СВЕТЛА КОЕВА, ИВЕЛИНА СТОЯНОВА,
РОСИЦА ДЕКОВА

БЪЛГАРСКО-АНГЛИЙСКИ-X+ ПАРАЛЕЛЕН КОРПУС

SVETLA KOEVA, IVELINA STOYANOVA,
ROSITSA DEKOVA

BULGARIAN-ENGLISH-X+ PARALLEL CORPUS

(Summary)

Following a short overview of the existing parallel corpora, which include Bulgarian, the paper focuses on the structure of the Bulgarian-English-X parallel corpus, the general principles of compiling and structuring parallel corpora, and the metadata used to describe the texts in them. Each subcorpus – Fiction, EU Law, News, Subtitles, Healthcare (administrative texts of the European Medicines Evaluation Agency) – is then presented in details: methods of compiling, languages included and the number of compiled words, the distribution of texts over thematic field, genre, and period of creation. Special attention is paid to the corpus annotation on various levels (morphological, morpho-syntactic, syntactic, and semantic), as well as to the potential of the annotated corpora. Due to the specificity of compiling this type of language resources (rather new for Bulgaria), the authors bring up also the question of the place of the parallel corpora with respect to the Copyright Law.

Keywords: Bulgarian, English, parallel corpus, annotation, matadata

1. Въведение

Корпусите може да съдържат текстове само от един език (или форма на съществуване на езика) или повече от един език. Това са съответно едноезичните и многоезичните корпуси. Многоезичните корпуси се делят на преводни (състоят се от преводни еквиваленти на даден оригинал или оригинали), паралелни корпуси (състоят се от преводни еквиваленти на даден оригинал или оригинали, които са съотнесени помежду си дума по дума и/или изречение по изречение – например многоезичният паралелен корпус от документи на европейския парламент JRC-ACQUIS) и съотносими корпуси (съвкупност от сходни по тематика текстове на повече от един език) – например преводът на новините в програмата на Българското национално радио „Христо Ботев“.

Паралелните корпуси са необходими за всички двуезикови или многоезикови приложения за компютърна обработка на езика: търсене и извличане на информация едновременно от документи на различни езици, резюмиране на съдържанието на документи на различни езици, и разбира се – автоматичен превод. Както и при едноезиковите корпуси, обемът на корпусите е от съществено значение – колкото са по-големи, толкова вероятността даден езиков феномен да се срещне с определена честота е по-голяма. При паралелните корпуси освен езиковите явления за даден език е важно да има достатъчно информация за възможните преводи на лексикално, граматично, стилистично равнище, както дори и за спецификите на превода при конкретни преводачи. Това налага изискването за големи по обем корпуси, което от своя страна прави задачата още по-трудна, тъй като трябва да се колекционират преводни документи на два или повече езика. Въпросите за балансираността и представителността на паралелните корпуси също са съотносими с проблемите при едноезичните – балансираността предполага

включването на текстове от различни жанрове и стилове, така че да се застъпи разнообразие от текстове по определена методология. Това, от своя страна, прави корпусите представителни или за определен стил, или за определен отрязък от време, или комбинирано – за езика. Следователно трябва да бъдат намерени не само документи, принадлежащи към дадена тематична област, но и техните преводни еквиваленти. При съставянето на паралелните корпуси може да се подхожда по няколко начина – да се търсят преводни еквиваленти за конкретни произведения, автори, преводачи – например за документите от Българския национален корпус, които са преводи на художествена литература, са потърсени техните оригинали на чужд език (основно английски). Друг начин да се събират паралелни документи от интернет – вече съществуващи сбирки, например по европейско право, или систематично да се колекционират документи, които се създават на два или повече езика – например новини. Тези начини са комбинирани при създаването ан Българско-английския-Х паралелен корпус.

2. Кратък преглед на съществуващите паралелни корпуси за български

Досега създадените паралелни корпуси за български са фокусирани предимно върху други славянски, балкански или западноевропейски езици. Корпусите, които се споменават тук (трябва да се има предвид, че е невъзможно да се изброят всички опити за създаване на паралелни корпуси), се разделят основно на две групи: корпуси, създадени в/или анотирани от Секцията по компютърна лингвистика към Института за български език на БАН и корпуси, създадени от други изследователски групи в България или чужбина.

Една от целите на краткосрочния европейски проект SEE-ERA NET Building Language Resources and Tools for Machine Translation focused on South Slavic and Balkan Languages (Tufiș et al. 2009) е създаването на паралелни корпуси. Корпусите са паралелни извадки от Acquis Communautaire (наречен SEE-ERA.net административен корпус – SenAC) за български, гръцки, румънски и словенски плюс чешки, английски, френски и немски и повестта на Жул Верн „Около света за 80 дни“, преведена на френски, немски, испански, португалски, италиански, румънски, руски, сръбски, хърватски, български, македонски, полски, словенски, унгарски и гръцки (наречен SEE-ERA.net литературен корпус – SenLC). SEnAC съдържа общо 60 389 преводни единици - изречения, а средният брой думи за всеки език в SEnLC е около 60 000. Избраните текстове са токънизирани, тагирани с морфо-синтактична информация, лематизирани и съотнесени на ниво изречение за двата корпуса, а за SenAC съотнасянето е извършено и на ниво дума.

Документите на Таймс за югоизточна Европа (SETimes) (паралелни преводи на осем балкански езика – албански, български, хърватски, гръцки, македонски, румънски, сръбски и турски) се събират от средата на декември 2010 (Tyers, Alperen 2010)[1]. Досега са събрани по около 800 файла за всеки от езиците. Корпусът се използва за машинен превод – представените резултати за български обаче са много по-лоши от очакваните. Посочената от авторите причина е значително по-малкият брой на съотнесени изречения за български, отколкото за останалите езици.

