ИСПОЛЬЗОВАНИЕ ЛАТЕНТНО-СЕМАНТИЧЕСКОГО АНАЛИЗА В ИССЛЕДОВАНИЯХ И МОДЕЛИРОВАНИИ КОГНИТИВНОГО РАЗВИТИЯ ДЕТЕЙ[1]

Соловьев А. Н. (lechat1@mail.ru)

Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

В данной работе представлены результаты исследования когнитивного развития детей дошкольного возраста (4–7 лет), основанного на методе латентно-семантического анализа (ЛСА). Экспериментальная часть состоит из трех тестов. В первых двух тестах продемонстрирована ассоциативно-семантическая связь между моделями ЛСА и ответами испытуемых. В третьем тесте показана возможность использования ЛСА для исследования мнемонических способностей у детей. Проведен сравнительный анализ результатов с моделью ЛСА, полученной на корпусе СМИ.

Ключевые слова: латентно-семантический анализ (ЛСА), ассоциативно-семантические связи, моделирование когнитивного развития детей

USING LATENT SEMANTIC ANALYSIS FOR SIMULATING OF CHILDREN'S COGNITIVE DEVELOPMENT

Solovyev A. (lechat1@mail.ru)

St. Petersburg State University, St. Petersburg, Russia

In the 20th century Noam Chomsky formulated the so-called Plato's problem: why is the amount of our knowledge much greater than we can extract from our everyday experience? For example, the vocabulary of preschool children (aged 6–7) averagely increases by 3–8 words every day, and not every word refers to any reality or action (for example, abstract concepts, words carrying "phatic" or uninformative assignment, etc.).

[1] Работа выполнена при финансовой поддержке Российского гуманитарного научного фонда (РГНФ, проект № 11-06-12042в, 2011–2013 гг.)

How does the child recognize each new meaning of the word and its relation to others, or why are new "meanings" formed? We propose a method to simulate associative-semantic relations between words. On the one hand. it eliminates rigid binding of a lexical unit to any cluster, and on the other it presents a complete system of relationships between words.

The paper presents the results of three experiments with cognitive development of 4–7-year-old children using a Latent Semantic Analysis (LSA) that permits comparisons of semantic similarity between pieces of textual information. We used a technique developed by G. Denhière and B. Lemaire. The principal distinctions of our research are that for the first time, the experiments were performed 1) on the Russian language; and on pre-school children. The children were grouped into two categories: 4–5 and 6–7 years, which corresponds to age variability of cognitive development.

Two experiments describe semantic and associative similarity between LSA models and the children's cognitive development. The third experiment describes using LSA to measure the children's semantic memory. The results are compared to children's model data and adults' model data. The computational models are built from the LSA of a multisource child corpus and of an internet mass media corpus.

Our findings confirm that: 1) LSA can be used to simulate a variety of children's cognitive processes; 2) LSA models represent the development of different age groups children's cognitive processes, in particular associative semantic processes and short-term and long-term memory work; 3) this method may be recommended for the comparative study of children's cognitive development, in particular, the development of associative-logical thinking, verbal discourse, the development of memory.

Keywords: Latent Semantic Analysis (LSA), semantic similarity, semantic association, simulating of children's cognitive development

Введение

Одним из вопросов, которым задаются великие мыслители человечества со времен Платона, является гносеологический вопрос о нашей возможности познаваемости мира. В XX веке Ноам Хомский [Chomsky, 1984] сформулировал так называемую проблему Платона (Plato's problem): почему объем знаний отдельного человека намного больше, чем он может извлечь из своего повседневного опыта? Иначе говоря, как информация, получаемая из последовательности относительно небольшой вариативности событий, может корректно использоваться и адаптироваться к потенциально бесконечному числу ситуаций?

Например, лексикон детей дошкольного возраста (6–7 лет) составляет 3,5–4 тысячи слов, в то время как у детей 10–11 лет этот показатель уже порядка 7–15 тысяч слов [Львов, 2002], т. е. в среднем ежедневно увеличивается на 3–8 слов. При этом не всегда денотат имеет свой строго определенный референт, или, другими словами — не каждое слово имеет соотнесение с реально существующими вещами или выполняемыми действиями (например, абстрактные понятия, слова, несущие «фатическую» или неинформативную нагрузку и пр.).

Возникает вопрос: как ребенок определяет каждое новое значение слова и его соотношение с другими значениями или почему образуются новые денотаты («смыслы») и как они соотносятся между собой?

Современные лингвистические и психолингвистические теории поразному отвечают на этот вопрос: генеративистский подход (см., например, Chomsky, 2002; Fodor, 2009) предполагает существование неких врожденных, а коннекционистский (например, Deacon 2003) — приобретенных языковых структур. Но вопрос остается открытым: что это за структуры, как они работают? Или: каковы перспективы моделирования когнитивных процессов?

Работу «смысловых» механизмов концептуально можно сравнить с процессами категоризации или кластеризации (например, [Соловьев, 2008; Черниговская, 2013 (стр. 28)]). При таком подходе возникает проблема определения изначальных концептов или первичных кластеров, их границ и их числа.

