КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА И ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА

УДК 81.32

ПРИМЕНЕНИЕ МЕТОДОВ DATA MINING ДЛЯ РЕШЕНИЯ ЛИНГВИСТИЧЕСКИХ ЗАДАЧ

О. В. Донина

Воронежский государственный университет

Поступила в редакцию 31.03.2017 г.

Аннотация. В статье рассматриваются возможности использования методов Data Mining (таких как кластеризация методом k-средних, построение самоорганизующихся карт Кохонена, проведение факторного анализа и построение дерева решений) для анализа результатов криптоклассного исследования – авторского метода изучения скрытых именных классов, базирующегося на данных электронных корпусов.
Ключевые слова: компьютерная лингвистика, Data Minig, компьютерно-когнитивное моделирование, цифровые гуманитарные науки, корпусные исследования, криптоклассный анализ.
Annotation. The paper, based on the data from the electronic corpora of twenty national variants of the English language, discusses the ways of using Data Mining methods (such as k-means clustering, Self-Organizing Map, decision tree and factor analysis) to analyze the results of the author's method of the study of the covert nominal classes – *cryptotypes*.
Keywords: computer linguistics, Data Minig, computer modeling, Digital Humanities, corpora studies, cryptotype.

ВВЕДЕНИЕ

В связи с всеобщей дигитализацией, наблюдаемой в современном информационном обществе, все большее распространение получают Digital Humanities (DH) или eHumanities, т. е. *цифровые гуманитарные науки*, являющиеся инновационной междисциплинарной сферой исследований, объединяющей методы гуманитарных, социальных и компьютерных наук с целью исследования возможностей применения новых цифровых технологий в гуманитарных науках. Качественный анализ данных в этих науках может быть усовершенствован, главным образом, благодаря доступным для исследований оцифрованным текстам. Стоит отметить доступность и технологичность полнотекстовых архивов (например, различных национальных корпусов), которые, вместо небольшой выборки, выполненной вручную ($n < 100$), позволяют проанализировать статистически представительное подмножество ($n > 1{,}000$) или целый корпус ($n > 100{,}000$). Подход компьютерной лингвистики позволяет исследователю работать с большими объемами данных и при этом уделять больше внимания лингвистическим деталям, моделируя и визуализируя полученные результаты.

В науках, где объект исследования недоступен непосредственному наблюдению, таких как языкознание, возникает необходимость в его моделировании с использованием различных средств визуализации исследуемого объекта. Более того, в связи со сложностью изучения такого динамичного и

многоаспектного явления как язык, целесообразно, как показали исследования последних лет, применение средств *когнитивного моделирования*. Под когнитивной моделью понимают «основной механизм, обеспечивающий обработку и хранение информации о мире в сознании человека» [1: 58]. Обзор современных *теорий когнитивного моделирования* в лингвистике был сделан Л. С. Абросимовой [2]. В обзоре рассматриваются успехи применения данного подхода в том числе в прикладной лингвистике (например, при создании искусственных машинных языков и для совершенствования автоматизированного перевода). В своей работе мы используем *компьютерное когнитивное моделирование* (ККГ) ненаблюдаемых объектов, применяя набор методов Data Mining, суть которых состоит в процессе обнаружения в «сырых» данных новых интерпретаций знаний, необходимых для принятия решений в различных сферах человеческой деятельности, при помощи методов математической статистики. Data Mining является мультидисциплинарной областью, возникшей и развивающейся на базе таких наук как прикладная статистика, распознавание образов, искусственный интеллект, теория баз данных.

МЕТОДИКА ИССЛЕДОВАНИЯ

Методика исследования языковых категорий в разных языковых средах базируется на трех составляющих: криптоклассный анализ, методы корпусной лингвистики и Data Mining. Анализу подвергались языковые категории, скрытые в английском языке. Это скрытые классы имен существительных, называемые *именными криптоклассами*, которые были описаны в работах О. О. Борискиной [3, 4, 5, 6, 7, 8].