Паралелният корпус с документи от сесиите на Европейския парламент за периода 1996-2009 (Koehn 2005) включва преводи на следните европейски езици: романски (френски, италиански, испански, португалски,

румънски), германски (английски, холандски, немски, датски, шведски), славянски (български, чешки, латвийски, литовски, полски, словашки, и словенски), угро-фински (естонски, унгарски, финландски), гръцки. Тестовете от извадките са съотнесени на ниво изречения за целите на статистическия машинен превод. Българско-английският паралелен корпус възлиза на 23 Mb, където българската част е съставена от 226 768 изречения и 6 011 944 думи.

В рамките на проекта Multext-East преводните еквиваленти на романа на Джордж Оруел „1984“ на шест езика (български, чешки, естонски, унгарски, румънски и словенски) са анотирани с информация за част на речта и съотнесени по изречения с английския текст (Dimitrova et al. 1998).

В резултат на друг проект е създадена двуезична колекция от текстове в областта на художествената литература и народното творчество на български и гръцки (Ghouli et al. 2009). Корпусът възлиза на около 700 000 токъна (по 350 000 за всеки от езиците): литературният субкорпус съдържа около 550 000 токъна, а субкорпусът за фолклор и легенди се състои от приблизително 150 000 токъна.

OPUS (Tiedemann 2009) е безплатна колекция от паралелни корпуси, състоящи се от преводни текстове в електронна форма, свободно достъпни в интернет. Текстовете са токънизирани и разделени по изречения и се разпространяват в xml формат. Някои от корпусите включват и български език:

- Корпусът на EMEA (Европейската асоциация по лекарствата) на 22 езика;
- Корпусът от новини на SETimes.com на 9 езика;
- Корпус от субтитри на филми на 54 езика, набавен от http://www.opensubtitles.org/.

Корпусите на OPUS са достъпни на следния адрес: http://opus.lingfil.uu.se/.

Съществуват и други проекти, които целят събирането и обработката на паралелни корпуси (включващи също и български) – например корпусът RuN (Grønn, Marijanovic 2010) – паралелен корпус, съставен предимно от норвежки и руски текстове, но наскоро разширен с паралелни текстове на други европейски езици, включително и български); Българско-полско-литовският корпус (Dimitrova et al. 2009); корпусът ParaSol (Waldenfels 2006), известен още като Регенсбургския паралелен корпус – паралелен с корпус от преводни и оригинални белетристични текстове на славянски и някои други езици.

Заключението, което може да се направи от този кратък и непълен преглед на съществуващите паралелни корпуси, включващи и български, е че корпусите са сравнително малки по обем, представят предимно административни или литературни текстове и са съставени от свободно достъпни от интернет текстове, а не са резултат от планирана стратегия за изграждане на балансиран и представителен паралелен корпус.

3. Структура на Българско-английския-Х паралелен корпус

Българско-английският-Х паралелен корпус съдържа художествена литература, административни текстове, новини, особен тип разговорна лексика – субтитри. Стремежът при неговото създаване е да се следва методологията на създаване на Българския национален корпус и организацията на жанровото разпределение в него. Българските текстове от

Българско-английския-Х паралелен корпус вече са част от Българския национален корпус.

3.1. Художествена литература

3.1.1. Начин на съставяне

Корпусът от текстове на художествена литература е съставен ръчно. Много от текстовете на български са взети от Българския национален корпус, а други са добавени в процеса на компилация. Някои от текстовете (предимно на английски, но така също и на немски, френски, руски и др.) са копирани от свободно достъпни страници в интернет (например страницата на проекта Гутенберг – http://www.gutenberg.org/wiki/Main_Page).

3.1.2. Период

Основно изискване за включването на даден текст в корпуса е той да е нов, т.е. да е създаден след 1945 година. Ако оригиналът е на друг език, изискването е преводът на текста на български да е направен след 1945 година, като е съблюдавано условието, оригиналните текстове да не са написани по-рано от 16. век (границата е поставена толкова назад във времето, за да бъдат включени творби на Шекспир, чиито преводи за български са направени след 1945 г.).

3.1.3. Езици, брой оригинални и преводни документи, изходен език

Паралелният корпус с художествена литература съдържа текстове на български, английски, немски, френски, а в бъдеще ще съдържа документи и на други езици. Изходният език и посоката на превода не са фиксирани, т.е. корпусът включва както преводи от английски на български, така и преводи от трети език както на английски, така и на български. По-голямата част от паралелните текстове (634 текста или 93,24% от всички файлове) са английски оригинали с превод на български. Една сравнително малка част (46 текста или 6,76% от всички файлове) са превод от трети език (например френски, немски, руски, италиански и др.) както на английски, така и на български. Освен българско-английския паралелен корпус са събрани и съпоставени текстове на немски и френски, които са значително по-малко и работата в тази насока тепърва се развива.

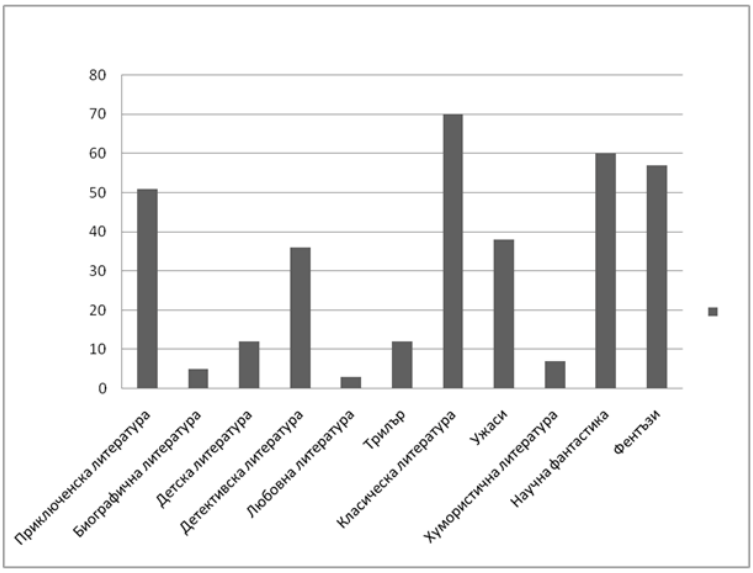
3.1.4. Брой думи

Българско-английският паралелен корпус с художествена литература се състои от 680 текста, които съдържат общо 39 590 472 думи за английски и 34 553 474 думи за български. Значително по-големият брой думи за английски е очакван, тъй като е свързан с някои морфо-синтактични особености на английския език (лексикално изразени пълен и кратък член при съществителните имена, както и аналитични форми на глаголните времена).

