https://doi.org/10.53656/bel2022-5-6-AARklbll

Methodology
Методика

КОРПУСНАТА ЛИНГВИСТИКА В ПРЕПОДАВАНЕТО НА БЪЛГАРСКИ ЕЗИК КАТО ЧУЖД

Доц. д-р Антония Радкова
Софийски университет „Св. Климент Охридски“,
Лисабонски университет

Резюме. Статията представя методи за използване на корпусни данни в обучението по български като чужд. Представени са използваните корпуси на българския език. Предложени са начини за инкорпорирането на корпусни данни и материал в следните функции: като източник на информация при планирането на обучението; при създаването на учебни материали; като учебен материал; като средство на използване на езика. Приложението им в обучението е демонстрирано с примери.

Ключови думи: обучение по български език; български език чужд; езикови корпуси; дигитални ресурси

Лингвистичните основи на преподаването на един език традиционно се основават на речниците, теоретичните и академичните граматики, теоретични изследвания, като монографии, статии и др. Немалко влияние оказват и преобладаващите в научната и преподавателската общност теории, както и личните виждания на преподавателя за същността на езика и оптималния начин за изучаването му. Към тях в последните години могат да бъде добавени и корпусните данни. Разликата между традиционните източници, като речници и граматики, и корпусните данни е, на първо място, инвариантността на първите и автентичността на вторите: корпусът ни дава информация за езика в естествената му употреба. Към това предимство трябва да се добавят и актуалността, съвременността на езика в корпусите, огромното количество езикова информация в тях – обем, който не може да бъде постигнат от един традиционен речник или граматика.

В последните години набира популярност обучението, базирано на данни – data driven learing (DDL). Изследователските методи на корпусната лингвистика могат да бъдат пренесени в обучението и превърнати в методически ре-

левантни и подпомагащи учителите, студентите и учениците (Pérez-Paredes, Harris and Moreno Jaén 2010).

Българската корпусна лингвистика също навлиза все по-успешно в сродните научни области. Данните от корпусите на българския език се използват в различни видове лексикографски, морфологични, стилистични, фразеологични и други области на лингвистиката. Появиха се и предложения за използване на корпусите като средство, подпомагащо изучаването на българския език като роден (Koeva, Leseva, Stoyanova, Todorova 2016). Все още данните от корпусите не са достатъчно активно използвани, а те биха могли да дадат много и актуална информация за езика, като допълнят , а в някои случаи дори и заместят досегашните традиционни източници на информация.

Целта на настоящото изследване е да представи подходи за използване на корпусни данни в обучението по български като чужд и да предложи методи за инкорпорирането им в качеството на източник на информация при планирането на обучението и при създаването на учебни материали и като учебен материал и средство на изучаване на езика.

Български корпуси

Преди да представим методите за инкорпориране на корпусни данни в преподаването на български език, ще представим накратко корпусите, които са използвани в това изследване.

На първо място, това е Българският национален корпус (БНК) – проект на Института за български език „Проф. Любомир Андрейчин“. Едноезиковата му българска част към 2014 година, ядрото на корпуса, съдържа приблизително 1,2 милиарда думи и над 240 000 текста. Материалите в Корпуса отразяват състоянието на българския език (предимно в неговата писмена форма) от средата на XX в. до наши дни. Корпусът е достъпен на http://dcl.bas.bg/bulnc/informatsiya/.

Другият използван в изследването корпус е WebCLaRK на проекта BulTreeBank. Той представлява електронен ресурс от синтактично анотирани текстове, който е в основата на формална компютърна граматика на българския език. Този проект предоставя достъп до различни езикови услуги, между които списък на 100 000 най-честотни български словоформи – BTB-FreqList и Stopword list, подредени съответно по честота на употреба и по азбучен ред. Тези ресурси са достъпни на http://webclark.org/.

