# За корпусите като езикови ресурси и тяхното приложение

Десислава Баева, Русе

*This article describes the corpora as language resources, the requirements they must satisfy and the main stages in their creation. The corpora are the tool for many linguistic researches and they are widely used in scientific and other fields - automatic translation of texts, synthesis and speech recognition, learning foreign languages. Focus of the article is the use of corpora in foreign language education.*

***Key words:*** *Corpus, Natural Language Processing, Corpora in the classroom*

В областта на обработка на естествени езици съществува силна комерсиализация на софтуерната индустрия, която предразполага динамичното развитие на тази научна област. Програмната реализация на различните видове езикови приложни продукти изисква лингвистична компетентност, което затруднява изработването и широкото им разпространение. Независимо от това езиковите ресурси намират все по-обширни области на приложение.

Терминът *езикови ресурси* включва множество електронни данни, отнасящи се и за писмената, и за устната реч. Тези данни варират от обикновени списъци до компютърни речници и граматики, терминологични бази от данни, корпуси, семантични мрежи и др. и играят важна роля за подготовката, обработката и управлението на информацията и познанията.

Популяризирането на корпусите като средство за множество лингвистични изследвания доведе до засиления интерес към тях, а от там и до обособяването на самостоятелната наука - *корпусна лингвистика.*

Според С.Коева [Коева 2009] съществува известна неяснота при определянето на относително сходни понятия като *текстов архив, библиотека от текстове, съвкупност от текстове и корпуси*. **Корпусите** се отличават по отношение на своето структуриране. Те представляват множество от текстове в електронна форма, които отговарят на определени тематични критерии и имат достатъчен обем (в съвременните ресурси броят на думите се измерва в милиони); дават възможност за извличане както на качествена, така и на количествена информация за езика.

**Репрезентативността** е признак, определящ обективността на корпусите. При експлоатирането им се избират максимални по обхват и разнообразие материали, с цел да се даде възможност за най-пълно описание на обективната картина на речевата практика. Именно тази характеристика различава лингвистичния корпус от случайния набор от текстове.

Предварителната обработка на текста обхваща различни нива от заложената в текста информация. Това определя множеството подзадачи за изпълнение преди текстът да придобие качества, необходими за включването му в корпус:

- ✓ Анотиране (лингвистична информация за частите на речта, синтактична структура и др.);
- ✓ токънизиране (разделяне текста на определени единици - фонеми, морфеми, лексеми, словосъчетания, изречения и др.);
- ✓ тагиране (приписване характеристики на всяка отделна единица - морфологични, синтактични, морфосинтактични, семантични и др. признаци);
- ✓ парсиране - морфологичен анализ, синтактичен анализ, разрешаване на различни езиково специфични явления като местоименни и неместоименни анафори, елипси и др. и на различните типове езикова многозначност.

Различните нива на анализ използват различни програми: морфологични анализатори (тагери), синтактични анализатори (парсери), разширители на анафори, анализатори на реторичната структура.[Коева 2009]

Работата на потребителите на корпусите се осъществява с помощта на специални програмни средства – «корпусни мениджъри», предоставящи разнообразни възможности за получаването на необходимата информация от корпуса.

**Корпусният мениджър** е специална система, включваща програмни средства за търсене на данни в корпуса, генериране на статистическа информация и предоставяне на резултатите на потребителя в удобна за работа форма. Резултатите от търсенето обикновено се представят във вид на конкорданс – списък с контекстите на думите, където търсената единица се представя в нейното лексемно обкръжение.

**Конкордансът се** явява едно от основните понятия в корпусната лингвистика. В общия случай под конкорданс се разбира първичният набор от необработени текстове. Съществуват специални програми за създаване на конкорданси на съответните корпуси, т.нар. конкодансери. Те позволяват да се установи честотата на употреба на различните езикови единици. Много от тях позволяват дори търсене в контекста по ключови думи или словоформи.

В световен мащаб корпусната лингвистика има сериозни постижения и вече създава свои традиции по отношение на някои подходи при използването на корпуси.

