

Корпусите в преподаването на чужди езици

Елена Тарашева

Търсенето в специализираната научна секция на Гугъл на термина “Чуждоезиково обучение” дава 41 080 публикации – книги и статии, публикувани между 2000 и 2008 година. 1010 от тях, или 24%, се позовават на работа с корпуси. В авторитетния обзор на чуждоезиковото обучение М. Савил-Тройки (2006-138) пише:

Определянето на конкретните потребности от чужд език за всяка една група учещи предполага идентифициране на подгрупи от лингвистични компоненти, свързани с дадена разновидност (или регистър), специфични за съответните ситуации. През последните години, по-голямата част от тази работа се извършва с използването на компютъризирани анализи на “корпуси”, или големи колекции от писмени и устни текстове. Такива корпусни изследвания могат да бъдат изключително полезни при определянето на относителната честотност на различни речникови единици и граматични структури като основа за това, което следва да бъде преподадено.

Настоящата статия обобщава възможностите, които корпусната лингвистика предлага на чуждоезиковото обучение. Дадени са конкретни примери от публикации и учебни помагала. Проблематизират се различни аспекти от използването на корпуси.

Подбиране на езиков материал за дидактични материали

През 2004 година Джон Синклер (2004:9) пише, че лингвистиката се формира на основата на неодстатъчни свидетелства, като това не е критика, а констатация. Така той преформулира свое предишно изказване от 1985, когато с по-малка доза политическа коректност възкликва: “Преподаваме английски в пълно неведение за огромно количество езикови факти”. С това твърдение се съгласява и един от най-авторитетните му критици Х. Уидоусън (2003:77).

Действително, в последните години голям брой дидактически пособия са изградени на основата на материал от корпуси. Една от най-авторитетните граматики на английския език, (Куърк, Р. и др. 1985), се основава на милионен корпус, съставен наполовина от устна и писмена реч и изграден от записи на телефонни разговори. Обособяването му като корпус следва създаването на граматиката (Куърк 1968), защото тогава събирането на материала се е смятало като част от създаването на граматиката и за него не се е предвиждал живот, независим от този на граматиката.

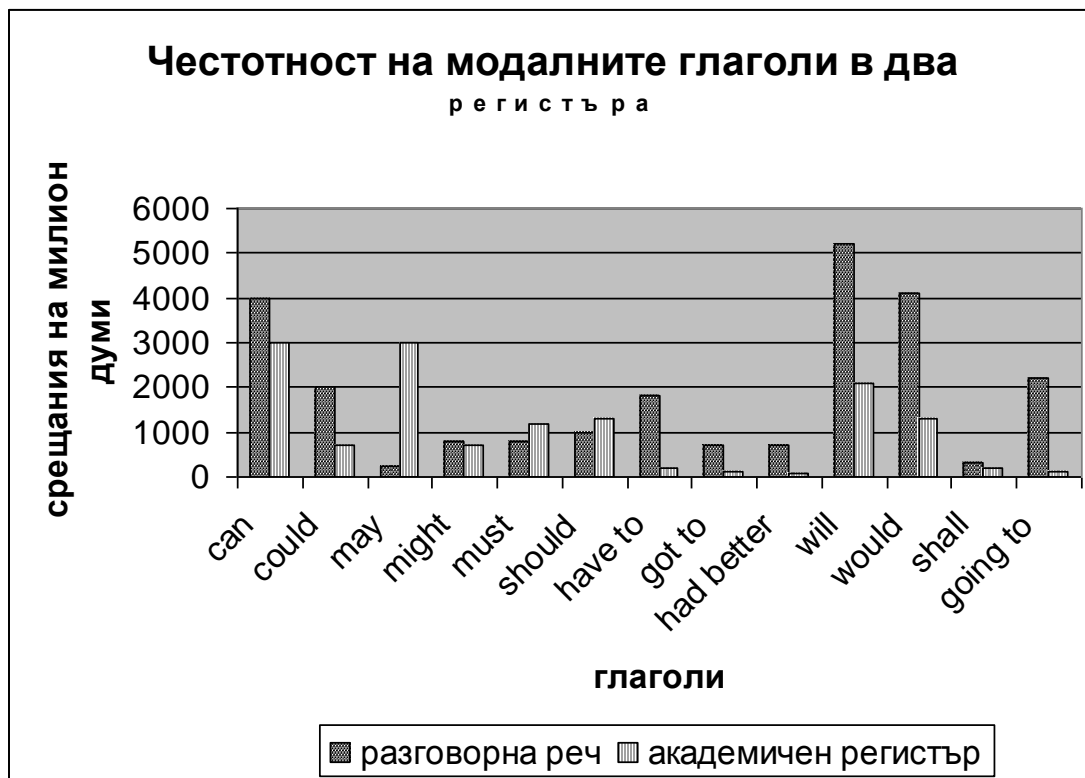
По-съвременен пример за дидактически пособия, основани на големи масиви езиков материал е проектът Кобилд, чиято продукция включва речници, граматики и множество помагала, познати и на нашия пазар. Наред с продуктите, които представят готови интерпретации на корпусния материал, съществува възможност за търсения в самия корпус. На адрес <http://www.collins.co.uk/Corpus/CorpusSearch.aspx> може да се направи справка,

