

СПОСОБЫ ВИЗУАЛИЗАЦИИ РЕЗУЛЬТАТОВ КРИПТОКЛАССНОГО ИССЛЕДОВАНИЯ

О. В. Дони́на

Воронежский государственный университет

Поступила в редакцию 20 апреля 2015 г.

Аннотация: в статье рассматриваются различные способы визуализации результатов криптоклассного исследования: криптоклассные портреты имен существительных, криптоклассные таксономии имен, лица Чернова для имен существительных английского языка.

Ключевые слова: корпусные данные, лингвостатистика, таксономия, именной криптокласс, лица Чернова.

Abstract: the paper describes various means of visual representation of cryptotype categorization of the English nouns, namely, noun profiles, noun taxonomies and Chernoff faces.

Key words: corpus data, linguistic statistics, noun taxonomy, noun cryptotype, Chernoff faces.

Визуализацию можно рассматривать как одно из наиболее перспективных направлений анализа данных, которое позволяет оценивать строение и состояние функциональной активности языковых единиц под новым углом зрения и обнаружить новые неожиданные связи и тенденции изменения элементов языка. Современные компьютерные технологии предлагают разнообразный инструментарий для обработки результатов, полученных в ходе лингвистического анализа. В рамках данной статьи будут рассмотрены возможности применения методов статистической обработки, а также способы визуального представления данных криптоклассного исследования метафорической сочетаемости абстрактной лексики (криптоклассные портреты, таксономии имен существительных, лица Чернова).

Криптоклассное исследование абстрактной лексики

Криптоклассное исследование абстрактной лексики предполагает изучение скрытых лексико-семантических классов имен существительных и их метафорическую сочетаемость [1–5]. Криптоклассом мы называем такой тип скрытой категории, где классная принадлежность имени выражена не эксплицитно (в случае отсутствия или стертости морфологических показателей класса), а имплицитно – в структуре предложения через классификатор (конструкцию и/или словоформу) [6; 7]. На данном этапе выделено 6 криптоклассов английского языка, имеющих соответствие в виде явных лексико-грамматических категорий других языков мира: Res Liquidae (эталон – ‘вода’), Res Acutae (эталон – ‘шип’), Res Filiformes

(эталон – ‘нить’), Res Rotundae (эталон – ‘мяч’), Res Parvae (эталон – ‘камень’), Res Longae Penetrantes (эталон – ‘копье’) [8; 9].

Для определения количественных параметров классной принадлежности имен мы использовали такие лингвостатистические показатели, как индекс разнообразия сочетаемости слова (далее – ИРа) и его показатель криптоклассной активности (далее – ПоКа) [9; 10]. Согласно гипотезе, выдвинутой и описанной в частности в работах [10; 11], индекс разнообразия сочетаемости имени показывает степень «прочности воспоминаний» этого имени обо всех прошлых контекстах, в которых оно участвовало согласно своей природе. Показатель криптоклассной активности имени в образовании словосочетаний (ПоКа) показывает коммуникативную значимость ассоциаций этого имени с определенным категоризирующим признаком в современном языке.

Корпусные данные криптоклассного распределения имени *ability*

Проиллюстрируем возможные способы визуализации результатов криптоклассного исследования на основе анализа лексемы *ability*. Согласно словарю Мюллера [12], *ability* – способность, возможность сделать что-л.; способность, ловкость; квалификация, умение; обыкн. pl. дарование; способность; ком. платежеспособность; юр. компетенция, правоспособность.

Материал исследования отобран из электронного корпуса Corpus of Global Web-Based English [13], который состоит из 1,9 миллиарда слов, содержащихся на 1,8 миллионах web-страницах с 340 000 сайтов в 20 различных англоговорящих странах (Великобритания (GB), США (US), Австралия (AU), Канада (CA),

Бангладеш (BD), Индия (IN), Пакистан (PK), Сингапур (SG), Ямайка (JM), Малайзия (MY), ЮАР (ZA), Гонконг (HK), Нигерия (NG), Ирландия (IE), Кения (KE), Шри-Ланка (LK), Гана (GH), Новая Зеландия (NZ), Филиппины (PH), Танзания (TZ)). GloWbE 2012–2013 – лингвистический ресурс, отличающийся широкой представленностью национальных вариантов английского языка, который вполне подходит для выявления различий между ними.

