Проблеми на изграждането на електронен корпус за лексикографски цели

Диана Благоева
Институт за български език „Проф. Л. Андрейчин“ – БАН

Проблемы создания электронного корпуса для лексикографических целей
В статье рассматриваются теоретические и практические проблемы, связанные с проектированием и созданием профилированного электронного корпуса болгарского языка для лексикографических целей. Обсуждается необходимость создания такого корпуса, представляются результаты и перспективы работы над корпусом. Представлены типологическая характеристика корпуса, объем, структура и состав корпуса, а также и уровни лингвистической обработки корпуса. Намечаются основные функции программного обеспечения для работы с корпусом.

Динамичното развитие на корпусната лингвистика през последните няколко десетилетия води до съществено изменение на изследователските практики и подходи в редица области на лингвистичната наука (Филмор 1992), в това число и в лексикографията. Използването на различни електронни ресурси, включително и корпуси, днес е стандарт в лексикографската практика. Добре проектираният и снабден с адекватни софтуерни средства електронен корпус е богата и лесно достъпна емпирична база за извличане на лексикографски релевантна информация и верифициране на основаващите се на езиковата интуиция лексикографски решения. Но корпусът е не само полезен инструмент в работата на лексикографите. В съвременния период той нерядко се превръща в интегрална част на лексикографския продукт (Захаров 2005: 226), срв. например корпусно базираните електронни речници. Може да се каже, че корпусният подход налага промени в традиционните лексикографски методи и познатите речникови модели.

Въпреки че за лексикографската работа успешно може се да използват съществуващите за съответните езици национални или други представителни корпуси с общо предназначение, както и някои специални (диахронни, диалектни, терминологични и пр.) корпуси, честа практика е създаването (обикновено в рамките на големи издателски или научни центрове) на профилирани лексикографски електронни корпуси, срв. например Longman Lancaster Corpus, Longman Learner's Corpus, Oxford English Dictionary corpus, корпусът на полското издателство PWN и пр. Това е така, защото спецификата на лексикографската дейност в повечето случаи налага конкретни изисквания към параметрите на използваните корпуси, напр.

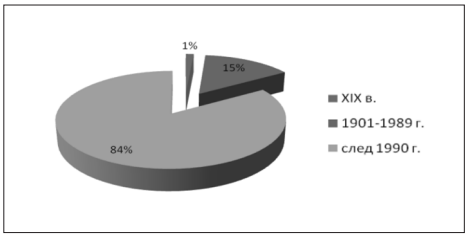
специфична хронологична рамка на включваните текстове, определен стилов, жанров и тематичен диапазон на текстовете, специален вид метаданни и тагиране и пр., вж. по-подробно Захаров 2005; Благоева, Колковска 2006 и др.

От няколко години в Института за български език се работи върху изграждането на профилиран корпус за целите на българската академична лексикография[1]. Неговото създаване беше наложително с оглед на това, че за разлика от повечето славянски езици българският език все още не разполага със свой национален електронен корпус, който да може да бъде използван за лексикографски цели. Тук ще бъдат представени накратко резултатите от досегашната работа върху корпуса с лексикографско предназначение и перспективите пред неговото развитие.

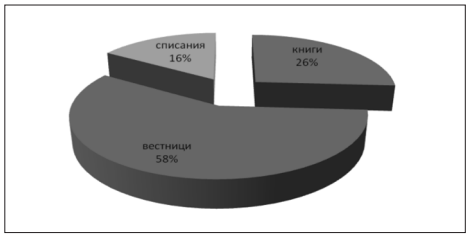
Изграждащият корпус е замислен преди всичко като средство за осъвременяване и обогатяване на емпиричната база и усъвършенстване на лексикографските методи при съставянето на многотомния академичен „Речник на българския език“ – проект с национална значимост, насочен към лексикографско описание на българския език от втората четвърт на XIX в. до наши дни (Благоева, Колковска 2007). От това се обуславят поважните типологични и структурни характеристики на корпуса, а също и особеностите на неговия състав и обем.

