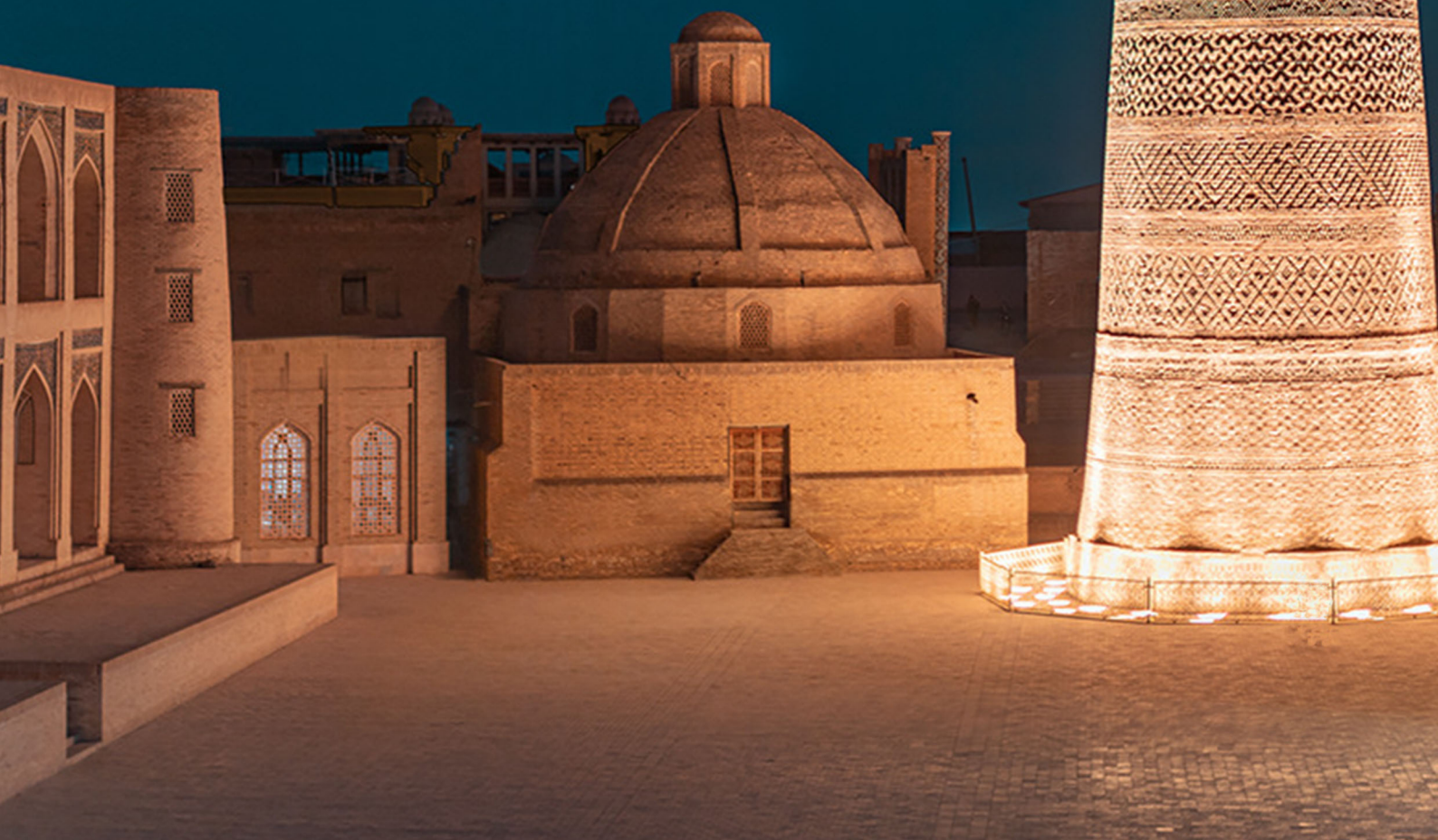
ISSN 3030-3370
№ 2,3 (9) 2025

AYNIY VORISLARI

XALQARO ELEKTRON JURNAL
INTERNATIONAL ELECTRONIC JOURNAL



ISSN 3030-3370

“AYNIY VORISLARI”
xalqaro elektron jurnal

«НАСЛЕДНИКИ АЙНИ»
международный электронный журнал

“AYNIY’S SUCCESSORS”
international electronic journal

№ 2,3 (9) 2025

Ахмеджанова Ситора Джураходжаевна,
преподаватель Бухарский государственного университета кафедры русского языка и литературы (Узбекистан)
E-mail: ***iamhappyandglad@gmail.com***

