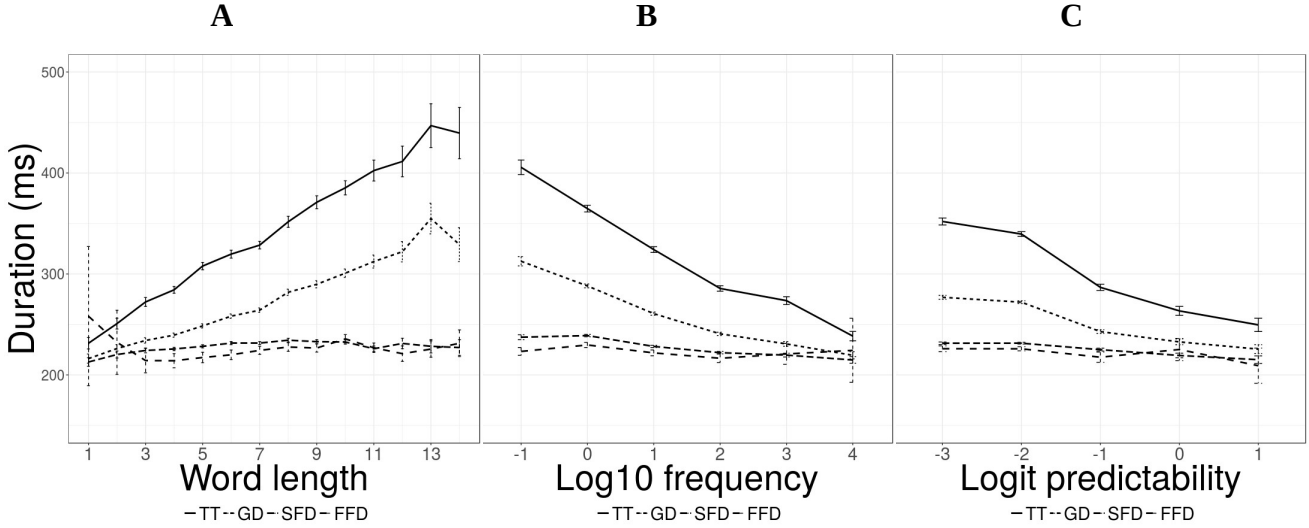

Figure 1. Average reading times as a function of word's length, frequency, and predictability.

One third of all the corpus words in the RSC were not fixated (34%), and this rate of skipping in Russian is consistent with 30–35% skipping rate reported for English (Rayner 1998). Half of the words were fixated once (56%), and the remaining 9% were fixated two or more times.

As in other alphabetic languages, the average saccade length in the RSC spans 8 character spaces, with the saccades landing mostly on the first half of the word and closer to the word center (0.43 of the word's length, where zero represents the beginning of the word).

Most of the basic effects reported in the Potsdam Sentence Corpus for German were also replicated in the RSC for Russian: in the analysis of target words (*N*=144), controlled for length and frequency, as frequency and predictability of the word increase, the reading times decrease (all measures), and as the target word length increases, the reading times also increase. The most notable difference between the two corpora with the respect to the target words is the influence of the square of word's length (which exaggerates the difference between short and long words): in German, increase in the squared word length leads to an increase in FFD, SFD, GD, and TT, while in Russian, increase in the squared word length leads to a slight decrease in FFD and SFD. At the moment we hypothesize that it has to do with inflectional morphology of Russian: longer words contain more inflectional morphemes that can be anticipated in the sentential context, and skilled readers take advantage of such anticipatory information by spending less time on longer words with inflectional morphemes.

The RSC is at present available form the first author upon request and will be freely downloadable online in the future.

The study was financially supported by the Russian Foundation for Humanities (RGNF, grant 17-34-01052a2).

References

Ablinger I., Huber W., Radach R. Eye movement analyses indicate the underlying reading strategy in the recovery of lexical readers //Aphasiology. – 2014. – Т. 28. – №. 6. – С. 640-657.

