



<https://doi.org/10.22363/2687-0088-30140>


Research article

## Discourse complexity in the light of eye-tracking: a pilot Russian language study

Svetlana TOLDOVA<sup>1</sup>  , Natalia SLIOUSSAR<sup>1,2</sup>   
and Anastasia BONCH-OSMOLOVSKAYA<sup>1</sup> 

<sup>1</sup>National Research University Higher School of Economics, Moscow, Russia

<sup>2</sup>Saint Petersburg State University, Sait-Petersburg, Russia

 [toldova@yandex.ru](mailto:toldova@yandex.ru)

### Abstract

The paper explores the influence of discourse structure on text complexity. We assume that certain types of discourse units are easier to read than others, due to their explicit discourse structure, which makes their informational input more accessible. As a data source, we use the dataset from the MECO corpus, which contains eye movement data for 12 Russian texts read by 35 native speakers. We demonstrate that the approach relying on elementary discourse units (EDUs) can be felicitously used in the analysis of eye movement data, since fixation patterns on EDUs are similar to those on whole sentences. Our analysis has identified EDU outliers, which show shorter time of first fixation than estimated. We arranged these outliers into several groups associated with different discourse structures. First, these are statements with nominal predicates that set exposition of the text or macroproposition and, following those, EDUs that elaborate on the previous statement and signal the beginning of the narrative. Second, they are EDUs that serve as the middle component of a listing or a group of coordinated clauses or phrases. The final group represents EDUs that are part of an opposition, contrast or comparison. Discourse analysis based on EDUs has never been applied to eye movement data, so our project opens many avenues for further research of complexity of discourse structure.

**Keywords:** *discourse, text complexity, eye movement, EDU, MECO corpus*

### For citation:


Toldova, Svetlana, Natalia Slioussar & Anastasia Bonch-Osmolovskaya. 2022. Discourse complexity in the light of eye-tracking: a pilot Russian language study. *Russian Journal of Linguistics* 26 (2). 449–470. <https://doi.org/10.22363/2687-0088-30140>



## Дискурсивная сложность в свете данных о движениях глаз при чтении: пилотное исследование на материале русского языка

С.Ю. ТОЛДОВА<sup>1</sup>  , Н.А. СЛЮСАРЬ<sup>1,2</sup> ,  
А.А. БОНЧ-ОСМОЛОВСКАЯ<sup>1</sup> 

<sup>1</sup>Национально исследовательский университет «Высшая школа экономики»,  
Москва, Россия

<sup>2</sup>Санкт-Петербургский университет, Санкт-Петербург, Россия  
toldova@yandex.ru

### Аннотация

В статье исследуется влияние структуры дискурса на сложность текста. Предполагается, что некоторые типы дискурсивных единиц читаются легче, чем другие, благодаря выраженной дискурсивной структуре, которая делает содержащуюся в них информацию более доступной для обработки. В качестве источника данных мы используем набор данных из корпуса МЕСО, который содержит данные о движении глаз для 12 русских текстов, прочитанных 35 носителями языка. В статье демонстрируется, что подход, основанный на элементарных единицах дискурса (ЭДЕ), может быть успешно использован для анализа данных о движении глаз, поскольку паттерны фиксации на ЭДЕ схожи с паттернами фиксации на целых предложениях. Проведенный анализ выявил выбросы ЭДЕ, которые показывают более короткое время первой фиксации, чем предполагалось. Они были разделены на несколько групп, связанных с различными структурами дискурса. Во-первых, это высказывания с номинативными предикатами, задающими экспозицию текста или макропропозицию, и следующие за ними ЭДЕ, развивающие предыдущее высказывание и сигнализирующие о начале повествования. Во-вторых, это ЭДЕ, которые служат средним компонентом перечисления или группы согласованных клаузул или фраз. Последняя группа представляет ЭДЕ, которые являются частью оппозиции, контраста или сравнения. Анализ дискурса на основе ЭДЕ никогда не применялся к данным движения глаз, поэтому наш проект открывает новые перспективы для дальнейшего исследования сложности структуры дискурса.

**Ключевые слова:** дискурс, сложность текста, движение глаз, ЭДЕ, корпус МЕСО

### Для цитирования:

Toldova S.Yu., Slioussar N.A., Bonch-Osmolovskaya A.A. Discourse complexity in the light of eye-tracking: a pilot Russian language study. *Russian Journal of Linguistics*. 2022. Vol. 26. № 2. P. 449–470. <https://doi.org/10.22363/2687-0088-30140>

## 1. Introduction

The paper explores the influence of discourse structure on text complexity. We assume that certain types of discourse units are easier to read than others, due to their explicit discourse structure, which makes their informational input more accessible. As a data source, we rely on the Multilingual Eye-movement Corpus, or MECO (Kuperman et al. 2022a, b). The first release of the corpus contains eye movement data from speakers of 12 languages reading 12 texts in their native language and 12 texts in English, as well as answering comprehension questions and passing several tests. We use a dataset with 12 Russian texts read by 35 native speakers.

Registering eye-movements as they unfold in real time, or eye-tracking, was shown to be a very precise and ecologically valid method to study reading (e.g. Rayner 1998, Rayner et al. 2012). However, so far there are not so many studies that analyze the influence of discourse structure on reading behavior. In the present paper, the discourse analysis method based on identifying elementary discourse units (EDUs) (Grosz & Sidner 1986, Mann & Thompson 1987, Polanyi 1988) is applied to eye-tracking data. Firstly, we demonstrate that using this method is very effective: fixation patterns on EDUs are similar to those on whole sentences. Secondly, we identify EDUs that are read significantly faster than expected (based on the estimates taking such parameters as word length into account). Then we analyze them qualitatively: we show that they form several groups associated with different discourse structures.

The first group includes statements with nominal predicates that set the exposition of the text or a macroproposition and, following those, EDUs that elaborate on the previous statement and signal the beginning of the narrative. The second group contains EDUs that serve as the middle component of a listing or a group of coordinated clauses or phrases. The third group includes EDUs that are part of an opposition, contrast or comparison. The main goal of our project is exploratory. We envision it as a pilot study that opens up many avenues for further research in the field of discourse structure complexity.

## 2. Background

### 2.1. Eye tracking studies

Let us start with several basic facts about human vision. We have high visual acuity only in the very center of the visual field. This area is called the fovea. Therefore, when we are reading, our eyes fixate on a word for a fraction of a second and then quickly move to the next word. During these movements, or saccades, no visual information is processed – this happens only during fixations. Some words may require more than one fixation, especially longer and less frequent ones, while the others may be skipped altogether. Short functional words are skipped regularly. About 10 – 15% of the saccades are regressive (Rayner 1998) – we return to what we have just read and then move forward again.