В табл. 1 отражены результаты криптоклассного исследования имени *ability*, проведенного на материале корпуса GloWbE. В первой строке перечислены сокращения национальных вариантов английского языка, далее построчно представлены количественные характеристики: индекс разнообразия сочетаемости имени *ability* (ИРа) и показатель криптоклассной активности слова (ПоКа) в образовании словосочетаний с классификаторами криптоклассов, в колонке Total представлены данные по всему корпусу GloWbE без специализации на национальные варианты.

Как видно из табл. 1, данное имя является метафоронимом четырех криптоклассов английского языка. Ниже наряду с примерами употребления имени в корпусе GloWbE (1–4) показаны классифицирующие эти классы имен структуры в виде коллострукций [14], характерных для *ability*.

1) *Res Parvae*. Характерные коллострукции: [throw the ability away/out], [take the ability (away/out/back)], [pick (up/out) the ability], [grasp the ability], [(have) a grasp of the ability], [cast the ability]. Например:

There is nothing more obnoxious to see super rich kids throwing away their abilities and potentials and relying on their dad's money.

2) *Res Filiformes*. Характерные коллострукции: [weave the ability], [twist the ability into (a knot)], [tie the abilities (up/with smth.)]. Например:

Their anguish floods through her, twisting her ability into a killing force, destroying the land and its people.

3) *Res Acutae*. Типичные коллострукции: [acute ability], [keen ability], [sharp ability]. Например:

From there he got involved in the Internet and has since shown a sharp ability to read online trends and take advantage of them.

4) *Res Liquidae*. Типичные коллострукции: [ability oozes], [ability flows]. Например:

You are not a bad speaker but your abilities flow from diligence not nature.

Криптоклассные портреты имени

Одним из способов представления результатов криптоклассного анализа выступают криптоклассные портреты имен [8; 15]. Вычисление меры соотносительности имени с каждым из криптоклассов английского языка и оценка активности имени по каждому криптоклассу позволяют осуществить сравнение разных криптоклассных проекций имени в рамках одного языка. Взаимодействие таких количественных характеристик имени, как ИРа и ПоКа, задает специфику представления имени в криптоклассах английского языка. Данные о распределении имени в криптоклассах, представленные средствами компьютерной графики, составляют его криптоклассный портрет. В данном случае криптоклассный портрет (рис. 1) был построен на основании данных всего корпуса GloWbE, без разделения на национальные варианты. Выделяются два криптоклассных портрета:

1) Q relative – строится на основании относительной частоты сочетаемости (по ИРа) (для удобства

Т а б л и ц а 1

Результаты криптоклассного исследования имени *ability*

	AU	BD	CA	GB	GH	HK	IE	IN	JM	KE	LK	MY	NG	NZ	PH	PK	SG	TZ	US	ZA	Total	
Res Acutae																						
<i>ИРа</i>	0,3	0,2	0,2	0,3	0	0,1	0,1	0,2	0,3	0	0,1	0,1	0,1	0	0	0,2	0,2	0	0,3	0,1	0,300	
<i>ПоКа</i>	0,1	0,4	0,2	0,1	0	0,17	0	0,12	0,25	0	0,1	0,17	0,3	0	0	0,2	0,2	0	0,1	0,2	0,116	
Res Filiformes																						
<i>ИРа</i>	0,3	0,1	0,1	0,4	0	0,13	0,1	0,13	0	0,1	0,1	0	0	0	0	0,1	0	0,3	0	0,625		
<i>ПоКа</i>	0	0,1	0	0	0	0,08	0,1	0,04	0	0,2	0,2	0	0	0	0	0,1	0	0	0	0	0,033	
Res Liquidae																						
<i>ИРа</i>	0,1	0	0,1	0,2	0	0,08	0,1	0	0	0	0	0	0,1	0	0	0	0	0	0,1	0	0,230	
<i>ПоКа</i>	0	0	0	0	0	0,08	0,1	0	0	0	0	0	0,1	0	0	0	0	0	0	0	0,013	
Res Parvae																						
<i>ИРа</i>	0,4	0,2	0,4	0,8	0,2	0,22	0,2	0,33	0,22	0,2	0,1	0,11	0,2	0,2	0,3	0,3	0,3	0,1	0,6	0,1	0,777	
<i>ПоКа</i>	0,8	0,6	0,8	0,8	1	0,67	0,8	0,84	0,75	0,8	0,7	0,83	0,6	1	1	0,8	0,7	1	0,9	0,8	0,836	

анализа представлен в виде столбчатой гистограммы) (рис. 1а);

2) *S relative* – строится по относительной активности имени в криптоклассе (по ПоКа) (представлен в виде круговой гистограммы) (рис. 1б).