3.1.5. Разпределение по жанрове

Текстовете са разпределени в 11 категории, като е следвана изцяло организацията при жанровото разпределение на Българския национален корпус (Koeva et al. 2010), а оттам и на Браун корпус: Приключенска литература; Биографична литература; Детска литература; Детективска литература; Любовна литература; Трилър; Класическа литература; Ужаси; Хумористична литература; Научна фантастика; Фентъзи.

Количественото разпределение на текстове във всяка от изброените области е представено в диаграмата във *Фигура 1*.



Фигура 1. Количествено разпределение на текстовете по области (в проценти)

3.2. Публицистика

3.2.1. Начин на съставяне

Корпусът от публицистични текстове включва новини и други текстове, публикувани на информационния сайт за Югоизточна Европа, http://SETimes.com. Корпусът е съставен автоматично с помощта на компютърна програма, която се свързва с архива на сайта, следва връзките от интернет страницата, докато достигне до страница с текст, и запазва съответните html файлове.

Новините в архива са организирани по година, месец и дата, като тази структура се пренася и в корпуса. При по-старите текстове името на файла съдържа информация за датата и автора или съставителя. При по-новите името на файла дава определение за жанра на текста и пореден номер.

Описанието на текстовете се извършва автоматично заедно с извличането на чистия текст от html файла. Зададени са някои данни за текста, например дата, заглавие, източник. Не е намерен начин за автоматично извличане на тематичната област от html файла, затова е разработен прост автоматичен метод за определяне на обща тематична област (Политика, Икономика, Социално дело, Военно дело, Култура и изкуство) въз основа на наличието на ключови думи в заглавието. По този начин на 29.1% от текстовете е приписана тематична област с много висока точност. Използвана е класификацията на тематичните области от Българския национален корпус (Koeva et al. 2010).

3.2.2. Период

Архивът на сайта съдържа текстове от октомври 2002 до днес. Това се съгласува с изискването за съвременност на включените текстове. Ежедневно на сайта се публикуват текущите новини и други информационни материали, които след това се съхраняват в архива. Корпусът се допълва ежемесечно с новопостъпилите в архива текстове.

3.2.3. Езици

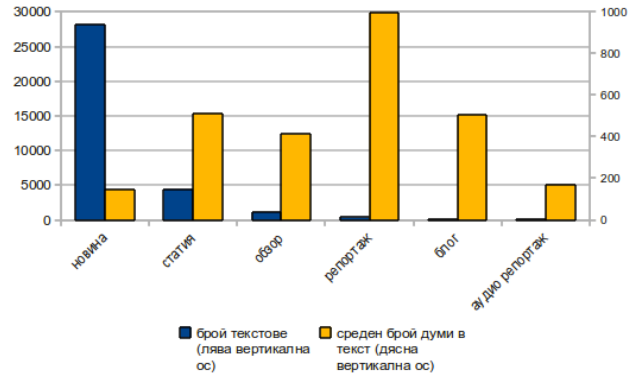
Новините са на всички балкански езици (албански, босненски, български, хърватски, гръцки, македонски, сръбски, румънски, турски) и английски. Не е посочен оригиналният език на текстовете, поради което тази информация не в включена в описанието на корпуса.

3.2.4. Брой думи

Българско-английският корпус се състои от 34 075 текста, които наброяват 7 393 317 думи за българската част и 7 245 294 думи за английската.

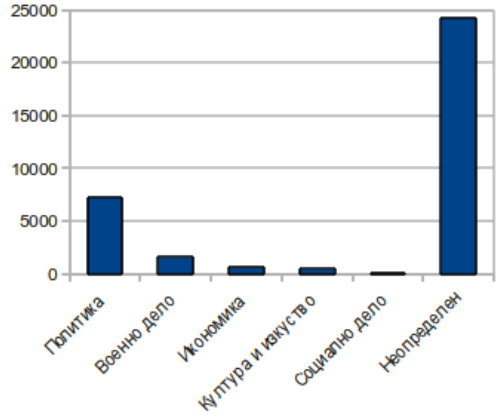
3.2.5. Разпределение по жанрове

В корпуса се съдържат текстове от следните жанрове: новина (кратък информационен текст), статия (по-дълъг текст с коментар и анализ на дадени събития), обзор (на новините за даден период), репортаж (обстойно описание на събития), блог (текст от новинарски блогове), аудиорепортаж (допълващ коментар към аудио репортаж). Класификацията на текстовете по жанрове и средният брой думи за текст са представени на *Фигура 2*.



Фигура 2. Класификация на публицистичните текстове (на български) по жанр

Разпределението на текстовете по тематична област е представено на *Фигура 3*.



Фигура 3. Класификация на публицистичните текстове (на български) по тематична област

3.3. Законодателство на Европейския съюз

3.3.1. Начин на съставяне

Текстовете от законодателството на Европейския съюз са набавени от сайта http://eur-lex.europa.eu/ и представляват актуалното законодателство към април 2011 г. Корпусът е съставен автоматично по методология, подобна на тази за съставянето на корпуса от новини. Тръгва се от началната страница

и връзките от нея се обхождат рекурсивно, докато се достигне до страниците, съдържащи търсените документите.

