УДК 81

КЛАСТЕРНЫЙ ПОДХОД В ЛИНГВИСТИЧЕСКОМ АНАЛИЗЕ (НА МАТЕРИАЛЕ КОРПУСНОЛИНГВИСТИЧЕСКОГО АНАЛИЗА ЗАИМСТВОВАНИЙ ИЗ АНГЛИЙСКОГО ЯЗЫКА (АНГЛИЦИЗМОВ) В НЕМЕЦКОМ ЯЗЫКЕ)

© Н. Х. Нургалиева

Башкирский государственный университет
Россия, Республика Башкортостан, 450074 г. Уфа, ул. Заки Валиди,32.
Тел: +7 (347) 273 67 78.
E-mail: uti-rosen67@yandex.ru

Статья посвящена общим принципам корпусной лингвистики и проблеме обработки лингвистических корпусных данных. Автор подробно рассматривает метод кластерного анализа с точки зрения применимости к квантитативным данным лингвистического корпуса и ставит вопрос о поиске новых закономерностей развития лексических единиц, исходя из корпуснолингвистического исследования англицизмов в немецком языке. Выводы, приведенные в статье, касаются возможности эффективного применения кластерного анализа в лингвистических исследованиях. Автор приводит собственные примеры кластеризации, указывая на необходимость более глубокого изучения особенностей данного метода с точки зрения лингвистической и статистической обработки языковых корпусных данных.

Ключевые слова: *англицизм, корпусная лингвистика, лингвистический корпус, кластер, кластерный анализ, дендрограмма, статистический метод.*

Возможность использования обширных электронных ресурсов значительно облегчила процесс сбора материала в лингвистических исследованиях. Разработкой общих принципов построения и применения больших корпусных массивов языковых данных занимается корпусная лингвистика – раздел прикладной лингвистики, получивший наиболее активное развитие в конце XX в.

Корпусная лингвистика основана на анализе больших корпусов текстов. Корпус (от лат. corpus – «тело, единое целое») – является собранием взаимосвязанных между собой языковых высказываний (письменных или устных), принадлежащих естественным коммуникативным ситуациям, то есть не созданных специально для лингвистического или какого-либо другого вида анализа. В. Захаров, известный представитель корпусной лингвистики в России, определяет лингвистический, или языковой, корпус текстов как «большой, представленный в электронном виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач» [1, с. 3].

Корпусная лингвистика призвана ответить на следующие вопросы:

1. Какие принципы лежат в основе устройства корпусов, как должна быть устроена стандартизованная разметка корпуса относительно различных языковых параметров (жанровая и стилевая разметка текстов, морфологическая разметка и т.п.). 2. Какие лингвистические и литературоведческие задачи можно решать с помощью корпусов. 3. Как пользоваться корпусами, включая специальные языки запросов к корпусам [2].

Корпусная лингвистика носит квантитативный характер. В ее основе лежит идея о том, что частота употребления языковой единицы является основой для утверждения о ее свойствах. При этом частота употребления языковой единицы рассматривается с двух сторон: с одной стороны, учитывается то, как часто отдельные морфемы, слова или грамматические конструкции встречаются в корпусе; с другой – факты употребления одних языковых единиц с другими или в составе определенной грамматической конструкции. Первый вид употреблений репрезентируется обычно в так называемых частотных списках (списках слов данного языка с информацией о частоте их встречаемости), второй – в так называемых конкордансах (от англ. concordance – «согласие, согласованность»). Под конкордансом понимаются все употребления заданного языкового выражения в контексте, возможно, со ссылками на источник. Конкорданс как средство корпуснолингвистического анализа может использоваться как при создании списка ключевых слов (что часто используется в научных публикациях), так и при исследовании частотности определенных словосочетаний и решения конкретных переводческих проблем.