Eye trackers record this complex pattern of fixations and saccades, and provide the researcher with many measures potentially reflecting different processing stages. These measures are usually defined at the word level: *skipping* (whether the word was fixated at least once or skipped); *first fixation duration* (the duration of the first fixation landing on the word); *gaze duration* (the summed duration of fixations on the word in the first pass, i.e., before the gaze leaves this word for the first time); *total fixation duration* (the summed duration of all fixations on the word); *number of fixations* on the word; *regression* (whether the gaze returned to the word after inspecting further text), etc. A detailed discussion of these and other measures can be found in (Boston et al. 2008, Clifton et al. 2007).

Eye tracking is widely used in psycholinguistics and other cognitive sciences to study different phenomena from low-level reading and viewing strategies to the most complex processes like decision-making. However, most studies are experimental: in psycholinguistics, they usually focus on the properties of preselected target words, presented in isolation or inside artificially constructed sentences and, sometimes, small texts. Recently, a number of eye-tracking corpora have been created for several languages, including English (Frank et al. 2013, Luke & Christianson 2018), German (Kliegl et al. 2006), Hindi (Husain et al. 2014), and Russian (Laurinavichyute et al. 2019). The Provo corpus (Luke & Christianson 2018) includes short text passages, while all other corpora mentioned above rely on individual sentences. There are also several corpora based on two languages, including the Dundee corpus (Pynte & Kennedy 2006) with texts in English and French and the GECO corpus (Cop et al. 2017) based on a novel by Agatha Christie (the English original and a Dutch translation). The motivation to create the MECO corpus (Kuperman et al. 2022a, b) used in the present study was to provide comparable data for a much larger set of languages and to do so using complete coherent short texts rather than sentences or text fragments. Among other things, this gives us a unique opportunity to study various text-level phenomena.

First of all, eye movement corpora have been instrumental in establishing the fundamental characteristics of eye movements within and across languages, the so called eye movement benchmarks. Three main parameters of words that influence eye movements were identified: frequency, length, and predictability. Other properties, like the age of acquisition of the word, the number of meanings or the morphological structure, may also play a role, but to a lesser extent (Clifton et al. 2007).

Studies of sentence-level phenomena mostly focused on such topics as syntactic ambiguity processing, which play a crucial role in developing syntactic parsing models. However, some basic generalizations have also been established: the first word of the sentence tends to have longer reading times, then the reader speeds up and slows down again at the last word (Just & Carpenter 1980). This final slowdown is associated with the so-called wrap up when the reader integrates all information presented in the sentence.

## **2.2. Discourse studies**

It has been shown by many researchers (e.g. Hasan & Halliday 1976, van Dijk 2019, Givon 1993) that various phenomena, like anaphora or connectives, cannot be described within an isolated sentence. One can easily distinguish a random sequence of sentences from a coherent text. Thus, it is assumed that discourse is organized in a kind of structure. There are different approaches to representing this structure, cf. the connective-based approach to the relation between discourse segments (Prasad et al. 2018) and the semantic-based approach within the Rhetorical Structure Theory, or RST (Mann & Tompson 1987).

Our research is based on the RST. Its basic assumption is that discourse is a set of nested discourse units up to elementary ones. Each discourse unit has to be related to another one. A set of relation types varies through different research groups. The basic set of relations was worked out by Mann and Thompson (1987). It resembles a set of clause types within a complex sentence, though it is bigger. The relations can be symmetric ('Join', 'Sequence') or asymmetric ('Cause-Effect', 'Purpose', etc.).

Consequently, a coherent text can be split into elementary discourse units, or EDUs (Grosz & Sidner 1986, Mann & Thompson 1987, Polanyi 1988). There are various approaches to EDU splitting depending on whether spoken or written discourse is analyzed, or whether prosodic, cognitive, semantic or pragmatic criteria for discourse segmentation are taken into account. Some approaches combine different dimensions for segmentation, e.g. prosodic and syntactic dimensions (Degand & Simon 2005), or semantic and prosodic dimensions (Kibrik et al. 2009). In the majority of cases, a discourse unit corresponds to a clause, which can be finite or non-finite. Semantically it denotes an event or a state of affairs. In addition to that, there are units larger or smaller than a clause (see Kibrik et al. 2009).

As we are dealing with written texts, we consider clauses as elementary discourse units, and not prosodic units, as in (Hirschberg & Litman 1993, Chafe 1994, Kibrik & Podlesskaya 2003). Structures smaller than a finite clause, such as nominalized constructions or infinitival clauses, can also be treated as EDUs (Carlson & Marcu 2001, Schauer 2000). For example, a preposition introducing a noun phrase can signal causal relations between its dependent expressed via nominalization and the rest of the clause, as in the following case: *iz-za Petinogo pozdnego vozvrashcheniya* 'due to Petya's late return'. At the same time, some EDUs can consist of two clauses. These are EDUs including sentential arguments and restrictive relative clauses. Appositive relatives are treated as separate EDUs.

### **2.3. Eye tracking studies of discourse-level phenomena**

The majority of eye-tracking studies of linguistic complexity are limited to within-sentence phenomena. Significantly fewer studies deal with discourse phenomena, though discourse coherence can influence sentence comprehension and hence reading performance.

One of the research questions is whether there is a difference in the reading behavior inside a discourse segment (a sentence, a paragraph, an intonational unit or a clause) and at a segment boundary. A great number of works focus on the so-called wrap-up effect briefly mentioned in the previous section (cf. Balogh et al. 1998, Hirotsu et al. 2006, Warren et al. 2009, among many others). The main claim of these studies is that clause or sentence final words are read slower than identical words within a clause.

Many works also investigate the start-up effect in the beginning of the clause and the general reading time dependence on the word position in a segment (e.g. Kuperman et al. 2010). In particular, it was found that sentence-initial words tend

to be processed slower (e.g. Gernsbacher 1990). Several experiments report readers' tendency to speed up as they proceed through a sentence (Aaronson & Ferres 1983, Aaronson & Scarborough 1976, Chang 1980, Ferreira & Henderson 1993). Another question is what types of segments (sentences or clauses) trigger these effects. It is also important whether the presence or absence of a comma can affect words reading parameters.

### 3. Data

#### **3.1. The dataset of eye movements and the procedure used to collect the data**

The dataset of eye movements used in the present study comes from the Multilingual Eye-movement Corpus, or MECO (Kuperman et al. 2022a, b). The goal of the MECO project was to collect comparable cross-linguistic eye-tracking data on reading. Native speakers of different languages who learnt English as their second language were recruited to read 12 short texts in their native language (L1) and 12 texts in English (L2). Participants whose native language was English read all 24 texts in their native language and served as the control group in some of the comparisons. After each text, there were two 4-alternative-forced-choice comprehension questions tapping into factual knowledge and inferencing.