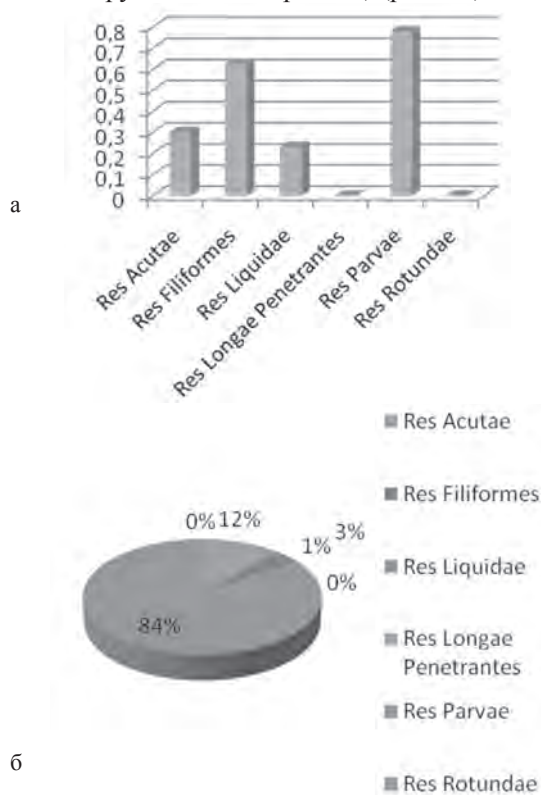


Рис. 1. Криптоклассный портрет имени *ability*: а) *Q relative* (данные по ИРа), б) *S relative* (данные по ПоКа)

Криптоклассные таксономии

На основании полученных в результате криптоклассного анализа данных мы можем построить

криптоклассные таксономии имен, которые способны наглядно отразить сходства и различия представления лексем как в одном языке, так и в разных, например, различных национальных вариантах английского языка. В работах [10; 11] описаны два типа таксономий: первая криптоклассная таксономия строится по показателю криптоклассной активности (ПоКа) (рис. 2), вторая криптоклассная таксономия строится по индексу разнообразия сочетаемости, с которыми имя встречается в корпусе (ИРа) (рис. 3).

Как видим, для лексемы *ability* во всех национальных вариантах английского языка в их современном состоянии приоритетным является криптокласс Res Parvae, при этом выделяются такие варианты, в которых это имя демонстрирует одинаковую последовательность криптоклассных предпочтений. Например, к одному таксону относятся американский, британский и австралийский варианты английского языка (рис. 2). В них лексема *ability* в первую очередь категоризируется как рукоятный предмет, во вторую – как шип, далее как нить, и наконец, как жидкое (та же последовательность характерна для корпуса GloWbE в целом без учета специализации на варианты – на рисунке обозначенного *Total*).

Во второй криптоклассной таксономии мы видим менее согласованную картину: так для двух вариантов английского языка криптокласс Res Parvae не является приоритетным (для носителей ямайского английского лексема *ability* в первую очередь категоризируется как нечто острое, а для говорящих на варианте английского в республике Шри-Ланка – как нитевидное).

Сопоставление данных двух таксономий дает возможность проследить эволюцию представлений об изучаемой лексеме [10; 16]: так, для англоговорящих жителей Ямайки признаком имени *ability*, «пришедшим на остров с британским английским», вы-