В данном исследовании используется метод, позволяющий моделировать ассоциативно-семантические связи между словами, что с одной стороны позволяет отказаться от жесткой привязки лексической единицы к какому либо из кластеров, а с другой представить целостную систему связей между словами.

1. Латентно-семантический анализ

Одним из методов, позволяющих продемонстрировать работу когнитивных механизмов, является метод латентно-семантического анализа (ЛСА). Еще в 1988 г. Вальтер Кинтч предложил интеграционную модель понимания [Kintsch, 1988], основанную на ассоциативных и семантических связях между лексическими единицами. Именно эта модель и была в дальнейшем реализована в методе латентно-семантического анализа.

Обычно ЛСА используется для выявления латентных (скрытых) ассоциативно-семантических связей между термами (словами, н-граммами) путем сокращения факторного пространства термы-на-документы.

Основная идея латентно-семантического анализа состоит в следующем: если в исходном вероятностном пространстве, состоящим из векторов слов, между двумя любыми словами из двух разных векторов может не наблюдаться никакой зависимости, то после некоторого алгебраического преобразования данного векторного пространства эта зависимость может появиться, причем величина этой зависимости будет определять силу ассоциатино-семантической связи между этими двумя словами.

Существуют три основных разновидности решения задачи методом ЛСА:

- сравнение двух термов между собой;
- сравнение двух документов между собой;
- сравнение терма и документа.

В нашем исследовании мы использовали все три разновидности в зависимости от поставленной задачи. Более подробно о методе ЛСА см. [Landauer, 1998; Соловьев, 2008].

Впервые ЛСА был применен для автоматического индексирования текстов и выявления ассоциативно-семантической структуры текста [Deerwester, 1990]. Затем этот метод был довольно успешно использован для представления баз знаний [Landauer, 1997]. Также метод ЛСА нашел широкое применение в построении когнитивных моделей понимания и формирования знания. В работах [Denhière, 2004; Lemaire, 2003] сделана попытка построить модель долговременной и эпизодической (кратковременной) памяти у детей разного школьного возраста на базе детских текстов. Авторы показали, что семантический анализ текстов методом ЛСА может прояснить как некоторые механизмы работы долговременной и эпизодической памяти, так и связного понимания текста.

Еще одно применение ЛСА нашел в моделях представления и проверки знаний. В вышеупомянутой работе [Landauer, 1997] ЛСА был исследован применительно к известной системе проверки знания английского языка TOEFL на студентах. Также данный метод зарекомендовал себя как эффективное средство проверки и оценочного предсказания для обучающего процесса [Wolfe, 1998]. С помощью ЛСА можно оптимизировать метод обучения, находя оптимальную зону в гауссовом распределении векторного пространства множества знаний. Этим же методом можно давать оценку полученных знаний: студенты, чьи предварительные знания недостаточно хорошо перекрываются семантическим векторным пространством текста, считаются недостаточно хорошо обученными данному предмету, и наоборот.

ЛСА является не единственным методом исследования ассоциативно-семантических связей в тексте. Для выявления лексической синонимии и поиска коллокаций широко используют метод взаимной информации (MI & PMI) и обобщенный ЛСА (GLSA), который представляет собой смесь методов взаимной информации и ЛСА [Matveeva, 2005]. Целесообразность использования того или иного подхода зависит от решаемой задачи. Например, при поиске синонимов методы PMI и GLSA демонстрируют большую точность, чем ЛСА, а в ассоциативных тестах успешнее ЛСА и обобщенный ЛСА [Budiu, 2007]. Следует отметить, что ЛСА является более универсальным методом для моделирования когнитивных процессов, т. к. результаты его работы зависят только от обучающего корпуса и самого процесса обучения (выбора сингулярных значений диагональной матрицы, способа формирования веса термов и пр.). При этом «меру ассоциативности» можно получить для любого слова, содержащегося в обучающем корпусе. MI, PMI и GLSA также зависят от обучающих корпусов, но исследования синонимичности или ассоциативности в рамках этих методов ограничивается предварительно составленными экспериментаторами списками: при их изменении или расширении требуется полный пересчет модели.

2. Эксперимент

В наших экспериментах использовалась методика, разработанная G. Denhière и B. Lemaire [Denhière, 2004, 2007; Lemaire, 2001, 2003]. Основным отличием нашего исследования является: 1) впервые исследование было проведено

на русскоязычном материале; 2) впервые эксперименты проводились на детях дошкольного возраста (4–7 лет). В связи с этим возник ряд трудностей: от полуручной работы по составлению корпусов детских текстов, соответствующим разным возрастам, и разработке необходимого для экспериментов программного обеспечения для ЛСА, до работы с самими детьми — очень сложными испытуемыми.

Дети были разделены на две категории 4–5 и 6–7 лет, что соответствует возрастным особенностям когнитивного развития (см., например, экспериментальные работы Г. Р. Добровой [Доброва, 2007] по исследованию усвоения детьми лексической семантики или исследования Т. В. Черниговской и Т. И. Свистуновой [Черниговская, 2008] организации ментального лексикона и формированию грамматических правил).

