Lexical, morphological and semantic predictability analysis in different age groups

A. Fayer-Maximova
HSE University
Moscow, Russia
aofaier@edu.hse.ru

P. Lazukova
HSE University
Moscow, Russia
pilazukova@edu.hse.ru

K. Vashpanova
HSE University
Moscow, Russia
kvvashpanova@edu.hse.ru

E. Klyshinsky
HSE University
Moscow, Russia
eklyshinsky@hse.ru

Introduction

As part of the study of the mental lexicon, there is a task to study lexical predictability. The purpose of this task is to assess how correctly native speakers predict the next word based on the left context, and it is not so much the specific word that is important as its semantic class or morphological characteristic. Despite the relevance of the topic, there have been practically no such studies over the past 15 years, and they mainly evaluate adult speakers. Almost all works about children are mostly about their readability and speech, which are not the same topics with predictability of words. In our topic area there are few works by Anastasiya Lopuhina, which we will describe in the literature part and from where we took experiment design with left context predictability.

In our work we compared lexical, morphological and semantic predictability among schoolchildren from 4th to 11th grade and adults. We compared both schoolers between themselves and children with adults. Also, we analyzed some factors other than school class and age that can affect word predictability: gender, native language, age of starting reading and amount of reading per week.

Literature review

The study of lexical predictability has garnered significant attention in recent years due to its implications for understanding language processing. Lexical predictability refers to the pre-activation of forthcoming words, including their meaning and, to some degree, their form [1]. Historically, this concept has been explored through various lenses, including psycholinguistics, cognitive psychology, and computational linguistics. Understanding lexical predictability is essential for advancing our knowledge of cognitive mechanisms underlying language comprehension. Of particular interest within the framework of this paper is the consideration of lexical predictability among school-age children. This interest is primarily motivated by our intention to identify the distinctive features of children's linguistic development. This literature review aims to synthesize current research on lexical predictability,

with a particular focus on its cognitive mechanisms and impact on children's reading comprehension.

Measuring lexical predictability can be approached through various methodologies, each offering unique insights into language processing. Eye-tracking technology is commonly used to assess how predictability influences reading behavior by monitoring eye movements, fixation durations, and saccades as readers encounter predictable and unpredictable words [2]. This method provides real-time data on cognitive load and processing efficiency. Another prevalent technique is the cloze test, where participants are asked to fill in blanks within a passage [3]. The predictability of a word is gauged by the frequency with which participants correctly anticipate it, reflecting their ability to use contextual cues. Additionally, advances in neuroimaging techniques, such as event-related potentials (ERPs) and magnetoencephalography (MEG), allow researchers to measure brain impulses associated with lexical predictability. These methods capture the neural responses that occur when the brain processes predictable versus unpredictable words, offering a window into the temporal dynamics and underlying neural mechanisms of language comprehension [4]. Each of these approaches provides complementary perspectives, enhancing our understanding of how predictability influences language processing.

The study of lexical predictability should include an examination of human brain activity during the information processing procedure as there are studies reporting a correlation between the unexpectedness of the word and the neural activity associated with word processing [5]. The N400 is an event-related potential (ERP) component that is critical for understanding the neural mechanisms underlying language comprehension, particularly in the context of lexical predictability.

N400 is a negative-going wave that peaks around 400 milliseconds after the presentation of a word or other meaningful stimulus, and it is typically larger for words that are less predictable or incongruent within a given context [4]. The N400 is important as it provides direct insight into the brain's real-time processing of meaning and context during language comprehension. By studying the N400, researchers can gain valuable information about how the brain integrates semantic information and predicts upcoming words.

As presented in Figure 1, the amplitude of N400 is greatest when the predicted word is least relevant in the given left context. The appropriateness of a word is determined by the speaker's linguistic experience.

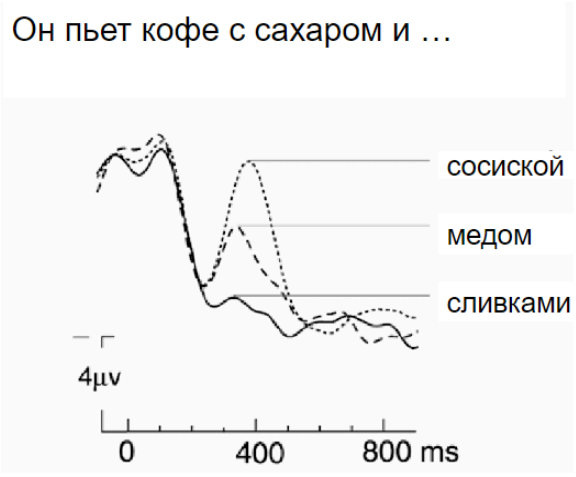


Fig. 1. N400 amplitude

This knowledge is essential for developing theories of language processing, improving educational strategies for reading comprehension, and designing interventions for

language-related disorders. According to [6] N400 is a marker of prediction and that prediction is an essential aspect of sentence comprehension. The existence of the N400 component proves the relevance of our study as it provides a rationale for the theory that predictability is a graded and probabilistic process [5].

Examination of brain activity and evidence of the importance of the N400 component in speech comprehension allows us to narrow the field of research to the developmental features of speech among children. The study was based on the research of Lopukhina [3] where a similar experiment was conducted. The key finding of the paper partially confirmed the hypotheses about the influence of literacy level in Russian on prediction ability - higher literacy level in Russian facilitated prediction, which was evident in the earliest stages of eye tracking. In our paper, we modify the purpose of the experiment to examine the differences in lexical predictability across various age groups.