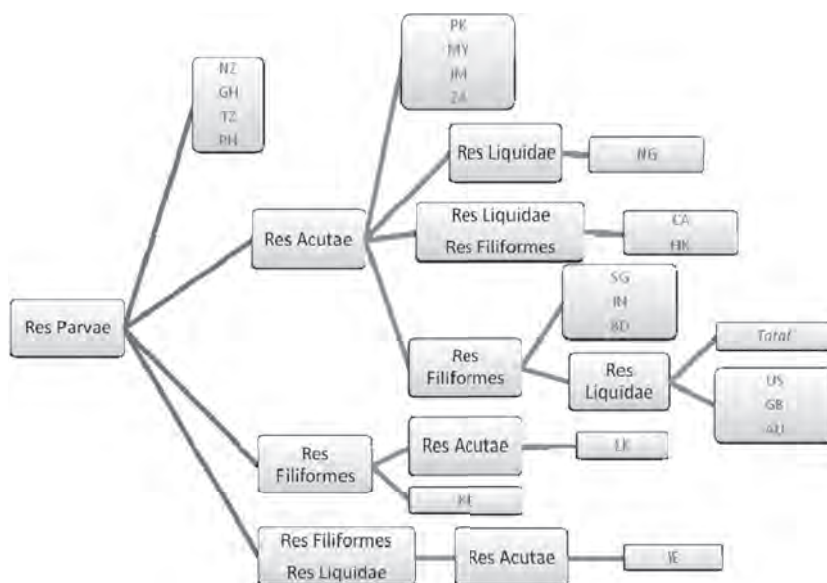


Рис. 2. Первая криптоклассная таксономия для имени *ability* (по ПоКа)

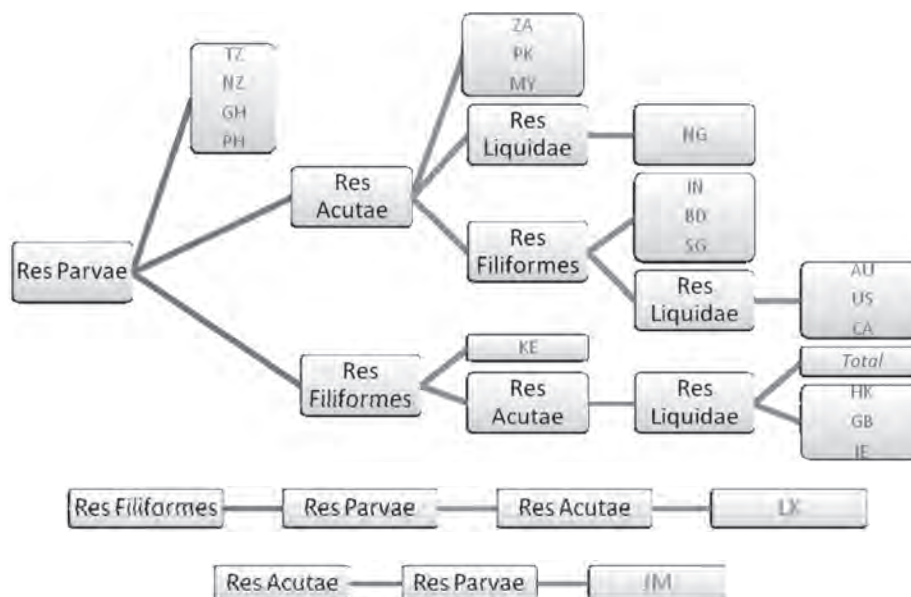


Рис. 3. Вторая криптоклассная таксономия для имени *ability* (по ИРА)

ступает признак «быть острым», что проявляется в максимальном разнообразии его сочетаемости с классификаторами-коллострукциями криптокласса *Res Acutae* (рис. 3), тогда как максимальная активность имени в образовании словосочетаний (его «сегодняшняя жизнь») протекает в классе *Res Parvae* (рис. 2). Таким образом, из приведенных для лексемы *ability* таксономий видно, что «современная жизнь» слова очень схожа для всех диалектов, в то время как «этимологическая память слова» (ИРА) характеризуется большей дифференцированностью.

Лица Чернова как способ представления результатов криптоклассного анализа

Еще один способ визуального представления данных основывается на методе Data Mining, т.е. процессе обнаружения в «сырых» данных ранее неизвестных нетривиальных практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности [17]. Data Mining – мультидисциплинарная область, возникшая и развивающаяся на базе таких наук, как прикладная статистика, распознавание образов, искусственный интеллект, теория баз данных и др. Образное, правополушарное восприятие, благодаря таким своим качествам, как целостность, непрерывность, семантическая мягкость, позволяет дополнить традиционное логическое, вербальное, левополушарное восприятие и тем самым повысить эффективность решения задач.