The first release of the MECO corpus includes data from 12 languages that differ typologically and orthographically and belong to different linguistic families: English, German, Dutch, Norwegian, Italian, Spanish, Russian, Greek, Turkish, Finnish, Hebrew and Korean. As a result, reading in different L1 could be compared, as well as the influence of different L1 on reading in English as L2. In addition to that, all participants filled in an abridged version of the Language Experience and Proficiency Questionnaire (LEAP-Q, Marian et al. 2007) and passed several tests in their L1 and in English, assessing vocabulary size, word and pseudoword naming, phonological/morphological awareness, and other component skills of reading. The goal was to evaluate how different skills influence reading in L1 and L2. Notably, only participants with an intermediate or advanced level of English as L2 were invited to take part in the study. In other words, the MECO project did not aim to assess reading in L2 at the stages when the learners could not read fluently.

Materials used in the MECO project were encyclopedia-style texts on a variety of topics including historical figures, events, and natural or social phenomena. Firstly, 24 English texts were created. They were loosely based on Wikipedia entries and contained 6–12 sentences (107–185 words). 12 texts were selected for the L2 part of the project, while the other 12 texts were used in the L1 part. Out of the latter 12 texts, five were translated into 11 languages. For the other 7, original texts on the same topic and in the same genre were created in 11 languages. In the present study, we use eye movement data for 12 Russian texts (101 sentences, 1831 words in total).

For each of the 12 languages, at least 45 participants were recruited. We analyze data from 35 native Russian speakers (25 female and 10 male, 20–31 years old). All participants signed an informed consent form before taking part in the study.

Eye-movements were recorded with an EyeLink 1000+ eye-tracker (SR Research, Kanata, Ontario, Canada) with a sampling rate of 1000 Hz. Each text appeared on a separate screen. Consolas font (16 points) was used. Participants were asked to read the texts silently and to press the space bar when they were ready to answer comprehension questions.

### **3.2. Discourse properties of the texts in the dataset: genre and discourse segmentation**

The texts in the MECO corpus are loosely based on Wikipedia entries. They have the same genre and overall structure, but are shorter: every text consists of approximately 10 sentences. The majority of them start with the introduction of a new entity or notion, providing a definition or a basic description of it. Then some narration about these entities follows. All these texts have clear-cut topics repeated throughout the text. Besides, as encyclopedic texts, they include join and elaboration rhetoric relations, other kinds of enumerations, and comparison and contrast relations.

For the subsequent analysis, we split the texts into elementary discourse units (EDUs). As a result, 283 EDUs have been identified in our dataset. As we noted in section 2.2, there are different approaches to this procedure. We used the rules adopted from Ru-RSTreebank with some revisions. The instruction is worked out for written texts. According to this approach, there are EDUs smaller than a clause, while some EDUs (relative and argument clauses with their matrix clauses) are larger than a clause and can contain a comma.

Besides, we introduced several rule revisions. According to the current work on the Russian spoken discourse, coordinate noun phrase constructions (enumerations, like *красные фрукты* ‘red fruits’, *некоторые овощи и даже ягоды* ‘some vegetables and even berries’) often have intonational phrase boundaries in speech (pauses, specific intonation patterns) after each member of the list. The texts in our dataset are not read out loud, and we have no opportunity to judge whether to identify a separate EDU in each particular case relying on their prosodic features. Therefore, we decided to split all NP lists into separate discourse units in our set: for example, (1) *красные фрукты* ‘red fruits’, (2) *некоторые овощи* ‘some vegetables’ (3) *и даже ягоды* ‘and even berries’.

To sum up, there are different types of enumerations in our dataset. Firstly, there are coordinate clauses (e.g. [*Он спросил,*] [*и она ответила*] [*He asked,*] [*and she answered*]) and coordinate clauses with the coordinate subject deletion (e.g. [*Он пришел*] [*и сказал*] [*He came*] [*and said*]’). Secondly, there are NP lists. In the former case, the EDUs are in the multinuclear ‘Join’ or ‘Sequence’ relations. In the latter case, they are in the ‘Enumeration’ or ‘Specification’ relations.

## 4. Experiments looking for correlations between discourse unit characteristics and eye tracking parameters

### 4.1. EDU boundaries

The first question that we tested was whether the eye-tracking data for reading coherent texts provide evidence for the relevance of units that are smaller than sentences. In other words, we aimed to check whether elementary discourse units singled out according to semantic and structural criteria differ in terms of the reading patterns. As was mentioned in the section 2.3, there is a tendency to read the first word in a sentence slower than the following words. Therefore, we tested the hypothesis that the first words in EDUs are read slower than others.

According to some eye-tracking studies (e.g. Hirotani et al. 2006), commas influence eye tracking parameters in a significant way. Thus, the effect of EDU boundaries, if we find it, may be a result of a strong correlation between the end of the EDU and the comma following it. Indeed, many intra-sentential EDUs are separated by commas in Russian. Therefore, we also checked that the effect we found is due to EDU boundaries and not to punctuation marks.

### 4.2. The first-word effect

#### 4.2.1. Data and procedure

As we demonstrated in section 2.1, eye-tracking research provides multiple measures that may be associated with different stages of text processing. In our pilot study, we confine ourselves to two parameters that are often selected as reflecting very early and advanced processing stages: the *first fixation duration* (the duration of the first fixation landing on the word) and the *total fixation duration*, or *total time* (the summed duration of all fixations on the word, including possible multiple fixations during the first pass and refixations following regressions if there were any).

Eye-tracking research usually focuses on the properties of separate words rather than larger units as a whole. To study the latter, we transformed selected measures to take into account the crucial factor of word length. Namely, we analyzed the relative first fixation duration (RFFD) and the relative total fixation duration (RTFD): the average first fixation duration per symbol and the average total fixation duration per symbol calculated for every word (token) in our texts. The longer the duration the slower the reading speed. These measures were used in different analyses that we conducted.

To test for start-up effects, we compared the first words in EDUs to the third words (because the start-up effect may cover not only the very first word, but also the second word in the unit). For this analysis, we filtered out EDUs that are shorter than four words or have no fixations on the first or on the third words. We also did not include sentences consisting of a single EDU. Then we calculated an average RFFD and RTFD for both positions for every participant. Using these average



values, we performed a paired two-sided t-test of the null hypothesis of independence for average RFFDs and RTFDs.