2.1. Материал

Для проведения экспериментов были собраны текстовые корпуса, соответствующие детям двух возрастных групп 4–5 и 6–7 лет, которые подверглись ЛСА.

Материал корпуса — детская литература для детей разного возраста, включая рассказы, сказки, детские энциклопедии, разговорный материал.

Общий объем корпуса составил:

- для детей 4–5 летнего возраста — около 2373 тыс. словоформ.
- для детей 6–7 летнего возраста — около 2416 тыс. словоформ.

В сумме около 4789 тыс. словоформ.

После предподготовки корпуса проводился латентно-семантический анализ с разными параметрами. Варьировались количество сингулярных значений диагональной матрицы, расчет весов термов, использование фонетических слов[2], разбиение корпуса на составные части (документы).

В итоге были получены несколько десятков моделей, из которых методом классификации[3] были выбраны две лучшие модели соответствующие двум возрастным группам.

Для сравнения экспериментальных данных с моделью ЛСА «взрослых» текстов был собран корпус СМИ, объем которого составил более 36 млн. словоформ. Источником для корпуса послужили различные официальные интернет-информационные агентства, такие как www.rbc.ru, www.utro.ru, www.rian.ru, www.interfax.ru и др.

Корпус был обработан по аналогичному алгоритму корпуса детских текстов, и были построены несколько моделей ЛСА с разными размерами векторного

---

[2] Фонетическое слово — знаменательная часть слова плюс клитика [Крылов, 2006].

[3] К сожалению, не существует хорошо разработанных методов определения качества получаемой модели: обычно эмпирическим путем отбирают модель, показавшую согласно дизайну эксперимента наилучшие результаты на тестовых данных. Мы использовали метод автоматической классификации текстов, тематически однородных с обучающей выборкой и заранее размеченных экспертом на классы. Для экспериментов отбиралась та модель, которая показывала наилучшие результаты классификации.

пространства и различным количеством факторов, из которых также методом классификации была выбрана наилучшая.

Сокращение сингулярных значений диагональной матрицы при ЛСА составило около 96% при разбиении текстового пространства на документы по 100 предложений. Несколько лучшие результаты показали модели, в которых использовались фонетические слова и веса рассчитывались на основе меры TFiDF.

2.2. Дизайн эксперимента

На основе собранного и обработанного материала был разработан дизайн эксперимента, состоящий из трех тестов:

- «словарный» тест;
- «ассоциативный» тест;
- тест «на память».

2.2.1. Тест первый: «словарный»

Задача: определить связь между семантическим словарем детей разного возраста (две группы) и ЛСА.

Метод: был составлен список из 20 вопросов, к каждому из которых (каждому ключевому слову) были написаны ответы по следующей градации: точный ответ, близкий, слабо связанный или несвязанный ответ. Испытуемые должны были расположить ответы в порядке точности.

Ключевые слова выбирались из корпуса с весами, полученными методом сравнения терма и документа на модели ЛСА.

Было выявлено количество правильных ответов на каждый вопрос, после чего были построены графики зависимости процента правильных ответов от степени точности ответа.

Аналогичные расчеты были проведены на модели ЛСА, полученной из корпуса текстов СМИ.

Пример. Слово: *фокусник.* Ребенку читали слово и просили дать его определение. «Я тебе назову слово, а ты скажи мне, что оно означает».

1. *показывает фокусы* (точный) = .593
2. *показывает сказки* (близкий) = .573
3. *показывает мультфильмы* (слабосвязанный или несвязанный) = .55

2.2.2. Тест второй: «ассоциативный»

Задача: выявить ассоциативные связи между словами у детей двух возрастных групп и сравнить их с результатами ЛСА.

Метод: испытуемым предлагалось 21 слово (стимул), 7 из которых являлись существительными, 7 — глаголами, 7 — прилагательными. Для каждого из них испытуемые называли от 3 до 5 синонимов-ассоциаций. Полученные частотные вектора объединялись и отсортировывались по частоте; после чего вектора сравнивались с ЛСА-вектором.

Стимулы выбирались из материалов полученного корпуса с величинами, полученными из ЛСА методом парного сравнения термов. Рассчитывалось скалярное произведение между каждым стимулом и каждым ответом на стимул, а также среднее как по каждой части речи, так и в целом по группе. Дополнительно к этому рассчитывалось скалярное произведение между стимулом и первыми тремя наиболее частотными ответами на него. Расчет производился на моделях ЛСА, соответствующих каждой возрастной группе.

Аналогичные расчеты проведены на модели ЛСА, полученной из корпуса текстов СМИ.

Пример. Стимул: *море*. Ребенку читали слово и давали задание подобрать к нему несколько ассоциаций. «Я тебе назову слово, а ты постарайся подобрать такие слова, которые с ним могут быть связаны».

Ассоциации приведены в порядке называния со значениями параметров сравнения с вектором ЛСА:

купаться = .827
киты = .729
волнуется = .613
медузы = .457

2.2.3. Тест третий: «на память»

Задача: исследование памяти с помощью моделей ЛСА.