например, дали е приемлива употребата *a quite* + прилагателно име, наред с *quite a* + прилагателно име, като се ползва специален език на търсенето по дадена част от речта и лексикална единица. Макар, че и двете колокации се срещат в корпуса, изглежда, че първата структура се свързва преобладаващо с прилагателното *different/различен*, докато втората – с квантификатора *few/доста*. Така корпусното търсене отговаря на отговора на учителя, че разликата между двете конструкции е в смисъла, а не в честотността или приемливостта. Този корпус е известен с името Банка на английския език (Хънстън 2002:16), съдържа писмена и устна реч от американски и британски английски, което дава възможност да се правят съпоставки. За разлика от други корпуси, той непрестанно се обновява, поради което се нарича мониторен и дава възможност за установяване на тенденции в развитието на езика.

Друг вид корпус представлява общият. Там броят на единиците е фиксиран, за да може да се пресъздаде реалното съотношение между говорна и писмена реч, между различните жанрове и участници в комуникативните актове в езика. Такива корпуси се създават или като референтни (виж ?? долу), или в опит да се съхрани и издигне авторитета на даден език. Пример за това е корпусът на шотландския език, достъпен на адрес <http://www.scottishcorpus.ac.uk/corpus/>. Той съдържа 4 милиона думи, 20 % от които са устна реч и присъстват като транскрибиран и аотиран текст, но също и като звукови и видеозаписи. Такъв род корпуси са предпочитани, тъй като съдържат информация за интонацията, жестовете и действията на участниците и така се описва по-пълно контекста на езикова употреба. Подобен проект съществува и за баския език <http://www.lrc.salemstate.edu/ask/basquecorpus/> и за други “малки езици”.

Третият вид корпуси – специализираните - събират материал от даден жанр, или регионална разновидност. Често срещана практика е на базата на общ корпус да се обособят специализирани части. Пример за това е корпусът Suzanne, <http://www.grsampson.net/RSue.html>, създаден от Сампсън в опит да опише граматичната система на английския език. Негов източник е корпусът Лондон-Осло-Берген, от който са избрани писмени текстове на американски английски. По-късно, от същия източник Сампсън екстрактира и един от най-популярните корпуси на говорната реч, наречен Christine и наличен на адрес <http://www.grsampson.net/RChristine.html>. Писмена реч на британски английски език е налична в третия корпус на Сампсън, LUCY, <http://www.grsampson.net/RLucy.html>. Така става ясно, например, че думата *but bag/чанта за носене през кръста* не се установява в американския корпус, за разлика от британския. За разлика от преобладаващите очаквания, честотността на произнасяне на съотносителния съюз *either/ нито* няма значими разлики в британския и американския английски и в двете регионални разновидности има двойно произношение, с почти идентична честотност на двете форми.

Един от водещите автори в областта на корпусната лингвистика, Дъглас Байбър препоръчва подборът на лексика и граматични структури за включване в учебни системи да става след анализ на честотността им. Освен това последователността на граматичните единици трябва да започва от най-високочестотните. Следната графика представя кои модални глаголи са най-високочестотни в разговорния английски език и в академичните текстове, според корпуса на Лонгман:



Както се вижда, учебниците по разговорен английски език могат да наблегнат на *will*, *would* и *can*, но помагалата за академичния регистър трябва да се фокусират върху *can* и *may*. Понататъчни различия възникват при анализиране на значенията на тези глаголи, както и в британски и американски английски. Затова Байбър препоръчва подборът и подреждането на материала да се подчини на изследване на конкретната честотност в разновидността на езика, която е от интерес за учещите.