In summary, the literature on lexical predictability underscores its significant role in language processing, impacting reading comprehension and cognitive mechanisms. Studies employing methodologies such as eye-tracking, cloze tests, and neuroimaging techniques have provided valuable insights into how predictability influences language understanding. Despite the wealth of research, there are still gaps in our knowledge, particularly regarding how lexical predictability operates across different developmental stages. In our paper, we will focus on comparing lexical predictability within different age groups to better understand these developmental differences. Future research should continue to explore these age-related differences in more detail and utilize advanced neuroimaging methods to further elucidate the neural underpinnings of lexical predictability. By addressing these gaps, we can better understand the complexities of language processing and enhance educational and clinical approaches to language-related challenges.

Respondents

As respondents, 501 native speakers of the Russian language were interviewed, from 8 years to 54 years old. There were 146 adults. Also 355 children under the age of 18 were schoolchildren from 2nd to 11th grade from three educational schools: Moscow State School 57, Moscow State School "Letovo" and Moscow State School 91. The specialization of the classes ranged from general education to humanities and mathematics. The youngest children answered the experiment with a little help from their parents, who read them sentences.

Grade	**Number of respondents**
2	7
3	1
4	15
5	34
6	50
7	41
8	59
9	59
10	72
11	17

Table 1. Grades of respondents

There were 316 women and 185 men in total. Also, the majority of respondents indicated Russian as their native language, 3 — Tartarian, 2 – Kyrgyz and 1 – Belarusian.

In addition to the above information, respondents noted the amount of time they spend reading weekly and the age at which they started reading. Both of these characteristics were required in the study to assess the respondent' level of reading literacy.

Amount of reading time per week (min)	**Number of respondents**
10	234
30	98
60	104
90	65
Over 90	150

Table 2. Amount of reading time (number of respondents)

Age at which reading was started	**Number of respondents**
1	4
3	71
4	116
5	157
6	100
7	28
8	2
9	3
10	1

Table 3. Age at which reading was started (number of respondents), 18 incorrect answers

Design of the experiment

The main goal of the experiment was the prediction of the word from the left, previous context. In total, the experimental material contained 35 sentences from Russian adolescent literature, unrelated to each other, see them in Appendix 1. Each respondent was provided with 10 random sentences. Starting from the first sentence, the respondent was first presented with an empty field where any word could be entered, and then it was necessary to guess the word based on the previous already known words — the left context, consisting first of 1, 2, etc. words.

For example, there was the first word «Дорога» (the road) and then the respondent should predict some word, probably verb. Then the respondent was given two words «Дорога вела» (the road led to) and should predict next word, relying on given left context. This procedure repeated until the end of the sentence and then the whole process began again with a new sentence.

In this experiment we analyze not only correctness of the predicted word form, but also its semantic and morphological correctness (if it was in the correct semantic class or had needed case, number etc.).

Results

Data preprocessing

For the first 20000 lines of the dataset misprints and misspellings were manually corrected, inadequate answers were excluded (humoristic, absurdist etc.).

After, for each answer lemma, grammar and semantics were automatically extracted with Python. First, the word was analyzed with pymystem3 with automatic disambiguation: lemma, POS tag and grammatical features were extracted to separate columns of the dataframe. In grammatical tags all possible sets of tags were listed. Secondly, semantics were extracted with ruwordnet package. It the semantics column all possible hierarchical classes chains were listed. Thirdly, lexical, lemmatic, part-of-speech, grammatical and semantic predictability was calculated.

By lexical predictability we understand the probability of a full word form being guessed. The response was coded as ‘1’ only in the case when it fully matched the target word orthographically, otherwise it was coded as ‘0’.

Lemmatic predictability is the probability of a lemma being guessed, even if the grammatical form of the answer differs from the target. The response was binary coded (‘1’—the guessed word lemma matched the target one, ‘0’—the guess did not match).

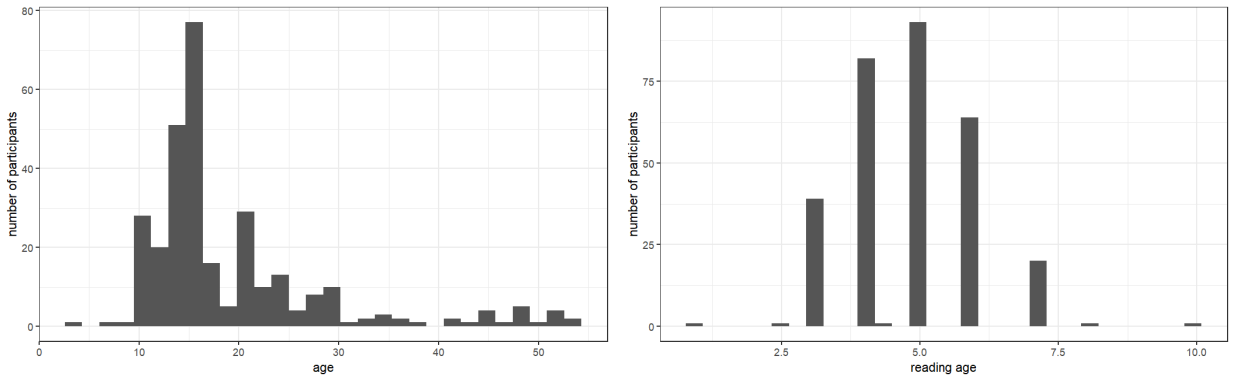
Part-of-speech predictability is the probability of the target part-of-speech tag being guessed. The response was binary coded (‘1’—the guessed word part-of-speech tag matched the target one, ‘0’—the guess did not match).

Grammatical predictability shows to what extent other than part-of-speech tag grammatical features of a word are guessed. The response was coded as the highest proportion of the target grammatical tags guessed among all possible tag sets for the answer given.