- По предназначение корпусите са създадени, за да дадат възможност за редица експериментални **научни изследвания** в различни области на езикознанието: в компютърната лингвистика, в лексикографията, за теоретични изследвания на определени лингвистични явления, за наблюдения върху особеностите на отделни области на езика, за извличане на примери за демонстрация и др. Популярни са за наблюдения върху честотата на употреба на думи или езикови конструкции, генериране на честотни списъци и др., а също и при формулирането на правописни и правоговорни правила.

- Широко **популярна сфера на приложение** на корпусите е и **автоматичният превод**. Той се базира на двуезични и многоезични корпуси от аналогични текстове, обхващащи милиони думи, с максимално съответствие между текстовете, т. нар. паралелни корпуси (текст оригинал - текст превод). Те са един вид експертни системи, помагащи да се намерят както типовите, така и частните проблеми на превода. Изгражда се модел на превода, като се прилагат статистически обучителни техники, които са с висок процент на резултатност. При автоматичния превод значението на думите се определя от контекста, в който се използват. Естествено все още има да се измине дълъг път преди да се достигне качеството на човешкия превод.

- **Синтезът на глас и разпознаването на реч** със средствата на компютърната техника очертават една научна област, където корпусбазираните подходи търпят бързо развитие през последните години. Процесите по възприемане, анализиране, обработване и възпроизвеждане на реч от компютъра във възможно най-естествения му вид (Natural Language Processing – NLP) са фокусирани върху множество етапи на анализ на информацията – фонетичен, фонологичен, прозодичен, лексикален, синтактичен, семантичен и хиперсинтактичен. Корпусбазираните техники са основно статистически модели, като всеки изследван параметър е функция на вероятностите за поява на дадено състояние.

Успехът на тези подходи е стимул за изследванията в областта на използването на емпирични техники за учене, включително и семантичен анализ - разкриване на смисъла на изказването [Цонева, Баева 2011].

В практиката иновативните NLP технологии имат следните характеристики:

- ✓ намират приложение при взаимодействието човек-компютър без използване на ръцете и зрението;
- ✓ при динамични гласови команди и неограничен речник, без необходимост от адаптиране с гласа на оператора;
- ✓ потребителски диалози, близки до естествения език и др.

• Прилагането на корпусите в **обучението по чужд и по роден език** е свързано с инициативи, които отскоро се появяват и разгръщат своя потенциал. Това е все още е нова и недоразвита област. Корпусното приложение в образователния процес започва след популяризирането им извън рамките на научните институти и придобива все по-широк спектър. Според Т. Ангелова [Ангелова 2011] основната функция на дигитално управляваното учене е да направи автентично езиковото обучение, като се използват данни от езикови корпуси за усъвършенстване на постиженията на обучаваните**.**

Необходимо е да се открият нови методически подходи и да се преодолеят редица педагогически бариери за интегрирането на тези езикови ресурси в сферата на образованието.

При Обучението по чужд език се акцентира върху формирането на чуждоезикова комуникативна компетентност, което предполага постепенен преход при овладяването на различните нива на комуникация и необходимост всяко речево или писмено изявление да е съобразено със езиковите и културните традиции на лингвистичните общества. Тази теза е в основата на формирането на най-актуалните методики за изпозване на корпуси при обучението по чужд и по роден език [Цонева 2002].

Развитието на интернет технологиите облекчава значително достъпа до корпуси с автентични текстове и позволява да се интензифицира използването им в обучението по чужд език. Към настоящия момент съществуват множество корпуси, които могат да бъдат изпозвани за анализ на словоупотребата или на граматическия строй на езика. При това ползвателите (ученици, студенти, учители) могат да осъществят своя избор на конкретен лингвистичен корпус в зависимост от поставените конкретни научни и образователни задачи.

Филолозите са единомислещи, че езиковото обучение трябва да бъде основано на богат репертоар от материали, което предполага използването на корпусите като атрактивна възможност, защото осигуряват реално и потенциално богати и интересни материали.

Целенасочена е работата по прилагането на специални текстове към големите корпуси, които позволяват моделирането на нови подходи и за утвърждаването на нови методи за обучение.

