ЦИФРОВИЗАЦИЯ ТЕЗАУРУСНОГО ПОДХОДА В ОБРАЗОВАНИИ

Ю. А. Алябышева[1], **А. Ю. Антонов**[2], **А. А. Веряев**[2]

[1] ***Алтайский государственный университет***
656049, Россия, Алтайский край, г. Барнаул, пр-т Ленина, д. 61

[2] ***Алтайский государственный педагогический университет***
656031, Россия, Алтайский край, г. Барнаул, ул. Молодежная, д. 55

Аннотация
В статье рассказано о том, что такое тезаурусный подход в образовании, обоснованы истоки данного подхода, предложена реализация отдельных направлений цифровизации тезаурусного подхода. Тезаурусный подход иллюстрируется двумя задачами: 1) выявление индивидуальных стилевых особенностей написания авторами текстов; 2) оценка результатов обучения (полуавтоматическая оценка текстов, порождаемых учащимися в ходе учебного процесса). Первая задача проиллюстрирована более подробно, в ней использована информационная мера Кульбака—Лейблера, знакомство с которой может быть полезным для преподавателей информатики и читателей журнала. Кроме того, расчеты, которые сделаны в первой задаче, могут быть воспроизведены даже школьниками в рамках, например, проектной работы. Во второй задаче использован менее чувствительный статистический метод Спирмена. Тезаурусный подход позволяет при использовании информационных методов рассмотреть ряд других задач, имеющих отношение к педагогическому процессу. В частности, основное направление работы по тезаурусному подходу видится в поиске и апробации алгоритмических предписаний количественного/численного анализа текстов, имеющих отношение к образованию или возникающих в образовательном процессе. К указанным выше текстам можно отнести корпус учебных и учебно-методических пособий, тексты отдельных учебных предметов, дисциплин, тем, рефераты, письменные работы, выпускные квалификационные работы студентов, сочинения школьников и др.

Ключевые слова: учебная деятельность, учебный текст, тезаурус, стилевые особенности текста, мера Кульбака—Лейблера, закон Ципфа, интеллектуальный анализ данных, статистический анализ данных, педагогическая стилеметрия, полуавтоматическая оценка текстов обучаемых.

DOI: 10.32517/0234-0453-2020-35-1-51-58
Для цитирования:
Алябышева Ю. А., Антонов А. Ю., Веряев А. А. Цифровизация тезаурусного подхода в образовании // Информатика и образование. 2020. № 1. С. 51–58.
Статья поступила в редакцию: 29 декабря 2019 года.
Статья принята к печати: 21 января 2020 года.
Сведения об авторах
Алябышева Юлия Анатольевна, канд. пед. наук, доцент кафедры информатики, Институт математики и информационных технологий, Алтайский государственный университет, г. Барнаул, Россия; alyabysheva_y@mail.ru; ORCID: 0000-0002-0619-9984
Антонов Александр Юрьевич, аспирант, кафедра теоретических основ информатики, Институт физико-математического образования, Алтайский государственный педагогический университет, г. Барнаул, Россия; sanya.a23@mail.ru; ORCID: 0000-0003-3240-095X
Веряев Анатолий Алексеевич, доктор пед. наук, профессор, зав. кафедрой теоретических основ информатики, Институт физико-математического образования, Алтайский государственный педагогический университет, г. Барнаул, Россия; veryaev_aa@mail.ru; ORCID: 0000-0002-4338-0811

1. Введение

Прежде чем приступать к изложению основного материала статьи, сделаем некоторые предварительные пояснения.

При написании текста перед авторами стояла проблема выбора названия статьи, которое в максимальной степени отражало бы ее содержание. Альтернативный вариант названию, приведенному в заголовке, был такой: «Цифровая трансформация тезаурусного подхода в образовании». Этот вариант мы сочли в меньшей степени отражающим суть того, что к настоящему времени сделано в рамках тезаурусного подхода. Термин «трансформация» предполагает, что все необходимое в рамках подхода уже сформулировано и предложено, и осталось только перевести сделанное в цифровую форму. Напрашивается аналогия с оцифровкой аналогового сигнала, который уже есть в наличии, и имеются алгоритмы и программное обеспечение, с помощью которых это можно сделать. Применительно к данной работе на самом деле всего этого нет, и работы по цифровизации тезаурусного подхода носят поисковый и творческий характер.

В статье представлены две модельные задачи и такой возможный для потенциального осуществления класс задач по тезаурусному подходу, который ранее в педагогической литературе не рассматривался. Термин «цифровизация» с этой точки зрения является более емким, при этом совершенно не очевидно, какая математика, какие алгоритмы будут выбраны для перевода тезаурусного подхода в цифровую форму.

2. Представления о тезаурусном подходе

Нужно отметить, что тезаурусный подход достаточно редко упоминается в качестве используемого в диссертационных исследованиях по педагогике.

Поэтому имеет смысл начать с пояснения ключевых идей подхода и анонсирования того класса задач, которые могут быть решены при использовании современных информационных технологий, ставших доступными в сфере образования.

Начнем с некоторых исторических замечаний.

В работах К. Шеннона была предложена достаточно простая коммуникационная модель, которая в настоящее время чаще называется моделью Шеннона—Уивера [1]. На рисунке 1 приведена ее графическая иллюстрация: по каналу связи от источника информации к получателю передаются сигналы, а шум в канале связи препятствует этому процессу.