Semantic predictability shows to what extent semantic classes of a word are guessed. The response was coded as the highest proportion of the target semantic classes guessed among all possible classes chains for the answer given and for the target word.

For each participant the mean value of each type of predictability was calculated. Before the analysis, we filtered only those participants who indicated Russian as their native language and made guesses for at least 20 words. The participants older than 19 years were automatically assigned “adult” (grade=12) instead of the grade they specified. The participants that indicated their age being <1 or >100 or reading age being <1 or >age were excluded from the analysis. Also participants from 1–3 grades were excluded since their number was insufficient (6 people in total).

The analyzed data statistics can be seen in Figure 2.



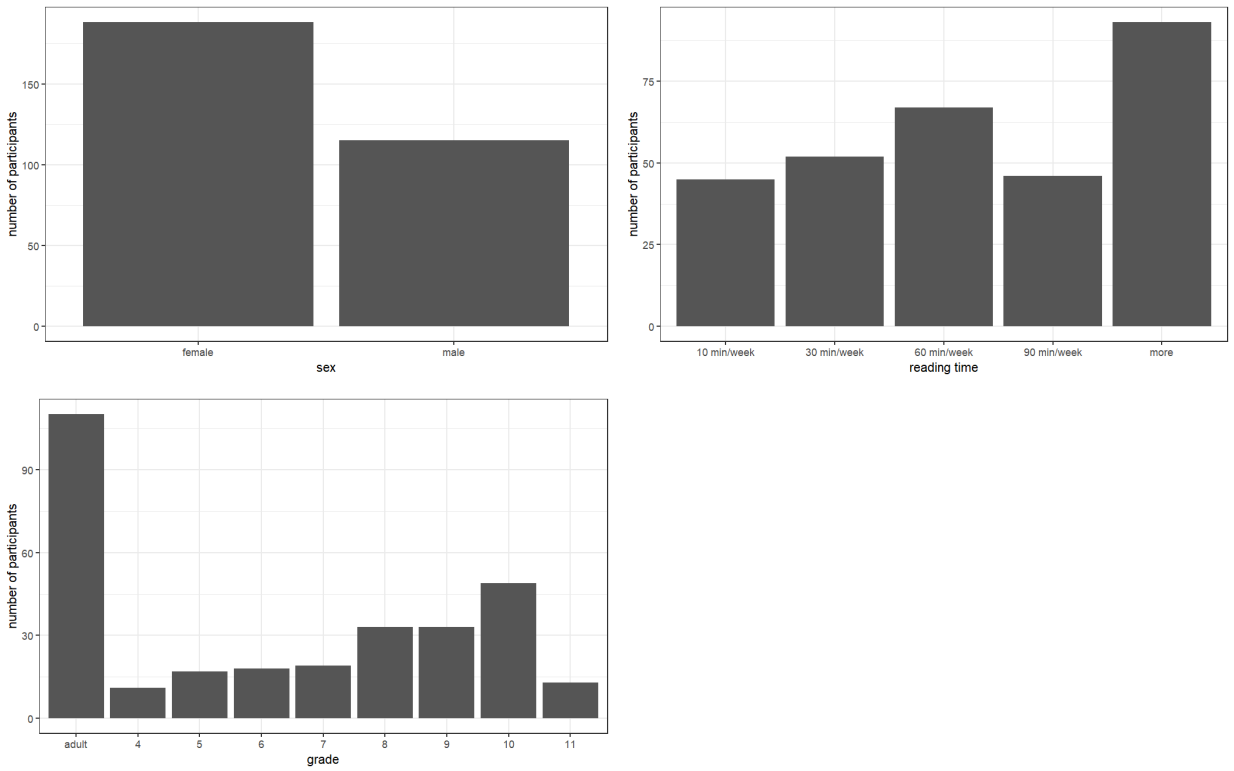


Fig. 2. Number of participants by age, reading age, sex, reading time and grade

Predictability on the full dataset

We observe that the least predictability is lexical, which is rather expected since it has the strictest criteria. Lemmas are predicted slightly better: this is probably caused by unpredictability of semantic grammatical features such as number. Semantic, on the other hand, is predicted much better, which may be a result of semantic accuracy calculation procedure (non-zero result can be obtained rather easily, since there are few higher level semantic classes). Part-of-speech and grammatical predictability are the highest, which agrees with the previous studies results [3]. Grammatical predictability may be higher than part-of-speech because of noun vs. adjective mixing, which often affects only part-of-speech tags, but not other grammatical features.