Текстовете в сайта са организирани по тематични категории и субкатегории на няколко нива. Името на всеки файл представлява уникалният CELEX индекс на документа. В корпуса файловете се записват със CELEX в съответната директория по език. Не се запазва структурата по категории на сайта, но информацията за тематичната принадлежност на всеки текст се включва в описанието на корпуса.

Описанието на текстовете е извършено автоматично, като от html кода на всяка страница е извлечена информация за дата, заглавие, тематична област, ключови думи. Системата от тематични области на законодателните текстове не съвпада със системата на Българския национален корпус, а е значително по-детайлизирана. Също така 99.7% от текстовете са причислени към повече от една тематична област. Поради това е извършена допълнителна класификация въз основа на наличната информация – първата посочена тематична област на всеки текст е приета за водеща и системата от тематични области е сведена до тази на Българския национален корпус. Допълнителната информация за оригиналните тематични области на текстовете се пази в описанието в поле domain_info (вж. 5).

3.3.2. Период

Корпусът съдържа текстове, създадени между 1958 и 2011 г., като 72.5% от текстовете са създадени след 2000 г. Не е известна годината на създаване и публикуване на отделните преводи. Някои от текстовете са изменени и редактирани по-късно, но това е направено или с отделни документи (които са включени самостоятелно в корпуса), или датата на промяната не е отразена или маркирана в документа и поради това не се включва в описанието.

3.3.3. Езици

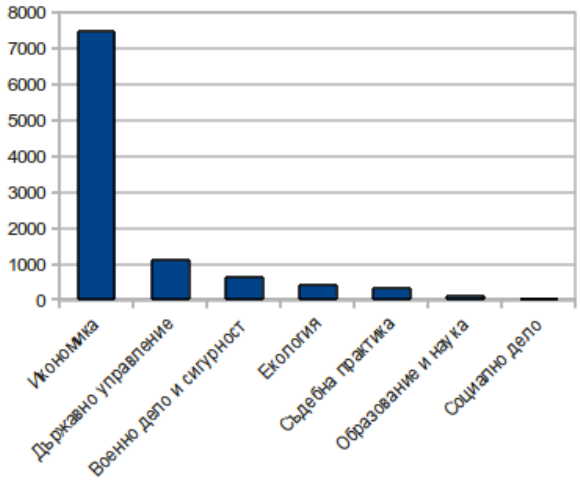
Сайтът съдържа текстове от европейското законодателство на 23 европейски езика. В корпуса са включени всички достъпни текстове на български език и наличните съответствия на останалите езици. Не е известен езикът на оригинала и кои текстове са оригинални и преводни, поради което тази информация не е включена в описанието. Различните езици не са равномерно представени, тъй като не всички текстове са преведени на всички езици.

3.3.4. Брой думи

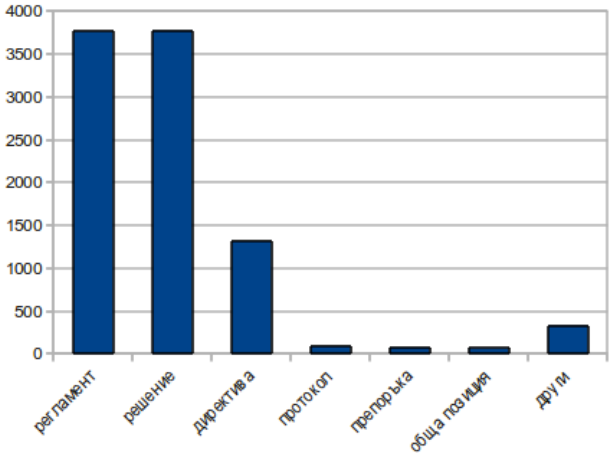
Българско-английската част от корпуса включва 9 856 текста, които наброяват 36 917 940 думи на български и 34 597 176 думи на английски.

3.3.5. Разпределение на текстовете

Корпусът е тясно специализиран и съдържа текстове от административния стил, които са неравномерно разпределени по тематична област и жанр. Преобладаващата част от текстовете принадлежи към област Икономика (*Фигура 4*). Най-застъпени са жанровете регламент, решение и директива (*Фигура 5*), което е очаквано за корпус от международни законодателни документи.



Фигура 4. Класификация на законодателните текстове (на български) по тематична област



Фигура 5. Класификация на законодателните текстове (на български) по жанр

3.4. Субтитри

3.4.1. Начин на съставяне

Корпусът се състои от субтитри на игрални, документални и анимационни филми. Той е част от колекцията от OPUS (Jörg Tiedemann, 2009) и е достъпен на адрес: http://opus.lingfil.uu.se/. Текстовете от колекцията на OPUS са в xml формат и съдържат определена анотация. За нашите цели файловете са обработени, като е оставен чистият неанотиран текст и са премахнати xml таговете.

Към момента е обработена и се използва версия 1.0 на корпуса. Отскоро е достъпна и версия 2.0, която съдържа по-голям брой текстове на повече езици. Обработването на този корпус остава като една от непосредствените бъдещи задачи.

Текстовете от корпуса бяха снабдени и с описание, което съдържа данни за текста, например заглавие, година на издаване и др. Описанието на текстовете е извършено автоматично, като е използвана информация от името на файла, което съдържа името на филма на английски, и от структурата от директории, която съдържа тематичната област на филма и годината на създаване. Използван е списък с 1857 заглавия на филми на български език и

техните английски съответствия, свободно достъпен в интернет, с помощта на който 47.5% от филмите бяха описани с българските им заглавия. Използваният списък обаче не е актуален, защото съдържа само филми, излезли преди 2004 г. Стилът на текстовете в корпуса е описан като Художествен/Разговорен, тъй като съдържат реч, но тя не е автентична, а авторска.