Alexeeva, S. V., Slioussar, N. A., & Chernova, D.A. StimulStat: baza dannyx, okhvatyvajushchaja razlichnye kharakteristiki slov russkogo jazyka, vazhnye dlja lingvisticheskix i psukhologicheskix issledovanij // Proceedings of the 21st International Conference on Computational Linguistics DIALOG. – 2015.

Ashby J., Rayner K., Clifton C. Eye movements of highly skilled and average readers: Differential effects of frequency and predictability //The Quarterly Journal of Experimental Psychology Section A. – 2005. – Т. 58. – №. 6. – С. 1065-1086.

Boston M. F. et al. Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus //Journal of Eye Movement Research. – 2008. – Т. 2. – №. 1.

Cop U. et al. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading //Behavior research methods. – 2016. – С. 1-14.

Crawford T. J. et al. The disengagement of visual attention in Alzheimer’s disease: a longitudinal eye-tracking study //ICT for assessment and rehabilitation in Alzheimer’s disease and related disorders. – 2016. – С. 40.

Demberg V., Keller F. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity //Cognition. – 2008. – Т. 109. – №. 2. – С. 193-210.

Engbert R. et al. SWIFT: a dynamical model of saccade generation during reading //Psychological review. – 2005. – Т. 112. – №. 4. – С. 777.

Gibson E. The dependency locality theory: A distance-based theory of linguistic complexity //Image, language, brain. – 2000. – С. 95-126.

Hale J. A probabilistic Earley parser as a psycholinguistic model //Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies. – Association for Computational Linguistics, 2001. – С. 1-8.

Husain S., Vasishth S., Srinivasan N. Integration and prediction difficulty in Hindi sentence comprehension: evidence from an eye-tracking corpus //Journal of Eye Movement Research. – 2015. – Т. 8. – №. 2.

Kliegl R. et al. Length, frequency, and predictability effects of words on eye movements in reading //European Journal of Cognitive Psychology. – 2004. – Т. 16. – №. 1-2. – С. 262-284.

Reichle E. D., Rayner K., Pollatsek A. Eye movement control in reading: Accounting for initial fixation locations and refixations within the EZ Reader model //Vision research. – 1999. – Т. 39. – №. 26. – С. 4403-4411.

Tiffin-Richards S. P., Schroeder S. Children's and adults' parafoveal processes in German: Phonological and orthographic effects //Journal of Cognitive Psychology. – 2015. – Т. 27. – №. 5. – С. 531-548.

АННОТАЦИЯ И КЛЮЧЕВЫЕ СЛОВА

Аннотация. We describe the Russian Sentence Corpus (RSC) that establishes benchmarks of eye movements in reading in Cyrillic. The RSC design follows the cross-linguistic protocol of the Potsdam Sentence Corpus for German (Kliegl et al. 2004). The RSC consists of 144 sentences that include target words of three parts of speech (i.e., nouns, verbs, and adjectives) and the corresponding eye-tracking while reading data from 96 young native speakers of Russian reading these sentences. The basic characteristics of eye movements while reading in Russian were described and compared to those of German. In general, the basic characteristics of eye-movements were similar across languages, although Russian manifests systematic differences in the way word length affects reading times, which we tentatively attribute to the morphological structure of Russian words.

Ключевые слова: eye-movements, eye-tracking, reading, Russian, corpus

Аннотация. Мы представляем Русский корпус предложений — первый корпус движений глаз при чтении на кириллице, дизайн которого повторяет дизайн Потсдамского корпуса предложений (Kliegl et al. 2004). Корпус состоит из 144 предложений, содержащих целевые слова трёх частей речи (существительные, прилагательные, глаголы), и данных движений глаз 96 взрослых носителей русского языка, читающих все предложения корпуса про себя. Мы описали базовые дескриптивные характеристики движений глаз при чтении на русском и сравнили их с данными немецкого языка. Данные двух языков оказадись похожи; самое существенное различие заключалось в характере взаимосвязи между длиной слова и временем его чтения. В данный момент мы объясняем это различие морфологической структурой слов в русском языке.

Ключевые слова: корпус, регистрация движений глаз, чтение, русский язык