Такъв род изследвания са многобройни. Лий (2006) установява, че учебниците в Хонг Конг настояват за предаване на нереалност да се използва формата на глагола *be/съм* за условно наклонение *were* вместо *was*, но Хонконгският корпус съдържа 1 такава употреба, а австралийският – 2. Това дава основание на авторката да критикува учебните системи в страната си, както и критериите за оценка на знанията по английски език.

Както признава един от съавторите на граматиката на Куърк и участник в изграждането на други 2 големи корпуса – Лондон-Линд и Международния корпус на английския език (ICE), Сидни Грийнбаум (1988:83), “основната функция на корпуса тогава (през 70-те години на ХХ в – Е.Т.) се състоеше в стимулиране на интуицията на лингвиста”. Това означава, че прескриптивното съдържание се диктува от усета на езиковеда за правилното, а корпусът служи за подбиране на подходящи примери. В това лингвистите днес виждат контраста между емпиричните и рационалистични подходи. Най-яркият представител на рационализма, Чомски, създава модел на езика, в който символно е кристализирала коректната форма. Емпириците, от своя страна, събират големи масиви езикова продукция, на базата на която описват езиковите явления. Но както емпириците не могат да правят обобщения без

абстрактните структури, създадени от рационалистите, така и рационалистите създават по-ефикасни езикови модели, като обобщават наблюденията на събрания от емпириците материал. Затова Манинг и Шутце (1999:4-7) говорят за сливане на двете течения в настоящето на корпусната лингвистика.

Макенери (2001:110) нарича търсенето на потвърждение за дадена хипотеза в корпусни данни *рационализъм-към-емпирицизъм*. Той описва опита на Сампсън да намери примери за формулите от генеративната граматика, като ръчно анализира именни фрази от корпуса Лондон-Осло-Берген. Заключение е било, че генеративната граматика не отразява същността на материала в корпуса, поради огромното разнообразие от фрази, генерирани от един модел. Същевременно Тейлър, Гровър и Бриско използвали автоматичен парсър за същия материал и установили, че той дава коректни резултати за близо 97% от именните фрази. Сега почти всички корпуси са снабдени с информация за това каква част на речта са думите, както и с най-разнообразни сведения за стилистичните, синтактичните и други функции на езика. Байбър (1998:10) предлага две насоки на използване на корпусната лингвистиката за дидактически цели – първата е да се започне с непредубедени описания на корпуси от езикова продукция и от анализа да изплуват рационалистични обобщения за езика (по-скоро – на конкретния регистър, както личи от понататъчните му публикации – Е.Т.). Втората насока е проверка на хипотези чрез корпусен материал, подобно на описаното от Макенери като рационализъм-към емпирицизъм.

Но каква информация може да поднесе една корпусно-базирана граматика, която не е налична в традиционните? Например, корпусно изследване (Хънстън и Франсис 2000:29) на фрази от типа

It + глагол връзка + прилагателно + подчинено изречение

показва, че те се срещат най-често с прилагателни, които означават някаква модалност, например *It is interesting that he comes late/ Интересно е, че все закъснява*. Така граматичната структура чрез наблюдения на високочестотните си появи в корпус от реална езикова продукция се свързва с определено значение. Уточнено е също кои подчинени допълнителни изречения се въвеждат с *that/че* и кои – с *to/да*. Има информация, че честотността на *It is surprising that / Изненадващо е, че* е по-голяма в отрицателната форма на глагола. Смята се, че корпусна информация от този род насочва към контекста на употреба, с което допринася за активизиране на познанията на учещите.

Много от корпусите са налични в световната мрежа. Един от уебсайтовете, подходящи за учещи обобщава структурите, в които се среща дадена дума и дава информация за най-честотното езиково обкръжение. Това е *Just the word* (<http://193.133.140.102/JustTheWord/>). Например, желаещият да установи с какви глаголи се среща името *conversation/разговор* установява следната картина (цифрите в скоби показват броя на срещания в корпуса):

engage in conversation (51)
have conversation (311)
have conversation about (11)

have conversation with (29)
hold conversation (67)
hold conversation with (7)
include in conversation (7)
involve in conversation (12)
keep conversation (38)
think about conversation (9)
think of conversation (8)

Най-често срещаният глагол е *have*, а предлозите *about* и *with* въвеждат темата на разговора или комуниканта. Учениците могат да видят примери за всяка структура:

YOU'RE having a conversation about traffic problems.