Метод: испытуемому в соответствии с его возрастной группой читался один из двух взятых из корпуса детских текстов типов рассказов. Испытуемый должен был пересказать его в нескольких предложениях. Пересказ записывался на диктофон и был транскрибирован. Пауза между прочтением и пересказом варьировалась от 15 минут до недели. Результаты пересказов сравнивались с ЛСА-моделью методом сравнения документов.

Проведены аналогичные расчеты на модели ЛСА, полученной из корпуса текстов на материале СМИ.

Для проведения тестов был разработано программное обеспечение: 1) программа для построения моделей ЛСА; а также 2) интерфейсная программа для представления результатов моделей латентно-семантического анализа.

2.3. Проведение экспериментов

В исследовании приняли участие 66 детей дошкольного возраста двух возрастных групп: 39 детей 4–5 лет и 27 детей 6–7 лет.

Запись проводилась в два этапа. На первом этапе нашего исследования были протестированы 31 ребенок (16 детей 4–5 лет и 15 детей 6–7 лет). На втором этапе дополнительно к предыдущему было записано еще 35 детей (23 ребенка 4–5 лет и 12 детей 6–7 лет). Основное отличие второго этапа от первого заключалось в том, что пауза между прочтением и пересказом в тесте «на память» составила сутки и более (на первом этапе этот промежуток был 15–30 минут, что не привело к разнице результатах). При этом в тесте

«на память» повторно удалось записать только 31 ребенка (22 — 4–5 лет и 9 детей 5–6 лет).

Тестирование детей проводили в домашних условиях в течение 1–2 часов на каждого ребенка.

В ходе тестирования с использованием «ассоциативного» теста не все дети (в особенности дети первой группы) понимали поставленную перед ними задачу, поэтому детям приводили несколько примеров или задавали наводящие вопросы. Дети называли от одной до четырех ассоциаций; к некоторым словам дети затруднялись подобрать ассоциации.

Тест «на память» вызвал наибольшие трудности, которые связаны с возрастными особенностями детей, а также проблемой обучения пересказу.

3. Результаты

Тест первый: «словарный»

В этом тесте считались средние значения скалярных произведений между словом-стимулом и первыми тремя словами-ассоциациями, последовательно сказанными испытуемыми. Значения усреднялись как по частям речи, так и учитывалось среднее в целом (см. Табл. 1 и Табл. 2).

Табл. 1. Результаты «словарного» теста для детей первой группы (4–5 лет): для каждого слова-ассоциации считалось скалярное произведение с соответствующей моделью ЛСА. В скобках приводится стандартное отклонение.

Стимул	Номер слова-ассоциации		
	1	2	3
Существительные			
голова	0,791615	0,808147	0,791750
дом	0,624179	0,616711	0,597182
лес	0,715897	0,707528	0,726091
подарки	0,530280	0,512679	0,564222
сад	0,710000	0,655471	0,588852
шкаф	0,532842	0,511568	0,468538
яйцо	0,527514	0,553444	0,590296
среднее по классу	0,633190	0,623650	0,618133
Глаголы			
бежать	0,631778	0,667519	0,696579
брать	0,470103	0,549844	0,478826
жить	0,801375	0,702257	0,734148
кричать	0,642892	0,678710	0,704526
научить	0,698686	0,591900	0,623450
плавать	0,533795	0,506622	0,499556
сесть	0,673263	0,630471	0,640913
среднее по классу	0,635984	0,618189	0,625428

Стимул	Номер слова-ассоциации		
	1	2	3
Прилагательные			
быстрый	0,449308	0,501156	0,447407
весёлый	0,621595	0,659667	0,667769
деревянный	0,694333	0,539829	0,618143
красный	0,578450	0,583789	0,571385
маленький	0,596769	0,603636	0,554720
простой	0,663917	0,679929	0,666706
смелый	0,597378	0,594100	0,549650
среднее по классу	0,600250	0,594587	0,582254
среднее по всем	**0,623141 (0,01980)**	**0,612142 (0,01540)**	**0,608605 (0,02310)**

Табл. 2. Результаты «словарного» теста для детей второй группы (6–7 лет): для каждого слова-ассоциации считалось скалярное произведение с соответствующей моделью ЛСА. В скобках приводится стандартное отклонение.

Стимул	Номер слова-ассоциации		
	1	2	3
Существительные			
волшебник	0,748926	0,629727	0,504533
город	0,731120	0,623320	0,613905
за́мок	0,799889	0,732538	0,632875
лето	0,726407	0,678231	0,645652
море	0,646160	0,618826	0,594952
семья	0,752556	0,600115	0,641920
шкаф	0,561593	0,552640	0,498381
среднее по классу	0,709521	0,633628	0,590317
Глаголы			
бежать	0,748905	0,678826	0,683632
болтать	0,732222	0,747269	0,737588
жить	0,785815	0,76284	0,758550
кричать	0,686423	0,697261	0,706385
плавать	0,727958	0,674833	0,566000
сидеть	0,583444	0,566583	0,736550
учить	0,634808	0,585875	0,551174
среднее по классу	0,699939	0,673355	0,677125
Прилагательные			
весёлый	0,765654	0,719708	0,731500
деревянный	0,591462	0,508080	0,610375
красный	0,605462	0,645304	0,589619
маленькая	0,557846	0,565500	0,642625
прекрасная	0,820074	0,772000	0,745381

