81`233

Анализ лексической, морфологической и семантической предсказуемости в разных возрастных группах

Вашпанова Кристина Викторовна[1] kvvashpanova@edu.hse.ru
Файер-Максимова Анастасия Олеговна[1] aofaier@edu.hse.ru
Лазукова Полина Игоревна[1] **(*)** pilazukova@edu.hse.ru
Скрипкарь Юлия Валерьевна[1] iskripkar@edu.hse.ru
[1]**Национальный исследовательский университет «Высшая школа экономики», Москва, Россия**

Аннотация

В статье исследуется лексическая, морфологическая и семантическая предсказуемость слов у школьников с 4 по 11 класс и взрослых носителей языка. Основное внимание уделяется влиянию возраста, класса обучения, времени, затрачиваемого на чтение, пола и возраста начала чтения на способность предсказывать слова в контексте. Результаты показывают, что грамматическая и частеречная предсказуемость имеют наибольшую точность, особенно у старших школьников и взрослых. Различия между возрастными группами наиболее выражены между младшими и старшими классами, тогда как старшеклассники мало отличаются от взрослых. Пол и возраст начала чтения оказывают незначительное влияние, хотя женщины в среднем демонстрируют лучшую предсказуемость. Представленные результаты могут быть полезны для психо- и нейролингвистики, а также для оценки языковых моделей в вычислительной лингвистике.

Ключевые слова

Возрастные различия в предсказуемости, лексическая предсказуемость, морфологическая предсказуемость, семантическая предсказуемость, когнитивные процессы в языке, языковая обработка, психолингвистика.

Lexical, morphological and semantic predictability analysis in different age groups

Kristina Vashpanova[1] kvvashpanova@edu.hse.ru
Anastasia Fayer-Maximova[1] aofaier@edu.hse.ru
Polina Lazukova[1] **(*)** pilazukova@edu.hse.ru
Julia Skripkar[1] iskripkar@edu.hse.ru
[1]**National Research University Higher School of Economics, Moscow, Russia**

Abstract

The article examines lexical, morphological, and semantic predictability among schoolchildren from grades 4 to 11 and adult native speakers. The study focuses on the influence of age, school grade,

reading time, gender, and age of starting reading on the ability to predict words in context. Results show that grammatical and part-of-speech predictability have the highest accuracy, particularly among senior students and adults. Differences between age groups are most pronounced between junior and senior students, while senior students differ minimally from adults. Gender and age of starting reading have minor effects, though women generally exhibit slightly better predictability. The findings have implications for psycholinguistics, neurolinguistics, and the evaluation of language models in computational linguistics.

Keywords

Age differences in predictability, lexical predictability, morphological predictability, semantic predictability, cognitive processes in language, language processing, psycholinguistics.

Введение

Изучение лексической предсказуемости — одно из направлений исследования ментального лексикона. В таких экспериментах изучается способность носителя языка предугадать слово на основании левого контекста, причём важно не столько само слово, сколько его семантический класс и морфологические характеристики. За последние 15 лет таких исследований было проведено немного, и все они проводились на взрослых испытуемых. Большинство работ про носителей-детей сводятся к изучению чтения и устной речи, что не пересекается с нашей темой. В нашем исследовании мы опирались на работы Анастасии Лопухиной [1], и сравнивали лексическую предсказуемость у детей с 4 по 11 класс и взрослых. Помимо класса и возраста испытуемых мы анализировали пол, опыт и время чтения в неделю.

Материалы и методы

Часто исследования предсказуемости осуществляются с использованием методов айтрекинга (eye-tracking) [1, 2], который позволяет отслеживать движения глаз для изучения когнитивных процессов, и анализа компонента N400 [3], связанного с обработкой семантической информации в мозге и регистрируемого через электрическую активность (ЭЭГ). Однако в данном исследовании применяется иной подход.

Для оценки предсказуемости на основании левого контекста мы использовали 35 несвязанных друг с другом предложений на русском языке. Каждому из испытуемых показывались 10 случайных предложений. В каждом из 10 предложений сначала нужно было написать первое слово, затем, увидев корректное первое слово, угадать второе, потом на основание первых двух слов — третье и так далее.

В этом эксперименте мы анализировали не столько корректность самого слова, сколько угадывание его семантического класса и морфологических характеристик (например, части речи, времени глагола, рода и числа существительных и прилагательных).

Всего в эксперименте принял участие 501 испытуемый, из них 146 взрослых и 355 детей с 4 по 11 класс из московских школ "Летово", №57 и №91. Специализация классов варьировалась от стандартной до гуманитарной и математической. Среди испытуемых было 316 женщин и 185 мужчин, для 6 из них русский был не родным языком.

Результаты

Анализ всего корпуса

Наименьшими значениями обладают лексическая и лемматическая предсказуемость, несколько лучше предсказывается семантика следующего слова. Наиболее успешно участники предсказывали часть речи и грамматические характеристики.