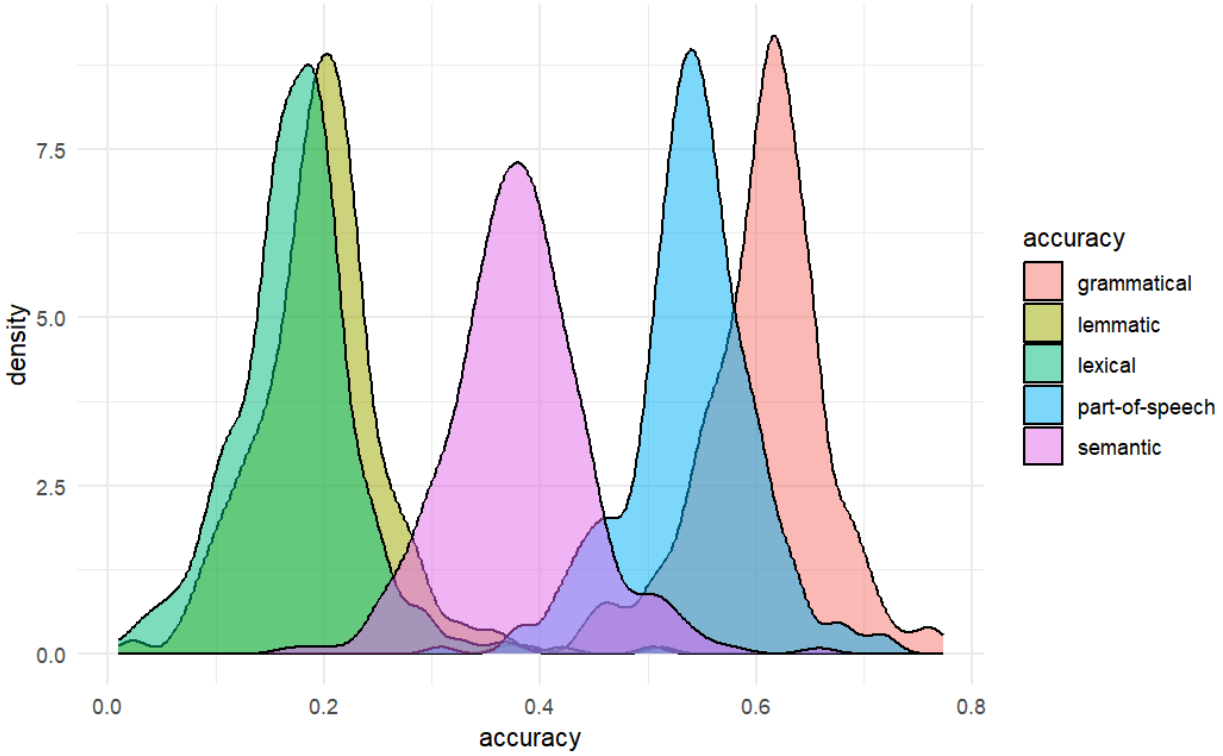
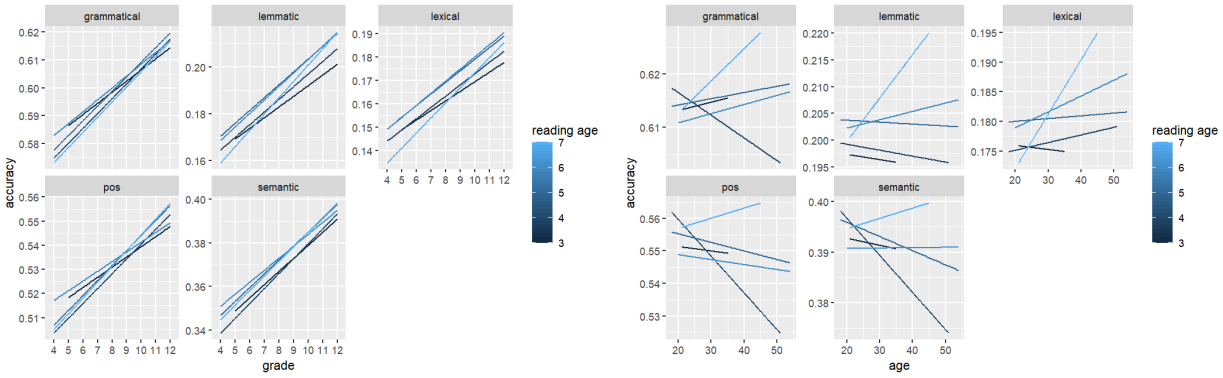


Fig. 3 Predictability density plot

We performed all data analysis in R (version 4.4.0; **R Core Team 2021**). We used Anova test implemented through R function *rstatix::anova_test* and Tukey test through *stats::TukeyHSD*.

First, we checked for possible interactions between predictors (age or grade and reading time, reading age, sex). Figure 4 shows that for age there are possible interactions with reading age and reading time for all types of predictability, and with sex for grammatical, lemmatic and lexical predictability. For grade, possible interaction with reading age is observed for lemmatic, lexical and part-of-speech predictability, with reading time – for all types of predictability except for lexical one, with sex – for lemmatic and lexical predictability.