В данном исследовании мы использовали метод Лиц Чернова, предложенный в 1973 г. известным американским математиком Г. Черновым (Н. Chernoff) [18] для иллюстрации тенденций в многомерных данных путем изображения их в виде пик-

тографии серии лиц. Метод считается одним из наиболее эффективных, поскольку из всех зрительных навыков у человека сильнее всего развита способность к восприятию лиц других людей. Особый участок коры головного мозга узнает лицо, определяет направление взгляда и т.д. Другие части мозга (миндалевидное тело и островковая доля) анализируют выражение лица, а участок в префронтальной зоне лобной доли и система мозга, отвечающая за чувство удовольствия, оценивают его красоту [19; 20]. Лица Чернова – одно из наиболее искусно разработанных средств визуализации, представляющее собой схему визуального представления мультимедийных данных в виде человеческого лица. Каждая часть лица: нос, глаза, рот – представляет собой значение определенной переменной, назначенной для этой части. Для каждого наблюдения рисуется «лицо», где относительные значения выбранных переменных представлены как формы и размеры отдельных черт лица (например, длина носа, угол между бровями, ширина лица) (рис. 4). Таким образом, наблюдатель может идентифицировать уникальные для каждой конфигурации значений наглядные характеристики объектов.

В рамках данного исследования чтобы построить Лица Чернова для *ability*, мы воспользовались программой Statistica, представляющей собой пакет статистического анализа, в котором реализованы все новейшие компьютерные и математические методы анализа данных. В табл. 2 представлено соответствие параметров Лиц Чернова и изучаемых нами показателей. Например, ширина лица соответствует индексу разнообразия сочетаемости имени *ability* с классификаторами криптокласса Нитевидное (*Res Filiformes*).

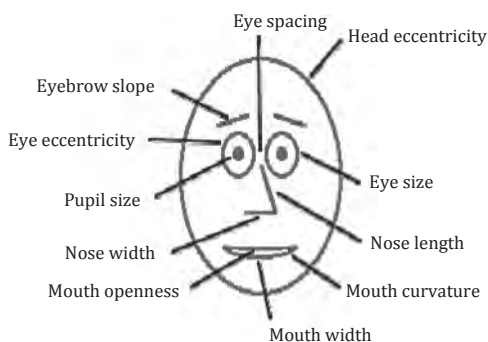


Рис. 4. Параметры для построения Лиц Чернова

Т а б л и ц а 2

Соответствие параметров Лиц Чернова и показателей исследования

Параметры	Показатели
Ширина лица	Res Filiformes – ИРа
Расположение уха	Res Longae Penetrantes – ИРа
Обвод лица	Res Longae Penetrantes – ПоКа
Форма: верхняя часть лица	Res Rotundae – ИРа
Форма: нижняя часть лица	Res Rotundae – ПоКа
Длина носа	Res Parvae – ИРа
Расположение рта	Res Liquidae – ПоКа
Изгиб рта	Res Acutae – ИРа
Размер рта	Res Filiformes – ПоКа
Расположение глаза	Res Liquidae – ИРа
Размер глаза	Res Acutae – ПоКа
Изгиб брови	Res Parvae – ПоКа

На рис. 5 приведен результат данного метода по данным ПоКа.