ОПРЕДЕЛЕНИЕ ОСНОВНЫХ ПОНЯТИЙ И ТИПОЛОГИЯ ЛИНГВИСТИЧЕСКИХ КОРПУСОВ

Аннотация. Корпусная лингвистика считается одной из наиболее перспективных и прогрессирующих сфер в изучении языка. Актуальность данной статьи заключается в огромном потенциале лингвистических корпусов, который еще не в полной мере осознан научным сообществом, хотя бы по тому, что текст – основной объект корпусной лингвистики – в разнообразных формах своей реализации является одной из основных составляющих языковой системы и мыслительно-речевой деятельности современного носителя языка. В статье раскрывается понятие “корпус”, приведена классификация корпусов текстов, подробно описывается каждая группа корпусов текстов, приводятся критерии лингвистических корпусов, объясняется понятие “разметка”, рассматриваются основные понятия корпусной лингвистики, методы и области ее применения. Описывается преимущество корпусов текстов в лингвистическом исследовании. Также в статье анализируется возникновение и становление корпусной лингвистики, приводится типология корпусов, описывается каждый тип корпуса в отдельности.

Ключевые слова: корпус, лингвистика, корпусная лингвистика, текст, семантическая, анафорическая, графематическая, репрезентативность, понятийный аппарат, вариация.

Annotatsiya. Korpus lingvistikasi tilshunoslikning eng istiqbolli va ilg‘or yo‘nalishlaridan biri hisoblanadi. Ushbu maqolaning dolzarbligi ilmiy hamjamiyat tomonidan hali to‘liq e’tirof etilmagan lingvistik korpusning ulkan salohiyatida, hech bo‘lmaganda, chunki korpus tilshunosligining asosiy ob'ekti bo‘lgan matn uning turli shakllarida til tizimining asosiy tarkibiy qismlaridan biri bo‘lib, zamonaviy ona tilida so‘zlashuvchining aqliy va nutqiy faoliyati. Maqolada “korpus” tushunchasi ochib berilgan, matn korpusining tasnifi berilgan, matn korpusining har bir guruhi batafsil tavsiflangan, lingvistik korpus mezonlari keltirilgan, “belgilash” tushunchasi tushuntirilgan va korpus lingvistikasining asosiy tushunchalari, uni qo‘llash usullari va sohalari ko‘rib chiqilgan. Tilshunoslik tadqiqotlarida matn korpusining afzalligi tasvirlangan. Shuningdek, maqolada korpus lingvistikasining paydo bo‘lishi va rivojlanishi tahlil qilingan, korpus tipologiyasi berilgan va korpusning har bir turi alohida tavsiflangan.

Kalit so‘zlar: korpus, lingvistika, korpus lingvistikasi, matn, semantik, anaforik, grafematik, reprezentativlik, kontseptual apparat, variatsiya.

Abstract. Corpus linguistics is considered one of the most promising and progressive areas in language study. The relevance of this article lies in the enormous potential of linguistic corpora, which has not yet been fully recognized by the scientific community, at least because the text – the main object of corpus linguistics – in its various forms of implementation is one of the main components of the language system and the mental and speech activity of a modern native speaker. The article reveals the concept of “corpus”, provides a classification of text corpora, describes in detail each group of text corpora, provides criteria for linguistic corpora, explains the concept of “markup”, and examines the basic concepts of corpus linguistics, methods and areas of its application. The advantage of text corpora in linguistic research is described. The article also analyzes the emergence and development of corpus linguistics, provides a typology of corpora, and describes each type of corpus separately.

Key words: corpus, linguistics, corpus linguistics, text, semantic, anaphoric, graphematic, representativeness, conceptual apparatus, variation.

Введение. Каждое исследование, проводимое языковедом, должно быть направлено, как минимум, на определенные стадии деятельности:

Подбор положений и базы категоризации исследуемых объектов.