Стимул	Номер слова-ассоциации		
	1	2	3
сильный	0,686038	0,627826	0,633056
смелый	0,577308	0,536920	0,500579
среднее по классу	0,657692	0,625048	0,636162
среднее по всем	**0,765654 (0,01980)**	**0,719708 (0,01540)**	**0,731500 (0,02310)**

Как показали вычисления, средние (по всем частям речи) значения скалярных произведений у детей второй группы больше по абсолютному значению и имеют более резкий спад от первой к третьей ассоциации, чем у детей первой группы (см. Рис. 1 и Рис. 2). Причем, у обеих групп класс существительных имеет более стабильную тенденцию к спаду.

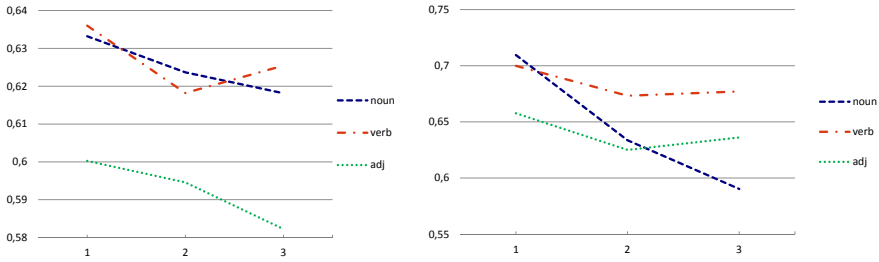

Рис. 1. Средние значения скалярных произведений для первых трех слов-ассоциаций с моделью ЛСА по частям речи для детей первой группы (4–5 лет) — слева и для детей второй группы (6–7 лет) — справа. По оси абсцисс — номера первых трех слов, по оси ординат — значения скалярного произведения.

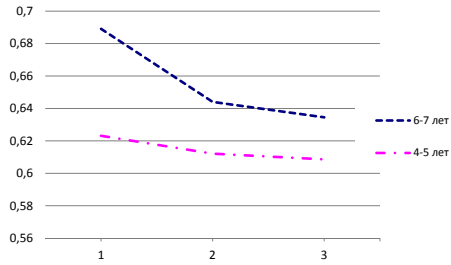

Рис. 2. Средние значения скалярных произведений для первых трех слов-ассоциаций с моделью ЛСА для детей первой группы (4–5 лет) и для детей второй группы (6–7 лет). По оси абсцисс — номера первых трех слов, по оси ординат — значения скалярного произведения.

Аналогичные расчеты были проведены на модели, полученной на текстах СМИ.

Табл. 3. Результаты «словарного» теста для детей первой группы (4–5 лет): для каждого слова-ассоциации считалось скалярное произведение с соответствующей моделью ЛСА, полученной из корпуса СМИ. В скобках приводится стандартное отклонение.

Стимул	Номер слова-ассоциации		
	1	2	3
существительные (среднее по классу)	0,295159	0,266668	0,232557
глаголы (среднее по классу)	0,383547	0,387877	0,372629
прилагательные (среднее по классу)	0,275848	0,220846	0,210666
среднее по всем	**0,318185 (0,051200)**	**0,291797 (0,032400)**	**0,271951 (0,015400)**

Табл. 4. Результаты «словарного» теста для детей второй группы (6–7 лет): для каждого слова-ассоциации считалось скалярное произведение с соответствующим вектором из модели ЛСА, полученной из корпуса СМИ. В скобках приводится стандартное отклонение.

Стимул	Номер слова-ассоциации		
	1	2	3
существительные (среднее по классу)	0,360405	0,333303	0,304718
глаголы (среднее по классу)	0,373503	0,369079	0,346461
прилагательные (среднее по классу)	0,300056	0,294374	0,291618
среднее по всем	**0,344654 (0,031500)**	**0,332252 (0,027500)**	**0,314266 (0,009200)**

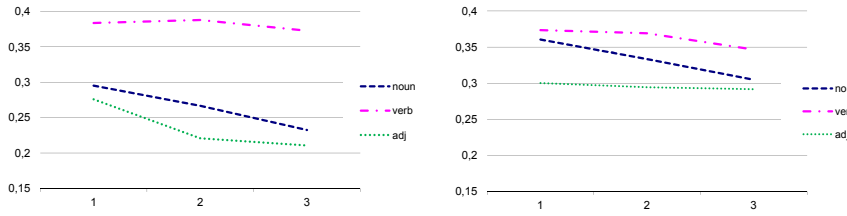

Рис. 3. Средние значения скалярных произведений для первых трех слов-ассоциаций с моделью ЛСА, полученной из корпуса СМИ, для детей первой группы (4–5 лет) — слева и для детей второй группы (6–7 лет) — справа. По оси абсцисс — номера первых трех слов, по оси ординат — значения скалярного произведения.