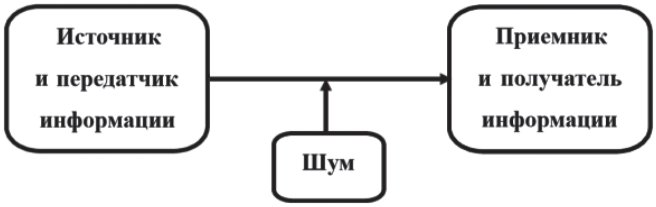

Рис. 1. Коммуникационная модель Шеннона—Уивера

Указанную модель практически сразу пытались приспособить к описанию и объяснению процесса обучения, объявляя педагогов источниками информации, а обучаемых — получателями информации. Однако не все так просто с применением этой модели в педагогике.

Во-первых, источник информации и получатель в учебном процессе постоянно меняются местами. Это замечание легко исправимо на рисунке расстановкой стрелок на изображении канала передачи информации и усложнением надписей на черных ящиках. Однако нужно отметить, что в учебном процессе могут меняться и педагогические позиции источника информации — учителя: он на время может влиться в группу обучаемых. При этом важно также, что взаимодействие идет не в системе «один к одному», а в системе «один ко многим», в группе многих могут организовываться коллективные субъекты учебной деятельности, и такие эффекты часто рассматриваются в педагогической литературе.

Во-вторых, если считать, что основным фактором, мешающим получению информации, является шум, то мы придем к абсолютизации педагогической установки: «Повторение — мать учения», чего не должно быть.

В-третьих, мы сознательно не указали на рисунке 1, что же передается по каналу связи: сигналы, данные, информация или знания. Вопрос этот чрезвычайно серьезный. Нельзя отождествлять то, что передается по каналам связи, с тем, что в конечном итоге формируется у получателя информации, однако часто это делается в ряде образовательных подходов. Очень часто на аналогичных рисунку 1 графических представлениях модели Шеннона—Уивера можно найти надпись, говорящую о передаче «сообщений». Различные ответы на вопрос: «Что передается?» приводят к разным подходам к организации педагогического процесса. В рамках тезаурусного подхода мы будем ориентироваться на то, что по каналам связи передаются «тексты», которые в процессе цифровизации нужно обрабатывать с учетом максимально возможной их семантизации. Здесь уместно вспомнить хорошо известную аббревиатуру «ЗУН», выступающую в качестве одной из составляющих целевых установок школьного или вузовского образования. В текстах (обучаемых и обучающих) не только репрезентируется знаниевый элемент, но и вербализованы умения, навыки, компетенции/компетентности, обсуждаемые в литературе УУД, личностные характеристики, опыт творческой и иной деятельности [2].

В-четвертых, в модели неявно предполагается, что все, что послано источником информации, может быть воспринято получателем, он к этому готов. На самом деле это не так, и процесс обучения состоит в том, что имеет место «перестройка» приемника информации, его возможностей. Психологи говорят, что формируются новообразования и, кроме того, меняются вероятностные связи в памяти приемника информации. А для того, чтобы учебная коммуникация была успешной, нужно также, чтобы и источник информации подстраивался под возможности ее получателя. Это замечание делает обоснованным введение понятия «тезаурус» в модель Шеннона—Уивера, подстраиваемую под нужды педагогики. Это и дает основание назвать подход, рассматриваемый в статье, тезаурусным.

Под ***тезаурусом*** в рамках данной работы будем понимать «множество смысловыражающих единиц некоторого языка с заданной на нем системой семантических отношений. Тезаурус фактически определяет семантику языка. <...> В широком смысле тезаурус интерпретируют как описание системы знаний о действительности, которыми располагает индивидуальный носитель информации или группа носителей. Этот носитель может выполнять функции приёмника дополнительной информации, вследствие чего изменяется и его тезаурус. Исходный тезаурус определяет при этом возможности приёмника при получении им семантической информации». Мы привели выдержки из статьи, написанной для Большой советской энциклопедии Ю. А. Шрейдером (см. также [3]). Именно Ю. А. Шрейдеру принадлежит введение в рассмотрение семантической информации, базирующееся на изменении (росте) тезауруса приемника информации [4].

Таким образом, одна из первых работ о роли тезаурусов была выполнена Ю. А. Шрейдером. Более подробно роль тезаурусов в педагогике отмечена и систематизирована в монографии Л. Т. Турбовича, вышедшей 50 лет назад [5], когда возможности автоматической обработки текстов были крайне ограничены. Центральной идеей в указанной работе является использование при рассмотрении педагогического процесса коммуникационной модели Шеннона—Уивера. Отметим еще раз ограниченный характер этой модели без учета изменений во времени характера работы и возможностей как источника, так и приемника информации. Тем не

менее Л. Т. Турбович обратил внимание на важность рассмотрения «сообщений» в этой модели, а следовательно, текстов, которыми обмениваются источник и приемник информации. На анализе этих текстов и будет центрироваться данная статья.

С одной стороны, нами взята за основу при рассмотрении тезаурусного подхода кибернетическая модель коммуникации, но, с другой стороны, она требует максимального привлечения положений семантики, процессов означивания, осмысливания текстов, чего в кибернетических моделях нет. В подтверждение сказанного достаточно вспомнить дискуссии и многочисленные предложения по введению синтаксических, семантических и прагматических способов измерения информации. Отсюда следует важный вывод: построение оцифрованного тезаурусного подхода в педагогике возможно только при применении достижений компьютерной лингвистики, с использованием данных, накопленных в корпусе современного русского языка (http://www.ruscorpora.ru/new/), такой характеристики, как TF-IDF (от *англ.* TF — term frequency, IDF — inverse document frequency) — статистической меры, используемой для оценки важности слова в контексте документа, а также других способов анализа текстов.