Разделение объектов по категориям сообразно с данной базой.

и толкование итогов разделения объектов по категориям, интерпретация оснований такого разделения [1, 29].

В то же время первая стадия этой деятельности предполагает существование исследуемых объектов, то есть приобретение практических данных для создания на заключительной стадии теории.

В наши дни корпусная лингвистика при подготовке и анализировании эмпирических данных получает широкое распространение, благодаря интенсивному росту информационных технологий.

Основная часть. Впервые о корпусной лингвистике стало известно в 60-х гг. XX в. Тексты формировались главным образом на базе английского языка, но вскоре стали появляться корпуса (в корпусной лингвистике принято использование формы множественного числа “корпуса”. См.: Толковый словарь русского языка под редакцией Д. Н. Ушакова: Корпус, мн. ч. корпуса) на материале и других языков. Тогда же в Брауновском университете США учеными У.Н. Френсисом и Г. Кучерой был составлен первый корпус текстов на электронном носителе [2], который состоял из 1 миллиона словоупотреблений (500 текстов по 2 000 слов в каждом). Также к нему были приложения в виде показателя частоты употребления слов по алфавиту и определенные данные статистики.

Корпус – это сборник текстов на одном или нескольких языках, которые связаны некоторыми характеристиками. В своём труде Л. Лемницер и Х. Цинсмайстер дали такое понятие корпусу: “Корпус представляет собой набор письменных или устных высказываний. Данные корпуса обычно оцифровываются, т.е. часто хранятся на компьютерах и машиночитаемы” [3, 7].
В то же время составные элементы корпуса – тексты – складываются из материалов, метаданных, которые представляют эти материалы, и из языковых обощений, которые эти материалы организовывают.

Как частный раздел языкознания корпусная лингвистика окончательно сложилась в конце XX в.

Корпусная лингвистика как отдельный раздел языкознания окончательно сформировалась в первой половине 90-х гг. XX в. Тогда же и начинает складываться понятийный аппарат [4, 37]. В частности, Дж. Синклер определяет понятие “корпус” следующим образом: “набор естественно встречающихся языковых текстов, выбранных для характеристики состояния разнообразия языка” [5, 171]. Здесь выделяется одно из ключевых положений при подборе текстов для составления корпуса – имеются в виду недоработанные тексты, другими словами, язык демонстрируется в том виде, в котором он был выражен (устная или письменная речь). К тому же, в корпусе предложены не реальные “шаблоны” и “положения” для верной организации сообщения, а максимально возможное число “вариаций” языка, хоть часть из них и располагаются не в центре системы языка. Дальше понятие “корпус” все больше уточняется: “Корпус – это совокупность текстов, предназначенная для какой-либо цели, обычно обучающей или исследовательской. Корпус – это не то, что говорит или знает говорящий, а нечто, созданное исследователем. Это запись производительности, обычно множества разных пользователей, предназначенная для изучения, чтобы мы могли сделать выводы о типичном использовании языка. Поскольку оно предоставляет методы наблюдения закономерностей того типа, которые уже давно были замечены литературными критиками, но не были выявлены эмпирически, компьютерное исследование больших корпусов, возможно, может предложить выход из парадоксов дуализма” [6, 239-240] (Перевод наш).

Мы предполагаем, что более или менее цельную формулировку понятия “корпус” возможно отыскать в работах В.П. Захарова. Лингвист описывает корпус как большой, изображаемый в электронном виде, организованный и спланированный, филологически внушительный конгломерат языковых данных, рассчитанный для разрешения определенных лингвистических вопросов и задач [7, 3]. Эту формулировку можно квалифицировать как “деятельностную”, по большому счёту объясняющую лингвистическую тенденцию организованных текстовых массивов.

Результаты и обсуждение. В результате, в любой из описанных формулировок понятия “корпус” отмечается следующее:

1) большое количество текстов должно быть представлено в электронном виде (в сети Интернет или на любом носителе);

2) языковой материал должен быть распределен для рассмотрения в лингвистических целях;

3) по итогам рассмотрения должен существовать способ разнообразного разделения полученных языковых данных (по теме, жанру, году создания и т.д.).