3.4.2. Период

Версия 1.0 от корпуса, която се използва в момента, съдържа субтитри на филми, създадени между 1927 и 2007 г., като 66.8% от филмите са излезли на екран след 2000 г., а едва 5.6% след 2005 г. Във версия 2.0 на корпуса на OPUS има включени повече по-нови филми от последните 10 години. Измежду най-старите филми са „Метрополис“, немски научнофантастичен филм от 1927 г., детективският филм „Малтийският сокол“ от 1941 г. и „Казабланка“ от 1942 г.

3.4.3. Езици

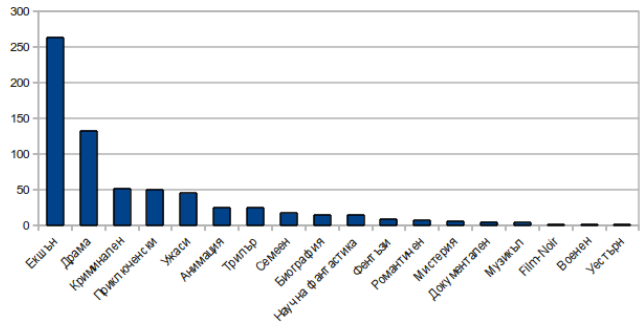
Корпусът от субтитри на OPUS – версия 2.0 включва текстове на 54 езика. Версия 1.0 включва текстове на 30 езика.

3.4.4. Брой думи

Българско-английската част от корпуса се състои от 497 текста, които наброяват 2 970 974 думи на български и 3 931 784 думи на английски. Прави впечатление, че за разлика от публицистичния и законодателния корпус, тук броят думи на английски език е значително по-голям от този на български. След проучване на двойки текстове със значителна разлика в броя думи е установено, че едната причина е съкращаване на репликите при превода от английски на български, а другата – наличието на аудиокоментари в английските субтитри (напр. [*Звъни телефон*]), които не са предадени на български. Това трябва да бъде взето предвид при подравняване на текстовете, тъй като затруднява задачата.

3.4.5. Разпределение на текстовете

Разпределението на субтитрите по тематичната област, към която се отнася филмът, е представено на *Фигура 6*.



Фигура 6. Класификация на субтитрите (на български) по тематична област

3.5. Здравеопазване – административни текстове на Европейската агенция по лекарствата

3.5.1. Начин на съставяне

Корпусът се състои от административни текстове, публикувани от Европейската агенция по лекарствата (EMEA – European Medicines Evaluation Agency). Корпусът е част от колекцията на OPUS (Tiedemann 2009) и е

достъпен на адрес: http://opus.lingfil.uu.se/. Текстовете са обработени по същия начин, както корпусите от субтитри. От текста са извлечени и някои основни метаданни, например заглавие на документа, жанр и година на издаване, които са добавени в описанието на корпуса.

3.5.2. Период

Корпусът включва текстове, създадени и публикувани от Европейската агенция по лекарствата между 1978 и 2009.

3.5.3. Езици и брой думи

Поради големия обем на корпуса, за момента сме се ограничили с обработването само на българско-английската част, която включва 1587 текста, наброяващи 12 586 236 думи за български и 9 735 375 думи за английски език.

3.5.4. Разпределение на текстовете

Текстовете в корпуса са разделени на следните тематични подобласти – Хуманна медицина (77.5% от текстовете); Ветеринарна медицина (13.6% текстове) и общи текстове (8.9%). Използван е автоматичен метод за определяне на жанра на текста по наличието на ключова дума в първите пет реда на файла – първото споменаване на жанр (от определен лист с предполагаеми жанрове) се приема за жанра на текста. По този начин 28.8% от текстовете са определени като доклад, а на останалите е приписан общият жанр на документа. По-голямата част от текстовете не съдържат определение за жанр. Преобладават текстове с характер на доклад, които дават информация за определени лекарства, което е свързано с дейността на Агенцията по лекарствата.

4. Общи принципи при съставяне и структуриране на паралелните корпуси

Основната цел при разработване на колекцията от паралелни корпуси, включващи и български език, е да се предостави достъп до налични паралелни ресурси, които да се използват за изследователски и образователни задачи. Понастоящем тези ресурси за български език са силно ограничени, а те предлагат широки възможности за анализи в областта на сравнителното езикознание, машинния превод и др.

При съставянето на корпусите не е търсена изчерпателност или представителност на дадени езикови явления или разновидности, както и не се вземат предвид изисквания за балансираност. Аткинс, Клийн и Остлър (Atkins et al. 1992) очертават четири типа колекции от текстове: текстов архив (колекция без определена структура); електронна текстова библиотека (колекция, следваща определени принципи, текстовете са в стандартизирана форма, но без да е приложена селекция); корпус (колекция, съставена за конкретни цели и съобразена с експлицитно зададени принципи за съставяне); подкорпус (част от корпус, обикновено статична, използвана за конкретни анализи). Въз основа на тази класификация, колекцията от паралелни корпуси на Секцията по компютърна лингвистика може да бъде определена като електронна текстова библиотека, включваща няколко отделни многоезични паралелни корпуса (вж. 3). Всеки отделен корпус е обособен основно по тематична област и източник (напр. Законодателство, Новини от http://SETimes.com и др.) и обхваща различен брой езици. В корпусите отделните езици не са равномерно представени – т.е. даден текст може да няма съответствие на всички езици.

Основен принцип при съставянето на корпусите е, че българската част е ръководеща – в корпуса влизат само текстове, които имат български съответствия (оригиналът е на български или има български превод). В някои случаи това значително ограничава броя на текстовете и прави задачата по-обозрима, което е съществено за голям многоезичен корпус (например корпусът със субтитри версия 2.0 на 54 езика наброява 1.4 милиона текста).

Българската част на всеки от паралелните корпуси влиза в състава на Българския национален корпус. Поради това е взето решение паралелните корпуси да бъдат снабдени с описание, съвместимо с това на БНК, което да използва същите елементи и класификационна схема (вж. 5).