Поради широките хронологични граници на лексиката, отразявана в Речника, се налага корпусът да обхваща текстови материали от три столетия. Ето защо от гледище на разделението синхронност – диахронност корпусът има хетерогенен характер. На сегашния етап са оформени три субкорпуса, които включват съответно: текстове от XIX в., текстове от началото на XX в. до 1989 г., текстове от 1990 г. насам (вж. Таблица 1 и Фигура 1). Чрез подялбата на субкорпуси се дава възможност за наблюдение и описание на лексиката и за откриване на лексикалните иновации в различни синхронни срезове. Например субкорпусът с текстови материали от края на XX и началото на XXI в. е създаден специално за изследване на засилените неологизационни процеси в този период и съставяне на неологичен речник. За решаването на такива конкретни лингвистични задачи при необходимост може да бъдат обособявани и други подразделения. По технически причини материалите от XIX в. и първата половина на XX в. се включват в нормализиран вид (което е в съответствие с приетата в Речника практика за нормализиране на привеждания илюстративен материал от източници със стара графика). Необходимостта от предварителна обработка затруднява набирането на текстове от посочения период, поради което техният брой засега е сравнително малък.

[1] Работата върху корпуса е финансирана от Фонд „Научни изследвания“ при МОН (проект ОХН1512/2005).



Фигура 1. Процентно отношение на текстовете в корпуса по хронологичен признак



Фигура 2. Процентно отношение на броя на текстовете в корпуса според източниците

Таблица 1. Разпределение на текстовете в корпуса по хронологичен признак

период	брой електронни документи	брой думи (в милиони)
XIX в.	60	2,5
1901–1989 г.	695	29
след 1990 г.	3937	80,5
общо	4692	112

Изгражда се корпус на писмения език. На сегашния етап се включват само електронни версии на печатни текстове – книги, вестници, списания (вж. Таблица 2 и Фигура 2). По-късно ще се пристъпи към оформяне на субкорпус с непечатни писмени текстове, извлечени от интернет страници, електронни писма, постинги и пр., които са източник на ценни сведения за състоянието на съвременната българска лексика. Не се предвижда обособяване на подкорпус с устна реч (най-вече поради затруднения от технически характер).

Таблица 2. Разпределение на текстовете в корпуса според източниците

източник	брой електронни документи	брой думи (в милиони)
книги	1243	56
вестници	2713	48,5
списания	736	7,5

Корпусът е от типа отворен, динамичен корпус. Планираният в началото минимален обем от 100 милиона думи вече е достигнат (вж. Таблица 3), но попълването на отделните субкорпуси продължава. Привличането на нови текстове в корпуса е от особена важност за лексикографската работа, тъй като осигурява възможност за по-широки наблюдения върху функционирането на лексикалните единици и проследяване на промените в лексикалната система във всеки един момент от нейното развитие, което при корпус от затворен тип е трудно постижимо. Необходимостта от използването в лексикографията на възможно най-големи по обем електронни

корпуси, отразяващи в достатъчна степен разнообразието от езикови регистри и езикови средства (включително и такива с рядка употреба), е подчертавана нееднократно в изследванията по корпусна лингвистика (Байбър 1990, Байбър 1993, Синклер 1991, Шимкова 2005 и др.).

Таблица 3. Обем на корпуса

общ брой токъни	112 052 456
общ брой типове	1 206 552
процентно отношение типове : токъни	1,07

Един от най-важните проблеми, възникващи при планирането и изграждането на електронни корпуси от различен тип, се отнася до репрезентативността на съответния корпус и критериите за подбор на включваните в него текстове. Въпреки широките научни дискусии в тази област не може да се каже, че съществува напълно задоволително решение и общоприета методика за класификация на текстовете и определяне на критериите за подбор. Когато става дума за представителен (национален) корпус, съществуват различни подходи: от пренебрегване на свойството представителност (Bank of English, Mannheimer Corpora) до установяване на детайлно процентно съотношение между отделните типове (стилове и жанрове) текстове въз основа на периодични социологически проучвания (Český národní korpus), вж. по-подробно Синклер 1993, Байбър 1993, Шимкова 2005, Чермак и др. 2006 и др. Важно е също така дали принципът за репрезентативност се отнася до отразяването в корпуса на всички съществуващи типове текстове или до разпределението на езиковите средства в съществуващите текстове (Байбър 1993, Шимкова 2005).

