Semantic clustering of early children's vocabulary

Valeriya Lelik
National research university "Higher school of economics"
Moscow, Russia
lelik_valeriya@mail.ru

Anastasiya Lopukhina
National research university "Higher school of economics"
Moscow, Russia
nastya.lopukhina@gmail.com

Abstract

This paper describes the development of vocabulary in one-to-three-year-old Russian-speaking children. Based on the previous experimental studies, we hypothesized that children should acquire novel words from dense semantic categories faster than from sparse categories. For the first time we applied a method of vector semantic analysis and clustering to the longitudinal corpora of two children. We found that all semantic clusters remained in the children's vocabulary as they grew older and most of clusters became larger. However, we did not find that children produced more words in originally dense semantic categories compared to originally sparse categories.

Keywords: vocabulary acquisition; child language corpora; vector models; semantics; early stages of lexical development

Семантическая кластеризация слов в речи детей на ранних этапах речевого развития

Валерия Лелик
Национальный исследовательский университет "Высшая школа экономики"
Москва, Россия
lelik_valeriya@mail.ru

Анастасия Лопухина
Национальный исследовательский университет "Высшая школа экономики"
Москва, Россия
nastya.lopukhina@gmail.com

Аннотация

В статье описывается развитие словарного запаса у русскоязычных детей в возрасте от одного до трех лет. Основываясь на предыдущих экспериментальных исследованиях, мы выдвинули гипотезу, что дети усваивают новые слова из тех семантических категорий, в которых уже содержится много слов. Впервые мы применили метод векторного семантического анализа и кластеризации к корпусам лонгитюдных записей двух детей. Мы обнаружили, что по мере взросления детей все семантические кластеры сохранялись, большинство из них увеличилось в объемах. Однако мы не обнаружили разницы в скорости роста изначально больших кластеров по сравнению с изначально маленькими.

Ключевые слова: формирование детского словаря; корпус детской речи; векторные модели; семантика; ранние этапы формирования словаря

1 Введение

Формирование словарного запаса ребенка обычно начинается в конце первого года жизни, когда ребенок начинает узнавать и порождать первые слова. Исследования на материале английского языка показывают, что в течение первых лет жизни словарный запас детей увеличивается в десятки раз (Frank et al., 2017; wordbank.stanford.edu). На формирование словарного запаса могут влиять, например, пол ребенка (Eliseeva & Vershinina, 2017), количество слов в речи взрослых (Fenson et al., 1994), а также особенности семантической структуры раннего словарного запаса

ребенка (Bergelson & Aslin, 2017; Borovsky et al., 2016; Mani & Ackermann, 2018). Влияние последнего фактора изучалось в контролируемых психолингвистических экспериментах и не проверялось на материале лонгитюдных корпусов детской речи.

В настоящей работе мы хотим проверить гипотезу о том, что семантическая структура раннего словарного запаса ребенка может предсказать, какие слова ребенок усвоит раньше: предположительно, дети быстрее усваивают слова из той семантической категории, в которой уже содержится много слов. Данная гипотеза нашла подтверждение в эксперименте с записью движений глаз у двухлетних англоязычных детей (Borovsky et al., 2016). Авторы обнаружили, что дети дольше смотрели на изображения, соответствующие новым словам из тех семантических категорий (например, животные, одежда, транспорт), в которых уже знали большое количество слов. Вероятно, изучение новых слов происходит «кластеризованным» способом: слова из категорий, которые уже были хорошо усвоены, облегчают усвоение новых слов, близких по смыслу с известными словами из этих категорий. Наше исследование впервые проверяет гипотезу о влиянии количества слов в семантических категориях на усвоение новых слов на материале лонгитюдных корпусов речи детей при помощи метода векторного семантического анализа и кластеризации.

Кластеризация – одна из задач машинного обучения, которая заключается в разделении некоторого набора данных на группы, объединенные общим признаком (Kaufinann, 2006). Кластеризация относится к методам обучения «без учителя»: это значит, что в задаче кластеризации нет обучающей выборки, то есть размеченных данных, которые модель бы могла запомнить и научиться на их примере выявлять некоторые закономерности. Метод кластеризации обычно применяется на данных, о структуре которых известно мало. В настоящем исследовании кластеризация будет применяться на материале двух новых корпусов детской речи, чтобы получить представление об их составе и структуре.