У читателя может возникнуть вопрос: почему речь в статье идет именно о подходе в педагогике, а не об отдельных методах рассмотрения педагогических явлений? Ответ достаточно прост. Легко увидеть, что применение модели Шеннона—Уивера в педагогике во всей ее полноте может повлиять на описание, а иногда и на определение всех составных частей педагогических систем (цель, содержание образования, формы, методы, средства обучения). Для этого нужно смотреть на модель Шеннона—Уивера с точки зрения разных временны́х масштабов: изучать и анализировать, какие тексты должны быть усвоены обучаемыми в течение всех лет обучения, или в рамках учебного курса, или в рамках цикла занятий. Тезаурусный подход позволяет создавать системы оценки качества образования, с его помощью можно пытаться организовывать индивидуальные образовательные маршруты обучаемых и т. д. Детальные обоснование и объяснение сказанного выходят за рамки настоящей статьи. Отметим, что в рамках тезаурусного подхода авторами ранее была разработана трехстадийная модель обучения разговорным темам в процессе преподавания английского языка [6].

Тезаурусный подход плодотворно использовался в лингвистической научной школе Ю. В. Рождественского (см., например, первый том его избранных работ [7]). Что касается более современных работ, то можно назвать ряд авторов, работавших в данном направлении: В. П. Вейдт, Т. А. Кувалдина, В. А. Луков, Л. А. Лукина, Н. В. Сидорова, Л. Ю. Монахова, Т. А. Новикова, В. А. Сидорина.

Зарубежные исследования по тезаурусному подходу в образовании ориентированы на методики изучения отдельных понятий, выявление наличия синонимов, антонимов в текстах. Это работа с отдельными словами и словоформами, а не с их статистическими характеристиками. Практически все отечественные и зарубежные работы по тезаурусному подходу, опубликованные за последние несколько десятков лет, не имеют отношения к цифровизации подхода.

3. Проблема цифрового представления текста

Ключевым моментом в цифровизации тезаурусного подхода является векторизация текста, его цифровое представление. Остановимся на этом вопросе подробнее, поскольку это позволит лучше понять две предлагаемые для иллюстрации задачи.

Существует несколько методов превращения слов в цифры. Делается это в рамках обработки естественного языка (Natural Language Processing, NLP). Далее мы будем использовать только представление текста в виде модели, получившей название «мешок слов» (bag-of-words). Суть ее состоит в том, что в тексте подсчитывают частоту, с которой встречаются слова. Как правило, при подсчетах выделяют словоформы, намного реже — словосочетания. Словоформы ранжируются по частоте (от наиболее часто встречающихся к наименее часто встречающимся). Модель порождает частотное распределение Дж. Ципфа (1948 год, его также называют гиперболическим ранговым распределением) [8]. Такие распределения часто возникают в том числе и в педагогике, однако редко используются для анализа педагогических процессов и явлений. На рисунке 2 приведено такое распределение.

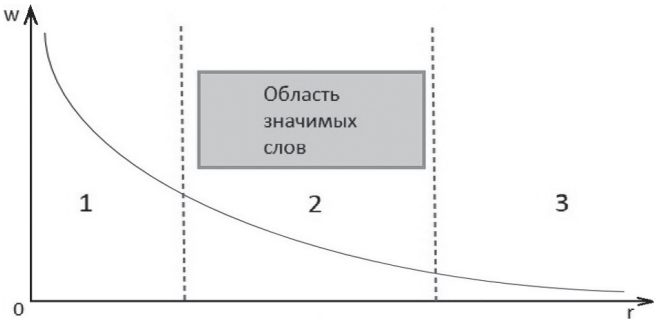

Рис. 2. Ранговое распределение словоформ/слов по частоте, с которой они встречаются в тексте. Здесь r — ранг слова, w(r) — количество слов ранга r, встретившихся в тексте. Эмпирическая зависимость на рисунке 2 описывается формулой: $w(r) = C/r$

Если положить $r = 1$, что соответствует наиболее часто используемому в тексте слову, то становится понятным смысл величины C. Слово с рангом 2 встретится в тексте с частотой в два раза меньшей: $C/2$, и т. д. Часто вместо графика, представленного на рисунке 2, для иллюстрации приводят его логарифмическую форму. Распределение на рисунке 2 приводит к хорошо известному принципу Парето — «80/20». Горизонтальная ось на рисунке репрезентирует тезаурус текста, который и породил гиперболическое распределение. Вопрос о выводе

или обосновании гиперболических распределений до настоящего времени открыт [9].

Не будем приводить в рамках статьи дополнительную информацию о законе Ципфа, точности его выполнения и т. д. Читателю, который заинтересовался этим вопросом, порекомендуем публикации [10, 11]. Первая из этих публикаций посвящена преимущественно наукометрии, вторая имеет отношение к описанию социальных явлений. Если кого-то заинтересуют ранговые распределения в технике, то можно порекомендовать познакомиться с сайтом и публикациями Б. И. Кудрина (http://www.kudrinbi.ru/). Материал статьи станет чуть более знакомым, если читатель вспомнит о так называемых облаках тегов, связанных с законом Ципфа.

На графике, приведенном на рисунке 2, принято выделять три области.