#### 4.2.2. Results and discussion

The t-test revealed a statistically significant difference in RFFD and RTFD between the first and the third word in the EDU ( $t = 8.59$ ,  $df = 32$ ,  $p < 0.001$  for RFFD;  $t = 4.21$ ,  $df = 32$ ,  $p < 0.001$  for RTFD). Average RFFDs for the words in different positions are also presented in Figure 1. The thin gray lines are for separate EDUs, the black dots represent an average RFFD for a position. The blue lines are model predictions for EDUs. We can see that there is a tendency for decreasing the relative first fixation duration while moving further away from the first word in the EDU. In addition to that, the t-test comparing the first word RFFD characteristics in an EDU vs. in a sentence revealed no significant differences between sentences and EDUs ( $t = 1.16$ ,  $df = 51$ ,  $p = 0.27$ ).

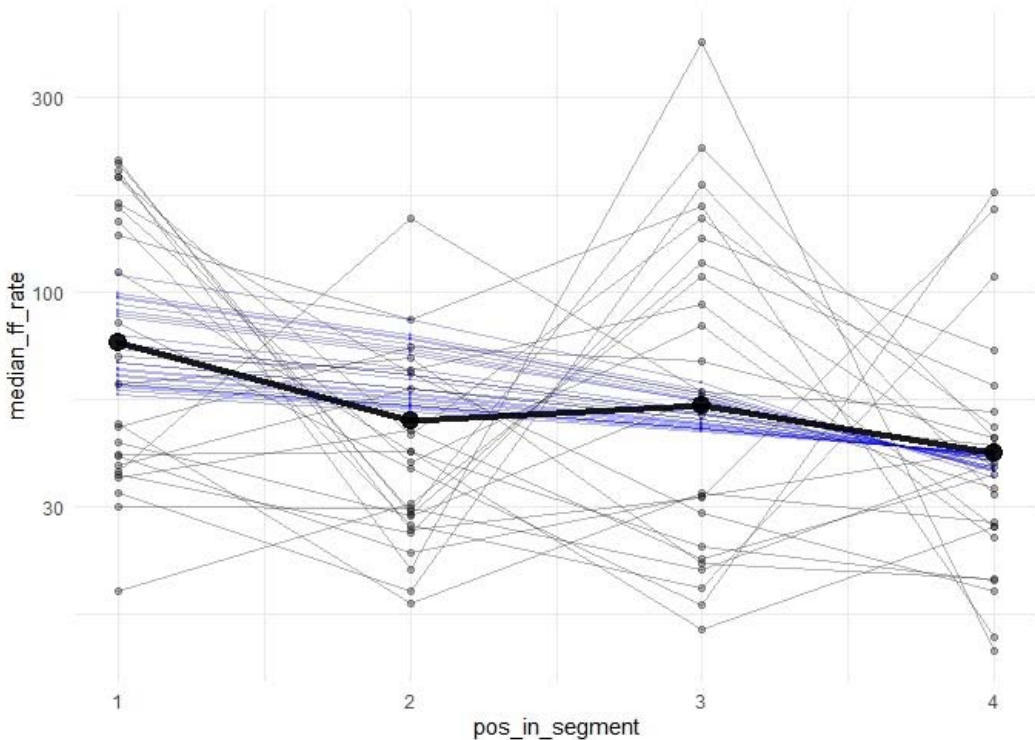


Fig. 1. The average RFFD plot for different word positions in the EDU

To conclude, our analysis provides additional evidence for the effect of the first word in a discourse segment on Russian real-text data. Though the particular patterns of fixation duration may vary greatly for different speakers and different EDUs (cf. gray lines in Fig. 1), the initial fixation on the first word in a segment is longer than on the following words. Moreover, this difference is significant

irrespective of the discourse segment level: a sentence vs. an EDU. We can conclude that for this effect, the EDU boundaries matter. Thus, these results confirm the role of EDU boundaries for reading a coherent text.

### **4.3. EDU boundaries vs. punctuation marks**

An alternative hypothesis is that the first-word effect is due not to the EDU boundaries, but to the punctuation marks. To test this hypothesis, we compared RFFDs for the first unskipped word in a segment under different conditions: (1) the first word of an EDU after a comma; (2) the first word after a comma inside an EDU; (3) the first word of an EDU not following comma, (4) the first word after a dot.

#### *4.3.1. Data and procedure*

Firstly, we checked the hypothesis of the independence of first word RFFD from the comma position (inside an EDU vs. before an EDU). Secondly, we checked whether there is a difference between EDUs after a comma and after another EDUs without a comma. To do so, we used linear mixed effect models (LMEs) in the R package *lme4* (Bates et al. 2015). Participants and words were treated as random effects. Finally, we performed a pairwise comparison of the four conditions enumerated above using the Tukey test. For the first analysis we selected EDUs after a comma and EDUs with a comma inside and detected the first unskipped word after the comma.

#### *4.3.2. Results*

The results for the two LME models are presented in Figures 2 and 3. Figure 2 confirms that the RFFD on a word after a comma is significantly longer when it is the first word in an EDU than when it is in a middle position. Figure 3 shows that there are no significant differences for the first words in an EDU preceded or not preceded by a comma.

We can conclude that EDU boundaries play a more important role for the RFFD than punctuation marks. Finally, we performed a pairwise comparison of all the four conditions (after a comma in the middle of an EDU, after a comma in the beginning of an EDU, after a dot in the beginning of an EDU, in the beginning of an EDU without any punctuation marks) using the Tukey method. The results are provided in Figure 4.

As Figure 4 shows, there is a difference in RFFD depending on the position of the word inside an EDU (in the beginning vs. in the middle). The dot vs. comma is a significant factor, but there is no statistically significant difference between the word in the beginning of an EDU preceded or not preceded by a comma. To sum up, our data shows the impact of EDU boundaries on reading parameters.

<i>Predictors</i>	<b>firstfix_rate</b>		
	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	35.20	30.61 – 39.79	<0.001
segment startTRUE	5.51	2.89 – 8.13	<0.001
<b>Random Effects</b>			
$\sigma^2$	511.18		
$\tau_{00}$ words.word	1002.53		
$\tau_{00}$ words.subid	32.53		
ICC	0.67		
$N_{\text{words.subid}}$	33		
$N_{\text{words.word}}$	317		
Observations	5627		
Marginal $R^2$ / Conditional $R^2$	0.004 / 0.671		

Fig. 2. The LME model for a word after a comma in the beginning vs. in the middle of an EDU

<i>Predictors</i>	<b>firstfix_rate</b>		
	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	39.58	34.63 – 44.52	<0.001
start after commaTRUE	0.65	-2.82 – 4.13	0.712
<b>Random Effects</b>			
$\sigma^2$	847.71		
$\tau_{00}$ words.word	1045.55		
$\tau_{00}$ words.subid	44.64		
ICC	0.56		
$N_{\text{words.subid}}$	33		
$N_{\text{words.word}}$	336		
Observations	6104		
Marginal $R^2$ / Conditional $R^2$	0.000 / 0.563		

Fig. 3. The LME model for the first word in an EDU preceded or not preceded by a comma