На настоящий момент выделено и описано 6 криптоклассов английского языка, имеющих соответствие в виде явных лексико-грамматических категорий других языков мира: криптокласс Res Liquidae («Жидкое», эталон – ‘вода’), Res Acutae («Остроконечное», эталон – ‘шип’), Res Filiformes («Нитевидное», эталон – ‘нить’), Res Rotundae («Круглое», эталон – ‘мяч’), Res Parvae («Рукоятное», эталон – ‘яблоко’), Res Longae Penetrantes («Длинно-тонкое стабильной формы», эталон – ‘палка’) [9, 10, 11, 12]. Например, в *именной криптокласс* английского языка «Жидкое» входят такие существительные как *water*, *blood*, *milk* и другие номинации объектов действительности, существующие в жидком состоянии. Эти существительные являются эталонами криптокласса. Вместе с тем, в этом криптоклассе представлены и существительные, обозначающие абстрактные понятия. Такие понятия как *жизнь*, *добро* или *страсть* не встречаются в жидком состоянии, но, тем не менее, человек часто категоризует их *по аналогии* с жидким. Таким образом, имена абстрактной семантики в метафорическом употреблении могут входить в выделенные в английском языке именные криптоклассы, так что в одном языковом классе сосуществуют имена конкретной и абстрактной семантики.

Криптоклассный анализ проводился на именах эмоционального состояния и чувственного переживания (таких как *гнев*, *страх*, *любовь* и т. д.) в 20-ти вариантах английского языка, представленных в корпусе М. Дэвиса [13]. Наряду с британским и американским, в нем представлены и редкие варианты (например, кенийский или танзанийский английский). Объем исследовательского корпуса, сформированный на основании результатов полуавтоматической обработки корпусных запросов, составил 65000 словоупотреблений. Ранее для визуализации результатов криптоклассного исследования мы использовали лица Чернова, построенные при помощи программы Statistica [14]. Но сейчас, учитывая довольно большой объем получившегося исследовательского корпуса и необходимость свести воедино разные по качеству параметры (а именно: 20 рассматриваемых вариантов английского языка, 23 имени эмоций и 6 именных криптоклассов), использование данного метода оказалось невозможным. На помощь в решении указанной задачи нам пришли методы Data Mining и компьютерно-когнитивного моделирования.

В первую очередь с целью определения значимости факторов и возможной редукции

входных параметров перед кластеризацией посредством программы Deductor Academic 5.3 был проведен факторный анализ, базирующийся на «варимакс» методе. В качестве входных данных выступали шесть выделенных на данный момент криптоклассов.

Следующим этапом стало проведение кластеризации методом k-means (k-средних), а также построение самоорганизующейся карты Кохонена, являющейся разновидностью нейросетевых алгоритмов. Кластеризация применяется для распределения множества объектов по классам, которые изначально не заданы, при этом наиболее часто используемым является алгоритм k-средних. Искусственные нейронные сети, возникшие в качестве модели биологической нервной системы, состоят из входного, скрытого и выходного слоев нейронов, причем для входного и выходного слоев параметры известны, в то время как в скрытом происходят неявные преобразования сигналов. В лингвистике нейронные сети используются в нейросетевых моделях языка, при машинном переводе, автоматической кластеризации лексики (карты Кохонена) и пр.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Посредством факторного анализа было показано, что в результате произведенного вращения в структуре факторного пространства не произошло существенных изменений (т. е. установленные автоматически факторы соответствуют шести криптоклассам на 96,09 % – 99,51 % (табл.1)). Таким образом, можно говорить об устойчивости и стабильности данных, что указывает на независимость факторов, отраженных в корреляционной матрице, т.е. была доказана высокая объяснительная значимость всех факторов (криптоклассов) и возможность использования их при дальнейшей кластеризации.

Посредством нейронной сети Кохонена были сформированы карты входов нейронов шести криптоклассов, т. е. внутренняя структура входных данных была визуализирована путём подстройки весов нейронов карт, где определенным цветом обозначены области, содержащие примерно одинаковые входы для анализируемых примеров. В результате векторного квантования отдельные варианты английского языка были сгруппированы в шесть кластеров, соответствующих географическим ареалам (рис.1): американскому (языковые варианты английского языка Америки и Канады), австралийскому (варианты Австралии и Новой Зеландии), европейскому (варианты Великобритании и Ирландии), азиатскому (варианты английского, на котором говорят в Индии, Сингапуре, Гонконге, Шри-Ланке, Бангладеш, Пакистане, Малайзии и на Филиппинах), африканскому (варианты английского, как официального государственного языка Ганы, Нигерии, Танзании, Кении и ЮАР) и карибскому (вариант английского на Ямайке).