2 Методы

Данными для работы послужили лонгитюдные записи речи двух детей – мальчика и девочки – и членов их семей, а также других взрослых (няни, друзей и др.), сделанные по протоколу Child Language Data Exchange System (CHILDES; childes.talkbank.org; MacWhinney, 2000) в 2016-2019 годах. Семьи, участвующие в исследовании, на протяжении нескольких лет раз в две недели в течение часа фиксировали повседневное общение с ребенком на видеокамеру. Корпус записей мальчика состоит из 42 аудиозаписей общей длительностью приблизительно 12 часов. На момент первой записи ребенку был 1 год 5 месяцев, на момент последней – 3 года. Для девочки имеется 246 записей длительностью примерно 32 часа. На момент первой записи девочке было 10 месяцев, на момент последней – 2 года 10 месяцев. Все видеозаписи были расшифрованы и затранскрибированы в системе CLAN (MacWhinney, 2014) на латинице.

Для автоматической обработки записей мы транслитерировали материал, для чего был написан код на языке Python. Программа принимает на вход путь к папке с записями на латинице и название файла для записи готового материала. Каждому латинскому символу (или сочетанию символов) была поставлена в соответствие буква русского алфавита, например, sh – ш (shkola – школа), shch – щ (shchi – щи). В каждой строке транслитерированного материала присутствует идентификатор говорящего и название записи. Транслитерированный csv файл подавался на вход другой программе для лемматизации и морфологического разбора. Он был выполнен с помощью модуля pymystem3 (https://yandex.ru/dev/mystem/), предназначенного для автоматической разметки частей речи и их грамматических характеристик. Далее мы использовали только леммы, которые были получены в результате разбора.

Оба детских корпуса были разделены на две части: речь ребенка и речь взрослого, обращенную к ребенку. В настоящем исследовании мы работали только с речью ребенка из обоих корпусов. Чтобы исследовать, как семантические категории на раннем этапе влияют на усвоение слов на более позднем этапе, речь ребенка была разделена на два подкорпуса для каждого участника. Для мальчика выделились следующие равные по времени периоды: 1 год 5 месяцев – 2 года 2 месяца (702 леммы), 2 года 2 месяца – 3 года (1277 лемм). Речь девочки была разделена на следующие периоды: 10 месяцев – 1 год 10 месяцев (816 лемм), 1 год 10 месяцев – 2 года 10 месяцев (2248 лемм). Во втором периоде количество слов в словаре ребенка ожидаемо увеличилось.

Для выявления семантических категорий мы использовали метод семантических векторов, который позволяет представить слова в виде векторов в многомерном векторном пространстве. Слова, близкие по смыслу, будут находиться в этом пространстве рядом. Таким образом, мы можем разбить векторное пространство на семантические кластеры, внутри которых будут находиться близкие по смыслу слова. В данной работе для построения семантических векторов слов была использована модель ruwikiruscorpora_upos_skipgram_300_2_2019, предобученная на Национальном корпусе русского языка (НКРЯ) и русской Википедии (Kutuzov & Kuzmenko, 2017). В модели уже содержатся вектора для слов из НКРЯ и Википедии, она была выбрана нами из-за большого объема обучающей выборки, что позволяет найти больше векторов для слов, содержащихся в наших корпусах. Стоит отметить, что в детских корпусах содержится много детских неологизмов, а также междометий. Для таких слов векторов в предобученной модели нет, поэтому мы убрали их из рассмотрения. Итоговый объем слов, для которых удалось найти вектора, следующий: 1 период корпуса мальчика – 304 слова, 2 период – 572 слова, 1 период корпуса девочки – 192 слова, 2 период – 1175 слов.

Следующим этапом анализа было определить количество кластеров, на которое следует разделить детскую речь. Для этого использовался метод иерархической кластеризации (Nielsen, 2016). Иерархическая кластеризация выделила в корпусах большое количество кластеров, однако качественный анализ показал, что некоторые кластеры состоят из 1 или 2 элементов, и такие кластеры было решено убрать из рассмотрения. В итоге метод иерархической кластеризации позволил выделить следующее количество кластеров: 1 период корпуса мальчика – 12 кластеров, 2 период – 30 кластеров, 1 период корпуса девочки – 13 кластеров, 2 период – 32 кластера. Этот метод был выбран исключительно для подбора количества кластеров, так как в нем невозможно зафиксировать рандомное значение, которое сделало бы результаты воспроизводимыми: при каждом новом запуске программы кластеры определяются по-разному (однако их количество колеблется в маленьком диапазоне). Иерархическая кластеризация была использована с параметрами по умолчанию: метод ближайшего соседа как алгоритм вычисления расстояния между кластерами, Евклидова метрика как метрика расстояния между наблюдениями, для параметра оптимального упорядочивания дерева было присвоено значение False.