Рассматривая первое, замечена возможность постоянного доступа текстов в электронном виде. Огромное количество корпусов текстов можно классифицировать на три значительные группы:

Свободно-доступные;

Частично-доступные;

Коммерческие.

К первой группе относятся достаточно малое количество из имеющихся на сегодняшний день корпусов текстов. Достаточно содержательным считается Национальный корпус русского языка, объём которого составляет более 500 миллионов слов.

К следующей группе относится большинство из имеющихся корпусов, тем не менее для решения определенных лингвистических задач данный частичный доступ считается вполне достаточным. Например, в Британском национальном корпусе итоги запроса составляют всего до 50 произвольных примеров, также нет большинства функций search-интерфейса, который предоставляется только вкупе с полноценной (и платной) версией корпуса.

Вместе с тем есть некоммерческая разновидность этого корпуса [8], которая становится доступной после простого процесса регистрирования. В данной вариации для поиска предлагается около 100 миллиона слов в текстах 1980-1993 гг.

А в третью группу определяют, к примеру, Банк английского языка (British National Corpus), в котором есть вариант предварительной бесплатной подписки на месяц для приобретения доступа в Collins Wordbanks Online [9], который содержит около 533 миллиона слов, и только потом следует купить коммерческую версию корпуса.

Другим значительным критерием языковедческого корпуса текстов считается присутствие или отсутствие разметки, потому что для решения языковедческих вопросов и задач наличия обычного конгломерата текстов недостаточно.

Разметка – это присвоение текстам и их элементам специальных меток: внешних, внеязыковых, системных и собственно лингвистических, которые описывают различные параметры элементов текста [7, 6]. В метаразметку входит не только информация о тексте, но и данные об авторе. Изучим собственно лингвистические типы разметки. Можно начать с разметки частей речи, являющейся часто встречающейся в существующих корпусах, вместе с тем принимается во внимание не только морфологические показатели, но и грамматические.

Разметка частей речи происходит при участии специальных программ автоматизированного морфоанализа. К примеру, в малой части Национального корпуса русского языка (6 миллионов словоупотреблений) осуществлены ручное устранение морфоомонимии и вспомогательная корректировка итогов процесса программы автоматического морфоанализа [10, 86].

В Мангеймском корпусе немецкого языка разметка частей речи наличествует в большинстве своем в под-корпусах текстов публицистики. В числе остальных типов разметки особенно необходимо обратить внимание на синтаксическую разметку, которая представлена не во всем конгломерате корпуса, а только в его незначительной части, ведь этот тип разметки, предполагающий определение синтаксической структуры для любого предложения, производится практически вручную и нуждается в значительных временных затратах. Также, в корпусе есть и другие типы разметки, к примеру, семантическая, просодическая, анафорическая, графематическая и т.д. Это по большей части способствует упрощению процедуре естественного сбора данных исследователем с учётом верно указанных условий поиска.

Тем не менее, чтобы разработанный корпус текстов удовлетворял разного рода лингвистическим задачам, которые стоят перед языковедом, он должен в свою очередь иметь еще как минимум два показателя.

В первую очередь, имеется в виду репрезентативность корпуса текстов. Кибрик А.Е., Брыкина М.М., Леонтьев А.П. и Хитров А.Н. считают, что репрезентативность позволено оценивать по преобразованию “относительной частоты” исследуемого факта при росте “выборки”. Если “относительная частота” факта от увеличения “каждого последующего фрагмента текста” не будет часто меняться, то это значит, что “корпус в целом репрезентативен” [11, 21]. Вместе с тем хоть и наблюдается недопустимость при данной формулировке репрезентативности определить связи со статистикой, акцентируется, что данное требование является обязательным, но все же неполным для установления репрезентативности корпуса текстов.

В основном, вопрос установления репрезентативности разных корпусов текстов считается до сих пор актуальным, однако, надо признать, недостаточно разработанным. Только репрезентативность трансформирует обычный комплекс различных текстов именно в корпус текстов, подходящий для осуществления лингвистического исследования. Вместе с тем, речевая деятельность человека до такой степени многообразна, что почти невозможно реально передать все имеющиеся вариации языка, вышеупомянутых ранее. По этой причине вопрос репрезентативности корпуса текстов считается больше вопросом из сферы беспристрастности любого научного исследования. В данном случае желательно полагаться на здравый смысл самого исследователя, когда имеется в виду пользовательский корпус (разрабатывается самим исследователем в соответствии с целями его исследования), или группы исследователей, когда имеется в виду создание корпуса, требующего масштабность языковых явлений, стилей, жанров и т. д. (к примеру, национального корпуса конкретного языка).