Част от корпусите са статични (напр. част от използваните корпуси от ОПУС), а друга част са динамични – публицистичният корпус се допълва ежемесечно с нови текстове, а законодателният корпус може да се актуализира периодично с новото актуално законодателство на Европейския съюз, което по своята същност е динамично.

Повечето паралелни корпуси (с изключение на корпусите с художествени текстове) са създадени и обработени автоматично. Някои от тях са получени след обработка на съществуващи корпуси (например от колекцията на OPUS), докато други се състоят от текстове, свободно достъпни в интернет (например новини от http://SETimes.com). Описанието на текстовете в паралелните корпуси в голямата си част също е изготвено автоматично. Понастоящем разполагаме с описание на българската и английската част на всеки корпус. В началото на файловете с българските текстове са добавени и метаданните за всеки текст.

5. Метаданни и описание на паралелните корпуси

Метаданните се дефинират като „данни за данните“, а конкретно в корпусната лингвистика метаданните представляват данни за текстовете, включени в даден корпус – като заглавие, автор, източник, жанр, тематична област и други. Наличието на метаданни и детайлното описание на текстовете е от особена важност. В (Уейн, 2005, Глава 3) се посочва, че данните от описанието на даден текст трябва да са достатъчно детайлни, така че да е възможно да се определи дали даденият лингвистичен ресурс е релевантен за конкретното изследване. Също така въз основа на описанието може да се правят анализи на представителността на корпусите за дадените цели, както и лесно и бързо може да се извличат подкорпуси по определени критерии (напр. по тематична област или жанр, по година и т.н.).

Като най-съществени се определят няколко основни типа данни при описанието на корпусните единици (Wynne 2005, Глава 3). Това са:

- Издателски метаданни: информация за източника на текста, начина на придобиване и добавяне в корпуса;
- Аналитични метаданни: информация, получена при лингвистичен анализ на текста – различни видове анотация;
- Описателни метаданни: класификационна информация за текстовете, която описва характеристиките на текста – стил, жанр, тематична област и др.;
- Административни метаданни: данни за положението на текста в самия корпус, начин на достъп (напр. път и име на файла), формат, дата на добавяне и др.

Както може ясно да се види от описанието на метаданните по-долу, паралелните корпуси на Секцията по компютърна лингвистика са съобразени с принципите и изискванията за наличие на тези основни типове метаданни, като те са разширени и допълнени с някои нови категории.

Описанието на по-голямата част от паралелните корпуси (с изключение на корпуса с художествена литература) е изготвено автоматично и следва установените принципи и класификационна схема на Българския национален корпус (Koeva et al, 2010). В текстовите файлове на български език са добавени и метаданните за текста. Описанието съдържа следните метаданни:

(1) *filename* – име на файла;
(2a) *author_info* – информация за автора (един, колективен, неизвестен);
(2b) *author* – име на автора;
(3a) *translator_info* – информация за преводача (един, колектив, неизвестен);
(3b) *translator* – име на преводача;
(4a) *text_info* – информация за текста (един, няколко);
(4b) *title* – заглавие;
(5a) *year_of_creation* – година на създаване на оригинала;
(5b) *publishing_date* – дата/година на публикуване на конкретния текст (превод, преработка и др.);
(6a) *source_type* – тип на източника (от интернет, от издателя и др.);
(6b) *source* – данни за източника, интернет адрес;
(7) *translated* – преводен или оригинален;
(8) *medium* – форма на текста (писмен, устен);
(9) *number_of_words* – брой думи;
(10) *style* – стил (административен, публицистичен, научен, разговорен, художествен);
(11a) *genre* – жанр, системата от жанрове е специфична за стила;
(11b) *genre_info* – конкретизация и бележки за жанра;
(12a) *domain* – тематична област, също зависи от стила;
(12b) *domain_info* – конкретизация и бележки за тематичната област;
(13) *notes* – общи бележки по текста;
keywords – ключови думи.

Както общата, така и детайлната информация, включена в описанието на метаданните, съответства на установените стандарти (Atkins et al. 1992; Burnard 2007) с малки промени, за да обхваща някои специфични жанрове и тематични групи. Описанието на всеки текст е изготвено с информацията, която се съдържа в оригиналния файл, придобит от интернет или от други източници (напр. OPUS) с разширение html или xml. В голяма част от тези текстове са маркирани някои основни данни, като заглавие, автор, дата на изготвяне, жанр, тематична област и др.

Не всички полета от описанието обаче е възможно да бъдат попълнени директно чрез извличане на данни от кода (html или xml). За някои полета бяха използвани допълнителни методи за приписване на стойност (например определяне на тематична област по ключови думи в заглавието). Също така описанието може да бъде допълвано и доразвивано ръчно.

Всеки файл се описва в съответствие с неговото съдържание и се назовава спрямо принадлежността на текста към съответния жанр и тематична група. По този начин всеки файл получава уникално наименование, което се състои от пореден номер на файла, малка буква, показваща дали текстът е преводен или оригинал, и поредица от главни букви, които да обозначат стила на текста (A-административен, B-научен, C-масмедия, D-художествена литература, E-неформален), неговия жанр (за всеки стил е въведен отделен списък с букви, обозначаващи различните жанрове в него), и неговата тематична принадлежност (за всеки стил е въведен отделен списък с букви, обозначаващи различните тематични полета в него). Освен това файловете се записват в директории и поддиректории, подредени по същия начин, за да се запази организационната йерархия и да се улеснят бъдещи търсения в системата.

В крайна сметка имената на паралелните текстове са почти еднакви, като се различават само по окончанието *en*, което обозначава, че текстът е на английски. Например, L000033bDAEen означава, че текстът е оригинален (b), написан е на английски, стилът му е художествена литература (D), жанр – приключенска литература (A), и е роман (E); докато паралелният български текст е обозначен L000033tDAE, където единствените разлики са буквата (t), за да маркира текста като преводен, и отсъствието на окончанието en, за да покаже че текстът е на български.