Българският Браун корпус е създаден в съответствие с методологията, разработена в Университета „Браун“. Той съдържа над 1 млн. думи към 2007 г. и включва текстове, създадени или публикувани като първо издание в периода 1990 – 2005 г. Предимство на Браун корпуса са повечето възможности за разширено търсене (Koeva, Obreshkov, Tinchev, Rizov 2007). За целите на обучението това означава възможност за намиране и анализиране например на определени форми като прилагателните, завършващи на -ски (градски, селски), глаголите с

представки изпона- (изпонастроиха), сродни думи, т.е. думи, съдържащи обща част, и др. Корпусът е достъпен на http://dcl.bas.bg/Corpus/home_bg.html.

Корпусът от устна българска реч съдържа записи от автентични диалози, които са транскрибирани според специално разработени системи, като се отчитат спецификите на спонтанната реч. Той може да бъде използван за подборка на примери за актуално общуване, особено полезни в обучението по български като чужд в България. Корпусът няма собствена търсачка, но може да се използва външна за търсене в базата данни на сайта. Данните са достъпни на http://bgspeech.net/.

Съществуват и други корпусни източници, като например българския WordNet, както и други ресурси, които могат да бъдат използвани като корпуси. Това е всяка база данни от текстове, за които са достъпни някакъв тип систематизиране и анализ, като например търсачка, сайт на медия, социална мрежа и т.н. От тях също, както и от посочените по-горе корпуси, може да се получи информация за честота на употребата на определена дума или израз, типична съчетаемост, регистри и контексти, в които те се употребяват, и др. Характерно различие на корпусите от тези ad hoc бази данни е, че първите дават и лингвистична информация, в някои от тях детайлизирана. Немаловажен е фактът, че за разлика от езика в интернет и в социалните мрежи езикът в българските корпуси е нормализиран, т.е. в корпусите са направени проверки за правописни и пунктуационни грешки.

Корпусите като теоретична и методическа основа на езиковото обучение

Възможностите за използването на корпусни данни при преподаването на български език като чужд са няколко. Първо, различни са субектите, които използват корпусите: това могат да бъдат както преподавателите, така и авторите на учебници, така и самите студенти или ученици. Второ, могат да бъдат разграничени тези четири типа на прилагането на корпусите в обучението – използване на корпусите като:

– източник на информация при планирането на обучението;
– източник на информация при създаването на учебни материали;
– материал за преподаване или учебен материал;
– средство за самостоятелно или в група изучаване на езика.

Първите три типа са използвани от обучаващите: преподаватели, автори на учебни програми, учебници, учебни материали, а последният – от обучаемите.

Корпусна информация при планирането на обучението

Всяко обучение започва с етапа на планиране. Това може да бъде създаване на учебна програма, учебен план или разпределение на учебния материал,

което се налага да прави всеки преподавател. Тук помощ могат да окажат корпусните данни, тъй като те дават информация за честотата, употребата в различни регистри на речта, актуални процеси в съвременния български език и др. Един от спорните въпроси при граматичния материал е начинът на въвеждане на българските времена. Докато за сегашно и бъдеще време е общоприето да се дават на начално ниво, за другите времена няма единно мнение. В съществуващите учебници по български език като чужд те се дават по различен ред: повечето, но не всички учебници въвеждат като първо аорист. Равнищата, на които се изучават аористът, перфектът и имперфектът, са в диапазон от A1 до B2.

Един от критериите за решаването на този проблем е честотата на използване на различните времена в речта. Корпусните данни са категорични, че най-често употребяваното след сегашно време е аористът. Според данните на БНК в съвременния български език аорист се употребява над 6 млн. пъти, имперфект има над 4 млн. употреби, перфектът се употребява над 1 млн. пъти и бъдеще време се употребява малко над 1 млн. пъти. Честотата на употреба на времената е важен, но не е единствен фактор при избора на последователността на въвеждането на времената, имат значение също така трудността (бъдещето време обичайно се изучава преди другите времена поради лесното образуване на формите и непротиворечивото значение), връзката с други глаголни категории като вид, сложност на формообразуването и употребата, синтактични особености на употребата (словоред) и др.