С. Грис, американский эксперт в области корпусной лингвистики, призывает правильно толковать понятие «частота» – не только как феномен употребления слова 100 или 1000 раз в исследуемом корпусе. Ученый считает, что и не частотные, единичные употребления, часто окказиональных образований, так же необходимо учитывать при корпуснолингвистическом анализе. Даже так называемую «не-встречаемость» (non-occurrence) С. Грис считает релевантной для корпуснолингвистического анализа. Таким образом, он выделяет три вида случаев, с которыми имеет дело корпусная лингвистика: 1. Наблюдаемая частота больше или равна нулю. 2. Наблюдаемая частота одного явления больше или меньше наблюдаемой частоты другого явления. 3. Наблюдаемая частота больше или меньше ожидаемого показателя [3, с. 17].

Корпус является в корпусной лингвистике исходной точкой, своеобразной базой лингвистических данных, на основе которой становится возможным описание и объяснение лингвистических феноменов. Корпусные исследования позволяют, используя статистические методы, сформулировать, подтвердить или опровергнуть некоторую гипотезу о том или ином языковом явлении на большом объеме материала. Несмотря на то что использование статистических методов в лингвистике находится на стадии развития, некоторые лингвисты применяют данные методы исследования для достижения определенных лингвистических целей. Подсчеты абсолютных частот и выведение из них посредством элементарных арифметических операций относительных частот и подобных же данных помогают проследить не только количественные изменения слов, но и закономерности их развития. В этом отношении наиболее удобным нам представляется кластерный анализ, получивший распространение благодаря развитию компьютерных технологий и формализованных программ расчетов.

Кластер (от англ. cluster – «пучок, скопление, концентрация») – это группа объектов, выделенная с помощью одного из методов кластерного анализа по формальному критерию их близости друг к другу [4, с. 56]. Главное назначение кластерного анализа – разбиение множества исследуемых объектов и признаков на однородные группы или кластеры. Это означает, что решается задача классификации данных и выявления соответствующей структуры в ней [5, с. 14].

Задачи кластерного анализа можно объединить в следующие группы:

1) Разработка типологии или классификации. 2) Исследование полезных концептуальных схем группирования объектов. 3) Представление гипотез на основе исследования данных. 4) Проверка гипотез или исследований для определения того, действительно ли типы (группы), выделенные тем или иным способом, присутствуют в имеющихся данных.

Выделение элементов с высокой степенью корреляции происходит в кластерном анализе на основе определения расстояния между объектами выборки. Расстоянием (метрикой) между объектами в пространстве параметров называется такая величина d_{ab}, которая удовлетворяет аксиомам:

A1. $d_{ab} > 0$, $d_{ab} = 0$

A2. $d_{ab} = d_{ba}$

A3. $d_{ab} + d_{bc} >= d_{ac}$

[5, с. 15].

Все методы кластерного анализа можно разделить на иерархические и неиерархические. Каждая из групп включает множество подходов и алгоритмов. Для решения лингвистических задач наиболее подходящим нам представляется иерархический кластерный анализ, суть которого состоит в последовательном объединении меньших кластеров в большие или разделении больших кластеров на меньшие. Преимущество этой группы методов в сравнении с неиерархическими методами – их наглядность и возможность получить детальное представление о структуре данных. В качестве правил выступают критерии, которые используются при решении вопроса о «схожести» объектов при их объединении в группу (агломеративные методы) либо разделения на группы (дивизимные методы) [6]. Агломеративные методы мы будем использовать в нашем исследовании. Эта группа методов характеризуется, таким образом, соответствующим уменьшением числа кластеров.

Входными данными для кластерного анализа является матрица N*M, где N– число рассматриваемых объектов, M – количество признаков. Общая идея алгоритма кластеризации состоит в следующем: на первом шаге все объекты разбиваются по отдельным кластерам (т.е. каждый кластер будет содержать ровно один элемент). На каждом последующем шаге по определенным формулам вычисляются расстояния между кластерами и два наиболее близких объединяются в один [7] (*рис.* 1).