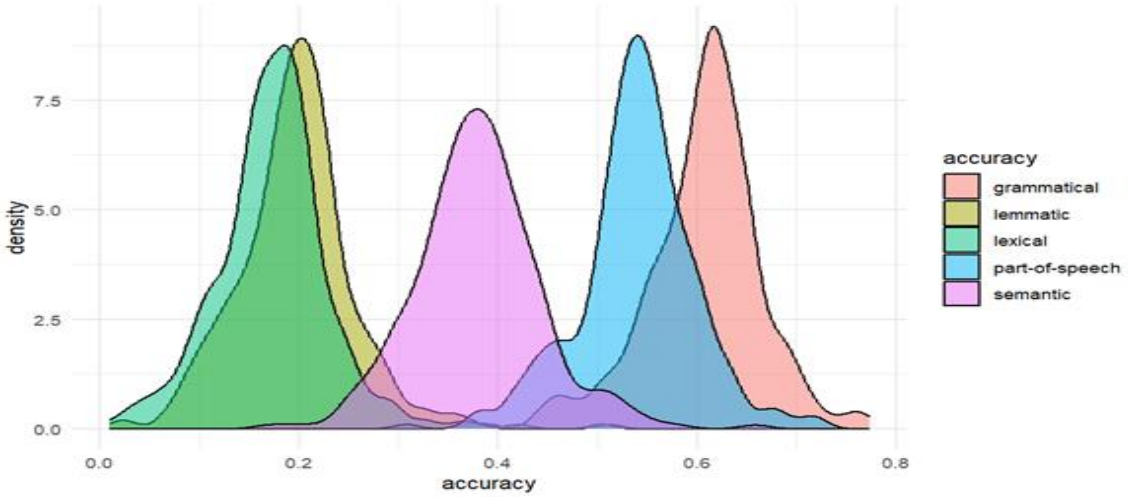

Рисунок 1. График плотности для разных типов предсказуемости

Поскольку возраст и класс в большой степени дублируют друг друга, раздельно для каждого из них в сочетании с остальными параметрами был проведён тест ANOVA для всех типов предсказуемости.

Таблица 1

Значимые параметры для разных типов предсказуемости

Предсказуемость	Вариант	Предиктор	DFn	DFd	F	p	ges
грамматическая	*класс*	класс	8	258	3.082	0.002	0.087
	возраст	возраст	1	289	4.142	0.043	0.014
		возраст:ВЧ	4	289	2.417	0.049	0.032
лемматическая	*класс*	класс	8	242	4.480	4.37e-05	0.129
		класс:ВЧ	30	242	1.521	4.60e-02	0.159
	возраст	возраст:ВЧ	4	289	2.465	0.045	0.033
лексическая	*класс*	класс	8	250	4.163	0.000108	0.118
	возраст	возраст	1	289	4.243	0.040	0.014
частеречная	*класс*	класс	8	250	3.100	0.002	0.090
		ВЧ	4	250	2.797	0.027	0.043
	возраст	возраст	1	290	4.395	0.037	0.015
		ВЧ	4	290	2.604	0.036	0.035

Предсказуемость	Вариант	Предиктор	DFn	DFd	F	p	ges
семантическая	*класс*	класс	8	280	3.568	0.000586	0.093
	возраст	возраст	1	290	6.000	0.015	0.020

По результатам теста мы наблюдаем средний эффект класса на лемматическую предсказуемость и малый эффект на остальные типы. Также средний значимый эффект на лемматическую предсказуемость оказывает взаимодействие класса и ВЧ. Малый значимый эффект возраста наблюдается для всех типов предсказуемости, кроме лемматической. Также можно видеть эффект взаимодействия возраста и ВЧ на лемматическую и грамматическую предсказуемость. Наблюдается малый эффект ВЧ на частеречную предсказуемость. Для пола и ВНЧ значимых эффектов не выявлено.

Анализ влияния класса

На рис. 2 наблюдается тенденция к увеличению точности предсказания от младших классов к взрослым.