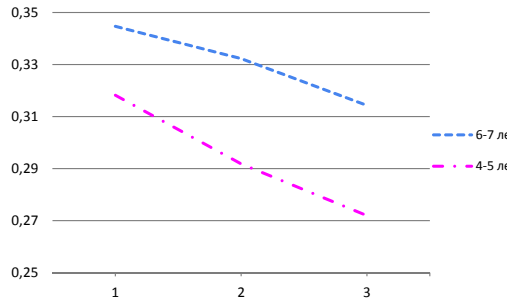

Рис. 4. Средние значения скалярных произведений для первых трех слов-ассоциаций с моделью ЛСА, полученной из корпуса СМИ, для детей первой группы (4–5 лет) — слева и для детей второй группы (6–7 лет) — справа. По оси абсцисс — номера первых трех слов, по оси ординат — значения скалярного произведения.

Как видно из рисунков 3 и 4, значения скалярных произведений у детей первой и второй групп также имеют в среднем тенденцию к уменьшению с каждым последующим ответом, при этом абсолютные значения скалярных произведений примерно в два раза меньше.

Частотный анализ слов-ассоциаций и значения скалярного произведения с моделью ЛСА нескольких первых наиболее частотных слов не показал стабильной зависимости.

Тест второй: «ассоциативный»

Для каждой группы (4–5 лет — 39 испытуемых и 6–7 лет 27 испытуемых) детей было подсчитано количество данных ответов на каждую группу ассоциаций: выбор одного из трех вариантов (точного, близкого и слабо связанного или несвязанного) означал единицу, остальное ноль (Табл. 5).

Табл. 5. Сумма выбранных вариантов ответов детей двух групп (в абсолютном и процентном соотношении); 1 — точный, 2 — близкий, 3 — слабо связанный или несвязанный ответ.

	4–5 лет		6–7 лет	
1	490	63,06%	424	78,37%
2	148	19,05%	99	18,30%
3	139	17,89%	18	3,38%
сумма	777	100%	541	100%

Анализ выбора вариантов показал, что дети второй группы значительно лучше выбирают наиболее близкий ассоциативный вариант (Рис. 5, 6).

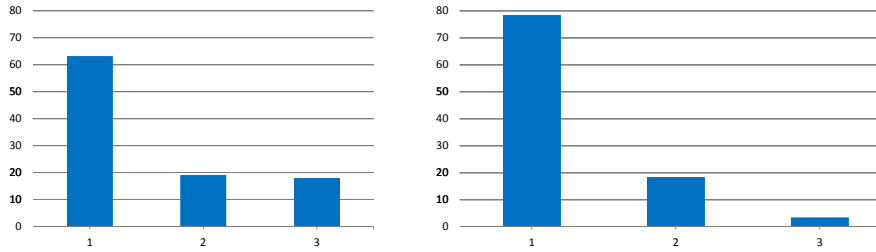

Рис. 5. Результирующие графики выбранных ответов детей первой (слева) и второй (справа) групп. По оси абсцисс — варианты ответа, по оси ординат — сумма ответов в процентном соотношении.

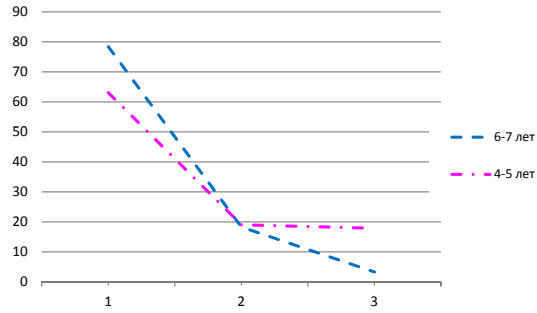

Рис. 6. Линии тренда ответов для детей первой и второй групп. По оси абсцисс — варианты ответа, по оси ординат — сумма ответов в процентном соотношении.

Для обеих групп испытуемых был посчитан коэффициент согласованности (Coeff. of Concordance) ответов. Для первой группы он составил 0,24853, для второй группы — 0,57919.

Далее были получены средние значения результатов значений косинуса между вопросом и ответом для каждой возрастной группы. Аналогичная процедура была проведена с результатами значений косинуса, полученных из модели СМИ.

Результаты в абсолютном и процентном отношении представлены в таблице № 6.

Табл. 6. Средние значения косинусов на «детских» и «взрослом» корпусах в абсолютном и процентном соотношении (в скобках)

Ответы	**1**	**2**	**3**
4–5 лет			
модель 4–5 лет	0,7708 (38,45%)	0,68705 (34,27%)	0,5467 (27,27%)
модель СМИ	0,5126 (34,0%)	0,4958 (32,89%)	0,49885 (33,10%)
6–7 лет			
модель 6–7 лет	0,81175 (37,75%)	0,73395 (34,13%)	0,6045 (28,12%)
модель СМИ	0,4813 (33,57%)	0,4826 (33,66%)	0,4699 (32,77%)

Как видно из таблицы, в отличие от «детских» моделей на «взрослой» модели значение угла косинуса почти не меняется.