Корпусна информация при създаването на учебни материали

Нерядко преподавателят трябва да направи избор как точно да въведе новия материал. Например числителните се въвеждат още в първите часове при изучаването на чужд език и са сравнително обемен и труден за усвояване материал. Българските числителни са сложни и поради близостта на формите на числителните от 11 до 19 с десетиците след 20 (*дванадесет* и *двадесет* се различават само с едно -на- в средата на думата). Освен това студентите се сблъскват с още една трудност: те трябва да усвоят за повечето числителни две паралелни форми (*дванайсет* и *дванадесет*) или дори три (*дванадесет, дванайсет, дванайсе*). Би било добре поне в началото на обучението да се избере и въведе само една от дублетните форми, а втората да се въведе по-късно. Но коя да изберем? Корпусните данни показват, че в писмената българска реч все още по-често се използват формите, завършващи с -надесет, а не приетите в устната реч форми на -найсе. По-долу са представени данни за честотата на употреба на различните форми на числителните 12 и 20. Номерът пред думата обозначава поредността сред всички словоформи в корпуса, напр. формата „дванайсет“ е 14 409-а поред.

Поредност по честота на употреба	Форма	Поредност по честота на употреба	Форма
10665.	дванадесет	2963.	двадесет
14409.	дванайсет	3223.	двайсет
36809.	дванадесетте	40060.	двайсетте
54771.	дванайсетте	42469.	двадесетте

Фигура 1. Поредност на формите на числителните 12 и 20

В Корпуса на българската устна реч данните са по-различни: *двадесет* има 5 употреби, всичките в медийни текстове, *двайсет* – 10, *двайсе* – 9, като последните две форми са употребени както в медийни, така и в разговорни текстове.

Вторият пример е при изучаването на формите на прилагателни със суфикс -ен. По-голямата част от тях имат суфиксно -е- във формата за мъжки род и нямат във всички останали форми: членувани форми, женски и среден род, множествено число. Според българските граматики това правило има няколко подправила (ще ги представим накратко): прилагателните за цвят, за материал, с ударение върху суфикса -ен и повечето относителни прилагателни нямат редуване -е/Ø. Като резултат за целите на лингвистичното описание на българския език имаме едно комплексно морфологично правило с разнородни (семантични, фонологични) и непълни условия, с изключения и с изключения на изключенията.

Част от тези изключения са важни, честотни и актуални за съответното ниво на изучаване на езика, част не са и могат да бъдат пропуснати. За това кои особености са по-важни и кои – по-маловажни, информация могат да дадат корпусните данни. Подобно търсене би могло да бъде направено с българския Браун корпус, който позволява търсене по част от дума, в случая със суфикс (-ен) или суфикс и флексия (-ена, -ени и т.н.), или с честотния речник FreqList100000 на BulTreeBank. Този списък е от 100 000 токъни, но от гледна точка на изучаването на езика като чужд това е ненужно голям обем. Затова създадохме FreqList1000, FreqList2000 и FreqList3000, които съдържат съответно 1000, 2000 и 3000 най-често срещани словоформи в българския език. Данните от списъците FreqList лесно могат да бъдат превърнати в обратен речник, който може да бъде ползван от преподавателя.

Сред 2000 най-употребявани словоформи се среща само едно прилагателно, което няма редуване на -е- в суфикса (*определен*), във всички останали случаи има редуване -е/Ø; сред 3000 прилагателните без редуване са шест. От методическа гледна точка на началния етап на обучението изключенията е по-подходящо да се дават не в граматичното правило, а лексикално: когато