Для разделения слов на кластеры с целью их последующего качественного анализа мы использовали метод k-means (MacQueen et al., 1967). Метод k-means позволяет зафиксировать рандомное значение, а значит, его результаты являются воспроизводимыми и разбиение на кластеры не меняется при каждом новом запуске кода. Здесь мы задавали только 1 гиперпараметр: количество кластеров, полученное с помощью метода иерархической кластеризации. Выбор метода кластеризации влияет на разбиение слов на семантические общности, но k-means, в отличие от иерархической кластеризации, при фиксировании рандомного значения позволяет получать одинаковые результаты при каждом новом запуске кода. Поскольку автоматические методы кластеризации не всегда позволяют получить интерпретируемые результаты, был проведен дальнейший отбор кластеров и слов в них вручную (убирались кластеры, состоящие только из междометий, кластеры, содержащие набор слов, которым сложно присвоить единое значение, внутри кластеров удалялись не связанные по смыслу с остальными слова).

3 Результаты

3.1 Корпус мальчика

В первый период (1 год 5 месяцев – 2 года 2 месяца) в результате кластеризации методом k-means выделились 7 кластеров (Таблица 1).

Кластеры 1 периода	Количество слов в 1 период	Слова 1 периода	Количество слов во 2 период
Глаголы, имеющие компонент значения	26	Полететь, побежать, убегать, лететь, подходить, плавать, ступать, догнать, пойти, водить, выбрасывать, выгонять, идти, упасть,	11 (-57%)

«перемещение в пространстве»		выпрыгивать, нести, отдавать, давать, открывать, отливать, вырывать, падать, выскочить, закрывать, выходить, гнать	
Животные	19	Медведь, зайчик, заяц, кошка, петух, петушок, птичка, кот, кролик, овца, собака, лис, лиса, лисичка, рыбка, мышь, лягушка, крокодил, волк	47 (+147%)
Предметы	17	Часы, рюкзак, игрушка, вилка, велосипед, ботинок, щетка, шорты, майка, шуба, ремень, пакет, пакетик, мяч, мячик, нож, палочка	38 (+111%)
Люди или принадлежность людям	16	Тетя, Я. (имя собственное), Л., А., баба, дед, дедушка, дядя, ляля, мамин, папин, папа, М., Н., мама, мальчик	28 (+64%)
Глаголы, не связанные с перемещением в пространстве	14	Становиться, помогать, получаться, захотеть, жить, работать, читать, рисовать, хотеть, говорить, быть, бояться, сделать, напрашиваться	33 (+135%)
Продукты	15	Чай, хлеб, кусок, хлопья, мандарин, лимон, салат, банан, каша, картошка, орех, ананас, желудь, сыр, ягода	29 (+92%)
Части тела	12	Палец, волос, коса, ухо, нога, пальчик, рука, нос, пасть, пупок, плечо, глаз	4 (-200%)

Таблица 1. Распределение слов по кластерам (корпус мальчика)

Во втором периоде (2 года 2 месяца – 3 года) корпуса мальчика было выделено 20 кластеров. По сравнению с первым периодом, во втором значительно увеличились кластер, объединяющий наименования продуктов (прирост составил 92%), кластер глаголов, не связанных с перемещением в пространстве (прирост 135%), кластер, объединяющий наименования людей (прирост 64%). Стоит отметить, что во втором периоде выделился еще один кластер глаголов, которые на предыдущем этапе могли бы быть отнесены в кластер глаголов, не связанных с перемещением в пространстве. Вероятно, значение стало более дифференцированным, выделилась новая семантическая группа глаголов, имеющих компонент значения разрушения. То же самое произошло и с кластером, объединяющим предметы: ко второму периоду он разделился на 3 кластера (предметы, природные объекты, предметы, имеющие характерную форму), можно считать, что суммарный прирост слов этой категории составил 111%. Более дифференцированным стало и значение кластера, объединяющего наименования животных: ко второму периоду он разделился на 3 кластера (животные; дикие животные; домашние животные + дикие животные с уменьшительно-ласкательным суффиксом) Суммарный прирост этой категории составил 147%. Однако два кластера, выделенные на первом периоде, ко второму периоду уменьшились: слова, обозначающие части тела (уменьшился на 200%), глаголы, имеющие компонент значения «перемещение в пространстве» (уменьшился на 57%).

3.2 Корпус девочки

В первый период (10 месяцев – 1 год 10 месяцев) выделились 6 кластеров (Таблица 2).