Тест третий: «на память»

В третьем тесте измерялась величина скалярного произведения текстов: реального и пересказанного испытуемым в разные моменты времени. На втором этапе в этом тесте был увеличен промежуток времени после первого пересказа до суток и более. Это было связано с тем, что промежуток записи в 15 минут не привел к значимой разнице при анализе данных.

В силу ряда причин, не зависящих от исследователей, повторно удалось записать не всех детей. В итоге в данном тесте были заново записаны 31 ребенок 4–5 лет и 10 детей 6–7 лет.

Результаты сравнения по мере косинуса двух текстов пересказа с источником представлены в табл. 7.

Табл. 7. Значения скалярного произведения исходного и пересказанных детьми текстов за разный промежуток времени

	повтор сразу	**через сутки и более**
4–5 лет		
модель 4–5 лет	0,482375	0,435139
модель СМИ	0,485271	0,467736
6–7 лет		
модель 6–7 лет	0,489400	0,488600
модель СМИ	0,512300	0,514500

Как видно из таблицы 7, значения меняются только у испытуемых первой группы как на «детской», так и на «взрослой» модели.

Табл. 8. Коэффициенты корреляции между результатами ответов сразу и через промежуток времени для испытуемых обеих групп

	коэф.корреляции
модель 4–5 лет	0,326179
модель СМИ	0,461313
модель 6–7 лет	0,350000
модель СМИ	−0,418414

В таблице 8 представлены результаты расчета коэффициента корреляции ($p < 0,05$) между результатами ответов двух этапов теста «на память» (сразу и с задержкой) для всех моделей ЛСА. Как видно из таблицы корреляция результатов лучше для моделей ЛСА, полученных на корпусе СМИ.

4. Выводы

Анализ результатов «ассоциативного» теста показал, что средние значения скалярных произведений первых трех ассоциаций у детей второй, более старшей группы больше по абсолютному значению и имеют тенденцию к более резкому снижению при ослаблении ассоциации (Рис. 2). По всей видимости, это связано с тем, что у детей 6–7 лет должна быть больше развита ассоциативность по сравнению с детьми 4–5 лет. Таким образом, можно сделать вывод, что соответствующие возрастам модели адекватно отражают ассоциативно-семантические связи для когнитивного развития детей обеих групп.

Распределение слов-ассоциаций по частям речи не выявило статистических особенностей в виду небольшой статистики (максимальная частота слова была около 6–7 для стимула, а зачастую и меньше, что недостаточно для статистического анализа, стабильность которого определяется десятками и сотнями повторений).

Сравнение с моделью СМИ показало похожие результаты. При этом абсолютные значения косинуса получились почти в два раза меньше, чем результаты, полученные на «детских» моделях. Это говорит о том, что ассоциативность «взрослой» и «детских» моделей имеют одинаковую тенденцию к ослаблению. При этом «взрослая» ассоциативность на «детских» корпусах ожидаемо ниже, т. к. смоделирована на других текстах. Другой особенностью данного сравнения является то, что если «детские» модели показали разную линейность (разную скорость ослабления ассоциативности) в зависимости от возрастной группы, то на «взрослом» корпусе кривые практически параллельны. Это подтверждает то, что соответствующие «детские» модели более корректно, чем «взрослая» модель, описывают ассоциативно-семантические связи для детей двух возрастных групп.

Результаты «словарного» теста уверено подтверждают результаты «ассоциативного» теста: модели ЛСА согласуются с данными когнитивного развития детей. У испытуемых старшего возраста более развиты ассоциативно-синонимические понятия. Это видно по линии тренда: у детей 6–7 лет более сильная зависимость от силы ассоциативности стимулов (более крутой спад линии тренда), в то время как у детей 4–5 лет связь менее сильна и нелинейна (Рис. 6). Этот результат подтверждает коэффициент согласованности: у детей 6–7 лет он примерно в два раза выше.

Сравнение полученных результатов с «взрослой» моделью ЛСА показало, что значения косинуса на модели СМИ практически не меняется от связанности ответа (Таб. 6). Причиной этого может быть то, что во «взрослом» корпусе отсутствует часть лексики «детских» корпусов, поэтому сравнение идет, в основном, по наиболее употребимым словам, которые примерно одинаково распределены в ответах.

В любом случае сравнение результатов ответов детей разного возраста как между собой, так и со «взрослой» моделью говорит о том, что 1) «детские» модели различаются по силе ассоциативности: у детей старшего возраста ассоциативность развита лучше; 2) модели взрослого и детского восприятия на лексическом уровне существенно различаются.

Исследование памяти (тест «на память») на данных моделях продемонстрировало результаты только для детей первой возрастной группы. Причем

это видно как на «детской», так и на «взрослой» модели, хотя и не столь выраженно. Результаты второй группы не показали изменений. Возможно, это связано с тем, что, во-первых, более старшие дети уже лучше запоминают тексты, а во-вторых, возможно это связано с неудачно выбранной методикой экспериментов: в виду объективных причин дети опрашивались в разные промежутки времени (на следующий день, через 5–7 дней), что не учитывалось при обработке; к тому же детей этой группы удалось записать в три раза меньше, чем первой. Исходя из этого, можно сказать, что метод ЛСА может быть использован при исследовании как кратковременной, так и долговременной памяти у детей, но это требует проверки дальнейшими экспериментами.