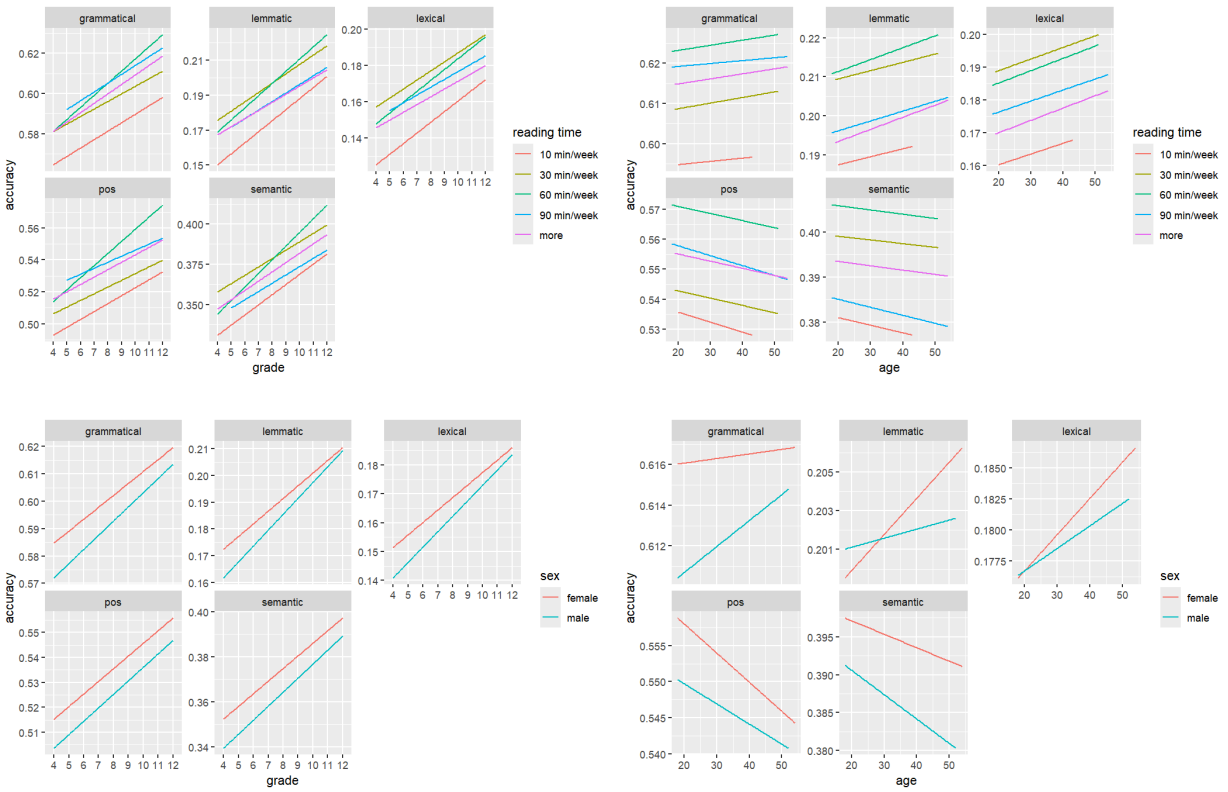


Figure 4. Interaction plots for grade and age

Then, taking into account the possible interactions observed, ANOVA test was conducted for each type of predictability for age and for grade separately. Results for predictors that proved to be statistically significant can be seen in Table 4.

Variant	Effect	DFn	DFd	F	p	p<.05	ges
gramatical							
grade	grade	8	258	3.082	0.002	*	0.087
age	age	1	289	4.142	0.043	*	0.014
	age:r_time	4	289	2.417	0.049	*	0.032
lemmatic							
grade	grade	8	242	4.480	4.37e-05	*	0.129
	grade:r_time	30	242	1.521	4.60e-02	*	0.159
age	age:r_time	4	289	2.465	0.045	*	0.033
lexical							
grade					0.00010		
	grade	8	250	4.163	8	*	0.118
age	age	1	289	4.243	0.040	*	0.014
part-of-speech							
grade	grade	8	250	3.100	0.002	*	0.090
	r_time	4	250	2.797	0.027	*	0.043
age	age	1	290	4.395	0.037	*	0.015
	r_time	4	290	2.604	0.036	*	0.035
semantic							

grade					0.00058		
	grade	8	280	3.568	6	*	0.093
age	age	1	290	6.000	0.015	*	0.020

Table 4. Significant predictors for predictability types

We observe that grade has a significant effect on all predictability types, with medium-size effect on lemmatic and lexical predictability and small-size effects for other types. We should also note that there is a medium-size effect of interaction between grade and reading time on lemmatic predictability. Small-size effects of age are observed for all predictability types except for the lemmatic one, however there is an effect of age and reading time interaction on it, which is also observed for grammatical predictability. Reading time itself only has small-size effect on part-of-speech predictability. No effects of sex and reading age are observed.

Difference between grades

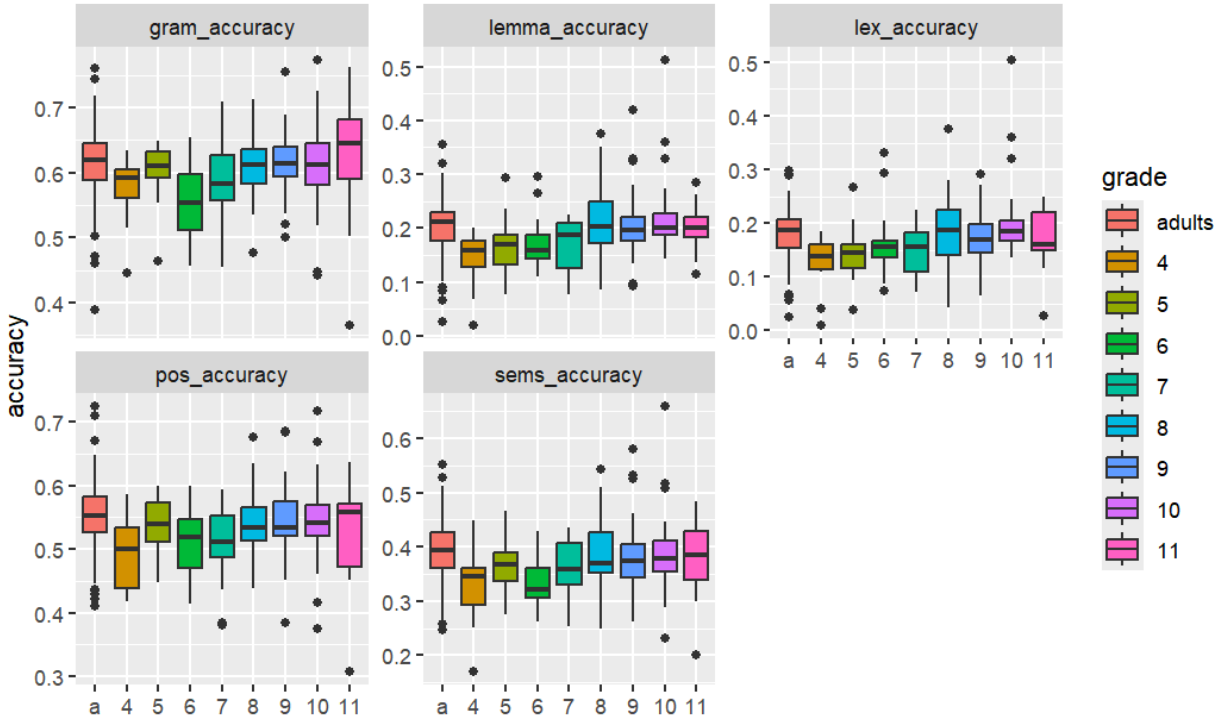


Figure 5. Accuracy by predictability type and grade (including adults)

In Figure 5 we observe that accuracy of prediction tends to grow with the higher grades up to the adults, although this tendency is not strict and is sometimes violated by the 5th and 6th grades.