Совпадение криптоклассной категоризации с географической подчеркивает важность ареального влияния, отмеченного в работе В. Н. Полякова и Е. И. Ярославцевой [15]. В рамках указанной работы (на материале БД

Таблица 1

Результаты факторного анализа

	Окончательные факторы (Варимакс метод)					
	Фактор 1	Фактор 2	Фактор 3	Фактор 4	Фактор 5	Фактор 6
Res Acutae			0,9951			
Res Filiformes		0,9841				
Res Liquidae						0,9609
Res Longae Penetrantes					0,9769	
Res Parvae				0,9851		
Res Rotundae	0,9849					

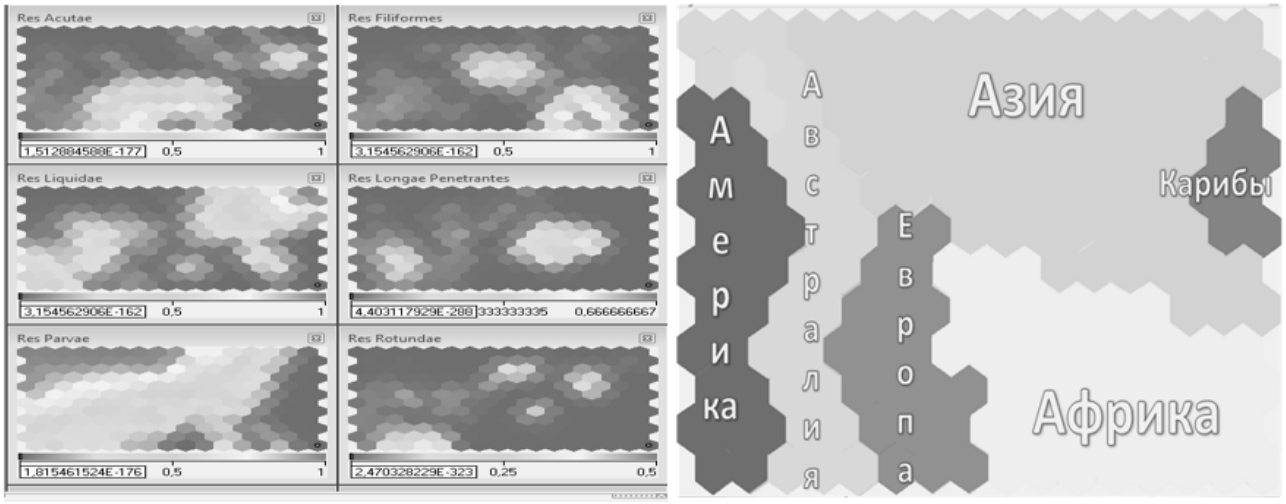

Рис. 1. Самоорганизующиеся карты Кохонена

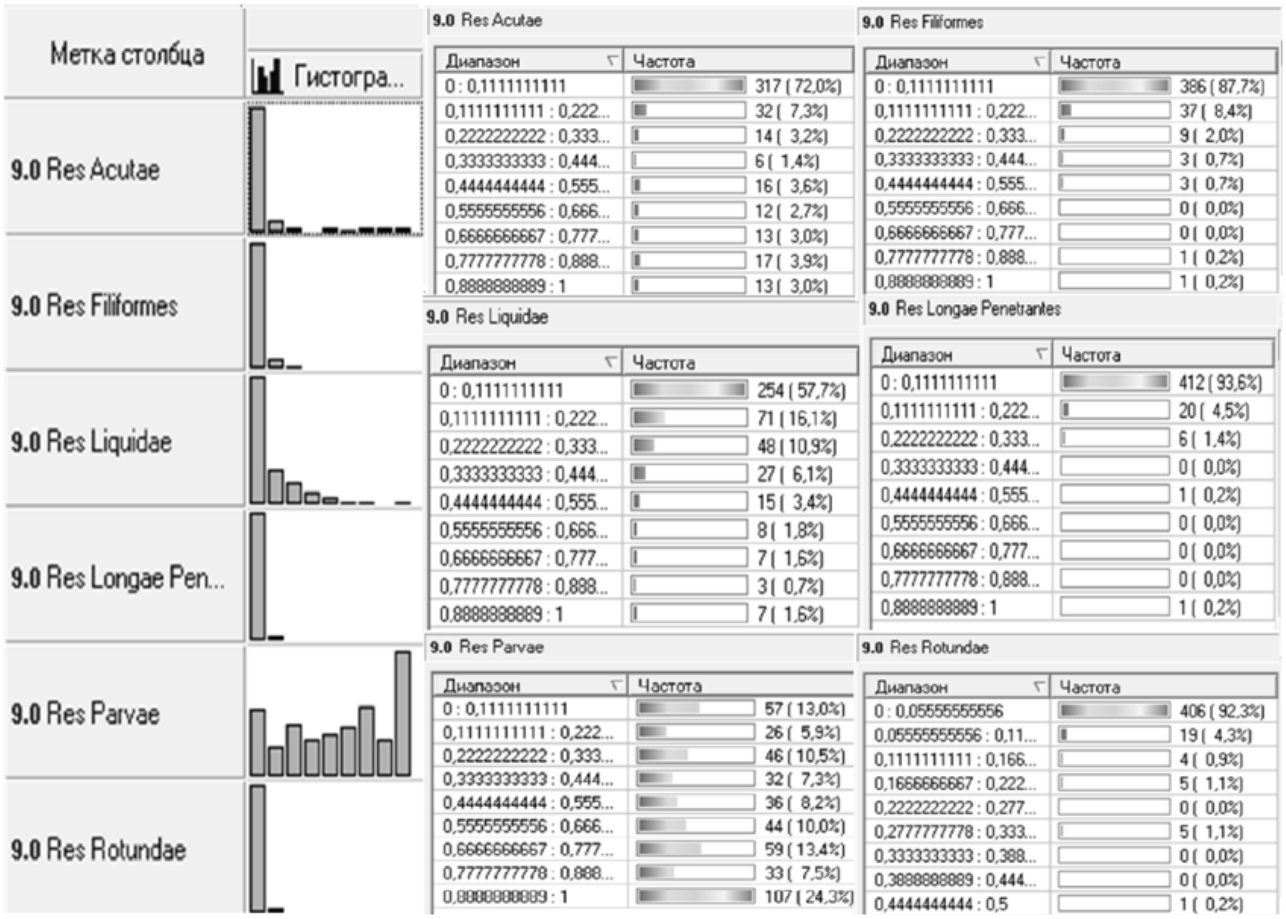

Рис. 2. Частотность криптоклассной представленности в исследовательском корпусе

«Языки мира» ИЯ РАН) изучается феномен типологического сдвига, сутью которого является то, что «языки в процессе ареальных контактов приобретают новые типологические черты и теряют часть существующих. При этом широко распространенные признаки имеют тенденцию к дальнейшему распространению, а низкочастотные – к вымыванию» [15: 114–115]. Данное явление объясняется именно при помощи ареальной гипотезы («сдвиг обусловлен естественными процессами, связанными с постоянными ареальными контактами между языками») [15: 113].

По данным криптоклассного анализа для всех 23 имен эмоций в 20 вариантах языка была рассчитана статистика по криптоклас-