Фактът, че корпусите съдържат основно текстови записи, както и тяхна контекстуализация, мотивира необходимостта от методическо посредничество. В учебния процес учителят трябва да напътства дедуктивния процес на анализ, извършван от обучаваните, а корпусът да изпълнява функцията на дидактичен инструмент.

Основните предизвикателства, поставени пред тази насока на развитие, са:

✓ Задължително е при работа с електронни езикови ресурси да се използват компютърни кабинети за учебни зали, което допълнително възпрепятства организацията на такъв тип упражнения.

✓ Отсъствието в учебните програми на време за подобен вид упражнения е друг практически проблем.

✓ Преподавателите филолози трябва да са с достатъчно висока компютърна грамотност, за да могат да ръководят такъв тип практическа работа;

✓ Обучаваните също имат необходимост от предварителна подготовка, за да се обучат как се работи с корпусни мениджери, а това е свързано със загуба на време, каквото няма заложено в учебните програми .

На фона на очертаните вече методологически предизвикателства може да се твърди, че интегрирането на корпусите в учебен контекст зависи от наличието на специфични малки адаптирани корпуси, които съдържат текстове, аналогични на традиционните корпуси и са съобразени с учебните планове [Бернардини 2004].

В заключение може да се обобщи, че електронният формат на езика крие много предизвикателства за изследователите от хуманитарните специалности. Електронните езикови ресурси осигуряват възможности за създаване и верифициране на изследователски хипотези в областта на лингвистиката, социо- и психолингвистиката философията на езика, сравнителната граматика и др. Огромният размер на съременните електронни текстове обаче изискава прилагането на специални компютърно подпомогнати методи за изследване.

За последните десетилетия с развитието на компютърните технологии корпусните данни се превърнаха в общодостъпен езиков ресурс. Факт, който би бил изключително полезен в случай на достатъчна компетентност за използване на корпусите и на способност за самостоятелна работа с тях. Това от своя страна говори за необходимостта от създаване на качествени и достъпни методически ръководства за тяхното приложение. В противен случай значителна част от корпусния потенциал, свързан с относително сложни или неочевидни методи на работа, ще остане неприложим за широкия кръг от потребители.

Корпусбазираните подходи са в основата на системите за автоматичен превод на текстове и на системите за разпознаване и синтез на глас. Всички те намират широко приложение в нашето съвремие: за автоматично резюмиране на големи по обем документи, категоризиране на текстове, автоматично отговаряне на въпроси, автоматично водене на протоколи от съдебни и парламентарни заседания, в помощ на хората със слухови и зрителни увреждания, в логопедични софтуерни системи и т.н.

**ЛИТЕРАТУРА**

**Ангелова 2011:** Ангелова, Т. Показатели за интегриране на информационните и комуникационните технологии в обучението по български език. - В: *Списание на Софийския Университет за електронно обучение, 2011/2,* електронен формат

**Бернардини 2004:** Bernardini, S. *Corpora in the classroom. An overview and some reflections on future developments*. In: Sinclair, John McH., ed. How to use corpora in language teaching. - Amsterdam [u.a.] : Benjamins, 2004. - VI, 307 S. - (Studies in corpus linguistics ; 12) ISBN 90-272-2283-5

**Коева 2009:** Коева, Св. - В: *Приложение на информационните технологии в работата на филолога и при изграждането на езикови ресурси*, Архимед, София, 54-75, 2009. I

**Цонева 2002:** Цонева, Д. Развитие на речевия слух. *Русенски университет*,2002, с.157

**Цонева, Баева 2011:** Цонева,Д., Д.Баева, За интегриране на образователната интерактивна мултимедия в чуждоезиковото обучение. – В: Традиция и модерност в българската наука и образование, кн.4, Силистра 2011, стр. 310-321

**За контакти:**

Десислава Баева, Катедра "Информатика и информационни технологии", Русенски университет "Ангел Кънчев", тел.: 082-888 214, e-mail: dbaeva@ami.uni-ruse.bg

**Докладът е рецензиран.**