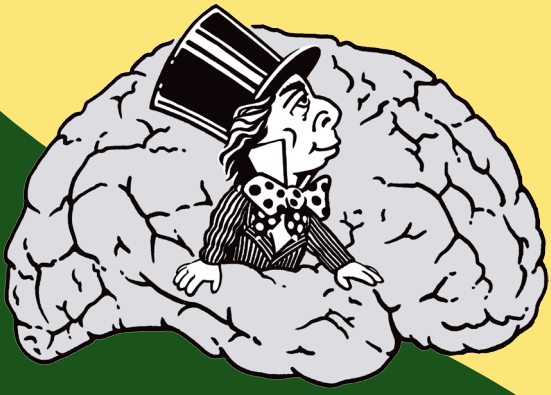
КОГНИТИВНАЯ НАУКА

В МОСКВЕ



НОВЫЕ ИССЛЕДОВАНИЯ

МАТЕРИАЛЫ КОНФЕРЕНЦИИ 2023

Под ред. Е.В. Печенковой, М.В. Фаликман, А.Я. Койфман

УДК 159.9
ББК 88.25
К57

Когнитивная наука в Москве: новые исследования. Материалы конференции 21 – 22 июня 2023 г. Под ред. Е.В. Печенковой, М.В. Фаликман, А.Я. Койфман. – М.: ООО «Буки Веди», Московский институт психоанализа. 2023 г. — 604 стр.

ISBN 978-5-4465-3880-5

УДК 159.9
ББК 88.25

ISBN 978-5-4465-3880-5

© Авторы статей, 2023

АНАЛИЗ ОШИБОК МОРФОЛОГИЧЕСКОГО АНАЛИЗАТОРА MYSTEM ПРИ РАБОТЕ С ЗАПИСЯМИ ДЕТСКОЙ РЕЧИ

В. П. Лелик*, Т. А. Еремичева, Д. А. Морозова, А. С. Сычева, К. З. Ревак, Н. Н. Псарева, И. А. Широков, С. В. Дорофеева
lelik_valeriya@mail.ru
Центр языка и мозга, Национальный исследовательский университет «Высшая школа экономики», Москва

Аннотация. Одно из условий эффективной работы автоматических морфологических анализаторов — это корректное распознавание незнакомых слов и успешное снятие морфологической омонимии. В настоящей работе оценивались результаты автоматической обработки расшифровок спонтанной детской речи с помощью морфологического анализатора MyStem. Материалами для исследования послужили лонгитюдные записи спонтанной речи двух детей-билингвов и их родителей, созданные по протоколу корпуса CHILDES. Общая длина записей составила 956 минут и 420 минут для каждого ребенка соответственно. В анализ вошли 12 828 строк, размеченных парсером. В результате проведенного исследования нам удалось определить частоту встречаемости случаев с морфологической неоднозначностью и с ошибками морфологического анализатора, а также предложить типологию таких ошибок и направления для возможного усовершенствования работы парсера MyStem.

Ключевые слова: детская речь, морфологический анализатор, морфологическая омонимия, автоматическая обработка текста, корпус CHILDES

Введение

Первичным этапом при лингвистической оценке текста является подготовка речевого материала для анализа. Для упрощения и ускорения данного процесса в компьютерной лингвистике используются автоматические морфологические анализаторы. Однако, хотя программы и в состоянии предоставить детальную морфологическую информацию о словоформах, составляющих текст, они все еще нуждаются в усовершенствовании: возникают как неоднозначности, связанные с лексической и синтаксической омонимией, так и ошибки в определении леммы, части речи и морфологических признаков слова (Kotelnikov et al., 2018). Так, несмотря на то, что современные анализаторы способны определять формы новообразованных слов, в случаях, когда эти слова нестандартны и не похожи на те, на которых происходило обучение, могут возникать ошибки (Зобнин, Носырев, 2015).

В данной работе мы концентрируемся на анализе ошибок одного из самых распространенных для русского языка морфоанализаторов — MyStem (Segalovich, 2003) — при обработке спонтанной детской речи, в которой, по сравнению с речью взрослых, присутствует больше грамматических ошибок, нестандартных форм и детских неологизмов (Красноперова, 2022). Мы исследовали количество и типы ошибок при автоматической разметке транскрипций детской речи на примере анализа существительных, для этого мы разработали типологию ошибок MyStem в работе с существительными. С определенными доработками эта типология может применяться и для классификации ошибок, которые допускают автоматические морфологические анализаторы при разметке других частей речи. В итоге мы предлагаем потенциальные пути уменьшения количества подобных неточностей.