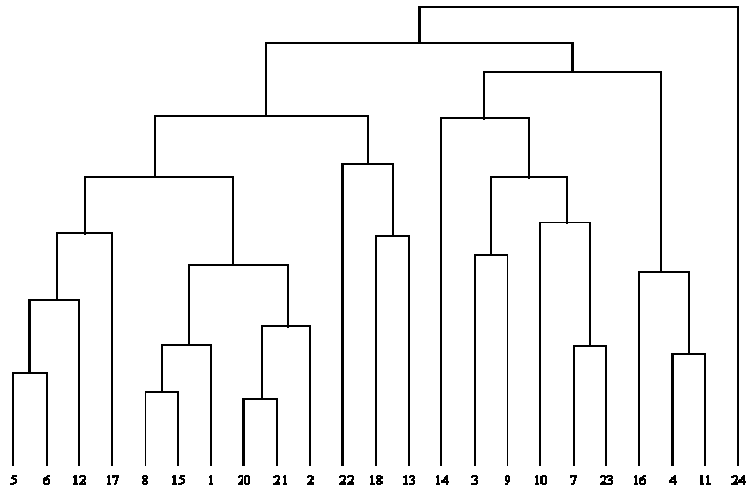

Рис.1. Модель объединения объектов выборки кластерного анализа в дендрограмме.

Процедура объединения кластеров (агломеративный метод) продолжается до тех пор, пока все они не объединятся в один кластер, содержащий все элементы. Процедура кластеризации визуализирована – можно пошагово следить за тем, какие кластеры образуются [6].

Результаты кластеризации представлены на дендрограмме (*рис*. 1, *рис*. 2). Дендрограмму можно определить как графическое изображение результатов процесса последовательной кластеризации, которая осуществляется в терминах матрицы расстояний [8]. Дендрограмму также называют древовидной схемой, деревом объединения кластеров, деревом иерархической структуры [5, с. 16].

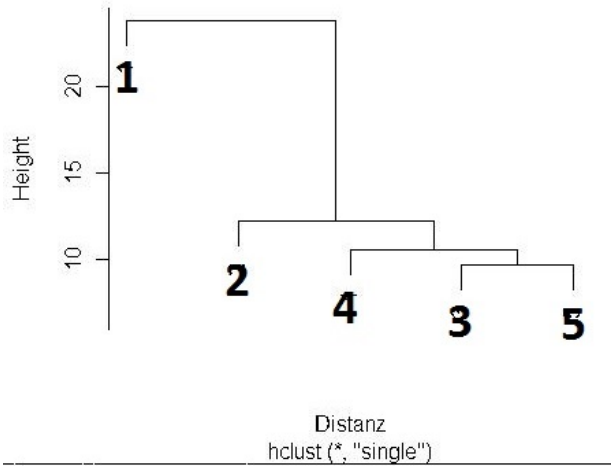

Рис.2. Объединение объектов выборки кластерного анализа в статистической программе R.

На *рис.* 2 изображена вертикальная дендрограмма, номера объектов выборки располагаются по горизонтали, результаты кластеризации (значения расстояний или сходства) – по вертикали.

Мы видим, что в начале анализа объединяются в один кластер объекты **3** и **5**, поскольку расстояние между ними самое минимальное и равно приближенно 7. Это слияние отображается на графике горизонтальной линией, соединяющей выходящие из соответствующих точек вертикальные отрезки. Далее, на втором шаге к этому кластеру, включающему в себя уже два объекта, присоединяется объект **4**. На следующем шаге происходит объединение объекта **2** и образовавшегося кластера, расстояние между которыми приближенно равно 12. 5. На последнем шаге происходит объединение объекта 1 с кластером из объектов **2, 3, 4, 5**. При наличии большего количества объектов мы бы рассматривали такое объединение как еще один кластер.

Общей проблемой в кластерном анализе является сложность определения «естественного» числа кластеров в модели. Конечное решение – сколько должно быть кластеров – принимается пользователем. Это число определяется в процессе разбиения множества на кластеры. Процесс его определения часто связан с нахождением баланса между решением задачи наиболее полного описания данных и ростом сложности модели [9].

Кластерный анализ в лингвистике широкого применения пока не нашел, но, по нашему мнению, данный статистический метод имеет большой потенциал для создания новых лингвистических открытий. К примеру, кластерный анализ может служить средством создания новых классификаций, определения моделей развития лексем и поиска общих закономерностей развития целых групп лексем.