It was probably in 1937 that he and Tolkien had a conversation about their distaste for much of what was being published at that time.

They had the same conversation the previous Friday afternoon, about the disappointing figures for Pringle's production in December.

So far as I remember, I had no conversation about me.

По същия начин може да се получи информация за прилагателните, с които най-често се среща името, както и за другите части на речта, с които се наблюдават високочестотни колокации. Те са обобщени според частите на речта, които участват по следния начин:

V obj N*, e.g. have conversation (311)
N* subj V, e.g. conversation take (40)
N* subj ADJ, e.g. conversation easy (10)
N* subj PREP, e.g. conversation about (8)
ADJ N*, e.g. private conversation (64)
NN*, e.g. telephone conversation (184)
N* PREP, e.g. conversation with (735)
N* V, e.g. conversation going (22)
N* N, e.g. conversation piece (13)
PREP N*, e.g. in conversation (855)
N PREP N*, e.g. topic of conversation (88)
ADJ PREP N*, e.g. deep in conversation (46)
article N*, e.g. our conversation (140)

Хънстън (2000:) предлага два дидактични метода, които директно използват корпуси. Първият е наречен взаимно обучение. При него студенти от различна националност се учат един друг на родния си език, като подкрепят интуицията си с данни от корпуси. Вторият – и по-интересен за българската действителност – се нарича учене, водено от данни. При него на учещите се задава работа върху корпус с цел изпълнение на дадена задача. Хънстън дава пример с работата си над есе на студент, в което той употребява израза *Under the influence of*

Halliday/под влияние на Халидей. Интуицията на преподавателката я кара да мисли, че този израз звучи странно, тъй като на английски, както и на български, преобладаващата употреба на *под влияние на* се свързва с негативния ефект, например от алкохола. Ученият и преподавателката проверяват в Британския национален корпус и установяват, че преподавателката е права – *influence/влияние* се употребява предимно в смисъл на негативно влияние. Малкият брой на срещания с неутрално и положително значение се обясняват с различни социолингвистични фактори. Така учещите се поставят в контакт с реална речева продукция и извършват свои открития с помощта на преподавателите си. Като “баща” на ученето, водено от корпусни данни се сочи Т. Джонс (1991). Според него основните дидактични стъпки са идентифициране – класифициране – обобщаване.

Ученето чрез използване на корпуси се утвърждава, най-вече като метод, който повежда учещия да открие езика като следва стъпките на учения, както твърди Лийч (1997:10-11):

“Критическото и аргументативното есе трябва да бъдат балансирани със задание, което подканва учещия да добие, организира и проучи данни от реална езикова продукция по свой избор. Вторият тип задание дава на учещия реално усещане, че прокарва пътя си на ‘изследовател’, като усеща свой личен принос, а не – че върви по стъпките на друг.”

Използването на корпуси стимулира учещия да поеме отговорност за процеса на учене, да бъде самостоятелен и креативен, а ролята на учителя става още по-творческа и отговорна. Работата с корпуси е евристична и потопява в реалния свят на научното многообразие и противоречия. Задачите Лийч класифицира като конвергентни и дивергентни. Пример за първия вид е описан от Бернадини (2004:26): италиански студенти по превод търсят значението на думата *staunch/мвърд* и достигат до единодушното решение, че думата се употребява само с идеологии, вярвания и т.н. Като дивергентна задача Бернадини описва търсенето на отговор на въпроса има ли разлика в употребата на *competence/компетенция* в единствено и множествено число. Корпусът не дава еднозначен отговор на така формулираното търсене, но учещата среща разнообразен речник, формулира и проверява хипотези, с което действително разширява знанията си.