Conditions	Estimate	SE	z	p
start_after_comma – comma_in_the_middle == 0	6.12	1.47	4.16	< 0.001***
start_after_dot – comma_in_the_middle == 0	12.87	2.12	6.07	< 0.001***
start_after_neither – comma_in_the_middle == 0	4.98	1.82	2.74	0.030*
start_after_dot – start_after_comma == 0	6.75	1.85	3.65	0.001**
start_after_neither – start_after_comma == 0	-1.15	1.45	-0.79	0.852
start_after_neither – start_after_dot == 0	-7.90	1.59	-4.98	< 0.001***

Fig. 4. Pairwise comparisons of the four conditions using Tukey method

#### 4.4. Outliers as a window onto discourse effects

##### 4.4.1. Finding outliers: sentences vs. EDUs

It is clear from previous studies that the variation in reading times is great both among participants and among sentences. As there could be a lot of low-level factors and discourse factors influencing the resulting average total fixation duration and other measures (different types of conjunctions, discourse topicality of a noun phrase, different types of rhetorical relations, the position of an EDU in the discourse tree etc.), we started with trying to find outliers in our dataset. Firstly, we performed an analysis to find the sentences that are read too slowly or too fast as compared to the average reading rate. Then we performed the same analysis to find outliers for EDUs.

To start with, we got the histograms and violin plots for sentences and EDUs. We calculated the average sentence duration rate, or ASDR, for all the sentences: the sum of total fixation durations of all unskipped words divided by the sentence length in symbols. Then, we used the standard deviation from the median ASDR per participant to calculate the deviation of the rate for every sentence.

We used the paired t-test with Bonferroni correction for multiple comparisons and identified 35 sentences for which the ASDR differs significantly from the estimated value ( $p < 0.05$ ). We also singled out 82 EDU outliers among 283 using the same method. Out of these EDUs, the majority (59) had *shorter* ASDR than expected. We look at this group below because it is large enough to observe some general tendencies. In order to formulate hypotheses for further research and to single out features that can subsequently be used for EDU classification, we started with a qualitative analysis of these outliers in the present paper.

##### 4.4.2. Qualitative analysis of EDU outliers

In our qualitative analysis we aimed to identify some recurring features in the discourse units that tend to be read with shorter total fixation durations. We analyzed 59 outlier EDUs and found that 47 of them belonged to three groups characterized by common semantic and syntactic properties and specific discourse

functions. These three groups may be loosely called “starters”, “contrasts” and “listings”.

In general, what they all have in common is an explicit discourse structure. This structure sets certain relations between the units of expressed information within the EDU and embeds it in a given way relative to the general representation of textual information. In other words, it is easy for the reader to immediately answer the question of what is being said here: the EDUs that we have classified as starters assert that a certain object belongs to a class of objects; the EDUs belonging to the contrast group introduce the opposite poles of a certain scale; the listing group includes EDUs that report different manifestations of the same property or situation. Looking at these results somewhat more broadly, we can conclude that the decrease in fixation times is somehow related to the idea of predictability. Predictability, which manifests itself here at a higher level, has been proven to have a significant effect on fixation patterns (e.g. Clifton et al. 2007).

Below we will examine each group separately. In addition to analyzing ASDR outliers, we will also pay attention to EDUs that have been completely skipped by three or more readers and have a structure similar to one of the groups. We will use the following notation to denote EDUs and the differences in fixation patterns: [...] is used to define the boundaries of EDUs, bold font represents EDUs characterized by shorter ASDR than estimated, crossed-out EDUs are those skipped by more than 3 readers.

#### 4.4.3. Starters

The texts of the MECO corpus have a similar composition, as they mostly describe natural, cultural or social phenomena. Loosely based on Wikipedia, they obey a common pattern, which is especially noticeable in the first sentences of these texts. 10 out of 12 first sentences have a similar syntactic structure: one or two joint EDUs consisting of a nominal predication with omitted copula, following the Russian grammar rules for the present tense. (1)–(3) show some examples of this type.

- (1) [**Янус – бог всех начинаний**] [*и переходов в древнеримской религии и мифологии.*]  
*[Yanus – bog vsekh nachinaniy] [i perekhodov v drevnerimskoj religii i mifologii.]*  
 ‘Janus is the god of all beginnings and transitions in the ancient Roman religion and mythology.’
- (2) [**Дегустация – это сенсорный анализ и оценка вина.**]  
*[Degustaciya – eto sensornyj analiz i ocenka vina.]*  
 ‘Tasting is a sensory analysis and evaluation of wine.’
- (3) [**Апельсиновый сок – это напиток, который получают из плодов апельсинового дерева.**]  
*[Apel'sinovyj sok – eto napitok, kotoryj poluchayut iz plodov apel'sinovogo dereva.]*  
 ‘Orange juice is a drink made from the fruit of the orange tree.’

One example (4) begins with an EDU clause with a light verb, and we have an example of a text (5) that begins with an EDU with verbal predication.

- (4) *[В профессиональном спорте допингом называют применение спортсменами запрещенных стимулирующих веществ.]*  
*[V professional'nom sporte dopingom nazyvayut primenenie sportsmenami zapreshchennyh stimulyruyushchih veshchestv.]*  
 ‘In professional sports, doping refers to the use of banned performance-enhancing substances by athletes.’
- (5) *[Всемирный день окружающей среды отмечают ежегодно пятого июня.]*  
*[Vsemirnyj den' okruzhayushchej sredy otmechayut ezhegodno pyatogo iyunya.]*  
 ‘The World Environment Day is celebrated annually on June 5.’

At the semantic level, 11 out of 12 sentences establish taxonomic relations by introducing the main topic of the text as belonging to a particular category. But only 6 sentences out of 12 (including sentence (5), which is semantically and syntactically different) are read with a significantly shorter ASDR. The other six share a common syntactic feature – they have a restrictive relative clause that is not split up into a separate EDU according to the RST marking rules. We can speculate that it is the restrictive relative clause that determines the longer fixation times. The tendency for restrictive relatives to increase reading times was previously noted in (Hirofani et al. 2006) with the following explanation: restrictive relative clauses belong to the assertive part of the sentence and contribute to the truth conditions of the sentence, they are part of focused material that is known to require more attention. The only exception to this observation in our corpus is (6), in which the restrictive nature of the relative clause is controversial.

- (6) *[Жест «шака» – это знак дружеских намерений, который часто связывают с Гавайями и сёрф-культурой.]*  
*[Zhest «shaka» – eto znak druzheskih namerenij, kotoryj chasto svyazyvayut s Gavajyami i syorf- kul'turoj.]*  
 ‘The “shaka” gesture is a sign of friendly intentions that is often associated with Hawaii and surf culture.’