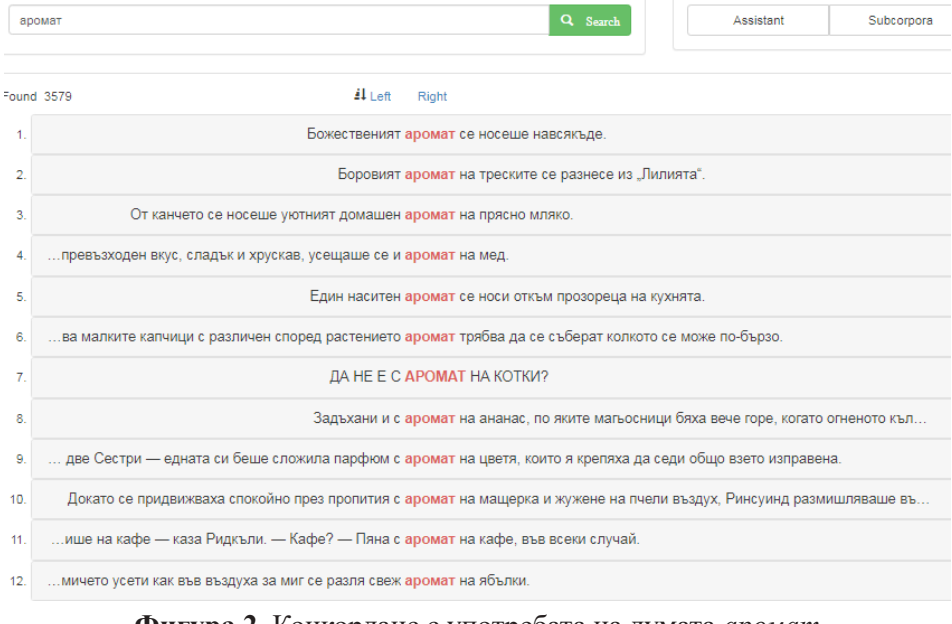
съответната дума се въвежда за изучаване, се обяснява особеността във формообразуването ѝ.

Корпусната информация като учебен материал

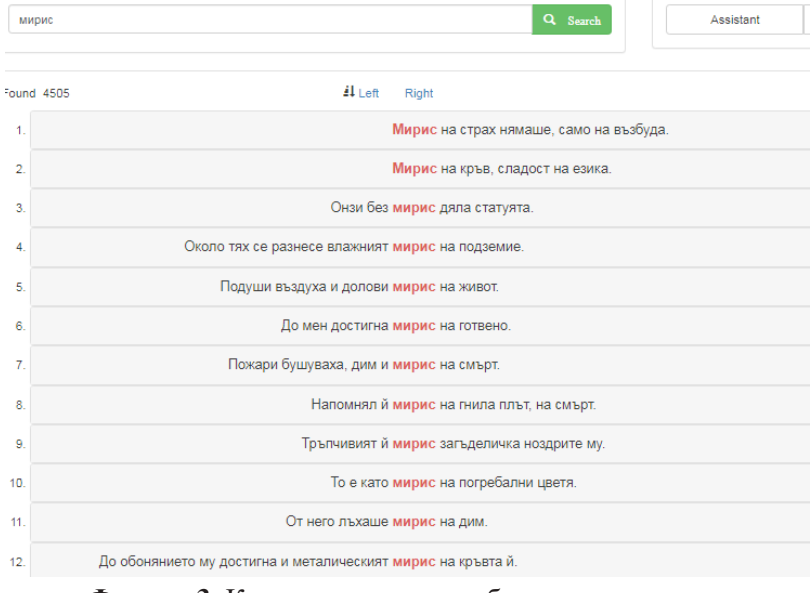
Примери от корпусите на българския език могат да се използват и като учебен материал по време на часовете или като задачи за самостоятелна работа. Възможностите за използване на корпуса в обучението по български език като роден вече предизвикват интереса на изследователите (Koeva, Leseva, Stoyanova, Todorova 2016).

Първият етап при въвеждането на корпусите е студентите и учениците да се запознаят с българските корпуси и сайтовете. Следващият етап е изграждането на корпусна грамотност, която изисква тренировка и известни технически умения. На този етап студентите знаят каква информация може да се намери в конкорданса или в корпуса, могат да правят базисни търсения и могат адекватно да разберат и интерпретират получените резултати.

Например базисно търсене в конкордански може да се използва за нагледна демонстрация за употреба на синоними в различни контексти и откриване на разлики в значенията им. По-долу са дадени примери с думите *мирис* и *аромат*. На практика студентите могат да видят много повече редове от конкорданса, отколкото са показани по-долу на фиг. 2 и 3.