Для проведения такого кластерного анализа, безусловно, очень важно подобрать статистическое программное обеспечение, отвечающее целям исследования. Мы остановились на открытой статистической программе R, созданной в 1992 г. на базе Оклендского Университета, Новая Зеландия. Программа доступна для скачивания по ссылке в интернете: www.r-project.org. Установить среду R можно на компьютере под управлением Windows, MacOS или Linux. R – язык программирования для статистической обработки данных и работы с графикой, хотя стандартная комплектация R не предполагает графического интерфейса, привычного для многих пользователей. Истоки языка программирования R лежат в языке программирования S, с которым у них очень много общего. Синтаксис языка достаточно прост и легок в изучении. На сегодняшний день написано более сотни книг по самым разным направлениям использования среды статистических вычислений R [10].

Статистическая программа R обладает функцией иерархического кластерного анализа, осуществляется которая с помощью команды hclust() и метода single (Single-Linkage). Рассмотрим пример кластерного анализа англицизмов в немецком языке. Для подготовки необходимых данных мы составили список англицизмов, содержащий 2 509 лемм. Под леммой в данном случае понимается базовая форма слова, зафиксированная в словаре; при таком анализе флективные формы обрабатываются как одно слово. С помощью проведения корпусно-лингвистического анализа мы вычислили частоты употребления слов в корпусе еженедельного немецкого журнала der Spiegel, включающего в себя статьи различных рубрик с 1947 по 2010 гг. Исследованный нами корпус содержит 307 107 статей с общим количеством слов 195 829 923. В результате квантитативного анализа корпуса мы получили временные ряды встречающихся англицизмов с частотами их употребления (*рис.* 3).

Absinthe	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Absinths	0	0	0	0	0	0.075561	0	0	0	0	0	0	0	0	0	0	0	0
Absorbers	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Accountants	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Account	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Acher	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Aches	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Acide	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Acquisitions	0	0	0	0	0.067758	0	0	0	0	0	0	0	0	0	0	0	0	0
Actionfilm	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Actionfilmen	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Actionfilmer	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Action	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Actor	0	0	0	0	0.067758	0	0	0	0	0	0	0	0	0	0	0	0	0.04468
Act	0	0	0	0	0	0	0	0	0	0.136115	0	0	0	0	0.055271	0	0	0.04468
Adapter	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Adapters	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Administration	0	0	0	0	0	0.075561	0	0	0	0	0	0	0	0	0	0	0	0
Adults	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Advanced	0	0	0	0	0	0	0.076712	0.075931	0.07168	0	0	0	0	0	0	0	0.052481	0.089359
Adventurer	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Adventures	0	0	0	0	0	0.075561	0	0	0	0	0	0	0	0	0	0.054412	0	0
Advertisements	0	0	0	0	0	0	0	0	0	0	0	0	0	0.331038	0	0	0	0
Advertiser	0	0	0	0	0.135516	0	0.076712	0.075931	0	0	0	0	0	0	0	0	0.052481	0
Aerobics	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Рис. 3. Матрица относительных частот употребления англицизмов.

Для статистического анализа полученные временные ряды преобразовали в векторную форму, то есть в матрицу N*M, где N – рассматриваемые англицизмы, M – их признаки, в данном случае частоты, встречающиеся в указанных годах. Матрица включает в себя 10 418 наблюдений, при этом учитывались леммы лексических единиц и их флективные формы. Абсолютные частоты были нами преобразованы в относительные – 1/ 100 000 токенов.

Применяя кластерный анализ к совокупностям временных рядов, мы можем выделить периоды схожести некоторых показателей и определить группы внутри временных рядов со схожей динамикой. Проведя кластерный анализ полученных данных, мы получили 67 кластеров (*рис*. 4). По оси X – объекты выборки, по оси Y – расстояние между ними.