Another type of EDUs that can be assigned to the starter category are EDU clauses that introduce macropropositions in the text in the sense of Van Dijk (2019). Here we observe two groups of cases. Firstly, there are clauses that introduce a new block of information the subject of which repeats the topical subject of the first sentence of the text. Secondly, they may be EDUs that follow the initial thematic sentence and serve as the first introductory piece of the narrative, elaborating on the points declared earlier. The beginning of the narrative may be marked by a tense change: from the present tense of the opening sentence to the narrative past tense.

#### 4.4.4. Contrasts

This category includes pairs of EDUs that express some sort of opposition.

- (7) [*Начиная с древних гонок на колесницах*] [*и до недавних скандалов в бейсболе и велоспорте*]  
[*Nachinaya s drevnih gonok na kolesnicah*] [*i do nedavnih skandalov v bejsbole i velosporte*]  
‘From the ancient chariot races to the recent scandals in baseball and cycling...’

Shorter fixations are characteristic either for both EDUs or only for the second element of the comparison. In some cases, as in (8), no specific fixation effects are statistically significant, but we observe skipping of the entire EDU by a certain number of readers.

- (8) [*что более дорогое вино будет обладать лучшими характеристиками,*] [*чем менее дорогое.*]  
[*что более дорогое вино будет обладать лучшими характеристиками,*]  
[~~*чем менее дорогое.*~~]  
‘...that a more expensive wine will have better characteristics than a less expensive one.’

We can conjecture that the shorter fixation effect is related to the lexical parallelism in these EDUs, such as antonymy.

- (9) [*Ворота его храма были открыты во время войны*] [*и закрыты в мирное время.*]  
[*Vorota ego hrama byli otkryty vo vremya vojny*] [*i zakryty v mirnoe vremya.*]  
‘The gates of his temple were open during the war and closed during peacetime.’

Moreover, in some cases the contrast is not expressed at the semantic level of the whole sentence, but only at the level of the individual lexemes, which form a kind of binary opposition.

- (10) [*допинг – явление не новое,*] [*а такое же древнее, как и сам спорт.*]  
[*doping – yavlenie ne novoe,*] [*a takoe zhe drevnee, kak i sam sport.*]  
‘...doping is not a new phenomenon, but is as old as sport itself.’
- (11) [*так как он смотрит и в будущее,*] [*и в прошлое*]  
[*tak kak on smotrit i v budushchee,*] [*i v proshloe.*]  
‘as he looks both to the future and to the past’

In addition, there are several cases in which we observe shorter fixations on a single EDU containing a lexical opposition represented by two opposite parameters or situations.

- (12) [*нет ни одного флага, у которого высота была бы больше ширины.*]

- [net ni odnogo flaga, u kotorogo vysota byla by bol'she shiriny.]*  
 ‘there is not a single flag that has a height greater than its width’  
 (13) *[также связывали с входом и выходом из дома.]*  
*[takzhe svyazyvali s vkhodom i vyhodom iz doma.]*  
 ‘also associated with entering and exiting the house’

We can assume that the binary scale manifested in the lexical structure supports the processing of EDUs. In any case, we do not see specific effects of this kind within the EDUs in which the objects of comparison cannot be contrasted by a unique parameter, such as the presence or absence of a quality or situation, see (14) and (15).

- (14) *[и включила в свои проекты не только сохранение природы.] [но и вопросы экологически безопасного развития.]*  
*[i vklyuchila v svoi proekty ne tol'ko sohranenie prirody.] [no i voprosy ekologicheski bezopasnogo razvitiya.]*  
 ‘...and incorporated into its projects not only the preservation of nature, but also issues of environmentally safe development’  
 (15) *[В некоторых странах номера регистрируются в едином реестре.] [в других реестры ведутся в отдельных штатах и областях.]*  
*[V nekotoryh stranah nomera registriruyutsya v edinom reestre.] [v drugih reestry vedutsya v otdel'nyh shtatah ili oblastyah.]*  
 ‘In some countries numbers are registered in a single registry, in others registries are maintained in individual states or provinces.’

Thus, we see in this group a compact discursive structure supported both at the syntactic level (by conjunctions) and at the lexical level. All these means contribute to building up the reader's expectations, which are expressed in the acceleration of information processing, especially in the second part of the contrast.

#### 4.4.5. Listings

Finally, the last category has proven to be the most numerous, as it includes groups of three or more EDUs that together form an enumeration. As we mentioned earlier, enumeration elements are defined as separate EDUs in our analysis. So far, we have distinguished three components in the enumeration list and, accordingly, three EDU types: the beginning, the middle, and the end. We observe that the middle component (or one of the middle components) tends to require less fixation time or is even skipped, as examples (16) and (17) illustrate.

- (16) *[Янус олицетворял золотую середину между варварством и цивилизацией.] [деревней и городом.] [юностью и зрелостью.]*  
*[Yanus olicetvoryal zolotuyu seredinu mezhdru varvarstvom i civilizaciej.] [derevnej i gorodom.] [yunost'yu i zrelost'yu.]*  
 ‘Janus represented the golden mean between barbarism and civilization, village and city, youth and maturity.’



- (17) [*например, его географическое происхождение,*] [*репутация*]  
 [*и прочие характеристики.*]  
 [*naprimer, ego geograficheskoe proiskhozhdenie,*] [*reputaciya*]  
 [*i-prochie-karakteristiki.*]  
 ‘such as its geographic origin, reputation, and other characteristics’

It can be assumed that in this case the discursive structure is graphically supported. It is interesting to note that, despite the fact that the comma generally slows down processing, in this case the placement of an EDU within two commas may be perceived by the reader as a way to save processing efforts by using the same cognitive schemas as in the preceding EDU.

We also found cases in which we have fixation acceleration on the first element of a list, but these lists are characterized by the fact that the preceding EDU contains a generalizing lexeme.

- (18) [*и с тех пор празднование сопровождается широкомасштабными кампаниями, посвященными важнейшим экологическим проблемам:*] [~~*загрязнению мирового океана,*~~] [*перенаселенности планеты,*] [*глобальному потеплению.*]  
 [*i s tekh por prazdnovanie soprovozhdaetsya shirokomasshtabnymi kampaniyami, posvyashchennymi vazhnejshim ekologicheskim problemam:*] [~~*zagryazneniyu mirovogo okeana,*~~] [*perenaselelnosti planety,*] [*global'nomu poteplenyu.*]  
 ‘and since then, the celebration has been accompanied by widespread campaigns on major environmental issues: ocean pollution, overpopulation of the planet, and global warming’

However, we do not observe this effect if there are only two EDUs in the list itself.