Таким образом, результаты нашего исследования показывают, что:

- Метод ЛСА может быть использован для исследований когнитивного развития детей.
- Используемые модели ЛСА отображают процессы развития когнитивного развития детей разных возрастных групп, в частности ассоциативно-семантические процессы и работу кратковременной и долговременной памяти.
- Данный метод рекомендуется использовать для сравнительного исследования когнитивного развития детей, в частности, развития ассоциативно-логического мышления, речевого дискурса, развития памяти.

Литература

1. *Доброва Г. Р.* О некоторых аспектах усвоения лексической семантики детьми 3–6 лет: влияние «нового знания» на речевое поведение // Возраст как фактор речевого поведения. Сборник статей. Пермь: Изд-во ГОВПО Пермского гос. ун-та, 2007.
2. *Крылов С. А.* Делимитация тактов в русском письменном тексте // Труды международной конференции «Корпусная лингвистика-2006». СПб.: Изд-во СПбГУ, 2006. — С. 54–55.
3. *Львов М. Р.* Основы теории речи. — М., 2002.
4. *Соловьев А. Н.* Язык, мышление и современные системы понимания речи // Вестник СПбГУ. Серия Биология (3). Вып. 1. СПб., Изд-во СПбГУ, 2008. С. 99–104.
5. *Соловьев А. Н.* Моделирование процессов понимания речи с использованием латентно-семантического анализа / Диссертация на соискание степени к. ф.н. СПбГУ, 2008.
6. *Черниговская Т. В.* Чеширская улыбка кота Шрёдингера: язык и сознание. — М.: Изд. Языки славянской культуры, 2013.
7. *Черниговская Т. В., Гор К., Свистунова Т. И.* Формирование глагольной парадигмы в русском языке: правила, вероятности, аналогии как основа организации ментального лексикона (экспериментальное исследование) // Когнитивные исследования. Сб. научн. трудов. Вып. 2. / Отв. ред. Т. В. Черниговская, В. Д. Соловьев. М.: Изд-во «Институт психологии РАН», 2008. С. 165–181.

8. *Budiu, R., Royer, C., & Pirolli, P. L.* Modeling information scent: a comparison of LSA, PMI and GLSA similarity measures on common tests and corpora. Proceedings of the 8th Annual Conference of the Recherche d'Information Assistée par Ordinateu. Paris, France, 2007– P. 314–332.
9. *Chomsky N.* Modular Approaches to the Study of the Mind. San Diego: San Diego State University Press, 1984.
10. *Chomsky N.* New Horizons in the Study of Language and Mind. Cambridge University Press, 2002.
11. *Deacon T. W.* Multilevel selection in a complex adaptive system: The problem of language origins. Weber B., Depew D. (eds.). Evolution and Learning: The Baldwin Effect Reconsidered. MIT Press. 2003.
12. *Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K., Harshman R.* Indexing by Latent Semantic Analysis // Journal of the American Society for Information Science. 1990. 41(6). — P. 391–407
13. *Denhière G., Lemaire B., Bellissens C., Jhean-Larose S.* A semantic space for modeling children's semantic memory // D. McNamara, T. Landauer, S. Dennis, W. Kintsch (eds.). The handbook of Latent Semantic Analysis. Mahwah: Lawrence Erlbaum Associates, 2007. — P. 143–165.
14. *Denhière G., Lemaire B., Bellissens C., Jhean-Larose S.* Psychologie cognitive et compréhension de texte: une démarche théorique et expérimentale // S. Porhiel, D. Klinger (eds.). L'unité texte. Pleyben: Perspectives, 2004. — P. 74–95.
15. *Fodor J.* Where is my mind? London Review of Books. Vol. 31, № 3, 12 February. — 2009.
16. *Landauer T. K., Dumais S. T.* A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge // Psychological Review. 1997. 104. — P. 211–240.
17. *Landauer T. K., Foltz P., Laham D.* An Introduction to Latent Semantic Analysis. Discours Processes, 25, 1998 — P. 259–284.
18. *Lemaire B., Bianco M., Sylvestre E., Noveck I.* Un modèle de compréhension de textes fondé sur l'analyse de la sémantique latente // La cognition entre individu et société (actes du colloque de l'ARCo) / H. Paugam Moisy, V. Nyckees, J. Caron-Pargue (eds.). Hermès, 2001. — P. 309–320.
19. *Lemaire B., Denhière G.* Cognitive Models based on Latent Semantic Analysis // Tutorial given at the 5th International Conference on Cognitive Modeling (ICCM'2003), Bamberg, Germany, April 9 2003. — P. 23–25.
20. *Matveeva, I., Levow, G., Farahat, A., & Royer, C.* Terms representation with generalized latent semantic analysis. In Proc. ranlp 2005.