Качественный анализ каждого кластера показал, что состав каждой группы является в высокой степени негомогенным с точки зрения принадлежности к частям речи, формы слов и их семантики. Рассмотрим, к примеру, кластер 14. Данная группа включает в себя следующие слова:

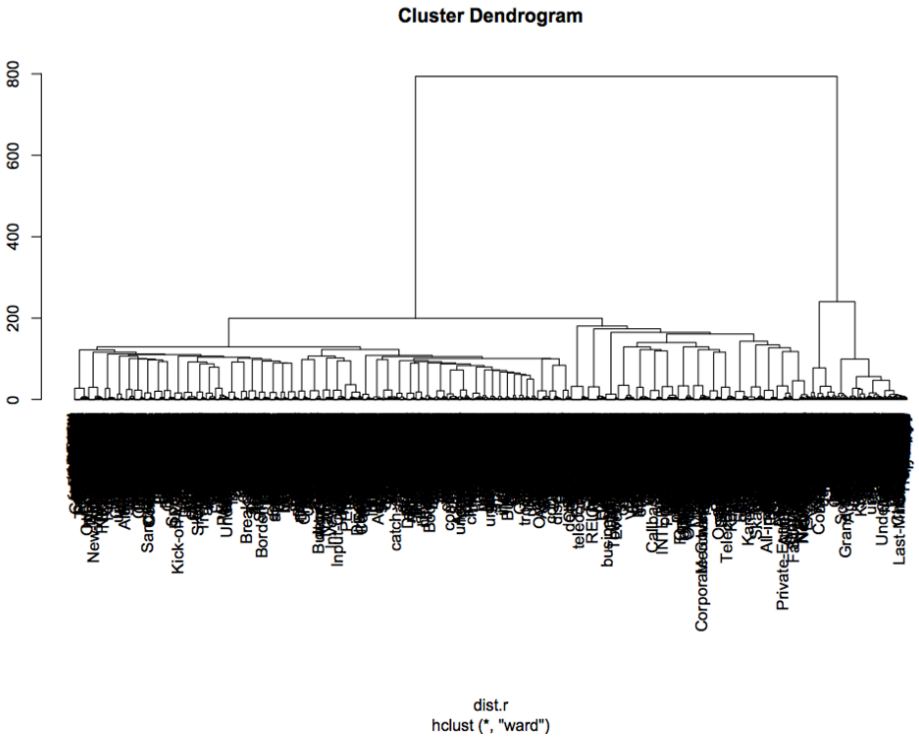

Рис. 4. Иерархический кластерный анализ 10 418 англицизмов в корпусе der Spiegel в статистической программе R.

Accounts, vegan, delete, BINGOS, Exploits, Casting, iPods, meet, Mainstream, Mails, Map, Hippie-, Jobs, Clevere, Demands, completed, freestyle, Haircut, indoor, rack, Traders, grabbing, Halfpipes, Tipi, Stresstest, Flyern, scannen, off-season, request, googeln, trades, supporters, Concept, Messages, perfect, habits, tasks, Errors, lifestyle, Sounddatei, Bankings, Guides, Bail-out, Audits, swap, Internet, failures, Onlineshops, Median, airlines, Einloggens, Screenshots, Offliner, iPad, smoking, nugget, Slacklinen, Slackliner, Backshop, Competitions, surfe, Dragstern, Karriere-Coaching, pleasures, Bollywood, Facelifts, PayPal, Celebrities, Skateboardfahrer, Hands, stress, bottom-up, Tokens, iPhone, iPhones, Spotmarkt, blockbuster, Blogger, Blog, Bloggers, fancy, Avatars, Fotoshooting, Nacktscanner, Touchscreens, Touchpads, Nollywood, shoppend, backup, BYD, maintain, mailen, Beamer, Clients, desks, gemailt, WLAN, cover-, PhD, Tasern, haircut, different, coaster, Loverboys, Snowboarden, Bonde, Cargohose, Smartphones, Smart-Phones, Smarts, stylisch.

Тем не менее, некоторые слова показывают сходные дистрибуции, например, такие как iPad, googeln, Blog, Smartphones, PayPal. Возникновение данных единиц относится к началу 2000 г., пика же их употребление достигает в 2008–2010 гг. (*рис*. 5).

По оси X – годы с 1948 по 2010, по оси Y– относительные частоты употребления.

Такое развитие слов является статистически эвидентым, так как сфера интернет- и компьютерных технологий получила в последнее десятилетие стремительное развитие. При этом значительная часть слов, обозначающих в немецком языке реалии развития компьютера, сети Интернет и техники в общем заимствована из английского языка.

Интересным нам показалось дистрибутивное соотношение лексем single (не замужем) и Powerfrau (сильная, уверенная в себе женщина) (*рис*. 6). По оси X – годы с 1984 по 2010, по оси Y – относительные частоты употребления.