Within the schoolchildren subgroup grade has a significant effect on all types of predictability: middle-size effects on lemmatic ($F(7, 185) = 3.933$, $p < .001$, ges = .13) and lexical predictability ($F(7, 185) = 3.461$, $p < .01$, ges = .116) and smaller size effects on grammatical ($F(7, 185) = 2.781$, $p < .01$, ges = .095), part-of-speech ($F(7, 185) = 2.072$, $p < .05$, ges = .073) and semantic ($F(7, 185) = 2.574$, $p < .05$, ges = .089) predictability. Pairwise Tukey test shows that there is difference in:

- lemmatic predictability between 4th and 8th (-0.066, p<.05), 9th (-0.064, p<.05), 10th (-0.075, p<.01) grade
- lexical predictability between 10th and 4th (0.073, p<.01), 5th (0.052, p<.05), 7th (0.049, p<.05) grade

- grammatical predictability between 6th and 9th (-0.062, p<.01), 10th (-0.057, p<.05) grade

We observe the difference between junior and senior classes, which may become more significant as we collect more data for younger participants. For example, a plot of Tukey test confidence intervals for grammatical predictability is shown in Figure 6.

Fig. 6. Plot of Tukey test confidence intervals for grammatical predictability

It seems that with more data for testing the results may become more confident for closer grades, although this hypothesis needs further investigation.

When compared to the adult group, only 4th (-0.060 for lemmatic, -0.055 for lexical, -0.060 for part-of-speech and -0.068 for semantic) and 6th (-0.061 for grammatical and -0.057 for semantic) grade have significantly lower predictability results (p<.05).

Predictability on the adult data

The grade factor being removed, the adult participants' data shows less homogeneity in predictors effects (see Table 5). Lexical and lemmatic predictability are mostly affected by reading time (middle size effect) with significant interaction between reading time and age. They are also influenced by sex and interaction between age and reading age (small size effects). Part-of-speech predictability is solely influenced by interaction of age and reading age, and semantic predictability is solely influenced by sex. No significant predictors are observed for grammatical predictability.

Predictor	DFn	DFd	F	p	p<.05	ges
lexical						
r_time	4	96	3.094	0.019	*	0.114
sex	1	96	6.795	0.011	*	0.066
age:r_tim						
e	4	96	2.953	0.024	*	0.110

age:r_age	1	96	4.552	0.035	*	0.045
lemmatic						
r_time	4	96	3.587	0.009	*	0.130
sex	1	96	8.844	0.004	*	0.084
age:r_time	4	96	2.942	0.024	*	0.109
age:r_age	1	96	7.086	0.009	*	0.069
part-of-speech						
age:r_age	1	96	5.254	0.024	*	0.052
semantic						
sex	1	96	8.846	0.004	*	0.084

Table 5. Predictors effect on adult participants' data

From Figure 6 we observe that male participants are less accurate in their predictions with significant difference ($p<.05$) for lexical, lemmatic and semantic predictability (significance level from Table 5).