Первая область соответствует так называемым стоп-словам. Их насчитывают несколько сотен. Первые двадцать стоп-слов в русском языке такие: «и», «в», «не», «на», «я», «быть», «он», «с», «что», «а», «по», «это», «она», «этот», «к», «но», «они», «мы», «как», «из». В английском языке самым популярным является артикль «the», который встречается в среднем в каждой последовательности из 16 слов. К стоп-словам относят предлоги, местоимения, междометия, союзы и пр. Слова из первой области распределения Ципфа говорят не о содержании написанного текста, а о том, как написан текст, о его стилевых особенностях. Мусорными стоп-слова можно называть только в контексте задач, связанных с поиском информации [12]. Область стоп-слов оказывается наиболее статистически устойчивой.

Вторая область на графике рангового распределения соответствует словам, которые несут основное смысловое содержание текста. Слова и обороты из этой области нужно использовать при поиске документов в сети Интернет с помощью поисковых систем.

Наконец, третья область соответствует так называемому хвосту распределения. Отметим, что, чем больше берется текст для анализа, тем сильнее «отрастает хвост». Количество уникальных слов в тексте связано с объемом текста и описывается эмпирической закономерностью, называемой законом Хипса [13].

4. Мера Кульбака—Лейблера

Перед тем как непосредственно перейти к обсуждению двух задач из тезаурусного подхода, познакомим читателя с используемым в первой задаче математическим аппаратом для расчетов. С нашей точки зрения, его роль выходит далеко за границы рассматриваемых задач, и аппарат может быть широко использован в исследованиях по педагогике. Речь пойдет об информационной мере Кульбака, более известной в литературе как «расходимость Кульбака—Лейблера».

Пусть заданы два вероятностных распределения на одном и том же множестве событий: $P(i)$, $Q(i)$. Переменная i нумерует события. Мы рассматриваем дискретный случай. В то же время все сказанное ниже может быть распространено и на континуальный случай. Меру, или расхождение, Кульбака—Лейблера распределения Q относительно P принято записывать как $D_{\mathrm{KL}}(P\|Q)$ и определять:

$$D_{\mathrm{KL}}(P\|Q) = \sum P(i) \log_2 \frac{P(i)}{Q(i)}. \qquad (1)$$

Первая работа, где введена предлагаемая к использованию мера Кульбака—Лейблера, была описана в 1951 году [14], см. также монографию [15].

Первый аргумент функционала (распределение P) обычно интерпретируется как истинное или постулируемое априори распределение, второй аргумент (распределение Q) — как проверяемое распределение. Суммирование в формуле идет по всем событиям i. Распределение $Q(i)$ служит приближением распределения $P(i)$. Значение функционала часто интерпретируют как расстояние между P и Q. Но нужно заметить, что $D_{\mathrm{KL}}(P\|Q) \neq D_{\mathrm{KL}}(Q\|P)$. Это означает, что «расстояния» от Q до P и от P до Q — разные, функционал (1) не является метрикой в пространстве распределений и не удовлетворяет неравенству треугольника. Этот факт нас вполне устраивает в силу семантики и интерпретации предлагаемых ниже вероятностных мер, поэтому симметризация выражений для расстояния между P и Q, как это часто делается, не требуется, однако наши геометрические образы и представления при интерпретации результатов перестают работать, и это нужно иметь в виду на последних стадиях визуализации и анализа результатов. Выбор основания логарифма в формуле непринципиален. Если основание равно 2, то безразмерное расстояние измеряют в битах. Расстояние Кульбака—Лейблера всегда неотрицательно и равно нулю в том случае, когда $P = Q$.

Обратим внимание на то, что мера P может быть отнесена в педагогике к тем частотам событий, что задаются учебниками, пособиями, тем, что ожидает от обучаемых учитель. А то, что реально наблюдается как результат обучения, описывается мерой Q. Таким образом, меру Кульбака—Лейблера можно использовать для оценки учебных достижений. Совокупность всевозможных вероятностных мер на разных этапах обучения можно рассматривать в качестве цифрового следа обучаемого. Объемность цифрового следа приводит к такому понятию, как «большие данные» (Big Data). Технологии с условным названием «большие данные», с нашей точки зрения, очень важны, их можно использовать в процессе цифровизации тезаурусного подхода. На это мы обращали внимание в своей обзорной статье [16].

Метод Кульбака—Лейблера на практике используют в ряде задач, например, при распознавании голоса, в медицине для диагностики патологических процессов, в психологии, в информатике для контроля сетевого трафика, распознавания образов, при изучении проблем информационного поиска и др. Часто мера Кульбака—Лейблера упоминается в работах по определению авторства текста или определению

его стилистических характеристик. В то же время нам не известны работы, в рамках которых авторы анализировали бы с помощью указанной меры входящие в тексты стоп-слова, а также не встречались работы, в которых использовался бы указанный метод в педагогике.

5. Первая иллюстративная задача

Используя информационную меру Кульбака—Лейблера, а также вероятностное распределение стоп-слов в текстах авторов, попытаться выявить количественные характеристики, имеющие отношение к авторским особенностям написания текста.

Данная задача имеет давнюю историю. Еще в начале прошлого века Н. А. Морозов (1915) пытался ввести так называемые лингвистические спектры, характеризующие авторский стиль написания текстов. С ним в полемическую дискуссию вступил академик А. А. Марков (1916). С тех пор было выпущено достаточно много работ, проведено множество исследований, защищено немалое количество диссертаций. Одна из последних работ, в которой обращается внимание на роль служебных слов (их авторы статьи называют скрепами) — это статья М. Ю. Михеева, Л. И. Эрлиха [17].