При лексикографските корпуси проблемът за репрезентативността зависи преди всичко от предназначението и степента на профилираност на конкретния корпус, тъй като различните типове и жанрове речници се нуждаят от различна по характер емпирична база. За целите на голям тълковен речник от типа на Речник на българския език е необходим голям по обем емпиричен материал, извлечен от максимално широк кръг стилове и жанрове, включително жанрове, които се смятат за силно маркирани и поради това неподходящи за включване в представителен референтен корпус (например драма и поезия). Съобразно с концепцията на Речника, която изисква регистриране не само на книжовна, но и на субстандартна (например общодиалектна) лексика, в корпуса се привличат и текстове, съдържащи субстандартни лексикални елементи, напр. публикувани фолклорни произведения.

Целесъобразно е в лексикографския корпус да бъдат включвани по възможност цели текстове, а не само извадки (samples) с определен брой

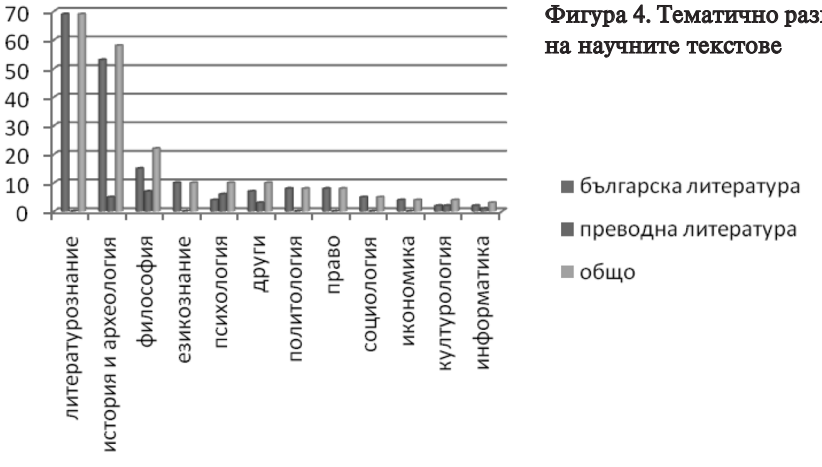
думи, каквато е практиката при референтни корпуси от рода на Brown corpus (за българския Браун корпус, следващ същия модел, вж. Стоянова и др. 2006). Основен довод в полза на привличането на цели текстове е това, че малко езикови явления имат равномерно разпределение в целия текст и поради това не може да се смята, че една или друга извадка ще има достатъчно представителен характер (Синклер 1991: 19). Това в пълна степен важи и по отношение на употребата на лексикалните средства на езика.

Съгласно с разпространената в литературата по корпусна лингвистика типологична стратификация на текстовете се разграничават два основни типа текстове: информативни (публицистика, научна литература, научно-популярна литература и пр.) и художествени (прозаични, поетични, драматични произведения, фолклорни произведения и пр.). Разпределението на текстовете в изграждания корпус по тип е представено в Таблица 4 и Фигура 3:

Таблица 4. Разпределение на текстовете в корпуса по тип

	брой електронни документи	**брой думи (в милиони)**
информативни текстове	3766	72
художествени текстове	926	40

Фигура 3. Процентно отношение на броя на текстовете в корпуса според типа им



Фигура 4. Тематично разпределение на научните текстове

Жанровото и тематичното разпределение на информативните и художествените текстове е представено по-долу в Таблица 5, Таблица 6, Таблица 7 и Фигура 4:

Таблица 5. Тематично разпределение на информативните текстове (периодични издания)