6. Анотация

Ползата от паралелните корпуси нараства значително, ако са анотирани, т.е. ако съдържат експлицитно представена допълнителна информация, която може да послужи за различни цели. Анотирането на даден текст представлява експлицитно представяне на езиковата интерпретация на отделните единици от текста. За анотация се използва маркиращ език, който е набор от конвенции за означенията на езиковата интерпретация.

Анотирането може да се направи автоматично или ръчно, в случая че анотацията е автоматична, тъй като става въпрос за големи по обем текстове, базира се на съществуващи езикови ресурси и програми за автоматична обработка на текста. Може да се отделят различни равнища на лингвистична анотация (Leech 1997: 8-15), например: морфологично, морфо-синтактично, синтактично, семантично и дискурсно (*EAGLES* 1996: 3), като анотираните корпуси обикновено се асоциират с повече от едно равнище.

На морфологично равнище всяка лексикална единица в текста може да се асоциира със своята основна форма – лема (лематизация). Анотацията на лемата за флективни езици като български, където преходен глагол от несвършен вид има петдесет и две синтетични форми, е важна част от компютърната обработка, за да може да се осъществява връзката при употребата на различните форми на дадена дума в различни контексти. Морфо-синтактичната анотация въвежда информация за граматичния клас на дадена лексикална единица и (възможно) съответните стойности на граматичните категории, характеризиращи единицата. Морфо-синтактичната анотация отстранява граматичната многозначност на омографите и се използва в голям брой приложения за компютърна обработка на езика: за определяне на коректната фонетична стойност на дадена форма, за съотнасяне на формите към правилната парадигма, за идентификация на съставните лексикални единици, за определяне на строежа на фразите, за съотнасяне с допустимо лексикално значение и т.н. Синтактичната анотация

най-общо може да представя граматичните зависимости между конституентите на словосъчетание или фразовата структура, в която конституентите се комбинират помежду си. При семантичната анотация се разграничават два типа (McEnery, Wilson 2001: 61-62) – представяне на семантичните релации между думите в текста (например анотация на семантичните роли на аргументите, на валентността на лексикалните единици, на реализацията на семантичните фреймове) или на семантичните свойства на думите (например анотация на значението на думите). Както се вижда, равнищата на анотация съответстват най-общо на равнищата на многозначност, от което следва, че една от основните области на приложение на анотираните корпуси е отстраняването на многозначността. Разбира се, анотацията е необходима и при еднозначните структури за съотнасянето им с коректните лингвистични стойности, но при многозначните структури анотацията показва не само коректните лингвистични стойности, но и избора им от множеството възможни.

Разделянето на корпуса на единиците, които го съставят, е най-ниското ниво на обработка на паралелния корпус: текстовете се разделят на единици (думи, изречения и параграфи), които се характеризират в зависимост от съставните си части. Най-общо токън се определя като поредица от символи между два отделящи символа (интервал, пунктуационен знак). Българският токънизатор е базиран на регулярни правила и разпознава последователности от букви, цифри, пунктуационни знаци, специални символи, комбинации от тях и празни символи. Също така се разпознават някои изрази, които означават дати, дроби, имейли, интернет адреси, съкращения и др.

Към това предварително равнище на обработка на текста принадлежи и автоматичното определяне на границите на изреченията. Определянето на границите на изреченията също е базирано на регулярни правила, но се използват и списъци от лексикални единици – лексикони (например за изброяване на съкращения, след които може да има/трябва да има главна буква/цифра и т.н.). Разпознаването и маркирането на единиците в текста е необходима предпоставка за повечето задачи, свързани с обработката на естествения език. Идентифицирането на границите на думите и на изреченията в много случаи включва отстраняване на многозначността при употребата на пунктуацията, т.е. кога даден знак означава край на изречение и кога – не.

За тагирането на българските текстове се използва статистически тагер, базиран на SVM (Support Vector Machine). Използва се SVMTool[2] (Gim'enez & M'arquez 2004), проект с отворен код за трениране на модели за тагери и прилагането им върху входен текст. SVMTool се характеризира с простота (лесно се конфигурира и тренира); гъвкавост (параметрите, с които работи, могат да се настройват, както и да се дефинират сложни параметри, включително н-грами по част на речта или класове на многозначност, също така може да се използва анализ на равнище изречение); преносимост (програмата е езиково независима); точност (в сравнение с други известни тагери има конкуриращa се точност); ефикасност (времето за тагиране зависи от параметрите, които са избрани при схемата за тагиране – при едностепенно тагиране от ляво на дясно прототипът на Пърл показва скорост на тагиране от 1500 думи на секунда, а версията на C++ – над 10000 думи на секунда). Изборът на стратегия за трениране е от особена важност (от ляво на дясно, от дясно на ляво или и двете, в случая от ляво на дясно, макар че се

твърди, че комбинацията на посоките дава най-добри резултати); дължината на контекста – колкото е по-голям, толкова са по-добри резултатите, в случая прозорец от седем елемента, дефиницията на параметрите (н-грами на думи или тагове или класове на многозначност, лексикализирани параметри като префикси, суфикси, главни букви и др.). При тагирането се използва специално редуциран тагсет.

Лематизацията е тясно свързана с тагирането по части на речта и включва приписването на лема, т.е. на основната форма при изменяемите думи, към всяка дума в текста след извършване на морфосинтактичния анализ, както и съответните граматични характеристики, с които се характеризира употребената форма на думата. Лематизаторът е базиран на електронния Граматичен речник на българския език (Коева 1998). Основната форма се определя, като се съотнася информацията, получена от тагера - част на речта и редуциран тагсет, с информацията в Граматичния речник за част на речта и пълен тагсет. На думите, които не се срещат в речника, се преписва редуцираният тагсет, получен от тагера. За анотиране на английските текстове е използвана системата Apache OpenNLP, която се разпространява безплатно от Apache Software Foundation и е достъпна на адрес: http://incubator.apache.org/opennlp/. OpenNLP представлява система от инструменти за обработка на естествени езици и включва инструменти за сегментация на изречения, за токънизация, за тагиране по част на речта, за маркиране на наименования, за синтактично парсиране и други. OpenNLP предоставя някои готови тренирани модели за извършване на гореизброените анализи, както и позволява трениране на собствени модели. Системата по същество е езиково независима и може да бъде използвана за произволен език.