В процессе получения вероятностных мер P и Q нами анализировались фрагменты художественных произведений объемом по 5000 слов. В работе [18] было показано, что такого количества слов достаточно для приемлемой классификации текстов. Для поиска частоты стоп-слов использовался сервис на сайте: https://webscript.ru

После обработки фрагментов литературных произведений было выбрано множество из 20 стоп-слов, которое присутствует во всех анализируемых текстах. Зная частоту употребления слова в выбранном множестве слов, несложно перейти к вероятностным мерам P или Q, что и было сделано. Отметим, что варьирование мощности выделяемого множества стоп-слов — предмет дальнейших исследований.

Приведем лишь часть результатов проделанного анализа.

Были проанализированы фрагменты произведений Л. Н. Толстого «Война и мир» (ВМ), «Анна Каренина» (АК), Ф. М. Достоевского «Преступление и наказание» (ПН), «Идиот» (И), Н. В. Гоголя «Мертвые души» (МД), «Вий» (В).

Далее приведены меры Кульбака—Лейблера, вычисленные по методу, изложенному выше:

$D_{KL}(ВМ\|АК) = 0{,}04586$; $D_{KL}(АК\|ВМ) = 0{,}03843$;
$D_{KL}(ПН\|И) = 0{,}04380$; $D_{KL}(И\|ПН) = 0{,}04034$;
$D_{KL}(МД\|В) = 0{,}04892$; $D_{KL}(В\|МД) = 0{,}04721$;
$D_{KL}(ВМ\|ПН) = 0{,}06075$; $D_{KL}(ПН\|ВМ) = 0{,}05782$;
$D_{KL}(ПН\|МД) = 0{,}06627$; $D_{KL}(МД\|ПН) = 0{,}05881$;
$D_{KL}(ВМ\|МД) = 0{,}08789$; $D_{KL}(МД\|ВМ) = 0{,}07574$.

Нетрудно заметить, что внутриавторские расхождения Кульбака—Лейблера меньше, чем межавторские. В процентном отношении это расхождение иногда достаточно значительное. Таким образом, мера Кульбака—Лейблера, подсчитанная на подмножестве стоп-слов, дифференцирует (различает) авторские стилевые особенности написания текстов. Это позволяет надеяться на выявление такой характеристики, как «стиль учебной деятельности», имеющей непосредственное отношение к педагогике.

Нами был реализован также и статистический метод различения текстов, основанный на использовании критерия Спирмена. Детали расчета здесь описывать не будем, поскольку статистический метод оказался менее чувствительным по сравнению с информационным.

Из результатов можно сделать ряд выводов.

Метод Кульбака—Лейблера достаточно чувствителен. Почему стоп-слова дают такой эффект — нужно дополнительно разбираться. Расхождение Кульбака—Лейблера описывает интегративный эффект и не требует индивидуальных сравнений по каждой «скрепе», как это сделано в работе М. Ю. Михеева и Л. И. Эрлиха [17]. Дополнительного исследования требует зависимость эффекта от объемов анализируемого текста. Метод открывает возможности цифрового изучения эволюции индивидуального стиля во времени, привязки стиля к тематике текстового произведения. Несложная математика позволяет надеяться на то, что исследования, аналогичные приведенному, могут реализовываться даже в школе в качестве проектной работы.

6. Вторая иллюстративная задача

Очередная задача тезаурусного подхода посвящена полуавтоматической оценке ответов обучаемых. Поскольку для этой задачи важно содержание текстов, для анализа и цифровизации выбирались слова преимущественно из области 2 распределения Ципфа, но, если эталонные учебные тексты требовали знания слов из областей 1 и 3, они вводились в рассмотрение. Тексты, которые анализировались в этой задаче, намного меньше по объему, чем в предыдущей задаче. Объем сданных учащимися текстов составлял примерно 300 слов. По этой причине порождаются ситуации, когда в вероятностной мере Q возникают события, для которых некоторые $Q(i)$ оказываются равными нулю. То есть отдельные слова/словоформы обучаемые просто не усвоили, забыли, не использовали в отчете или ответе. В выражении для расходимости Кульбака—Лейблера для таких слов возникает расходимость (нули в знаменателях), которую нужно обрабатывать особо. Мы же стремились реализовать некоторый унифицированный подход. По этой причине мы отказались от использования метода Кульбака—Лейблера и решили считать эффекты, опираясь на статистический критерий Спирмена. Этот метод, как показывают наши оценки, является менее чувствительным, но тем не менее в данной задаче он приводит к положительным, нужным нам результатам.

Следует отметить, что процедура оценивания учебных достижений является достаточно сложной,

противоречивой, несущей заметный элемент субъективности. Не случайно обучаемые предпочитают компьютерные формы контроля, считая их более объективными. Педагоги-методисты для упрощения процедур оценивания вводят всевозможные нормы-образцы, пытаются классифицировать задания по уровню сложности, вводят градации, отражающие и учитывающие трудность и сложность заданий.

Если обратиться к хорошо известной таксономии Блума целей обучения (сейчас принято выделять две таксономии Блума: старую, от 1956 года, и модифицированную, сделанную учениками Блума), то можно обнаружить, что первые три цели («знание», «понимание», «использование») достаточно легко репрезентируются в заданиях, которые затем могут быть подвергнуты автоматической или компьютерной проверке. Совсем иное — это проверка достижения трех других целей обучения, которые в рамках таксономии фиксируются как умение обучающимися проводить «анализ», «синтез», «оценку» ситуаций (профессиональных, квазипрофессиональных, жизненных и пр.). Задачи, отражающие первые три цели обучения, проверяются на умении выстраивать цепочки логических умозаключений, алгоритмически мыслить, фреймировать свою деятельность. Совсем другое дело — решение задач, отражающих следующие три цели обучения. Здесь, как правило, обучаемый для проверки предоставляет текст, который является слабо структурированным объектом, анализировать который сложно.