Както се вижда, граматичната и лексикалната информация се сливат. Една от централните фигури в корпусната лингвистика, Джон Синклер в съавторство с Картър (2004:165) развива основната си теза, че разделянето на граматика и лексика е плод на остарелите методи на лингвистиката от периода преди възникване на корпуси. Структурата на езика, според авторите е предимно граматична, но със задължителен лексикален компонент, а разглеждането да компонентите поотделно не позволява на езиковеда да види важни зависимости. Затова ключово понятие в корпусната лингвистика е мотив (*pattern*). Хънстън (2000:37) го дефинира като “думи и структури, асоциирани с дадена дума, които допринасят към значението и”. Мотивите се срещат често в прекъснатата или непрекъснатата група и поддържат едно основно семантично ядро.

Корпусите дават възможност за най-пълноценно обхващане на факторите в дискурса – участници, роли, обстановка, канал на общуване. Една от особеностите на различните професионални общности се състои в дискурса, който използват. Според Суейлс (1990:26) “една дискурсна общност развива своя специфична лексика”, както и свои специфични жанрове и граматични структури. Затова специализирани корпуси от продукцията на определени общности – академични есета, медицински диалози, статии по инженерни науки и т.н. са изходна база за учебници по специализиран език. Механичната представа за тези корпуси е като източник на текстове за учебни помагала. Но по-важно е да се погледне на методологичната страна – а именно, кои единици са количествено по-забележими в съпоставка с общите корпуси, представителни за езика, които споменахме по-горе като референтни. Затова повечето изследователи съпоставят данните си с нормативен корпус – за английски език като такъв се смята Британският национален корпус (БНК) (достъпен през своя уебсайт www.natcorp.ox.ac.uk).

Например, Бонди (2001:135) събира малък корпус от уводни уроци в учебници по икономика и статии в специализирани журнари (около 60 000 думи) и проследява честотността и функционирането на лексикални единици. За *show/показвам* установява честотност от 0.13 %, но съпоставката с БНК показва голяма разлика в полза на срещанията в специализирания корпус 52.7, изчислена чрез съпоставка на честотностите в двата корпуса с помощта на един от най-популярните инструменти за обработка на електронни корпуси Уърдсмит (??). Функциите на глагола също се различават от описаните за БНК – в икономическите текстове той служи за подчертаване на информация (рематизация на следващата информация). Тези и подобни заключения, резултат от корпусни операции и анализ на дискурса, помагат на Бонди да създаде дидактични материали за обучение на италиански икономисти в писане на икономически текстове на английски език.

Разбира се, създаването на корпус не може да бъде хаотично. Трябва съставителят да се постарее да постигне представителност за даден жанр. Сред “малките мръсни” начини за създаване на педагогически корпус Трибъл (2001:381) посочва тристепенна процедура. Първо се установяват текстове, които имат общи функционални характеристики, т.е. изпълняват една и съща комуникативна функция: в неговия пример – информират за определен вид курсове. Текстовете се подлагат на структурен анализ с цел потвърждаване на хипотезата, че всички те представляват еднотипни текстове – с еднаква цел и структурни части. След това се аотират поне по два показателя: части на речта и структурна организация. С това корпусът е готов за употреба, като търсене може да се извършва както на лексикална единица, така и по част на речта, структура или функционална принадлежност. Отдава се голямо значение на факта, че някои структури са типични за уводните части, а други – за заключителните (Майк Скот, Хоуи??). Такава информация е важна особено за обучение по писане. Според Трибъл, единственият начин учещите да бъдат научени да пишат в определен жанр е да се срещнат с максимален брой текстове от съответния жанр – което се постига със специалните корпуси. Създаване на корпуси по автоматичен начин може да стане и чрез уебсайта CLAWS.

Корпусите предлагат възможност за учене не само на езикови, но и на културни характеристики на текстовете. Например, студентите, на които предстои академичен обмен, могат да научат повече за академичната структура и практики в американските университети от специализирания корпус на американски академичен жанр Mi-Case. Представянето на корпуса дава възможност за търсене на текстове по академична позиция, роден език и степен на включеност на участниците в общуването, както и по жанр, факултет, тематика и интерактивност. Така може да се намери текст на академична консултация между афроамериканец и арабин, единият – преподавател, другият – студент, където водещата роля има студентът, на тема отлагане на домашна работа. Естествено, като всеки корпус, възможни са търсения на конкретни думи и фрази, или на части на речта, тъй като корпусът е анотиран.