Значительным условием при обозначении корпуса является также и простота его эксплуатации, иначе говоря, корпус должен быть оснащен специализированной системой поиска, которая должна быть (в идеале) довольно понятна в достаточной степени и проста в эксплуатации. Эксплуатация Национального корпуса русского языка или Британского национального корпуса (Банка английского языка) значительных проблем, чего не скажешь о поисковой системе Мангеймского немецкого корпуса. Мы считаем, что корпус не должен отнимать много времени, которое нужно для поиска определенного явления, и не должен предлагать хитрую методику поиска, так как изучение его базовых пунктов требует от исследователя в некоторых случаях чисто технических и математических знаний.

Корпус и его типы. В числе имеющегося разнообразия исследовательских корпусов в отдельных случаях очень трудно ориентироваться, потому как цели и задачи, поставленные перед языковедом, нередко отождествляются в общем, но в частных отраслях и сферах различаются. Начальный этап, осуществляемый исследователем при изучении исследуемых “объектов” – это верный выбор надлежащего корпуса. Всё многообразие действующих корпусов устанавливается разнообразием “исследовательских и практических задач, для решения которых они создаются” [7, 12].

1. Устные, письменные, смешанные.

Устный корпус – систематизированный комплекс речевых отрывков, оснащённый программными возможностями доступа к ним [12, 71-72]. Впервые устные корпуса стали функционировать в 80-х годах XX века на основе американского английского языка. Затем появляются особые координационные центры, которые собирали, хранили, распространяли и создавали устные корпуса. К примеру, LDC (Linguistic Data Consorcium) [13], CSLU (Center for Spoken Language Understanding) [14], ELRA (European Language Resources Association) [15].

Большая часть из действующих корпусов принадлежат к письменным или смешанным (к примеру, доступная часть Мангеймского корпуса немецкого языка) [16], все же часть лингвистически размеченных устных текстов даже в смешанных корпусах достаточно мала относительно всего конгломерата корпуса (очень часто это национальные корпуса определенного языка, например: русского языка [17], английского языка [18]).

2. Одноязычные – двуязычные / полиязычные.

Выделяются две группы одноязычных корпусов:

– корпуса, охватывающие весь язык,

– корпуса, охватывающие только язык для определённых целей.

К примеру, Corpus of Early English Medical Writing (CEEM) [19] – корпус текстов с медицинским содержанием на английском языке от 1375 до 1750 гг., объем которого составляет примерно 1,5 миллиона слов. В нём имеются теоретические работы, справочники, стихотворные тексты на медицинские темы.

В двуязычных и полиязычных корпусах тексты могут быть предложены сопоставимо или параллельно. Например, в 1992 г. возникла Европейская корпусная инициатива (European Corpus Initiative (ECI)) – международная организация, которая занимается составлением большого полиязычного корпуса для научно-исследовательских целей [20]. В настоящем сопоставимом корпусе имеются не только тексты европейских языков, но и тексты на русском, турецком, китайском, японском и многих других. Их объём составляет более 98 миллионов слов. Такого вида корпус считается коммерческим. Корпуса параллельных текстов предназначены, прежде всего, для сопоставительного анализа текстов в направлении “оригинальный – переведённый” для обучения способам, методам и приемам перевода. Например, European Parliament Proceedings Parallel Corpus 1996-2011 [21], в котором представлены параллельные тексты заседания Европейского парламента на разных европейских языках с переводом на английский.

3. Синхронный – диахронный.

Синхронные корпуса предусматривают репрезентацию текстовых данных для исследования системного состояния языка в конкретный период времени. Так, в некоммерческой версии Британского национального корпуса [22] есть только тексты периода с 1980 по 1993 гг.

Для исследования исторического развития определённого языкового явления или всей языковой системы в общем существуют диахронные корпуса. К примеру, Thesaurus Indogermanischer Text- und Sprachmaterialien [23], здесь представлены индогерманские тексты разных эпох.