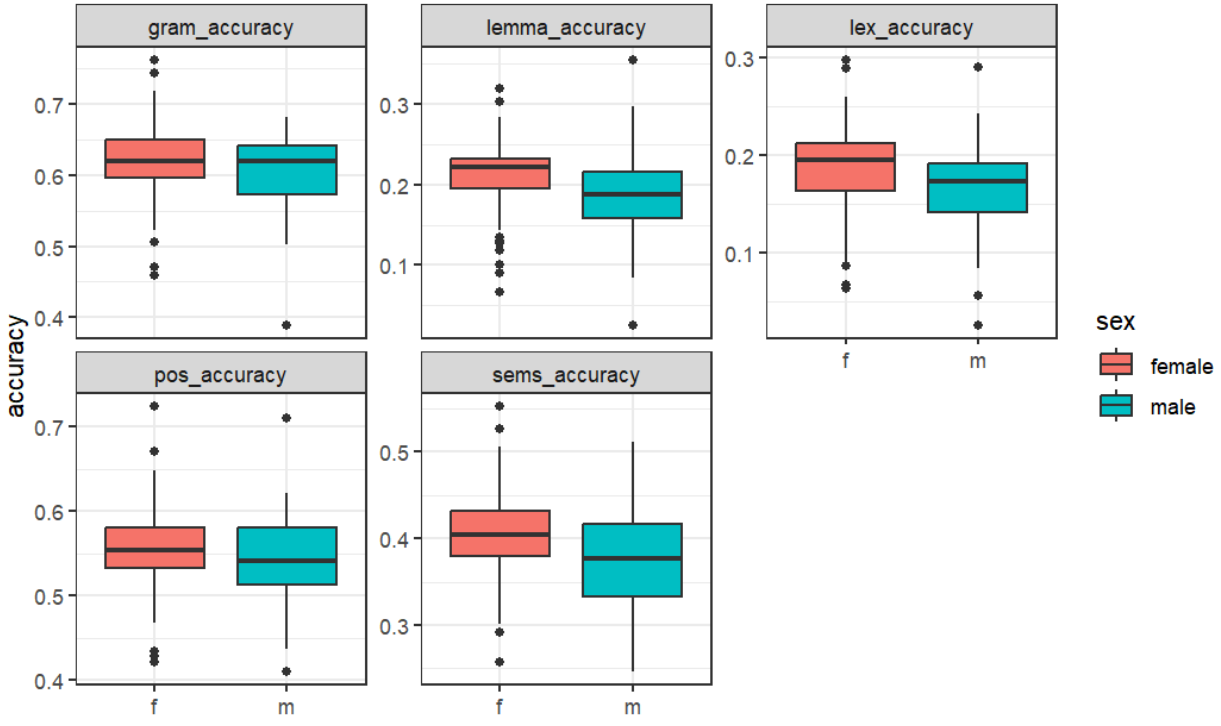


Fig. 6. Difference in predicting accuracy between men and women

For reading time, Tukey test show significant difference ($p<.05$) between 10 min/week and 30 min\week and between 10 min/week and 90 min/week in lemmatic predictability (-0.057 and -0.055 respectively), and between 10 min/week and 30 min/week in lexical predictability (-0.050).

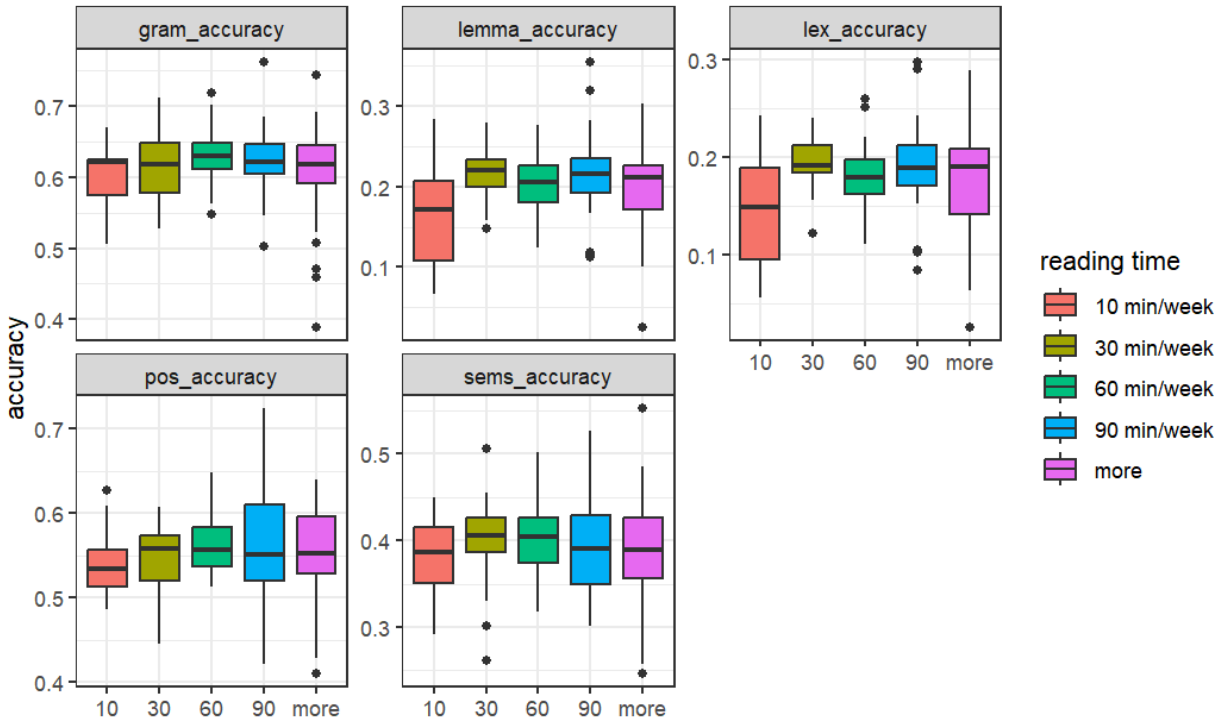


Fig. 7. Difference in predicting accuracy between reading time levels

The interaction effects mentioned above can be seen in Figure 8, although data for reading age may be insufficient or unbalanced.

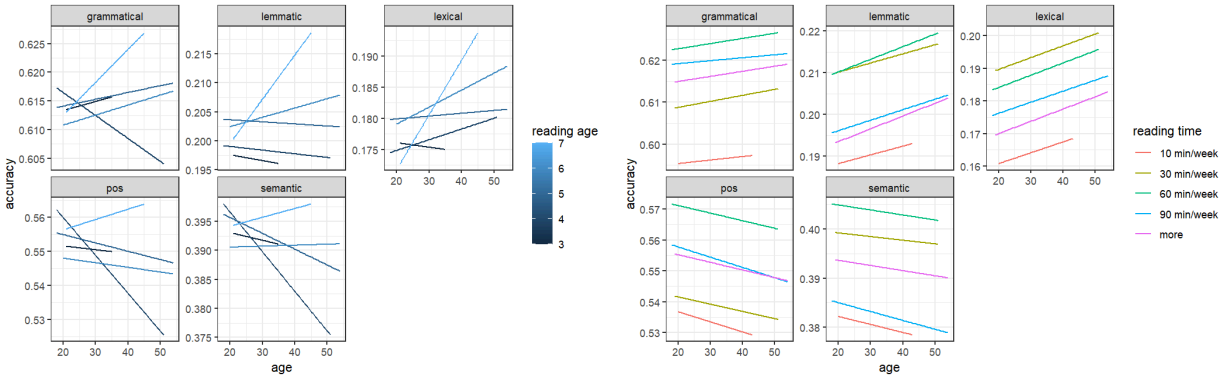


Fig. 8. Interaction effects between age and reading age and between age and reading time on adult participants data

Discussion

In this paper we have analyzed various types of predictability in different grades and ages and factors that may affect it. Grammatical and part-of-speech predictability types showed the highest accuracy, which is consistent with previous studies in this area [3]. this results from the fact that the previous word (esp. verbs, prepositions and adjectives) usually give let to determine case and to some extent number and gender of the upcoming word. Part-of-speech may have lower accuracy due to the fact that it is not always easy to determine the upcoming syntactic group members (ex. we can hardly ever guess whether the upcoming noun group starts with a noun or an adjective), though we expect some other grammatical features (ex. case).