Мы не претендуем на то, что предлагаемая процедура полностью решит проблему оценивания неструктурированного текста. Она всего лишь может помочь преподавателю более результативно разобраться в работах обучаемых, чтобы в дальнейшем обратить внимание на выявленные слабые или сильные стороны работ. Ситуация напоминает ставшую общепринятой практику проверки сдаваемых текстов на заимствования. После того как такая проверка проведена, все равно приходится разбираться с тем, свой ли текст заимствован, много ли источников проработал автор или списал из одного места, например, реферата с одноименным названием. Такая автоматическая проверка работ на заимствования очень полезна в работе учителя и преподавателя вуза. Также может оказаться полезной предварительная оценка текстов по предлагаемому алгоритму.

Суть алгоритма состоит в следующем.

На первом этапе ранжируются по частоте слова из эталонного текста, затем строится второй проранжированный ряд с использованием ответа, представленного обучаемым. После согласования (сопряжения) второго ряда с первым происходит вычисление ранговой корреляции Спирмена [19].

Чем ближе коэффициент корреляции к 1, тем «адекватнее» проверяемая работа эталонной. Последнее утверждение выступало в качестве гипотезы в настоящем исследовании. Мы проверили это утверждение, сравнив оценки, выставленные педагогом при традиционном чтении и анализе текстов, с набором коэффициентов ранговой корреляции Спирмена, полученных на первом этапе оценивания за эти же работы. Проверка осуществлялась с использованием того же критерия ранговой корреляции Спирмена. Поскольку набор оценок учителя ограничен («отлично», «хорошо», «удовлетворительно»), а группы школьников составляли от 10 до 15 человек, в этом случае вычисление ранговой корреляции Спирмена было более сложным, учитывающим наличие одинаковых рангов в рядах [19]. В первом случае при осуществлении расчетов сопряжение последовательностей рангов определялось словами, во втором случае сопряжение последовательностей осуществлялось перечнем учащихся. (Не будем на страницах журнала приводить математические выкладки и расчетные формулы.)

Эталонный текст был взят из учебника В. В. Артемова по учебной дисциплине «История» (для всех специальностей СПО). Учебник актуален, был издан в 2018 году в соответствии с ФГОС СПО для всех специальностей. После семантического анализа, проведенного с помощью ресурса «Адвего — SEO-анализ текста» (https://advego.com/text/seo/), была выявлена плотность ключевых слов, так называемого семантического ядра, которое мы проранжировали по частоте употребления, предписав каждому слову порядковое число. Обучающиеся выполняли письменные задания в рамках изучения материала учебной темы, собранные работы были оценены и иерархично выстроены в зависимости от отметки. Отметим, что гипотеза, о которой говорилось выше в тексте, в основном подтверждается. Корреляции между оценками учителя и автоматической «оценкой» текстов получаются выше 0.85.

7. Заключение

Таким образом, в настоящей статье продемонстрированы основные положения тезаурусного подхода в педагогике. Показано, как можно осуществлять цифровизацию этого подхода. В качестве примера разобраны две модельные задачи. Результаты позволяют надеяться на дальнейшие продвижения в деле цифровизации подхода, в основе которого рассматривается «учебный текст».

Список использованных источников

1. *Shannon C. E., Weaver W.* The mathematical theory of communication. Urbana: University of Illinois Press, 1971. 144 p.

2. *Веряев А. А.* Отражение в лексиконе учащихся их образованности, личностных качеств и характеристик. Возможности исследования // Школьные технологии. 2010. № 5. С. 123–133.

3. *Шрейдер Ю. А.* Тезаурус в информатике и теоретической семантике // Научно-техническая информация. 1971. № 3.

4. *Шрейдер Ю. А.* Об одной модели семантической информации // Проблемы кибернетики. 1965. № 13. С. 233–240.

5. *Турбович Л. Т.* Информационно-семантическая модель обучения. Ленинград: ЛГУ, 1970. 133 с.

6. *Антонов А. Ю., Веряев А. А.* Использование пакета CmapTools и концепт-карт в процессе обучения англий-

скому языку // Преподаватель XXI век. 2017. № 1-1. С. 9–19. http://prepodavatel-xxi.ru/sites/default/files/PXXI_2017-1-1.pdf

7. *Рождественский Ю. В.* Философия языка. Культуроведение и дидактика. Т. 1. М.: Грантъ, 2003. 239 с.

8. Закон Ципфа. https://ru.wikipedia.org/wiki/Закон_Ципфа

9. *Маслов В. П., Маслова Т. В.* О законе Ципфа и ранговых распределениях в лингвистике и семиотике // Математические заметки. 2006. Т. 80. № 5. С. 718–732. DOI: 10.4213/mzm3081

10. *Яблонский А. И.* Модели и методы исследования науки. М.: Эдиториал УРСС, 2001. 400 с.

11. *Хайтун С. Д.* Количественный анализ социальных явлений. М.: КомКнига, 2010. 280 с.

12. *Маннинг К. Д., Рагхаван П., Шютце Х.* Введение в информационный поиск. М.: Вильямс, 2011. 528 с.