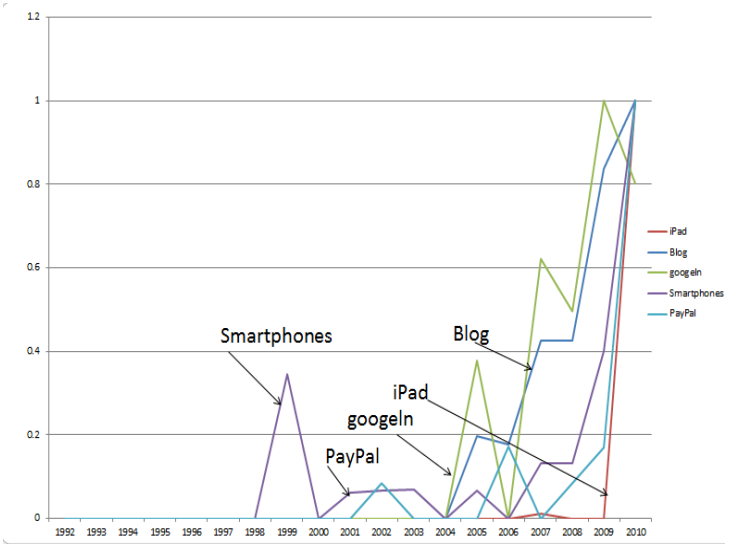

Рис.5 Временные дистрибуции лексем iPad, googeln, Blog, Smartphones, PayPal с первого момента возникновения в корпусе (1998 г.)

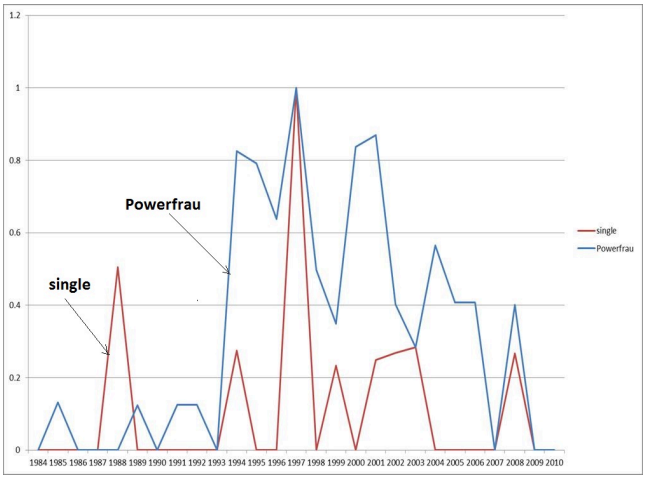

Рис .6 Временные дистрибуции лексем single и Powerfrau с первого момента возникновения в корпусе (1984 г.)