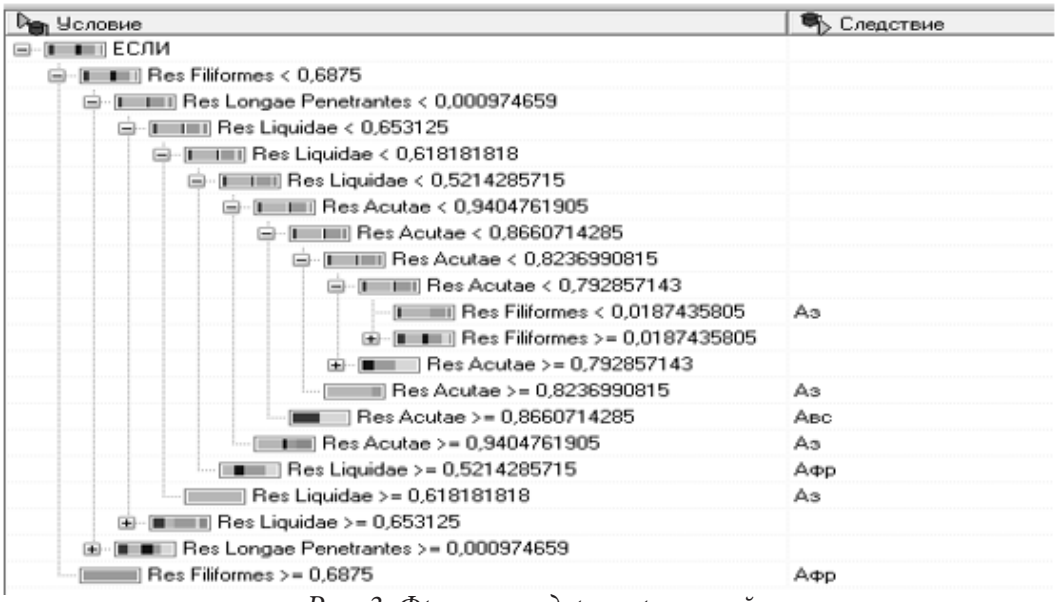

Рис. 3. Фрагмент дерева решений

сам, которая отразила общие тенденции, характерные для исследовательского корпуса в целом (рис. 2). Полученные данные показали, что в 92,3 % случаев доля представленности криптокласса «Круглое» не превышает 5,6 %, что является самым низким значением среди криптоклассов, т.е. данный класс наименее представлен в сформированном корпусе. В 93,6 % примеров представленность криптокласса «Длинно-тонкое стабильной формы» находится в диапазоне от 0 % до 11,1 %, т. е. он выступает предпоследним по представленности. Четвертым по распространенности оказывается криптокласс «Нитевидное»: в 96,1 % случаев доля представленности данного криптокласса варьируется от 0 % до 22,2 %. Криптокласс «Остроконечное» является третьим по частотности: в 79,3 % случаев его доля представленности не поднимается выше 22,2 %, но при этом в 3 % примерах он является одним из наиболее представленных со значением показателя криптокласснной активности некоторых имен эмоций от 88,9 % до 100 %. На втором месте находится криптокласс «Жидкое», доля представленности которого в 94,2 % случае располагается в диапазоне от 0 % до 55,6 %. Лидирующим же в превалирующем большинстве словоупотреблений оказывается криптокласс «Рукоятное», доля представленности которого в 55,2 % случаев превышает 55,6 %.

Следующим шагом нашего исследования было создание правил на основе дерева решений, позволяющих написать компьютерную программу, способную устанавливать ареальную принадлежность варианта языка. Деревья решений являются методом автоматического анализа данных, формирующим последовательную структуру правил, где каждому объекту соответствует единственный узел, дающий решение. Результаты подобного анализа могут быть представлены как в виде иерархии (рис. 3), так и в виде набора правил, описывающих классы. В перспективе благодаря составленным правилам и описаниям кластеров появляется возможность проследить динамику влияния вариантов языка/языковых ареалов друг на друга, проведя аналогичное исследование через 10–15 лет.

ЗАКЛЮЧЕНИЕ

Таким образом, в рамках статьи была рассмотрена возможность использования различных методов Data Mining для решения лингвистических задач. Так, при помощи указанных методов была выявлена когнитивная общность эмоциональных переживаний для всех носителей английского языка независимо от ареала его использования. Помимо этого, были выявлены языковые ареалы, а также установлено, что национальные вари-

анты английского языка внутри каждого ареала демонстрируют во многом аналогичную, но не идентичную картину категоризации эмоций. Содержательная специфика ареалов была определена посредством анализа профилей кластеров. Было установлено, что наиболее представленным криптоклассом во всех ареалах выступает «Рукоятное», для ареалов Австралии, Америки, Африки и Карибского бассейна вторым по значимости при языковой категоризации эмоций выступает криптокласс «Жидкое», а для ареалов Азии и Европы – «Остроконечное». На основе дерева решений были разработаны правила, позволяющие в перспективе написать компьютерную программу, способную устанавливать ареальную принадлежность варианта или диалекта языка.