13. Закон Хипса. https://ru.wikipedia.org/wiki/Закон_Хипса

14. *Kullback S., Leibler R. A.* On information and sufficiency // The Annals of Mathematical Statistics. 1951. Vol. 22. No. 1. P. 79–86. DOI: 10.1214/aoms/1177729694

15. *Кульбак С.* Теория информации и статистика. М.: Наука, 1967. 408 с.

16. *Веряев А. А., Татарникова Г. В.* Educational Data Mining и Learning Analytics —направления развития образовательной квалитологии // Преподаватель XXI век. 2016. № 2-1. С. 150–160. https://drive.google.com/file/d/0Bww8v66PaPfWSFpSbHlqQTJ0aHM/view

17. *Михеев М. Ю., Эрлих Л. И.* Идиостилевой профиль и определение авторства текста по частотам служебных слов // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2018. № 2. С. 25–34.

18. *Шевелев О. Г.* Разработка и исследование алгоритмов сравнения стилей текстовых произведений: автореф. дис. ... канд. тех. наук. Томск, 2006. 18 с.

19. *Гмурман В. Е.* Теория вероятностей и математическая статистика. М.: Высшая школа, 2003. 479 с.

DIGITALIZATION OF THE THESAURUS APPROACH IN EDUCATION

Yu. A. Alyabysheva[1], **A. Yu. Antonov**[2], **A. A. Veryaev**[2]

[1] ***Altai State University***
656049, Russia, Altai Krai, Barnaul, prospect Lenina, 61

[2] ***Altai State Pedagogical University***
656031, Russia, Altai Krai, Barnaul, ul. Molodezhnaya, 55

Abstract

The article describes what the thesaurus approach in education is like, its origins are justified, and it also suggests the implementation of certain areas of digitalization of the thesaurus approach. The thesaurus approach is illustrated by two tasks: 1) the identification of individual stylistic features of authors' texts; 2) assessment of learning outcomes (semi-automatic assessment of texts generated by students during the educational process). The first task is illustrated in more details, in it the Kullback—Leibler information measures are used, familiarity with which, in our opinion, will be useful to informatics teachers and readers of the journal. In addition, the calculations made in the first task can be reproduced even by school students in the framework of, for example, project activity. In the second task, the less sensitive Spearman statistical method was used. The thesaurus approach allows using information methods to consider a number of other tasks related to the pedagogical process. In particular, the main direction of work on the thesaurus approach is seen in the search and testing of algorithmic prescriptions for the quantitative/numerical analysis of texts related to education or arising in the educational process. The aforementioned texts include the corpus of educational and teaching aids, texts of individual academic subjects, disciplines, topics, essays, written works, final qualifying works of students, essays made by school students, etc.

Keywords: educational activity, educational text, thesaurus, stylistic features of text, Kullback—Leibler measure, Zipf's law, data mining, statistical data analysis, pedagogical stylemetry, semi-automatic assessment of student texts.

DOI: 10.32517/0234-0453-2020-35-1-51-58

For citation:

Alyabysheva Yu. A., Antonov A. Yu., Veryaev A. A. Tsifrovizatsiya tezaurusnogo podkhoda v obrazovanii [Digitalization of the thesaurus approach in education]. *Informatika i obrazovanie — Informatics and Education*, 2020, no. 1, p. 51–58. (In Russian.)

Received: December 29, 2019.

Accepted: January 21, 2020.

About the authors

Yulia A. Alyabysheva, Candidate of Sciences (Education), Associate Professor at the Department of Informatics, Institute of Mathematics and Information Technologies, Altai State University, Barnaul, Russia; alyabysheva_y@mail.ru; ORCID: 0000-0002-0619-9984

Alexander Yu. Antonov, postgraduate student, the Department of Theoretical Foundations of Informatics, Institute of Physical and Mathematical Education, Altai State Pedagogical University, Barnaul, Russia; sanya.a23@mail.ru; ORCID: 0000-0003-3240-095X

Anatoly A. Veryaev, Doctor of Sciences (Education), Professor, Head of the Department of Theoretical Foundations of Informatics, Institute of Physical and Mathematical Education, Altai State Pedagogical University, Barnaul, Russia; veryaev_aa@mail.ru; ORCID: 0000-0002-4338-0811

References

1. *Shannon C. E., Weaver W.* The mathematical theory of communication. Urbana, University of Illinois Press, 1971. 144 p.

2. *Veryaev A. A.* Otrazhenie v leksikone uchashhikhsya ikh obrazovannosti, lichnostnykh kachestv i kharakteristik. Vozmozhnosti issledovaniya [Reflection in the vocabulary of students of their education, personal qualities and characteristics. Research opportunities]. *Shkol'nye tekh-*

nologii — School Technologies, 2010, no. 5, p. 123–133. (In Russian.)

3. *Schreider Yu. A.* Tezaurus v informatike i teoreticheskoj semantike [Thesaurus in informatics and theoretical semantics]. *Nauchno-tekhnicheskaya informatsiya — Scientific and Technical Information*, 1971, no. 3. (In Russian.)

4. *Schreider Yu. A.* Ob odnoj modeli semanticheskoj informatsii [About one model of semantic information]. *Problemy kibernetiki — Cybernetics Problems*, 1965, no. 13, p. 233–240. (In Russian.)

5. *Turbovich L. T.* Informatsionno-semanticheskaya model' obucheniya [Information-semantic model of learning]. Leningrad, LSU, 1970. 133 p. (In Russian.)