- (19) [*Все страны требуют регистрационных знаков для таких наземных транспортных средств,*] [*как легковые и грузовые автомобили,*] [*а также мотоциклы.*]  
 [*Vse strany trebuyut registracionnyh znakov dlya takih nazemnyh transportnyh sredstv,*] [*kak legkovye i gruzovye avtomobili,*] [*a takzhe motocikly.*]  
 ‘All countries require registration plates for land vehicles, such as light and cargo vehicles, and also motorcycles.’

The last element of the list usually does not show a shorter fixation: probably the list is “assembled” as a whole at this moment. This hypothesis needs to be further tested using regression and saccade analyses. The exceptions are the examples with extremely short and very homogeneous lists.

- (20) [*Наиболее популярные цвета, используемые для государственных флагов,*] – [*красный,*] [~~*белый,*~~] [~~*зеленый*~~] [~~*и синий.*~~]  
 [*Naibolee populyarnye cveta, ispol'zuemye dlya nacional'nyh flagov,*] – [*krasnij,*] [~~*belyj,*~~] [~~*zelenyj*~~] [~~*i-sinij.*~~]

‘The most popular colors used for national flags are red, white, green, and blue.’

It is interesting that not only nominal enumerations, but also enumerations of situations expressed by verbal predicates with arguments behave as listings, as in example (21).

- (21) [*Более того, некоторые производители выпаривают сок, [и затем снова добавляют в него воду] [или разбавляют водой заранее изготовленный концентрат.]*  
*[Bolee togo, nekotorye proizvoditeli vyparivayut sok,] [i zatem snova dobavlyayut v nego vodu] [ili razbavlyayut vodoj zaranee izgotovlennyy koncentrat.]*  
‘Moreover, some producers evaporate the juice and then add water to it again, or dilute a pre-made concentrate with water.’

We suppose that the cognitive mechanisms that determine how lists are read, processed, and remembered within a coherent text require, in principle, require a separate study because they may work differently from the processing of narrative fragments. Not coincidentally, longer lists usually require special formatting with a separate line for each element and bullet points to be better understood and remembered. Perhaps the fact that middle elements tend to be skipped or have shorter fixation times has a significant impact on the processing of the entire list and requires further investigation using other eye tracking measures, such as regression probabilities.

## 5. Conclusion

In this paper, we tried to identify different ways in which discourse structure affects texts complexity. To do so, we analyzed eye movement data of 35 readers for a collection of 12 Russian texts from the MECO project (Kuperman et al. 2022a, b). The analysis of eye movements is the most precise method that can be used to assess the complexity of the text for a reader. Moreover, it provides the researcher with many measures potentially reflecting different processing stages that can be used to study linguistic phenomena from the word level up to the whole text level.

However, this richness made our task more challenging. The influence of low-level factors, primarily the length, frequency and predictability of individual words, tends to obscure the effects of the higher-level factors, and their contribution cannot be measured directly. In the present paper, we came up with an approach that let us overcome this problem and identify several types of EDUs that are read significantly faster than expected. We hypothesized that this can be explained by their discourse properties, but could only prove this by qualitative analysis on the dataset we analyze in the present paper.

We view this as the first step that opens multiple avenues for further research. Firstly, we plan to test the hypotheses we formulated on other sets of eye-tracking data, both from Russian and from other languages. Some predictions may

eventually be tested experimentally. Secondly, we plan to explore other eye-tracking measures: at first, using the same qualitative approach that we adopted in the present study and then validating the emerging generalizations on other data sets.

### Acknowledgements

We would like to thank Varvara Magonmedova for her help on data preprocessing and Ivan Pozdnyakov for his help in statistical processing of the data in R.

The work has been supported by the Ministry of Science and Higher Education of the Russian Federation within Agreement No 075-15-2020-793.

### REFERENCES

- Aaronson, Doris & Steven Ferres. 1983. Lexical categories and reading tasks. *Journal of Experimental Psychology: Human Perception and Performance* 9(5). 675.
- Aaronson, Doris & Hollis S. Scarborough. 1976. Performance theories for sentence coding: Some quantitative evidence. *Journal of Experimental Psychology: Human Perception and Performance* 2(1). 56.
- Balogh, Jennifer, Edgar Zurif, Penny Prather, David Swinney & Lisa Finkel. 1998. Gap-filling and end-of-sentence effects in real-time language processing: Implications for modeling sentence comprehension in aphasia. *Brain and Language* 61(2). 169–182.
- Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1). 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Boston, Marisa Ferrara, John T. Hale, Reinhold Kliegl & Umesh Patil. 2008. Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research* 2. 1–12.
- Carlson, Lynn & Daniel Marcu. 2001. Discourse tagging reference manual. *ISI Technical Report ISI-TR-545* 54. 56.
- Chafe, Wallace. 1994. *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. University of Chicago Press.
- Chang, Frederick R. 1980. Active memory processes in visual sentence comprehension: Clause effects and pronominal reference. *Memory & Cognition* 8(1). 58–64.
- Clifton, Charles Jr., Adrian Staub & Keith Rayner. 2007. Eye movements in reading words and sentences. In R. van Gompel (ed.), *Eye movements: A window on mind and brain*, 341–372. Amsterdam, Netherlands: Elsevier.
- Cop, Uschi. 2017. Presenting GECO: An eye-tracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods* 49. 602–615.
- Degand, Liesbeth & Anne-Catherine Simon. 2005. Minimal Discourse Units: Can we define them, and why should we. *Proceedings of SEM-05. Connectors, Discourse Framing and Discourse Structure: From Corpus-Based and Experimental Analyses to Discourse Theories* 477. 65–74.
- Frank, Stefan L., Irene Fernandez Monsalve, Robin L. Thompson & Gabriella Vigliocco. 2013. Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods* 45. 1182–1190.
- Ferreira, Fernanda & John M. Henderson. 1993. Reading processes during syntactic analysis and reanalysis. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale* 47(2). 247.