СПИСОК ЛИТЕРАТУРЫ

1. *Маслова В. А.* Введение в когнитивную лингвистику / В. А. Маслова - Litres, 2016. - 399 с.

2. *Абросимова Л. С.* Словообразование в языковой категоризации мира / Л.С. Абросимова. – Ростов-на-Дону, 2015. – 328 с.

3. *Борискина О. О.* Языковая категоризация стихий / О. О. Борискина // Филология и культура. Тезисы II-й Международной конференции. Институт языкознания РАН, Тамбовский государственный университет им. Г. Р. Державина. – 1999. – С. 149–157.

4. *Борискина О. О.* Криптоклассы первостихий как элемент онтогностического описания языка / О. О. Борискина // Проблемы лингвистической прогностики. Сб. научн. трудов под редакцией А. А. Кретова. – Воронеж, 2000 – С. 121–126.

5. *Борискина О. О.* Национально-специфическое языковое сознание и заимствованное слово / О. О. Борискина // Межкультурная коммуникация и проблемы национальной идентичности. Сб. научн. трудов. – Воронеж, 2002. – С. 406–410.

6. *Борискина О. О.* Моделирование синтагматической динамики слова / О. О. Борискина // Вопросы когнитивной лингвистики. – Тамбов. 2008. – № 3. – С. 57– 64.

7. *Борискина О. О.* Криптоклассные проекции мира непредметных сущностей: опыт криптоклассного анализа словосочетаемости / О. О. Борискина // Вестник Воронеж. гос. ун-та. Сер. Лингвистика и межкультурная коммуникация. – 2009. – № 1. – С. 32– 37.

8. *Борискина О. О.* Объяснение необъяснимого или о мотивации немотивированного / О. О. Борискина // Вестник Санкт-Петербургского университета. Серия 9 «Филология. Востоковедение. Журналистика». – 2010. – № 1. – С. 95–100.

9. *Boriskina O. O.* An Algorithm for Analysis of Distribution of Abstract Nouns in Cryptotypes / O.O. Boriskina, T. Marchenko // Proceedings of the 2010 International Conference on Artificial Intelligence, ICAI 2010. – 2010. – P. 907–913.

10. *Кретов А. А.* «Полёт мысли» и методика исследования криптоклассов / А. А. Кретов, О. О. Борискина, Н. Васильева // Вестник Воронеж. гос. ун-та. Сер. Лингвистика и межкультурная коммуникация. – 2004. – № 1. – С. 61–65.

11. *Донина О. В.* Место именного криптокласса «Res Acutae» в криптоклассных таксономиях английских непредметных имен / О.В. Донина // Дайджест – 2013 дипломные работы студентов факультета РГФ ВГУ. Воронежский государственный университет, Факультет романо-германской филологии. – Воронеж, 2013. – С. 43–51.

12. *Донина О. В., Борискина О. О.* Эмотивная лексика в аспекте ареальной вариативности / О.В. Донина, О.О. Борискина // Вестник Воронеж. гос. ун-та. Сер. Лингвистика и межкультурная коммуникация. – 2016. – № 4. – С. 1–45.

13. *Davies Mark.* Corpus of News on the Web (NOW): 3+ billion words from 20 countries, updated every day / M. Davies. – Provo, 2013. – Mode of access: http://corpus.byu.edu (12.02.2017).

14. *Донина О. В.* Способы визуализации результатов криптоклассного исследования / О.В. Донина // Вестник Воронеж. гос. ун-та. Сер. Лингвистика и межкультурная коммуникация. – 2015. – № 3. – С. 105–112.

15. *Поляков В. Н.* Квантитативные закономерности типологического сдвига в языках Евразии (на материале БД «Языки мира» ИЯ РАН) / В. Н. Поляков, Е. И. Ярославцева // Ученые записки Казанского государственного университета. Гуманитарные науки. – Казань: Казанский государственный университет, 2008. – Т. 150, кн. 2. – С. 97–118.

Донина О. В. – преподаватель кафедры теоретической и прикладной лингвистики, факультет романо-германской филологии, Воронежский государственный университет.
E-mail: olga-donina@mail.ru

Donina O. V. – Lecturer, Department of Theoretical and Applied Linguistics, Romance and Germanic Philology Faculty, Voronezh State University.
E-mail: olga-donina@mail.ru