тематична област	брой електронни документи
общество и политика	1792
икономика	596
изкуство и култура	338
военно дело	139
език и литература	73
здраве и медицина	55
компютри	55
лов и риболов	35
педагогика	25
музика	16
стил на живот	16
философия	12
спорт	7
история и археология	5
право	3
социология	2
други	280

Таблица 6. Жанрово-тематично разпределение на информативните текстове (литература)

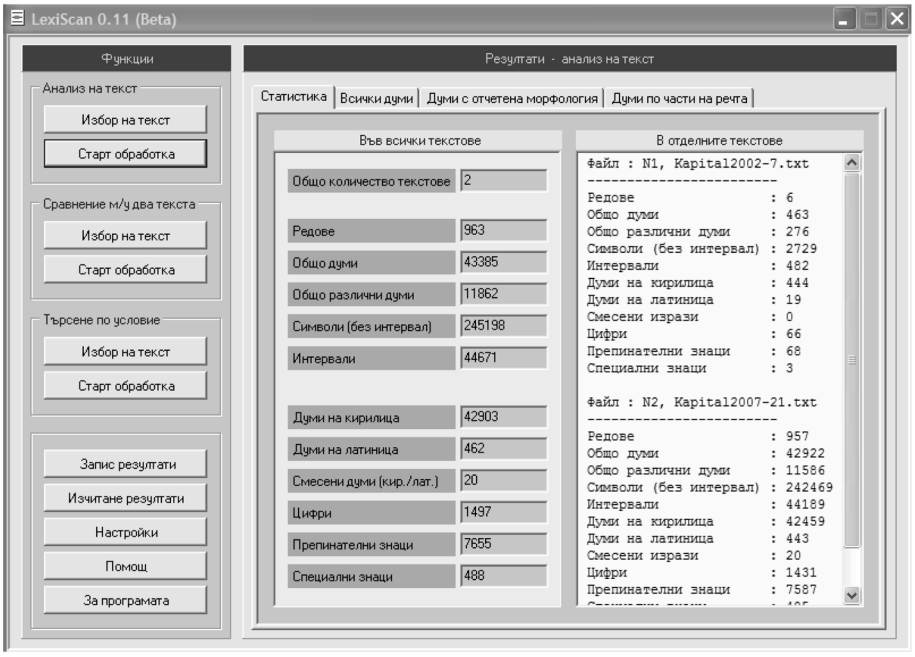
жанр / тематична област	Българска литература (брой електронни документи)	Преводна литература (брой електронни документи)	Общ брой
наука	178	24	202
документална и биографична литература	20	15	35
мемоарна литература	26	5	31
религия	5	6	11
научно-популярна литература	5	2	7
краезнание	5	5	
здраве	4	4	
кулинария	2	1	3
публицистика	3	3	
други	4	2	6