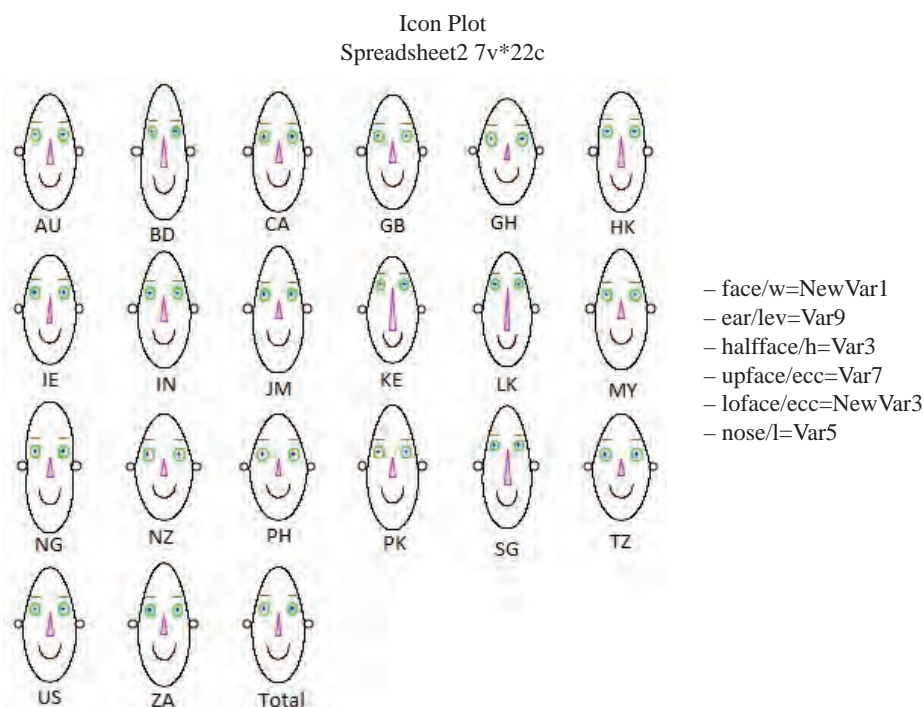


Рис. 5. Лица Чернова для *ability* по данным показателя активности слова в образовании метафорических словосочетаний в двадцати вариантах английского языка

Чтобы определить, насколько мы можем полагаться на данный метод в наших исследованиях, мы провели ранжирование пиктографиков в соответствии с данными коэффициента корреляции Пирсона (табл. 3) [21]. Данный коэффициент, подробно рассмотренный в нашей работе [22], определяет характер и тесноту статистической связи и вычисляется при последовательном сопоставлении каждого национального варианта английского языка с данными всего корпуса GloWbE.

Аналогичная методика построения Лиц Чернова и вычисление коэффициента корреляции Пирсона была произведена для данных по индексу разнообразия сочетаемости лексемы *ability* (рис. 6, табл. 4).

Сопоставив статистические показатели и данные визуализации, можно сделать вывод, что визуальное представление данных соответствует статистическим расчетам (табл. 3, 4).

Для того чтобы понять, насколько похожи криптоклассные представления об исследуемой лексеме в двадцати исследуемых национальных вариантах английского языка, также можно рассчитать коэффициент конкордации рангов Кендалла, который позволяет определить тесноту корреляционной связи между произвольным числом ранжированных признаков [22; 23]. В нашем случае коэффициент конкордации для индекса разнообразия составил $W = 0,78$; для показателя криптоклассной активности $W = 0,8$; для общей картины, включающей и ИРа и ПоКа, $W = 0,76$. Такие результаты во всех трех случаях говорят о высокой тесноте связи между

Ранжирование Лиц Чернова в соответствии с коэффициентом корреляции Пирсона (ПоКа)

Национальные варианты английского языка											
Коэффициент корреляции Пирсона (r)	0,898004039	0,926716521	0,960270211	0,974876387	0,977470257	0,98516924	0,98630911	0,99022105	0,990368913		
Тип связи	Сильная / тесная										
Нац. вар-ты англ. яз.											
r	0,991216467			0,991483406	0,996550063	0,997521943	0,999772514	0,999806869	0,999808897	0,999825	
Тип связи	Сильная / тесная										