Методика

Материалы. Материалами для работы послужили данные на русском языке из корпуса CHILDES (MacWhinney, 2000), который содержит записи спонтанной речи детей и их родителей. Расшифровки включали в себя полную фиксацию речи (например, в том числе транскрибировались междометия, оговорки и т. п.). Разметка выбранных расшифровок была проведена с использованием программы MyStem — морфологического анализатора для русского языка. В ходе разметки были определены части речи и соответствующие морфологические характеристики для каждого слова. В данной работе оценивались результаты морфологической разметки имен существительных. Транскрибированные кириллицей иностранные слова, которые употребляли в своей речи участники-билингвы, не были включены в анализ, так как для них не ожидалась корректная морфологическая разметка. Всего в анализ вошли 12 828 строк, полученные в результате автоматической разметки с использованием MyStem (версия pymystem3).

Участники. В данной работе мы использовали расшифровки лонгитюдных записей спонтанной речи двух детей-билингвов. Нам были доступны записи речи одного ребенка в возрасте от 2 лет до 3 лет 2 месяцев и еще одного ребенка в возрасте от 2 лет 8 месяцев до 3 лет 2 месяцев. Помимо детской речи, мы также работали с речью взрослых людей, зафиксированной в данных записях: в корпусе первого ребенка содержатся также данные речи 7 взрослых, в корпусе второго ребенка — 4 взрослых. Общая длительность записей составила 956 минут и 420 минут соответственно.

Процедура. После ручной обработки случаев с морфологической неоднозначностью и исправления ошибок, допущенных парсером MyStem, для каждого слова были зафиксированы наличие/отсутствие ошибки морфоанализатора и наличие/отсутствие неоднозначности. Ошибки парсера определялись путем проверки леммы и автоматического морфологического разбора для каждого существительного. В результате были обработаны все материалы: обоих детей и взрослых из соответствующих корпусов. По результатам первого этапа работы была составлена типология наиболее частотных ошибок (см. табл. 1). При составлении классификации выделялись как общие, так и детальные типы ошибок (см. табл. 2).

Таблица 1. Классификация и частотность ошибок при работе парсера MyStem с расшифровками детской речи

Общий тип ошибки	Общий тип ошибки (пояснение)	Случаи с ошибками парсера MyStem
PART_OF_SPEECH	Неверное определение части речи	55.0% (1164 случая)
PROPNAME	Ошибка при разметке имен собственных	14.5% (306 случаев)
CHILDWORD	Ошибка при разметке существительных, придуманных ребенком	4.9% (103 случая)
NORM_LEMMA	Неверное определение леммы у существительного	4.7% (99 случаев)
CHILDFORM	Ошибка в употреблении, допущенная ребенком	3.9% (83 случая)
ANIM	Неверное определение одушевленности	3.9% (83 случая)
NORM_FORM	Ошибка при разметке формы существительного	3.0% (65 случаев)
CHILDISHWORD	Ошибка при разметке эвфемизмов деликатной темы	2.5% (52 случая)
GENDER	Неверное определение рода	2.4% (51 случай)
DIM	Ошибка при разметке слова в диминутивной или аугеминативной форме	1.5% (32 случая)
NUM	Неверное определение числа	1.0% (21 случай)
ADULTFORM	Ошибка в употреблении, допущенная взрослым	0.9% (19 случаев)
HOMONYMY	Отсутствие омонимии	0.6% (13 случаев)
UNSORTED	Без определенного тега	0.6% (12 случаев)
NONSTANDFORM	Ошибки при разметке слов в нестандартной грамматической форме	0.4% (9 случаев)

Анализ. Анализ результатов, полученных в ходе обработки морфологической разметки, проводился в среде программирования R версии 4.0.2 (R Core Team, 2022).

Результаты

Количество случаев, где парсер допускал ошибки, составило 16.5% ($N = 2115$) от общего количества вхождений (словоформ в корпусах). Количество случаев с неоднозначностью составило 45% ($N = 5772$) от общего количества вхождений. В среднем на 1000 вхождений в данной выборке

встречается 162 случая с ошибкой парсера (*M* = 162, *SD* = 60.4) и 444 случая с неоднозначностью (*M* = 444, *SD* = 104.6). В случаях с неоднозначностью было отмечено 17% вхождений с ошибкой парсера (*N* = 985), в случаях без неоднозначности — 16% (*N* = 1130). В табл. 1 представлены результаты распределения ошибок, допущенных парсером, по общим типам разработанной нами классификации ошибок. Наиболее частотными оказались ошибки в определении части речи (55% от общего количества), поэтому был проведен детальных анализ ошибок этого типа. В табл. 2 представлены три наиболее частотных типа ошибок, которые парсер допускал при определении части речи.