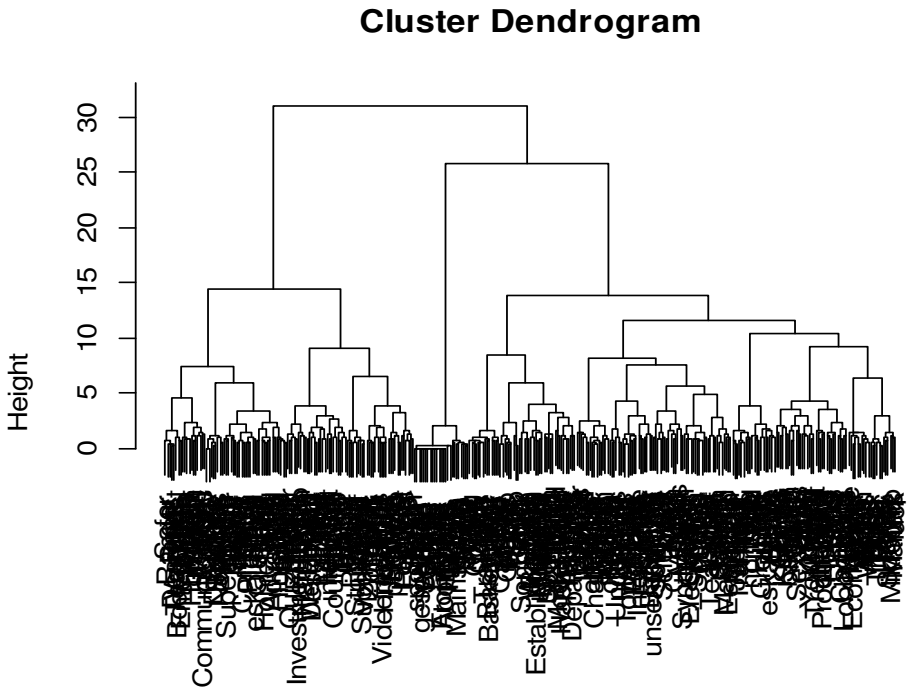

Рис. 7. Иерархический кластерный анализ 332 англицизмов в корпусе der Spiegel в статистической программе R

На графике можно проследить явные сходства в употреблении данных лексем на протяжении определенных периодов. По нашему мнению, такие корпуснолингвистические данные могут стать эмпирической базой для глубоких социолингвистических исследований; такие языковые факты могут быть объяснены с учетом большого количества социальных факторов.

Кластерный анализ такого большого количества англицизмов показал, что вычислить значения близкого сходства при большом количестве слов достаточно сложно – из-за попытки достичь полной схожести мы, тем самым, увеличиваем число кластеров, что еще больше затрудняет работу из-за необходимости анализа каждого конкретного кластера.

Попробуем проанализировать другие, менее объемные данные. Матрица для новой выборки включает в себя 332 наблюдения, не ориентированных на какой-либо конкретный семантический, грамматический или прочий критерии. После проведения кластерного анализа мы получили 10 кластеров (*рис*. 7).

Анализ показал, что слова, принадлежащие ко второму кластеру, содержат наибольшее количество одно- или двусложных слов; при сравнительном анализе всех кластеров вышло, что данная группа содержит наибольшее количество кратких форм слов – 24% (6 из 25). Это такие слова, как: ICE, Bossi, SMS, Soaps, Smog, Video. (К кратким формам слов мы относим как различные аббревиатуры и сокращения элементов в составе составных слов (например, E-Mail = (E-(lectronic) Mail), так и некоторые суффиксальные образования, такие, как Pulli, Profi и т.д.). Для сравнения данные кластеров приведены в *табл.* 1.

Типичная дистрибуция кластера 2 за временной промежуток 15 лет с момента первого употребления представляет собой следующий график (*рис*. 8).

По оси X – 15 лет с момента первого возникновения слова в корпусе, по оси Y – относительные частоты употребления.

Таблица 1

Распределения полных и кратких форм слов в кластерах 1–10

Номер кластера	Всего слов	Полная форма	Краткая форма
1	21	18	2 (9.5%)
2	24	18	6 (24%)
3	56	54	2 (3.5%)
4	70	68	2 (2.8 %)
5	32	31	1 (3.1%)
6	14	14	0 (0%)
7	34	34	0 (0%)
8	18	17	1 (5.5%)
9	32	30	2 (6.3%)
10	27	27	0 (0%)

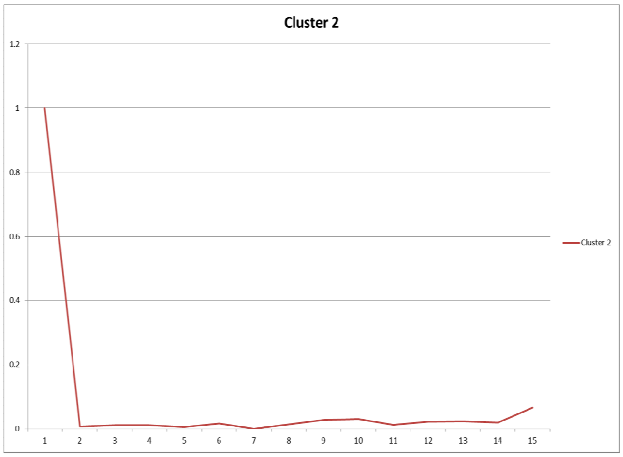

*Рис.*8 Временная дистрибуция средних значений кластера 2

Значит ли это, что фактор длины слова влияет на частоту его употребления? Воспринимаются ли говорящими более короткие слова как современные, «модные» и употребляются ли тенденциозно, постепенно теряя свою актуальность? Без серьезных статистических исследований ответить на этот вопрос невозможно. Необходимо провести целый ряд статистических процедур для определения релевантности данного фактора, для этого необходима также более обширная выборка для исследования (лучше всего из нескольких корпусов). Кроме того, необходим глубокий социолингвистический анализ элементов каждого из кластеров.