As for the accuracy predictors, grade, and to a lesser degree age and the amount of time a person spends reading for themselves proved to be the most significant. Grade significantly affects all types of predictability, while reading time is significant for part-of-speech predictability in general and for lexical and lemmatic predictability among adults. Age greatly loses its significance when only the adult group is analyzed, which implies that there is no much predictability development after school. The difference between grades is most prominent in comparing junior classes (4-7) to senior ones (9-10), while senior classes do not differ significantly from the adults. This tendency was also noted in previous research [7].

The least significant factors are sex and age, at which the person started reading, although they still show some significance on the adult data. We may say that women predict the upcoming word (the word form, lemma and semantics) slightly better than men.

Also, we discussed vocabulary size of the respondents and reading comprehension, as possible factors of correct predictability. However, we suppose it is difficult to estimate them remotely, so for now these factors haven't been used.

It should be mentioned that we interviewed scholars from advanced schools, however, there were respondents from different classes: mathematical, liberal arts and historical and also common classes without any concrete direction of study. So we assume that results are representative. Nevertheless, we plan to widen the range of respondents by interviewing scholars from non-advansed schools, from other regions and possibly non-native speakers as a different type of respondents.

In the future research more data on younger classes (1-3) should be collected, although this may be difficult due to the technical reasons. We assume that younger classes will confirm our results that the older respondents are, the higher the success rate of the predictability. Also we should conduct a more detailed analysis of which grammatical and semantic features are predicted better and what features are predicted the most often, at least in the context of our stimuli. This future research may be helpful not only for the fields of psycho- and neurolinguistics, but also for computational linguistics, and language models assessment in particular.

Literature

1. Kochari, A. R., & Flecken, M. (2019). Lexical prediction in language comprehension: A replication study of grammatical gender effects in Dutch. *Language, Cognition and Neuroscience*, *34*(2), 239-253.
2. Lopukhina, A., Lopukhin, K., & Laurinavichyute, A. (2021). Morphosyntactic but not lexical corpus-based probabilities can substitute for cloze probabilities in reading experiments. Plos one, 16(1), e0246133.
3. Parshina, O., Lopukhina, A., & Sekerina, I. A. (2022). Can heritage speakers predict lexical and morphosyntactic information in reading?. *Languages*, *7*(1), 60.
4. Sekerina, I. (2002). The Method of Event-Related Potentials in American Psycholinguistics and Its Application to Word Order in Russian. In Proceedings of the International Conference Dialogue (Vol. 2002).
5. Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension?. *Language, cognition and neuroscience*, *31*(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299
6. Szewczyk, J. M., & Schriefers, H. (2018). The N400 as an index of lexical preactivation and its implications for prediction in language comprehension. *Language, Cognition and Neuroscience*, *33*(6), 665–686. https://doi.org/10.1080/23273798.2017.1401101

7. Герен, С., & Лопухина, А. (2020). Предсказуемость слов в контексте. [PowerPoint slides]. Google Presentations. https://docs.google.com/presentation/d/1WnBPGILpWdb1FOf_xDAWZnVhInzQ5eD3248Cn2n4fGs/edit#slide=id.p1

Appendix 1

The stimulus sentences

1. Брошенный мальчиком снежок попал в окно второго этажа.
2. В вопросе командира ясно слышался упрёк солдату.
3. В воскресенье вся дружная семья поехала на дачу.
4. В грязной воде микробы размножаются особенно быстро.
5. В доме лесника охотники нашли крупу, сухари и спички.
6. В каждом углу комнаты сидели по две кошки.
7. В магазине Андрей купил молоко, сметану, творог.
8. Вася любил сгущенку, особенно с чаем.
9. Вдалеке за рекой виднелись крыши старинного замка.
10. Девочка никак не могла вынуть соринку, попавшую в глаз.
11. Дорога вела в глухой лес, петляя по склонам.
12. Ей никак не суметь испечь такой торт самой.
13. Коробку с подарками украшал бант огромного размера.
14. Мне было лень идти сметать снег, лежавший на машине.
15. На диване лежало покрывало ярко-зеленого цвета.
16. На кустах, росших по берегам реки, появились листочки.
17. Недалеко был сложен стог сена, рядом стояли грабли.
18. Обещают, что в этом году лето будет жарким.
19. Они заметили вагон красного цвета и переглянулись
20. Они зашли к маминой подруге, которая жила рядом.
21. От смерти его спасла собака, которая приносила ему еду.
22. Перед поворотом мопед затормозил и остановился.
23. Петя никак не мог доделать домашнюю работу.
24. Причиной аварии был мобильный телефон, который отвлекал водителя.
25. С самой первой страницы история захватывает читателя.
26. Сломанную вчера розетку сумел починить только электрик.
27. Среди всякого сора девочка нашла разноцветные камушки.
28. Старые ступеньки лестницы скрипели и выглядели ненадежными.
29. Там, недалеко от кухонной двери, сидел лис с треугольной мордой.
30. У них был уютный, спокойный дом, крепкая семья.
31. Что используют для этой прически, фен или ещё что?
32. Чтоб было удобнее, поправь ремешок своего рюкзака.
33. Я купил специальную мазь, которая помогает при ожогах.
34. Я отдал последнюю монету, найденную в кармане.
35. Я увидел осу, летавшую вперед и назад по комнате.