4. Неразмеченные – размеченные.

Неразмеченный корпус – это конгломерат текстов, содержащий конкретное число упоминаний необходимого компонента. Вместе с тем итоги поиска, предлагаемые в неразмеченных корпусах, могут быть применены в языковедческих исследованиях, но только с точки зрения статистики.

Размеченные корпуса (морфологически, синтаксически и т.д.) считаются многофункциональными, так как дают намного больше возможностей для осуществления лингвистического анализа.

Заключение. Итак, корпус – это представленный в электронном виде, как правило, размеченный для языковедческого анализа, снабжённый относительно несложной в эксплуатации поисковой системой представительный конгломерат неотредактированных текстов, которые представляют по возможности большое число языковых вариантов.

В годы возникновения корпусной лингвистики проблем компьютеризации в этом направлении не обозначалось, и “исследователи указывали на возможность” проигнорировать вариативность языка, а именно “территориальному, социальному, возрастному, гендерному” и т.п. языковому разграничению [24, 76-77]. В наши дни, игнорируя его, мы намеренно ущемляем себя разнообразными рамками при исследовании текстов конкретного языка, что подвергает сомнению объективность такого рода исследования. С возникновением электронных корпусов разнообразие форм существования языка стало более показательным, средства и возможности изучения данных языка возросли. В современном лингвистическом корпусе есть сотни миллионов словоупотреблений, а то, что благодаря электронному корпусу итоги примеров словоупотреблений можно получить невообразимо быстро, в значительной мере облегчает задачу языковедам. Показанная типология корпусов, не претендуя на масштабность, представляет нам реальное разнообразие корпусов текстов и позволяет сориентироваться в нем для дальнейшего проведения научного исследования.

ЛИТЕРАТУРЫ:

1. Мельников Г.П. Системная типология языков: принципы, методы, модели / РАН. Ин-т языкознания. – М.: Наука, 2003. – С. 29.

2. Lemnitzer L., Zinsmeister H. Korpuslinguistik: Eine Einführung. – Tübingen, 2006. – 272 с.

3. Ахмеджанова С. Д. Русский и узбекский языки в сопоставительном аспекте //Ответственный редактор, 2023. – С. 37.
4. Sinclair J. McH. Corpus, Concordance, Collocation. Describing English language. – Oxford: Oxford University Press. – 1991. – 178 c.
5. Stubbs M. Words and phrases: corpus studies of lexical semantics. – Oxford, 2001. – P. 239-240.
6. Захаров В. П. Корпусная лингвистика. – СПб., 2005. – 334 с.
7. Ахмеджанова, С. (2023). Специфика лакунарности в русской лингвистике. Евразийский журнал социальных наук, философии и культуры, 3(3), 84-88. извлечено от https://in-academy.uz/index.php/ejsspc/article/view/11205.
8. Кибрик А.Е., Брыкина М.М., Леонтьев А.П., Хитров А.Н. Русские посессивные конструкции в свете корпусно-статистического исследования // Вопросы языкознания, 2006. Вып. 1. – С. 21.
9. Кривнова О. Ф. Области применения речевых корпусов и опыт их разработки // Тр. XVIII Сессии Российского акустического общества РАО. – Таганрог, 2006. – С. 23-34.
10. Плунгян В. А. “Интегрум» и Национальный корпус русского языка в лингвистических исследованиях // Integrum: точные методы и гуманитарные науки. – М., 2006. – С. 76-77.
11. http://www.ldc.upenn.edu
12. http://www.cslu.ogi.edu
13. http://www.elra.info
14. http://www.ids-mannheim.de
15. http://www.ruscorpora.ru
16. http://corpus.byu.edu/bnc
17. http://www.helsinki.fi/varieng/CoRD/corpora/CEEM/index.html
18. http://www.elsnet.org/resources/eciCorpus.html
19. http://www.statmt.org/europarl
20. http://corpus.byu.edu/bnc
21. http://titus.uni-frankfurt.de
22. http://corpus.byu.edu/bnc/
23. http://corpus.leeds.ac.uk/protected/
24. http://www.collinslanguage.com/content-solutions/wordbanks