Таблица 7. Жанрово-тематично разпределение на художествените текстове

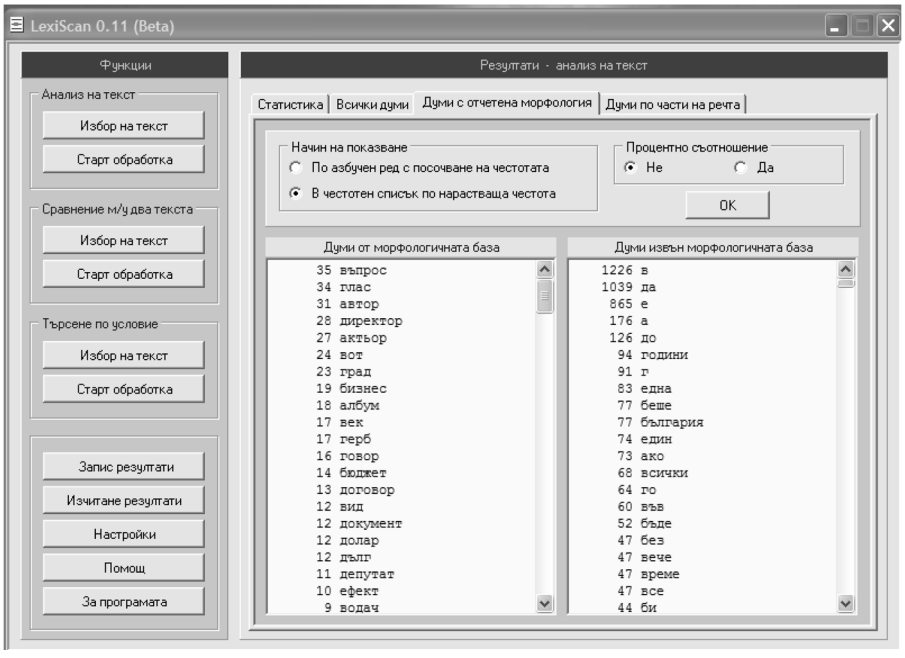
жанр / тематична област	Българска литература (брой електронни документи)	Преводна литература (брой електронни документи)	Общ брой
романи и повести			
исторически	13	10	23
криминални	16	31	47
приключенски	1	38	39
фантастични	13	43	56
фентъзи		17	17
любовни		2	2
други	219	76	295
разкази	7	8	15
поезия	245	4	249
фолклорни произведения			
народни песни	55		55
народни приказки	2		2
пословици и поговорки	1		1
драматични произведения	12	2	14
литература за деца			
проза	40	21	61
поезия	23		23
други	52	8	60

Предстои работа за постигане на балансираност на корпуса, разбирана по Кенеди 1998: 62 като пропорционално отношение между отделните части и елементи на корпуса.

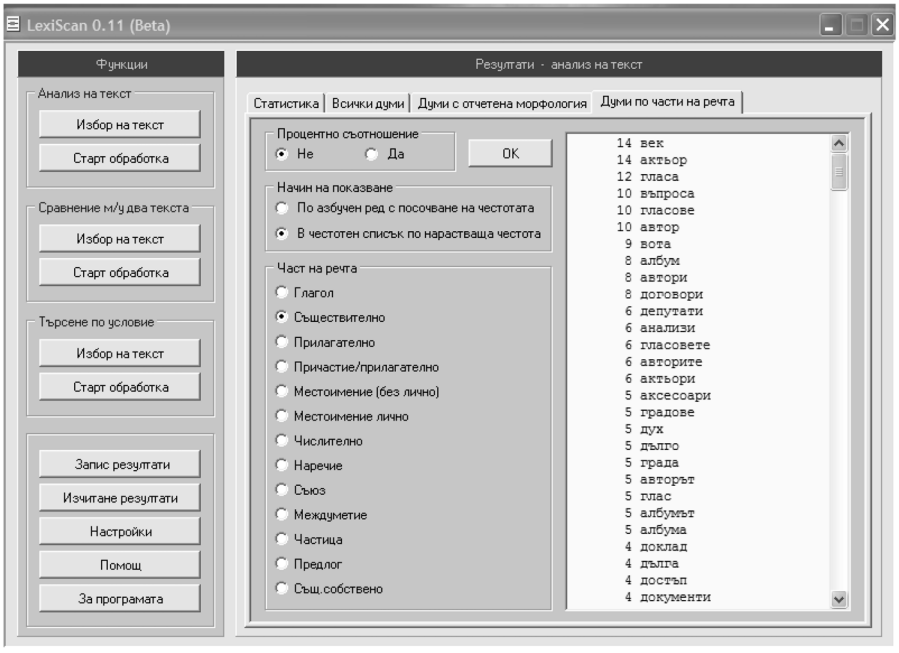
На сегашния етап лингвистичната обработка на корпуса се ограничава с токънизация и лематизация. Избран е релационен (data-driven) и контекстно свободен подход (Михайлов 2002), при който лематизацията се извършва чрез морфологична база данни, интегрирана в пакета от софтуерни средства за работа с корпуса. Базата съдържа около 500 морфологични таблици и списъци с над 70 000 основи, с помощта на които се осъществява автоматично генериране на повече от 700 000 словоформи (по този начин се покрива ядрената част на съвременната българска лексика). Включен е и списък с около 15 000 собствени имена. За всяка от лемите се генерира и при поискване се извежда словоизменителна таблица, която може да бъде редактирана и допълвана (например с остарели или диалектни форми, каквито поради хетерогенния характер на корпуса присъстват широко в някои типове текстове и е необходимо да бъдат разпознавани и лематизирани). Таблицата съдържа и допълнителни полета, в които се дават