- Gernsbacher, Morton A., Varner R. Kathleen & Mark E. Faust. 1990. Investigating differences in general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 16(3). 430.
- Givón, Talmy. 1993. Coherence in text, coherence in mind. *Pragmatics & Cognition* 1(2). 171–227.
- Grosz, Barbara J. & Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics* 12. 175–204.
- Halliday, M. A. K. & Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman.
- Hirotoni, Masako, Lyn Frazier & Keith Rayner. 2006. Punctuation and intonation effects on clause and sentence wrap-up: Evidence from eye movements. *Journal of Memory and Language* 54(3). 425–443.
- Hirschberg, Julia & Diane Litman. 1993. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics* 19(3). 501–530.
- Husain, Samar, Shruvan Vasisht & Narayanan Srinivasan. 2014. Integration and prediction difficulty in Hindi sentence comprehension: Evidence from an eye-tracking corpus. *Journal of Eye Movement Research* 8. 1–12.
- Just, Marcel A. & Patricia A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review* 87. 329–354.
- Kibrik, Andrej A. & Vera I. Podlesskaya. 2009. *Night Dream Stories: Corpus Study of Russian Discourse*. Moscow: Yazyki slavyanskikh kultur. (In Russ.)
- Kliegl, Reinhold, Antje Nuthmann & Ralf Engbert. 2006. Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General* 135. 12.
- Kuperman, Victor. 2010. The effect of word position on eye-movements in sentence and paragraph reading. *Quarterly Journal of Experimental Psychology* 63(9). 1838–1857. <https://doi.org/10.1080/17470211003602412>
- Kuperman, Victor. 2022a. Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO). To appear in *Behavior Research Methods*.
- Kuperman, Victor. 2022b. Text reading in English as a second language: Evidence from the Multilingual Eye-Movements Corpus (MECO). To appear in *Studies in Second Language Acquisition*.
- Laurinavichyute, Anna K., Irina A. Sekerina, Svetlana Alexeeva, Kristine Bagdasaryan & Reinhold Kliegl. 2019. Russian Sentence Corpus: Benchmark measures of eye movements in reading in Russian. *Behavior Research Methods* 51. 1161–1178.
- Luke, Steven G. & Kiel Christianson. 2018. The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods* 50. 826–833.
- Mann, William C. & Sandra A. Thompson. 1987. *Rhetorical Structure Theory: A Theory of Text Organization*. Los Angeles: University of Southern California, Information Sciences Institute.
- Marian, Viorica, Henrike K. Blumenfeld & Margarita Kaushanskaya. 2007. The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research* 50. 940–967.
- Podlesskaya, Vera I. & Andrej A. Kibrik. 2003. Methods of oral speech corpora research: Discourse transcription development experience. *Proc. of Cognitive Modeling in Linguistics*. Varna, Bulgaria.
- Polanyi, Livia. 1988. A formal model of the structure of discourse. *Journal of Pragmatics* 12. 601–638.

- Prasad, Rashmi, Bonnie Webber & Alan Lee. 2018. Discourse annotation in the PDTB: The next generation. *Proceedings 14th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*. 87–97.
- Pynte, Joël & Alan Kennedy. 2006. An influence over eye movements in reading exerted from beyond the level of the word: Evidence from reading English and French. *Vision Research* 46. 3786–3801.
- Rayner, Keith. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin* 124. 372.
- Rayner, Keith. 2012. *Psychology of Reading*. New York, NY: Psychology Press.
- Schauer, Holger. 2000. From elementary discourse units to complex ones. *1st SIGdial Workshop on Discourse and Dialogue*. 46–55.
- Van Dijk, Teun A. 2019. *Macrostructures: An Interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition*. Routledge.
- Warren, Tessa, Sarah J. White & Erik D. Reichle. 2009. Investigating the causes of wrap-up effects: Evidence from eye movements and E-Z Reader. *Cognition* 111(1). 132–137.

### Dictionaires and internet resources / Интернет-ресурсы

MECO. <https://meco-read.com> (accessed 26.05.2022).

Ru-RSTreebank. <https://rstreebank.ru/> (accessed 26.05.2022).

#### Article history:

Received: 20 October 2021

Accepted: 01 February 2022

#### Bionotes:

**Svetlana Yu. TOLDOVA** holds a Ph.D. in Philology and is Associate Professor of the School of Linguistics at the Faculty of Humanities and Head of Formal Linguistics Lab at the National Research University “Higher School of Economics”. Her research interests include NLP, corpus linguistics, discourse analysis and linguistic typology.

#### Contact information:

*e-mail*: [toldova@yandex.ru](mailto:toldova@yandex.ru)

room A-114, Building 1, 21/4 Staraya Basmannaya Ulitsa, Moscow, 117218, Russia

ORCID: 0000-0002-5777-9161

**Natalia A. SLIOUSSAR** is Doctor Habil., Associate Professor of the School of Linguistics at the Faculty of Humanities, Associate Professor of the Department of the Problems of Convergence in Natural Sciences and Humanities at St. Peterburg State University. Her research interests embrace psycholinguistics, syntax and morphology.

#### Contact information:

*e-mail*: [slioussar@gmail.com](mailto:slioussar@gmail.com)

room 518, Building 1, 21/4 Staraya Basmannaya Ulitsa, Moscow, 117218, Russia

ORCID: 0000-0003-1706-6439

**Anastasia A. BONCH-OSMOLOVSKAYA** is an Associate Professor of the Faculty of Humanities at the Higher School of Economics, Moscow, Russia. She holds a PhD in Language Theory. Her research interests include digital humanities, computational linguistics, corpus linguistics, theoretical linguistics and quantitative methods in linguistics.

**Contact information:**

room 521, Building 1, 21/4 Staraya Basmannaya Ulitsa, Moscow, 117218, Russia

*e-mail:* abonch@gmail.com

ORCID: 0000-0001-5826-8286

**Сведения об авторах:**

**Светлана Юрьевна ТОЛДОВА** – кандидат филологических наук, доцент Школы лингвистики факультета гуманитарных наук НИУ ВШЭ, заведующая научно-учебной лабораторией по формальным моделям в лингвистике. Ее научные интересы включают компьютерную лингвистику, корпусную лингвистику, анализ дискурса, лингвистическую типологию.

**Контактная информация:**

*e-mail:* toldova@yandex.ru

Россия, 117218, Москва, Старая Басманная ул., д. 21/4, стр. 1, каб. А-114

ORCID: 0000-0002-5777-9161

**Наталья Анатольевна СЛЮСАРЬ** – доктор филологических наук, доцент Школы лингвистики факультета гуманитарных наук НИУ ВШЭ, доцент кафедры проблем конвергенции естественных и гуманитарных наук, старший научный сотрудник института когнитивных исследований СПбГУ. В сферу ее научных интересов входят психолингвистика, морфология и синтаксис.

**Контактная информация:**

*e-mail:* slioussar@gmail.com

Россия, 117218, Москва, Старая Басманная ул., д. 21/4, стр. 1, каб. А-114

ORCID: 0000-0003-1706-6439

**Анастасия Александровна БОНЧ-ОСМОЛОВСКАЯ** – доцент факультета гуманитарных наук университета «Высшая школа экономики» (Москва), кандидат филологических наук по специальности «Теория языка». Сфера ее научных интересов включает цифровую гуманитаристику, компьютерную лингвистику, корпусную лингвистику, теоретическую лингвистику, количественные методы в лингвистике.

**Контактная информация:**

Россия, 117218, Москва, Старая Басманная ул., д. 21/4, стр. 1, каб. 521

*e-mail:* abonch@gmail.com

ORCID: 0000-0001-5826-8286