Кластеры 1 периода	Количество слов в 1 период	Слова 1 периода	Количество слов во 2 период
Люди и принадлежность людям	14	Баба, человечек, ребенок, папа, мама, ляля, дед, няня, папин, Н., дядя, мамин, Дж., друг	63 (+350%)
Предметы	13	Кораблик, звезда, ключ, качели, яма, эхо, носок, небо, желудь, камешек, шарик, клей, телефон	50 (+284%)

Продукты	11	Картошка, салат, мандарин, хлеб, варить, молоко, рыба, еда, банан, сыр, ягода	35 (+218%)
Действия	11	Дарить, лить, давать, упасть, пойти, идти, открывать, рисовать, хотеть, вытаскивать, спать	86 (+681%)
Животные	4	Собака, медведь, олень, бабочка	66 (+1550%)
Части тела	3	Губа, ухо, пупок	28 (+833%)

Таблица 2. Распределение слов по кластерам (корпус девочки)

Ко второму периоду (1 год 10 месяцев – 2 года 10 месяцев) значительно увеличился кластер, объединяющий людей (прирост 350%), кластер продуктов (прирост 218%), кластер животных (прирост 1550%), кластер частей тела (прирост 833%). Кластер, объединяющий действия, ко второму периоду разделился на 2 кластера: глаголы, не связанные с перемещением в пространстве и глаголы, имеющие компонент значения «перемещение в пространстве». Суммарный прирост составил 681%. Кластер предметов также разделился на 2 кластера: предметы и предметы, имеющие характерную форму (суммарный прирост 284%).

4 Обсуждение

Как и ожидалось, все кластеры, выделившиеся в течение первого периода в речи обоих детей, сохранились и во втором периоде. Большинство из них ко второму периоду увеличились в размерах, а значение некоторых кластеров стало более дифференцированным, в результате чего один кластер разделился на несколько. Это говорит о том, что структура более раннего периода развития словарного запаса ребенка во многом определяет структуру более позднего периода: семантические категории, появившиеся рано, становятся основой для развития семантических категорий в более старшем возрасте. Можно заметить, что максимальный прирост у обоих детей произошел в группе, объединяющей названия животных. Это может быть связано с типом наших данных: записи обычно производились в игровой обстановке, и названия животных, с большой вероятностью, обозначают игрушки или картинки в детских книгах. Предположение о том, что изначально маленькие кластеры будут расти медленнее изначально больших, не подтвердилось: практически все кластеры, выделившиеся в первом периоде, значительно увеличились ко второму в обоих корпусах. Мы не обнаружили более стремительного роста изначально больших кластеров. Это может говорить о том, что развитие словарного запаса зависит скорее не от размера семантических категорий в более раннем периоде, а от того, какие категории в принципе выделяются в более ранний период. Наши результаты не подтверждают выводы, высказанные в (Borovsky et al., 2016; Mani & Ackermann, 2018), что, вероятно, связано с методологией исследований. Метод семантических векторов и кластеризации действительно может использоваться для анализа корпусов детской речи и делать вклад в понимание того, как формируются речевые навыки детей.

Библиография

[1] Eliseeva M. B., Vershinina E. A. Gendernye osobennosti rechevogo i kommunikativnogo razvitiya detej 8–18 mesyacev (na materiale russkogo yazyka) //Acta Linguistica Petropolitana. Trudy instituta lingvisticheskih issledovanij. – 2017. – Т. 13. – №. 3.

[2] Bergelson E., Aslin R. N. Nature and origins of the lexicon in 6-mo-olds //Proceedings of the National Academy of Sciences. – 2017. – Т. 114. – №. 49. – С. 12916-12921.

[3] Borovsky A. et al. Lexical leverage: Category knowledge boosts real-time novel word recognition in 2-year-olds //Developmental science. – 2016. – Т. 19. – №. 6. – С. 918-932.

[4] Fenson L. et al. Variability in early communicative development //Monographs of the society for research in child development. – 1994. – С. i-185.

[5] Frank M. C. et al. Wordbank: An open repository for developmental vocabulary data //Journal of child language. – 2017. – Т. 44. – №. 3. – С. 677.

[6] Mining W. I. D. Data mining: Concepts and techniques //Morgan Kaufinann. – 2006. – Т. 10. – С. 559-569.

[7] Kutuzov A., Kuzmenko E. WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. In: Ignatov D. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science, vol 661. Springer, Cham, 2017.
[8] MacQueen J. et al. Some methods for classification and analysis of multivariate observations //Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. – 1967. – Т. 1. – №. 14. – С. 281-297.
[9] MacWhinney B. The CHILDES Project: Tools for Analyzing Talk // 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates. 2000.
[10] MacWhinney B. The CHILDES project: Tools for analyzing talk, Volume II: The database. – Psychology Press, 2014.
[11] Mani N., Ackermann L. Why do children learn the words they do? //Child Development Perspectives. – 2018. – Т. 12. – №. 4. – С. 253-257.
[12] Nielsen F. Hierarchical clustering //Introduction to HPC with MPI for Data Science. – Springer, Cham, 2016. – С. 195-211.
[13] Pymystem3 // URL: URL: https://yandex.ru/dev/mystem/