Кластерный подход позволяет проводить межъязыковые обобщения исследуемого понятия на основе сопоставления совокупностей признаков, которыми обладают объекты в составе одной группы. Разработаны пакеты компьютерных программ, реализующих процедуры кластерного анализа. Хотя существующие пакеты обладают большой мощностью и универсализмом, но достаточно сложны для использования. Это такие программы, как R, MatLab, SPSS, Statistica и др. С другой стороны, всегда необходимо учитывать недостатки кластерного анализа. Во-первых, данный вид анализа может давать неустойчивые кластеры. Поэтому очень важно внимательно отнестись к подготовке исходной выборки, чтобы кластеры были «чистыми». Кроме того, необходимо уделить особое внимание таким характеристикам, на основе которых проводится кластеризация, так и методу, эффективному в конкретной области исследования. Результаты классификации необходимо перепроверять на других примерах, то есть сравнивать результаты классификаций.

Во-вторых, кластерный анализ реализует индуктивный метод исследования от частного к общему. В идеальном случае выборка для классификации должна быть очень большая, неоднородная [11]. Все гипотезы, полученные в результате кластерного анализа, необходимо перепроверять. Не решена так же проблема точного определения числа кластеров [12, с. 139].

И, наконец, одна из самых важных проблем – проблема интерпретации результатов кластеризации. Безусловно, кластерный анализ предполагает более глубокий качественный анализ каждой из групп. Необходимо еще раз подчеркнуть роль статистических методов в лингвистических исследованиях. В свете научно-технической революции и растущего объема информации (в том числе корпусов текстов) лингвистика, несмотря на сложность и многогранность объекта исследования, нуждается в разработке новых методологических основ комплексного квантитативно-системного подхода.

ЛИТЕРАТУРА

1. Захаров В. П. Корпусная лингвистика: Учебно-методическое пособие. СПб: Санкт-Петербургский государственный университет, 2005. 46 с.
2. Архипов А. В. Корпусная лингвистика // Фонд знаний Ломоносов. URL: http: // http://www.lomonosov-fund.ru/enc/ru/encyclopedia:01210:article.
3. Stefan Th. Gries. Language and Linguistics. // Language and Linguistics Compass. 2009, vol. 3, №5. P. 1225–1241.
4. А. А. Грицанов, В. Л. Абушенко, Г. М. Евелькин, Г. Н. Соколова, О. В. Терещенко. Социология: Энциклопедия // Минск: Книжный Дом, 2003. 1312 с.
5. Н. Н. Буреева. Многомерный статистический анализ с использованием ППП “STATISTICA”. Нижний Новгород: ННГУ, 2007. 114 с.
6. Сотник С. Иерархическое группирование. URL: http://sotnyk.com/Articles/AILectures/htm/gl3_11.htm.
7. Соловьев В. Д. Кластерный анализ многофакторных лингвистических понятий. URL: http://www.dialog-21.ru/digest/archive/2000/?year=2000&vol=22725&id=6555.
8. Кластерный анализ в задачах социально- экономического прогнозирования. URL: http://www.roman.by/r-95658.html.
9. Гончаров М. Кластерный анализ. URL: http:// www.businessdataanalytics.ru/ClusterAnalysis.htm.
10. Обзор статистических программ. URL: http:// www.sciencefiles.ru/section/46/.
11. Попов О. Кластерный анализ. Просто о сложном. URL: http://psystat.at.ua/publ/1-1-0-18.
12. Fisher D. H. Knowledge acquisition via incremental conceptual clustering // Machine Learning. 1987, vol.2, issue: 2.

Поступила в редакцию 14.04.2013 г.