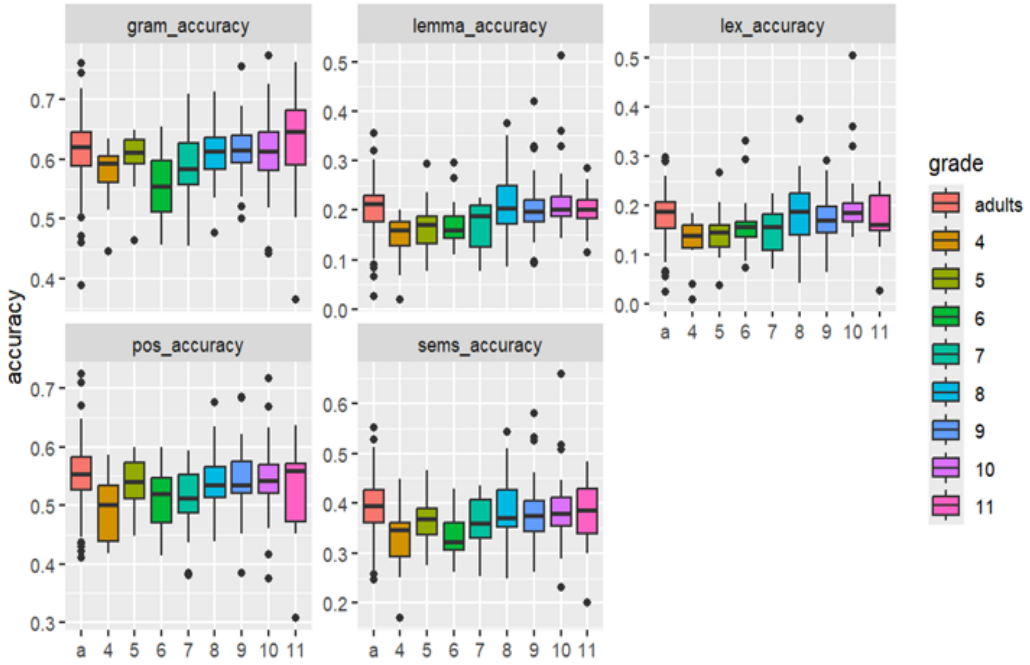

Рисунок 2. Предсказуемость в зависимости от класса

На школьном подкорпусе наблюдается значимый эффект (F(7, 185), p<.05) класса на все типы предсказуемости: средний для лемматической (.13) и лексической (.116) и малый для остальных (грамматическая: .095, частеречная: .073, семантическая: .089). Парный тест Тьюки выявил значимую (p<.05) разницу в лемматической предсказуемости между 4 и 8, 9, 10 классами (-0.066, -0.064, -0.075 соответственно), лексической между 10 и 4, 5, 7 классами (0.073, 0.052, 0.049 соответственно) и грамматической между 6 и 9, 10 классами (-0.062, -0.057 соответственно). Значимо меньшая (p<.05), чем у взрослых, предсказуемость наблюдается у 4 (лемматическая, лексическая, частеречная, семантическая) и 6 (грамматическая, семантическая) классов.

Анализ данных взрослых

Таблица 2

Значимые параметры для разных типов предсказуемости на взрослых

Предсказуемость	Предиктор	DFn	DFd	F	p	ges
лексическая	ВЧ	4	96	3.094	0.019	0.114
	пол	1	96	6.795	0.011	0.066
	возраст:ВЧ	4	96	2.953	0.024	0.110
	возраст:ВНЧ	1	96	4.552	0.035	0.045
лемматическая	ВЧ	4	96	3.587	0.009	0.130
	пол	1	96	8.844	0.004	0.084
	возраст:ВЧ	4	96	2.942	0.024	0.109
	возраст:ВНЧ	1	96	7.086	0.009	0.069
частеречная	возраст:ВНЧ	1	96	5.254	0.024	0.052
семантическая	пол	1	96	8.846	0.004	0.084

На данных взрослых участников для лексической и лемматической предсказуемости выявлены средний эффект ВЧ со значимым взаимодействием между ВЧ и возрастом и малые эффекты пола и взаимодействия между возрастом и ВНЧ. На частеречную предсказуемость значимо влияет только взаимодействие между возрастом и ВЧ, на семантическую – только пол. Для грамматической предсказуемости значимых эффектов не обнаружено.

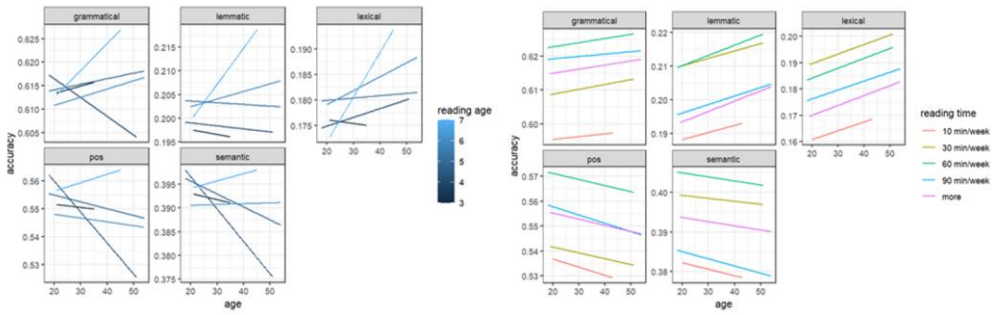

Рисунок 3. Взаимодействие между возрастом и ВНЧ и между возрастом и ВЧ на взрослых

У мужчин в среднем ниже точность предсказания, чем у женщин. С увеличением ВЧ точность предсказания имеет тенденцию расти, тест Тьюки выявил значимые различия (p<.05) между 10 и 30 и между 10 и 90 мин/неделю для лемматической предсказуемости (-0.057, -0.055 соответственно) и между 10 и 30 мин/неделю для лексической (-0.050).