6. *Antonov A. Yu., Veriaev A. A.* Ispol'zovanie paketa CmapTools i kontsept-kart v protsesse obucheniya anglijskomu yazyku [CmapTools and concept-maps in the process of learning English]. *Prepodavatel XXI vek— Teacher of the 21st Century*, 2017, no. 1-1, p. 9–19. (In Russian.) Available at: http://prepodavatel-xxi.ru/sites/default/files/PXXI_2017-1-1.pdf

7. *Rozhdestvensky Yu. V.* Filosofiya yazyka. Kul'turovedenie i didaktika. T. 1 [Philosophy of language. Cultural studies and didactics . Vol. 1]. Moscow, Grant", 2003. 239 p. (In Russian.)

8. Zakon Tsipfa [Zipf's law]. (In Russian.) Available at: https://ru.wikipedia.org/wiki/Закон_Ципфа

9. *Maslov V. P., Maslova T. V.* O zakone Tsipfa i rangovykh raspredeleniyakh v lingvistike i semiotike [On Zipf's law and rank distributions in linguistics and semiotics]. *Matematicheskie zametki — Mathematical Notes*, 2006, vol. 80, no. 5, p. 718–732. (In Russian.) DOI: 10.4213/mzm3081

10. *Yablonsky A. I.* Modeli i metody issledovaniya nauki [Models and Methods of Science Research]. Moscow, Ehditorial URSS, 2001. 400 p. (In Russian.)

11. *Haitun S. D.* Kolichestvennyj analiz sotsial'nykh yavlenij [Quantitative analysis of social phenomena]. Moscow, KomKniga, 2010. 280 p. (In Russian.)

12. *Manning C. D., Raghavan P., Schutze H.* Vvedenie v informatsionnyj poisk [Introduction to information retrieval]. Moscow, Vil'yams, 2011. 528 p. (In Russian.)

13. Zakon Khipsa [Heaps' law]. (In Russian.) Available at: https://ru.wikipedia.org/wiki/Закон_Хипса

14. *Kullback S., Leibler R. A.* On information and sufficiency. *The Annals of Mathematical Statistics*, 1951, vol. 22, no. 1, p. 79–86. DOI: 10.1214/aoms/1177729694

15. *Kullbak S.* Teoriya informatsii i statistika [Information theory and statistics]. Moscow, Nauka, 1967. 408 p. (In Russian.)

16. *Veryaev A. A., Tatarnikova G. V.* Educational Data Mining i Learning Analytics — napravleniya razvitiya obrazovatel'noj kvalitologii [Educational Data Mining and Learning Analytics — directions of the educational qualitology development]. *Prepodavatel XXI vek — Teacher of the 21st Century*, 2016, no. 2-1, p. 150–160. (In Russian.) Available at: https://drive.google.com/file/d/0Bww8v66PaPfWSFpSbHlqQTJ0aHM/view

17. *Mikheev M. Yu., Erlikh L. I.* Idiostilevoj profil' i opredelenie avtorstva teksta po chastotam sluzhebnykh slov [Idiostyle profile and definition of authorship of the text according to the frequencies of service words]. *Nauchno-tekhnicheskaya informatsiya. Seriya 2: Informatsionnye protsessy i sistemy — Scientific and technical information. Series 2: Information processes and systems*, 2018, no. 2, p. 25–34. (In Russian.)

18. *Shevelev O. G.* Razrabotka i issledovanie algoritmov sravneniya stilej tekstovykh proizvedenij: avtoref. dis. ... kand. tekh. nauk [Development and research of algorithms for comparing styles of textual works. Cand. eng. sci. diss. author's abstract]. Tomsk, 2006. 18 p. (In Russian.)

19. *Gmurman V. E.* Teoriya veroyatnostej i matematicheskaya statistika [Theory of probability and mathematical statistics]. Moscow, Vysshaya shkola, 2003. 479 p. (In Russian.)

НОВОСТИ

В России разработали нейросервис для «усиления» интеллекта человека

«Лаборатория знаний» объявила о запуске в промышленную эксплуатацию нейротехнологического сервиса Neuroangel. Он предназначен для повышения продуктивности и усиления интеллектуальной деятельности человека и команд. Команда специалистов «Лаборатории знаний» разрабатывала нейротехнологический сервис с 2016 года.

Neuroangel — это программно-аппаратный комплекс, состоящий из нейроинтерфейса, который снимает сигналы мозга, и программного обеспечения, оценивающего психоэмоциональное состояние человека и дающего рекомендации по его коррекции. Возможны различные варианты его применения. Например, нейросервис помогает найти свое продуктивное состояние для каждого вида деятельности и быстро входить в него для выполнения конкретных задач. Также технология позволяет оценить эффективность обучения, оценивая вовлеченность, уровень концентрации и другие параметры у студентов во время лекции. Кроме этого, с помощью Neuroangel можно проанализировать метакомпетенции сотрудников, оценить эффективность работы команды и каждого ее участника.

В декабре 2019 года «Лаборатория знаний» внедрила нейротехнологический сервис в образовательный процесс Московской финансово-юридической академии (МФЮА). Вуз планирует использовать технологию для повышения качества преподавания в аудитории и онлайн, а также для создания и реализации индивидуальных образовательных траекторий.

Александр Макаров, генеральный директор «Лаборатории знаний» отметил, что в течение последнего года была собрана обширная обратная связь от потенциальных и реальных заказчиков о технологии и вариантах ее применения и наибольший интерес нейротехнологический сервис вызвал у HR-сообщества и высших учебных заведений.

(По материалам CNews)