Фигура 5. Статистически анализ на текстове с *LexiStat*



Фигура 6. Честотен списък на думи, генериран с *LexiStat*



Фигура 7. Честотен списък на думи от зададена част на речта, генериран с *LexiStat*

препратки към словообразователни или графични варианти на лемата (напр. *препускам* и *препущам*, *приветствам* и *приветствувам*), които според концепцията на Речника се представят в обща заглавка, или към нейни остарели, диалектни или други форми, които се посочват в справочен отдел в края на речниковата статия (напр. *подкреплявам* към *подкрепявам*). Така се осигурява възможност при търсене в корпуса с една заявка да се извеждат и контексти с варианти на лемата или с нейни некнижовни форми, което улеснява лексикографа.

Пакетът от софтуерни средства за работа с корпуса *LexiScan* съдържа два базови модула: програма за статистическа обработка на текстове *LexiStat* и програма за пълнотекстово търсене *LexiSearch*. На сегашният етап е разработен първият модул, който има следните функции: статистически анализ на текст или масив от текстове (вж. Фигура 5), установяване на честотата на срещанията на зададени думи и словоформи, генериране на честотни списъци по нарастваща или намаляваща честота (вж. Фигура 6 и 7), сравняване на словния състав на текстове с отчитане на изключенията или пресичанията.

Азбучните честотни списъци се използват за допълване и обновяване на словника на Речника и за изготвяне на словници на други речници.

Функцията за сравняване на словния състав на текстове е насочена към установяване на липсващи в словника единици. Чрез нея новодобавяните текстове се сравняват с вече изготвения списък на всички словоформи в корпуса и се генерира списък на думи и форми, отсъстващи в корпуса. С помощта на тази функция може да се извършва и автоматично откриване на кандидати за лексикални неологизми, което е от значение при съставяне на словник за неологични речници.

Чрез модула *LexiSearch*, който е в процес на разработка, ще се осъществява пълнотекстово търсене на думи, словоформи или фрази, както и сортиране на откритите срещания на съответната единица според граматичните характеристики (част на речта) на думите в нейния ляв или десен контекст. Така лексикографите ще разполагат със систематизирана информация за употребата и съчетаемостта на търсената дума, което значително ще улесни работата им.

На следващ етап от работата върху корпуса се предвижда той да бъде морфологично и синтактично анотиран. Наистина според един от основоположниците на компютърната лексикография – Дж. Синклер (Синклер 1992: 385–386), анотирането на корпусите води до загуба на информация (най-вече поради факта, че нерядко се предпоставят решения на базата на синтактични или други теории с дискусионен характер), поради което за лексикографски цели най-подходящи са чистите (неанотирани) корпуси. Колкото по-високо е нивото на лингвистична обработка на корпусите[2] обаче, толкова по-ефективно може да се използват мощните съвременни инструменти за извличане на лексикографски релевантна информация от типа „lexical profiling software“ (Килгариф и др. 2002).

Създаваният за целите на българската академична лексикография електронен корпус е надежден и полезен инструмент в работата по съставяне, редактиране и преработка на различни видове речници на българския език.

Цитирана литература

БАЙБЪР 1990: D. Biber. Methodological issues regarding corpus-based analyses of linguistic variation. Literary and Linguistic Computing 5, 1990, pp. 257–269.

БАЙБЪР 1993: D. Biber. Representativeness in corpus design. Literary and Linguistic Computing 8, 1993, pp. 243–257.