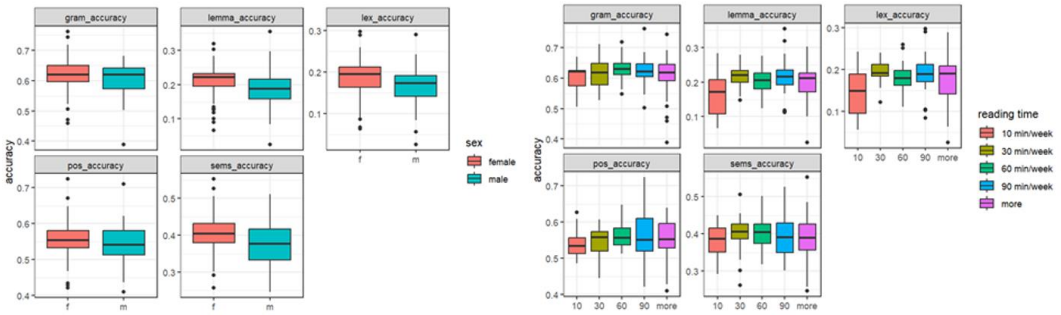

Рисунок 4. Влияние пола и ВЧ на разные типы предсказуемости

Обсуждение полученных результатов

Мы показали, что различные характеристики следующего слова предсказываются с разной успешностью. Наиболее высоки грамматическая и частеречная предсказуемость, что согласуется с прошлыми исследованиями [4]. Меньшие значения частеречной предсказуемости могут быть вызваны сложностью предсказания конкретного члена синтаксической группы, несмотря на возможность угадать его грамматические характеристики: например, мы вряд ли можем уверенно сказать, начнётся ли именная группа с существительного или прилагательного, но их падеж мы можем определить по глаголу.

Наибольшим влиянием на предсказуемость обладает класс, а в меньшей степени – возраст и ВЧ. Класс влияет на все типы предсказуемости, ВЧ же значим для грамматической, а среди взрослых – также лексической и лемматической. Во взрослой группе влияние возраста падает, что может говорить о слабом развитии способности предсказания после школы. Наиболее яркая разница в предсказуемости наблюдается между 4-7 и 9-11 классами, последние же мало отличимы от взрослых. Эта тенденция также отмечается в предыдущих исследованиях [5]. Вероятно, с увеличением объёма данных для младших классов можно будет получить больше значимых различий между младшими и старшими классами.

Наименьшей значимостью обладают пол и ВНЧ, хотя и имеют некоторое влияние на взрослых данных. Можно сказать, что женщины несколько лучше предсказывают следующее слово, чем мужчины.

В последующих исследованиях было бы полезно включить в факторы влияния размер словарного запаса, а также более подробно проанализировать предсказуемость конкретных грамматических и семантических характеристик. С точки зрения расширения датасета, стоит собрать данные 1-3 классов, а также расширить набор школ, регионов и, возможно, включить данные не-носителей.

Выводы

В данной работе мы показали, что наиболее успешно предсказываются грамматические характеристики слова, несколько меньше его часть речи, далее его семантика, и в наименьшей степени лемма и конкретная форма.

Наибольшее влияние на предсказуемость оказывает класс обучения с тенденцией к увеличению предсказуемости к старшим классам. Также на предсказуемость влияют возраст и время, которое человек тратит в неделю на чтение для себя, в то время как пол и возраст начала чтения играют наименьшую роль.

СПИСОК ИСТОЧНИКОВ

[1] 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.

[2] Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension?. *Language, cognition and neuroscience*, *31*(1), 32–59.

[3] 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.

[4] Parshina, O., Lopukhina, A., & Sekerina, I. A. (2022). Can heritage speakers predict lexical and morphosyntactic information in reading?. *Languages*, *7*(1), 60.

[5] 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

REFERENCES

[1] 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.

[2] Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension?. *Language, cognition and neuroscience*, *31*(1), 32–59.

[3] 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.

[4] Parshina, O., Lopukhina, A., & Sekerina, I. A. (2022). Can heritage speakers predict lexical and morphosyntactic information in reading?. *Languages*, *7*(1), 60.

[5] Geren, S., & Lopuhina, A. (2020). Predskazuemost' slov v kontekste. [PowerPoint slides]. Google Presentations. https://docs.google.com/presentation/d/1WnBPGILpWdb1FOf_xDAWZnVhInzQ5eD3248Cn2n4fGs/edit#slide=id.p1