Фигура 2. Конкорданс с употребата на думата *аромат*



Фигура 3. Конкорданс с употреба на думата *мирис*

От показаните примери се вижда, че думата *аромат* се използва по-често в позитивен контекст (мирише на нещо хубаво), докато думата *мирис* може да се използва както в позитивен, така и в негативен контекст (може да мирише както приятно, така и неприятно). Към тези думи на по-напреднал етап на обучението може да се добавят и *миризма*, *дъх*, *ухание*. След работата с примерите от конкорданса студентите могат да сверят изводите си с речниковите данни, според които аромат е приятна миризма, а мирис/миризма – физично свойство на вещество, което действа върху обонянието и предизвиква приятно или неприятно усещане. Положителната страна на използването на конкорданса е в запознаването с употребата на думата в речта.

Не всички синоними, антоними и други семантични отношения могат да бъдат въведени по този начин, понякога детайлите в семантиката не са ясно забележими в контекстите на употреба, например разликата между глаголите *обичам* и *харесвам* (*обичам кафе/харесвам кафе*) или между наречията *често* и *обикновено*.

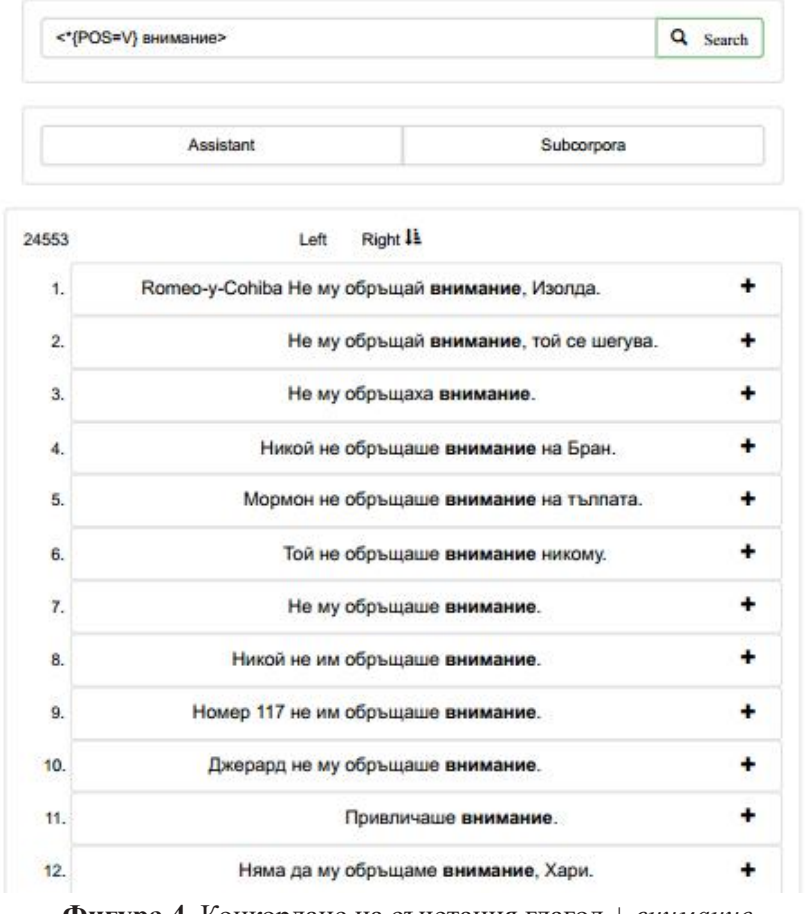
В заключение можем да обобщим, че в преподаването на български като чужд корпусните данни могат да предоставят автентична и актуална информация за:

– лексикална честотност;
– семантични отношения (синонимия, антонимия и др.);
– съчетаемост на лексиката и колокации;

– честота в употребата на граматични категории;
– формообразуване и употреба на словоформите на думите;
– словообразуване, сродни думи;
– употреба в различни регистри, стилове, жанрове и др.