Таблица 2. Классификация и частотность ошибок в определении части речи (категория PART_OF_SPEECH из таблицы 1) при работе парсера MyStem с расшифровками детской речи

Детальный тип ошибки	Детальный тип ошибки (пояснение)	Случаи с ошибками парсера MyStem
INTJ_ZV_NOUN	Междометия-звукоподражания определяются как существительные	32.9% (383 случая)
INTJ_NOUN	Иные междометия определяются как существительные	26.8% (312 случаев)
INTJ_CH_NOUN	Междометия, придуманные ребенком, определяются как существительные	18.4% (214 случаев)

Обсуждение и выводы

В подавляющем большинстве случаев (55%) ошибка MyStem в разметке имен существительных была связана с неправильным определением части речи. Чаще всего анализатор ошибочно размечает как существительное междометия: это могут быть как звукоподражания (например, «кар», «ням») или придуманные ребенком междометия (например, «онь», «атя»), так и этикетные междометия (например, «спасибо», «пожалуйста»). Хотя MyStem был обучен на словарях русского языка, достаточно часто (6.3% случаев) он относит к списку существительных частотные в языке слова служебных частей речи — предлоги (например, «с», «в»), союзы (например, «и», «а») и частицы (например, «да», «окей»). Менее частотные слова других частей речи, определенные как имя существительное, в большинстве случаев — следствие опечаток, допущенных при расшифровке записей, или ошибок произношения у детей. Например, распознанные как существительные глагол «сьела» (вместо «съела»), наречие «тожа» (вместо «тоже») и т. д.

Были обнаружены также ошибки при определении парсером леммы (4.7% случаев), одушевленности объекта (3.9% случаев), рода (2.4% случаев) и числа (1% случаев). В результате рассмотрения этих случаев можно предположить следующее: словари, которые использует MyStem, несколько ограничены в примерах, имеющих только форму множественного числа; в существительных среднего рода, оканчивающихся на букву «о»; а также в лексике, связанной с животными. Кроме того, морфоанализатор неверно считывает некоторые формы местного падежа, а также может ошибочно определить лемму в случае лексической омонимии (например, «полк» и «пОлка). Также частотными явля-

ются ошибки при разметке имен собственных (14.5% случаев от общего количества), которые включают в себя перечисленные выше проблемы (неверное определение леммы, одушевленности, рода и числа). Возможно, это объясняется относительной уникальностью имен собственных, у которых встречаются нестандартные формы употребления.

В качестве возможных направлений работы для повышения эффективности анализатора MyStem можно рассматривать следующие. Во-первых, обновление словарей с указаниями частей речи — в некоторых случаях возникали ошибки при разметке довольно частотных слов других частей речи, которые определялись анализатором как существительные. Во-вторых, пополнение словаря списком имен собственных с различными формами словоизменения. В-третьих, расширение словаря путем добавления в него специфических «детских» версий слов, которые широко используются детьми на ранних этапах речевого развития, а также в речи взрослых, обращенной к маленьким детям.

Литература

Зобнин А.И., Носырев Г.В. Морфологический анализатор MyStem 3.0 // Труды института русского языка им. В.В. Виноградова. 2015. № 6. С. 300 – 310.

Красноперова Е.С. Актуальная грамматика детской речи: корпусные исследования // Филологический класс. 2022. Т. 27. № 1. С. 87 – 100.

Kotelnikov E., Razova E., Fishcheva I. A close look at Russian morphological parsers: Which one is the best? // Artificial Intelligence and Natural Language. AINL 2017. Communications in Computer and Information Science, vol 789 / A. Filchenkov, L. Pivovarova, J. Žižka (Eds.). Springer, Cham, 2018. P. 131 – 142. https://doi.org/10.1007/978-3-319-71746-3_12

MacWhinney B. The CHILDES project: Tools for analyzing talk: The Database.Mahwah, NJ: Lawrence Erlbaum Associates, 2000. URL: https://childes.talkbank.org/.

R Core Team R: A language and environment for statistical computing. R Foundation for Statistical Computing.Vienna, Austria: 2022. URL: https://www.R-project.org/.

Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine // Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications. MLMTA'03, June 23 – 26, 2003. Las Vegas, Nevada, USA: MLMTA, 2003. P. 273. URL: https://tech.yandex.ru/mystem/.

ANALYSIS OF ERRORS OF THE MORPHOLOGICAL ANALYZER MYSTEM APPLIED TO THE CHILDREN'S SPEECH RECORDINGS

V. P. Lelik*, T. A. Eremicheva, D. A. Morozova, A. S. Sycheva, K. Z. Revak,
N. N. Psaryova, I. A. Shirokov, S. V. Dorofeeva
lelik_valeriya@mail.ru
Center for Language and Brain, HSE University, Moscow

Abstract. Some of the important conditions of the effectiveness of morphological analyzers are the correct recognition of unfamiliar words and successful morphological disambiguation. In this work, we evaluated the results of automatic processing of children's spontaneous speech using the morphological analyzer MyStem. We analyzed the longitudinal spontaneous speech recordings of two bilingual children and their parents created ac-

cording to the CHILDES protocol. The total length of the recordings was 956 minutes and 420 minutes, respectively. The analysis included 12,828 lines from the transcripts tagged by the parser. Based on the results of the research, we were able to determine the frequency of cases with morphological ambiguity and morphological analyzer errors, and we furthermore suggest a typology of such errors and some possible ways of improving the work of the MyStem parser.

Keywords: children's speech, morphological analyzer, morphological homonymy, automatic text processing, CHILDES corpus