БЛАГОЕВА, КОЛКОВСКА 2006: Д. Благоева, С. Колковска. Някои проблеми на корпусната лексикография. В: Светът на речника. Светът в речника. Юбилеен сборник,

[2] За съществуващите програмни средства за обработка на корпуси на български език вж. Коева и др. 2006 и цитираната там литература.

посветен на 70-годишнината на чл.-кор. д.ф.н. Емилия Пернишка. Велико Търново, 2006, с. 113–121.

БЛАГОЕВА, КОЛКОВСКА 2007: Д. Благоева, С. Колковска. Електронен корпус за целите на „Речник за българския език“ – състояние и перспективи. В: Лексикографията и лексикологията в съвременния свят. Велико Търново, 2007, с. 277–286.

ЗАХАРОВ 2005: V. Zakharov. Russian Historical Corpora of the 18th and 19th Centuries. – Computer Treatment of Slavic and East European Languages. Bratislava, 2005, pp. 220–228.

КЕНЕДИ 1998: G. Kennedy. An Introduction to Corpus Linguistics. London, 1998.

КИЛГАРИФ И ДР. 2002: A. Kilgarriff, M. Rundell. Lexical Profiling Software and Its Lexicographic Applications: a Case Study. In: Proceedings of the 10th Euralex International Congress. Copenhagen, 2002, pp. 807–818.

КОЕВА И ДР. 2006: Sv. Koeva, Sv. Leseva, I. Stoyanova, E. Tarpomanova, M. Todorova. Bulgarian Tagged Corpora. In: Proceedings of the Fifth International Conference Formal Approaches to South Slavic and Balkan Languages. Sofia, 2006, pp. 78–86.

МИХАЙЛОВ 2002: М. Н. Михайлов. Контекстно-свободная лемматизация как временное решение насущных проблем. В: АЛФАВИТ: филологический сборник. Смоленск, 2002. http://www.uta.fi/~mihail.mihailov/research/morfology.htm

СИНКЛЕР 1991: J. M. Sinclair. Corpus, Concordance, Collocation. Oxford, 1991.

СИНКЛЕР 1992: J. M. Sinclair. The Automatic Analysis of Corpora. In: J. Svartvik (ed.). Directions in Corpus Linguistics. Berlin and New York, 1992, pp. 379–397.

СИНКЛЕР 1995: J. M. Sinclair. Corpus typology: a framework for classification. In: G. Melchers, B. Warren (eds.). Studies in Anglistics. Stockholm, 1995, pp. 17–33.

СТОЯНОВА И ДР. 2006: I. Stoyanova, Sv. Koeva, Sv. Lesseva. Applying and analysing Brown corpus model for Bulgarian, Presentation at The Third Inter-Varietal Applied Corpus Studies (IVACS) group International Conference on „LANGUAGE AT THE INTERFACE“ 23rd-24th June 2006, Nottingham, UK. http://www.nottingham.ac.uk/english/IVACS/ivacs-stoyanova-koeva-lesseva.ppt

ФИЛМОР 1992: Ch. Fillmore. „Corpus Linguistics“ vs. „Computer Aided Armchair Linguistics“. In: J. Svartvik (ed.). Directions in Corpus Linguistics. Berlin and New York, 1992, pp. 35–60.

ЧЕРМАК И ДР. 2006: F. Čermák, M. Křen. Large Corpora, Lexical Frequencies and Coverage of Texts. In: Corpus Linguistics 2005 Vol. 1, No 1, (eds.) P. Danielsson, M. Wagenmakers, Proceedings from The Corpus Linguistics Conference Series (Birmingham, July 14–17). Birmingham, 2006. http://www.corpus.bham.ac.uk/PCLC.

ШИМКОВА 2005: М. Шимкова. Репрезентативность корпуса как лингвистическая проблема. В: MegaLing'2005. Прикладная лингвистика в поиске новых путей. Материалы международной конференции. Симферополь, 2005. http://korpus.juls.savba.sk/publications/block1/2005-simkova-reprezentativnost%20korpusa/2005-simkova-reprezentativnost%20korpusa.pdf