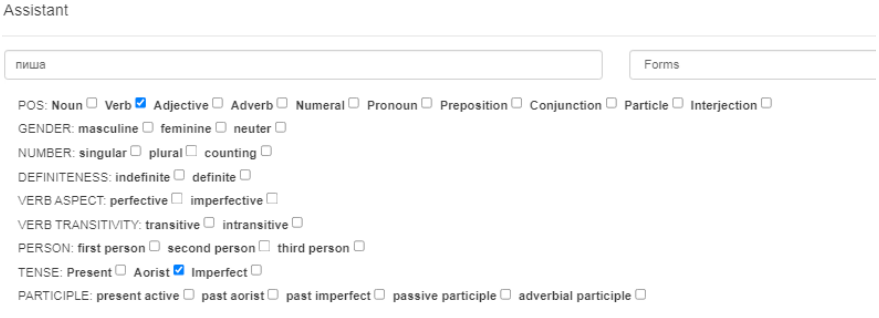
Използване на корпусите при изучаване на езика от студенти и ученици

Използването на корпусите като запознаване или упражняване на употребата на езика има определени ограничения. Студентите би трябвало да имат вече известни познания в българския език, за да могат адекватно да разберат и да използват информацията от корпусите. Подходящото равнище за въвеждане на корпусите е най-рано в края на ниво А2 или в началото на В1.

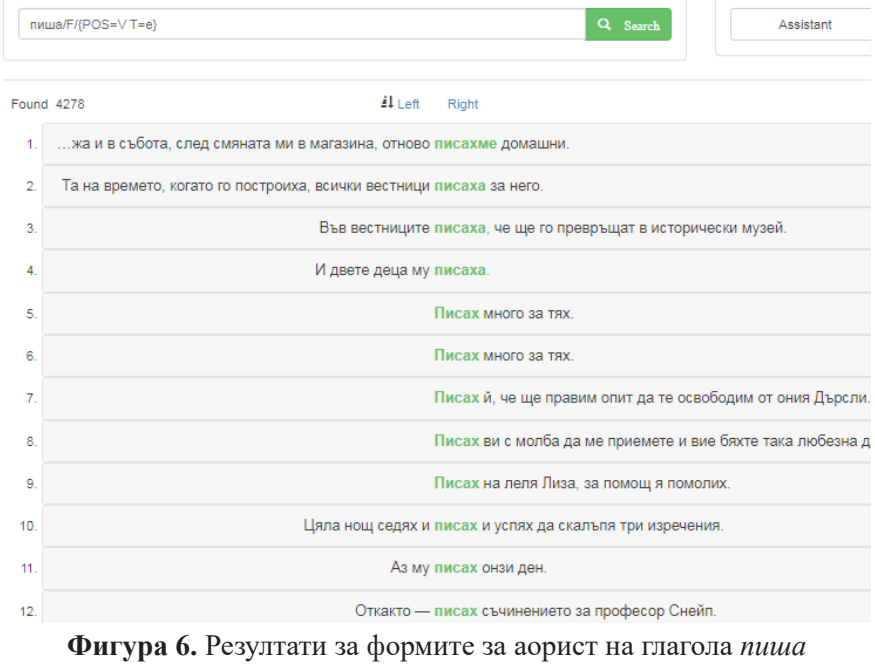


Фигура 4. Конкорданс на съчетания глагол + *внимание*

Целта е като краен резултат студентите да придобият корпусна компетентност и да могат да я използват в зависимост от своите езикови потребности и за търсене на информация. Техническите аспекти при използването на корпуса не са никак малко и не са лесни за хора, които нямат склонност към такъв тип дигитална дейност. Но дори и базисните търсения в корпуса могат да допринесат за намирането на адекватна информация за българския език.