Корпусите не винаги съдържат продукцията на носители на речта. Съставени са и корпуси от продукцията на учещи езика. Тяхната роля в изучаването на междинния език нараства, както става ясно от Sage???. Танко??? анализира употребата на конектори в писмени работи на напреднали по английски език унгарски студенти. В съпоставка с писмени работи на английски студенти, количеството и същността на конекторите очертава различни тенденции в двете групи изследвани обекти, което дава основания за изводи по повод на обучението на унгарските студенти. Корпусът CHILDS ??? дава възможност за изследване не само на писмената, но и на устната продукция. В него са събрани записи на френски, холандски и норвежки ученици, изучаващи английски език. Публикувани са изследвания на паузите, интонационните контури, реализацията на фонемите и т.н.

Корпуси са създадени и за целите на оценка на работата на преподавателите. Малък целеви корпус за обективизиране на наблюдения на уроци, интервюта с преподаватели и ученици, бележки на наблюдаващи търси тенденции в преподаването на френски език на английски ученици.

Паралелни корпуси

???Модерната корпусна лингвистика обаче предпочита обратното – да не се предсказват закономерности, а да се остави материалът да ги изведе на базата на изследваната езикова продукция. Макар проверката на хипотези да е допустим метод в корпусната лингвистика, непреднамереният анализ често води до изненадващи, но реалистични резултати.

Цитирани източници:

Байбър и други (1998) Douglas Biber, Susan Conrad, Randi Reppen *Corpus Linguistics: Investigating Language Structure and Use* CUP 1998

Бернардини (2004) S. Bernardini “Corpora in the Classroom” in Sinclair, J. (ed.) *How to use corpora in Second Language Teaching* John Benjamins Publishing 15 - 39

Бонди 2001 Bondi, Marina “Small Corpora and Language Variation” (135-174) in Mohsen Ghadessy, Alex Henry, Robert L. Roseberry (ed) *Small Corpus Study in ELT: Theory and Practice* Amsterdam: John Benjamins Publishing Company

Грийнбаум 1988 Greenbaum, Sydney *Guide to English Usage* Longman Harlow: 1988

- Джонс (1991) Tim Johns “Should you be persuaded: Two samples of data-driven learning materials” in: *English Language Research Journal* 1991- 4, pp1-16
- Куърк и други 1985 Quirk, R., S. Greenbaum, G. Leech, and J.Svartvik, *A Comprehensive Grammar of the English Language* Harlow: Longman
- Куърк 1968 Quirk, R. *The Survey of English Usage Corpus*
<http://www.ucl.ac.uk/english-usage/>
- Лий (2006) Jackie Lee “Subjunctive were and Indicative was: a corpus analysis for language teachers and textbook writers” in: *Language Teaching Research* 10.80. Sage Publications
- Лийч (1997) Geoffrey Leech “Teaching and Language Corpora: A Convergence” in A. Wichmann, S. Fligelstone, A.M. McEnery and G. Knowels (ed.) *Teaching and Language corpora* pp 1-23 London Longman
- Макенери 2001 McEnery, Tony and Andrew Wilson *Corpus Linguistics: An Introduction* Edinburgh University Press 2001
- Манинг и Шуце 1999 Christopher Manning and Hinrich Schutze *Foundations of Statistical Natural Language Processing* Cambridge Massachusetts: The MIT Press 1999
- Савил-Тройки (2006) Saville-Troike, Muriel *Introducing Second-language Acquisition* Cambridge: CUP 2006
- Синклер (2004) Sinclair, John and Ronald Carter *Trust the Text: Language, Corpus and Discourse* Routledge: London:2004
- Трибъл 2001 Tribble, Christopher “Small corpora and teaching writing Towards a corpus-informed pedagogy of writing” (381 – 408) in Mohsen Ghadessy, Alex Henry, Robert L. Roseberry (ed) *Small Corpus Study in ELT: Theory and Practice* Amsterdam: John Benjamins Publishing Company
- Хънстън и Франсис (2000) Susan Hunston, Gill Francis *Pattern Grammar A corpus-Driven Approach to the Grammar of English* John Benjamins Publishing Company 2000
- Хънстън 2002 Hunston, S. *Corpora in Applied Linguistics* Cambridge: Cambridge University Press
- Уидоусън (2003) Widdowson, H. *Defining Issues in English Language Teaching* OUP 2003