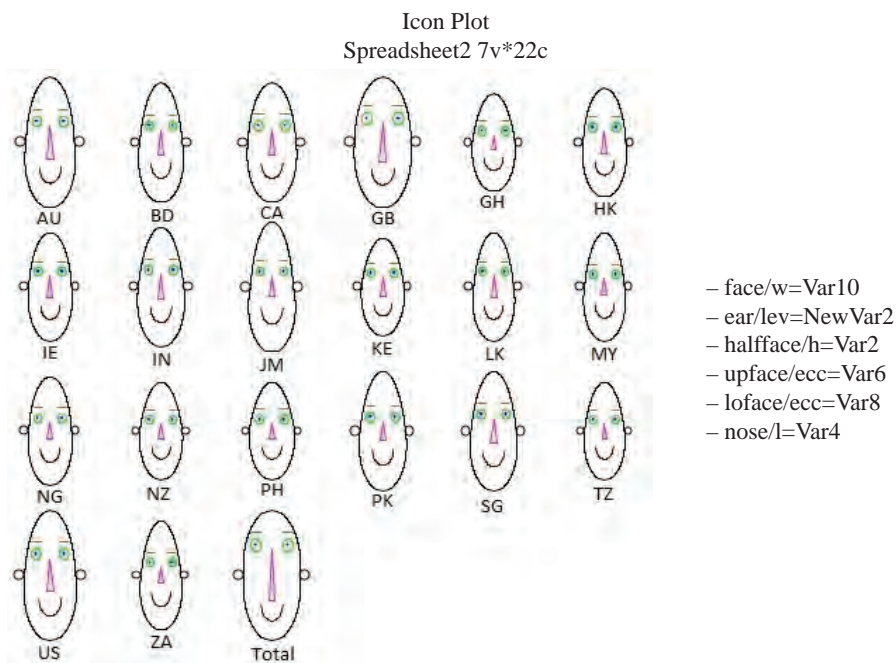


Рис. 6. Метод Лиц Чернова для данных ИРа

рассматриваемыми национальными вариантами английского языка по отношению к исследуемой лексеме.

Далее мы построили Лица Чернова для данных криптоклассного анализа лексемы *ability*, полученных при учете всех двенадцати показателей: индекса разнообразия и показателя криптоклассной активности для 6 выделенных криптоклассов, и объединили визуально сходные Лица в пять групп (табл. 5).

В первой группе оказались национальные варианты (TZ, GH, NZ, PH), в которых существительное абстрактной семантики является метафоронимом только одного криптокласса Res Parvae (рукоятное) (ПоКа_{ResParvae} = 1). Во второй группе представлены национальные варианты английского языка Кении и Шри-Ланки, которые выступили «соседями» в первой криптоклассной таксономии (см. рис. 2), а также показали близкий коэффициент корреляции Пирсона (по

Таблица 4

Ранжирование Лиц Чернова в соответствии с коэффициентом корреляции Пирсона (ИРа)

Нац. вар-ты англ. яз.												
Коэф. корреляции (r)	0,429586217	0,551666951		0,637324293	0,650563108	0,694851947			0,79859206			
Тип связи	Умеренная	Средняя							Тесная			
Нац. вар-ты англ. яз.												
r	0,839048648	0,852493529	0,867876513	0,906844663	0,90745054	0,913767773	0,950446139	0,964290632				
Тип связи	Сильная / тесная											

Таблица 5

Группировка Лиц Чернова для данных криптоклассного анализа имени ability

№	Национальные варианты английского языка				
Группа 1					
Группа 2					
Группа 3					
Группа 4					
Группа 5					

ПоКа) (см. табл. 3): $r(KE) = 0,96$; $r(LK) = 0,975$, что обусловлено сходными показателями криптоклассной активности данных национальных вариантов для двух криптоклассов: ПоКа_{Res Parvae} KE = 0,8; ПоКа LK = 0,73 и ПоКа_{Res Filiformes} KE = 0,2; ПоКа LK = 0,18. Аналогичные закономерности и соответствия с криптоклассными таксономиями (см. рис. 2, 3) и корреляционными связями (см. табл. 3, 4) прослеживаются и в других образовавшихся группах, что объясняется одинаковыми или близкими значениями изучаемых показателей (например, в

третьей группе для всех представленных национальных вариантов (IE, HK, NG, MY, ZA) ИРа_{Res Acutae} = 0,1. Показательно, что в одной группе оказались представители бывшей Британской Индии (Индия, Пакистан, Бангладеш) (группа 4), похожие «лица» имени ability наблюдаем в языках первых доминионов Великобритании (Канада, Австралийский Союз), в американском и британском английском (группа 5).

В рамках данной статьи были рассмотрены возможности применения различных методов работы с

данными криптоклассного анализа: построение криптоклассных портретов, построение криптоклассных таксономий, корреляционные связи, когнитивное компьютерное моделирование; было показано, какие обширные просторы открывает для гуманитариев использование современных технологий и математического аппарата.

ЛИТЕРАТУРА

1. Кретов А. А. «Полет 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.

2. 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. – С. 907–913.

3. 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.