Фигура 5. Заявка за търсене на формите за аорист на глагола *пиша*



Фигура 6. Резултати за формите за аорист на глагола *пиша*

Например, ако студентът се затруднява с кой точно глагол се използва думата *внимание – давам, плащам, вземам* или др., много лесно би могъл да открие тази информация в БНК. За това е необходимо да избере от асистент „Глагол“ и да добави думата „внимание“. Конкордансът на БНК показва, че глаголът, който често се употребява, е „обръщам“.

Базисно търсене в корпуса може да се използва и за граматически справки: например какви са миналите форми на глагола *пиша*.

Използването на корпусни данни е много полезно при обучението на студенти на напреднало равнище, които са потенциалните преводачи от и на български език, а ползата от корпусите при превод е неоценима. Те могат да използват корпусите за търсене на информация, за сверяване на своята езикова интуиция, за запознаване с реалното автентично използване на дадена дума, форма или израз в българския език. Корпусът също така може да даде информация за стила и регистъра, в който се употребява дума или израз, например какви са разликите в употребата на *карам кола* и *управлявам автомобил*.

Заключение

Българските корпуси ще продължават да се развиват, данните в тях ще стават по-пълни и все по-потребителски ориентирани, т.е. достъпни и лесни за употреба – това показва развитието на корпусната лингвистика. Методиката на обучение по български като роден и като чужд може да бъде обогатена с данните за езика от корпусните изследвания, с което да се повишат качеството на обучението и интересът към изучаването на езика. От полза не само за лингвистите, а и за преподавателите и учителите по български език е да бъдат запознати с възможностите на корпусите и начините за тяхното използване в практиката на обучението по български език.

ЛИТЕРАТУРА

КОЕВА, С., ЛЕСЕВА, С., СТОЯНОВА, И., ТОДОРОВА, М., 2016. Езиковите технологии и ресурси – нови перспективи в обучението по български език (българската лексикално-семантична мрежа Булнет и Българският национален корпус). *Български език и литература*, 58 (4), 377 – 393.

ТИНЧЕВ, Т., КОЕВА, С., РИЗОВ, Б. & ОБРЕШКОВ, Н., 2017. Система за разширено търсене в корпуси. *Литературата* (2), 99 – 115.

PÉREZ-PAREDES, P., 2010. Corpus linguistics and language education in perspective: Appropriation and the possibilities scenario. In Harris, T., & Moreno Jaén, M. (Eds.), *Corpus linguistics in language teaching.* Frankfurt, Germany: Peter Lang, 53 – 73.

REFERENCES

KOEVA, S., LESEVA, S., STOYANOVA, I., TODOROVA, M., 2016. Ezikovite tehnologii i resursi – novi perspektivi v obuchenieto po balgarski ezik (balgarskata leksikalno-semantichna mrezha Bulnet i Balgarskiyat natsionalen korpus). *Balgarski ezik i literatura*, 58 (4), 377 – 393.

TINCHEV, T., KOEVA, S., RIZOV, B.& OBRESHKOV, N., 2017. Sistema za razshireno tarsene v korpusi. (2), 99 – 115.

PÉREZ-PAREDES, P. (2010). Corpus linguistics and language education in perspective: Appropriation and the possibilities scenario. In Harris, T., & Moreno Jaén, M. (Eds.), *Corpus linguistics in language teaching* (pp. 53 – 73). Frankfurt, Germany: Peter Lang.

CORPUS LINGUISTING IN TEACHING BULGARIAN AS A FOREIGN LANGUAGE

Abstract. The article presents some possible approaches of the implementation of corpus data in teaching Bulgarian as a foreign language. An incorporation of corpus data and material can be made on different phases: planning phase; development of teaching materials; source for teaching or learning; tool to study the language. Proposals are illustrated with examples.

Keywords: Bulgarian language – teaching and learning; Bulgarian as a foreign language; corpus linguistic; digital resources

✉ **Dr. Antonia Radkova, Assoc. Prof.**
Web of Science Researcher ID: AAY-9249-2020
Faculty of Slavic Philologies
University of Sofia
Sofia, Bulgaria
Faculdade de Letras
Universidade de Lisboa
Lisbon, Portugal
E-mail: antoniar@edu.ulisboa.pt