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

5. Борискина О. О. Объяснение необъяснимого или о мотивации немотивированного / О. О. Борискина // Вестник С.-Петербург. ун-та. Сер. 9, Филология. Востоковедение. Журналистика. – 2010. – № 1. – С. 95–100.

6. Борискина О. О. Метафоронимы : в поисках толкования / О. О. Борискина // Вестник Воронеж. гос. ун-та. Сер.: Лингвистика и межкультурная коммуникация. – 2011. – № 1. – С. 63–68.

7. Борискина О. О. Выявление скрытой категориальности имен / О. О. Борискина // Вестник Воронеж. гос. ун-та. Сер.: Лингвистика и межкультурная коммуникация. – 2011. – № 2. – С. 15–18.

8. Борискина О. О. Теория, методология и опыт познания скрытой категориальности языка : дис. д-ра филол. наук / О. О. Борискина. – Воронеж, 2011.

9. Борискина О. О. Английский криптокласс «ResLongae», или и все-таки они колются / О. О. Борискина, О. Дони́на // Проблемы лексико-семантической типологии : сб. науч. трудов. – Воронеж, 2011. – С. 84–95.

10. Борискина О. О. Познание скрытой категориальности в языке : теория и методология / О. О. Борискина

// Когнитивные исследования языка. – 2014. – № 16. – С. 112–120.

11. Борискина О. О. К вопросу об «агрегатном состоянии» эмоций / О. О. Борискина // Среди нехоженых путей : сб. науч. ст. к юбилею А. А. Кретова / под ред. И. А. Меркуловой, К. М. Шилихиной. – Воронеж, 2012. – С. 281–292.

12. Мюллер В. К. Англо-русский словарь / В. К. Мюллер. – 24-е изд. – М. : Русский язык, 1995.

13. Davies M. Corpus of Global Web-Based English. – Mode of access: <http://corpus.byu.edu/glowbe/>

14. Борискина О. О. Классифицирующие структуры в криптоклассном исследовании / О. О. Борискина // Вестник Воронеж. гос. ун-та. Сер.: Лингвистика и межкультурная коммуникация. – 2010. – № 2. – С. 28–33.

15. Борискина О. О. Еще один подход к изучению метафорики политического дискурса / О. О. Борискина // Политическая лингвистика. – 2010. – № 1 (31). – С. 153–158.

16. Дони́на О. В. Место именного криптокласса «Res Acutae» в криптоклассных таксономиях английских непредметных имен / О. В. Дони́на // Дайджест–2013 : дипломные работы студентов факультета РГФ ВГУ / отв. ред. Н. А. Фененко. – Воронеж, 2013. – С. 43–51.

17. Зенкин А. А. Когнитивная компьютерная графика / А. А. Зенкин. – М. : Наука, 1991. – 192 с.

18. Chernoff H. The Use of Faces to Represent Points in K-Dimensional Space Graphically / H. Chernoff // Journal of the American Statistical Association, 1973. – № 68 (342). – P. 361–368.

19. Sinha P. Face Recognition by Humans : Nineteen Results All Computer Vision Researchers Should Know About / P. Sinha, B. Balas, Y. Ostrovsky, R. Russell // Proceedings of the IEEE. – November 2006. – Vol. 94, No. 11. – P. 1948–1962.

20. Atkinson R. L. Hilgard's Introduction to Psychology. History, Theory, Research, and Applications / R. L. Atkinson et al. – 13th ed. – 2000.

21. Pearson K. Notes on regression and inheritance in the case of two parents / K. Pearson // Proceedings of the Royal Society of London. – June 20, 1895. – № 58. – P. 240–242.

22. Дони́на О. В. Криптоклассные данные для определения меры языковой эквивалентности / О. В. Дони́на // Вестник Воронеж. гос. ун-та. Сер.: Лингвистика и межкультурная коммуникация. – 2015. – № 1. – С. 108–110.

23. Kendall M. Rank Correlation Methods / M. Kendall. – Charles Griffin & Company Limited, 1948.

Воронежский государственный университет

*Дони́на О. В., аспирант кафедры английского языка
в профессиональной международной деятельности
E-mail: olga-donina@mail.ru*

Voronezh State University

*Donina O. V., Post-graduate Student of the English
Language in the Professional International Activity Department
E-mail: olga-donina@mail.ru*