На английските текстове от всеки паралелен корпус е приложена следната обработка:

- Сегментиране по изречения;
- Токънизация;
- Тагиране по части на речта.

За нашите цели са използвани предоставените от Apache готови модели за английски език. За тагирането по части на речта OpenNLP ползва системата от тагове на Penn Treebank, която може да се намери на адрес: http://www.comp.leeds.ac.uk/amalgam/tagsets/upenn.html .

7. Авторски права

Законът за авторското право и сродните му права гласи:

Чл. 24. (*Изм. - ДВ, бр. 77 от 2002 г., в сила от 01.01.2003 г.) (1) Без съгласието на носителя на авторското право и без заплащане на възнаграждение е допустимо:*

9. (изм. – ДВ, бр. 99 от 2005 г., в сила от 10.01.2006 г.) възпроизвеждането на вече публикувани произведения от общодостъпни библиотеки, учебни или други образователни заведения, музеи и архивни учреждения, с учебна цел или с цел съхраняване на произведението, ако това не служи за търговски цели;

11. предоставянето на достъп на физически лица до произведения, намиращи се в колекциите на организации по смисъла на т. 9, при условие че се извършва за научни цели и няма търговски характер;

Създаването и използването на паралелните корпуси не нарушава Закона за авторското право и сродните му права, тъй като: (1) включването на текст в корпуса не е преиздаване; (2) корпусите не се използват с комерсиална цел, а изключително за изследователски и учебни цели; (3) в описанието на всеки текст са включени библиографски данни за автора и изданието (или електронния източник) на текста, ако такива са били достъпни; (4) при ползване на други корпуси е посочен източникът и са спазени изискванията за цитиране на съставителя.

8. Заключение

Многоезиковите паралелни корпуси са много богати на информация и показват как дадена езикова система си взаимодейства с друга езикова система при превод. Корпусите са маркирани с различен тип лингвистична анотация, която е въведена чрез автоматично определяне на границите и вида на единиците от текста, автоматично определяне на частите на речта и основните форми и паралелно съотнасяне на многоезиковите корпуси по думи и изречения.

Паралелните корпуси на два или повече езика са изключително полезен езиков ресурс особено при вероятностните подходи за обработка на езиковите данни, които се използват при задачи като извличане на информация, категоризация на документи, автоматичен превод и др. Тъй като компютърната обработка на езика не се ограничава до определен тип тематични области, задачата е да се създадат паралелни корпуси между български и основните европейски, южнославянски и балкански езици, които да са структурирани, така че пропорционално да съдържат различни типове документи.

ЛИТЕРАТУРА

Коева 1998: *Коева, С.* Граматичен речник на българския език. Описание на концепцията за организацията на лингвистичните данни. – БЕ № 6, с. 49-58.

ATKINS AT AL 1992: *Atkins, Sue, Jeremy Clean and Nicholas Ostler*. Corpus Design Criteria. – In: *Literary & Linguistic Computing*, Vol. 7 (1992), pp. 1-16 .

BURNARD, L. 2007 (ed) Reference Guide for the British National Corpus (XML Edition) http://www.natcorp.ox.ac.uk/docs/URG/

DIMITROVA AT AL. 1998: *Dimitroova L., Ide, N., Petkevic, V., Erjavec, T., Kaalep, H.J., Tufiş, D.* In Christian Boitet and Pete Whitelock (eds.), *Proceedings of the Joint 17th International Conference on Computational Linguistics 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 1998)*, Montreal, Canada, August 1998. Morgan Kaufmann Publishers. pp. 315-319

DIMITROVA ET AL. 2009: *Dimitrova, L., Koseska, V., Roszko, D., Roszko, R*. Bulgarian-Polish-Lithuanian Corpus – Current Development. – In: *Proceedings of the International Workshop "Multilingual resources, technologies and evaluation for Central and Eastern European languages" in conjunction with International Conference RANPL'2009*. Borovec, Bulgaria, 17 September 2009.

EAGLES 1996B: *Recommendations for the Morphosyntactic Annotation of Corpora. EAGLES Document EAG-TCWG-MAC/R*. Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale.

GHOULI AT AL. 2009: *Ghouli V., Simov, K., Glaros, N., Osenova, P.* A web-enabled and speech-enhanced parallel corpus of Greek-Bulgarian cultural texts. In: – *EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*, pp. 35-42.

GIM'ENEZ, J. & M'ARQUEZ, L. 2004. Svmtool: A general pos tagger generator based on support vector machines. In Proceedings of the 4th LREC Conference.

GRØNN, A. & MARIJANOVIC, I. 2010. (eds.). Russian in Contrast, Oslo Studies in Language 2(1). pp. 1-24.

KOEVA ET AL. 2010: *Koeva, S., Blagoeva, D., Kolkovska, S.* Bulgarian National Corpus Project, In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), pp. 3678-3684.

KOEHN, PH. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation, Proceedings of MT Summit, pp. 79-86.

LEECH, G. 1977. Introducing corpus annotation. – In: *Garside R., Leech, G., McEnery, A. M.* (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman.

MCENERY, A. M., WILSON, A. 2001. Corpus Linguistics. An Introduction. Edinburgh, Edinburgh University.

TYERS, F. M. & ALPEREN, M. S. 2010. "South-East European Times: A parallel corpus of the Balkan languages". – In: *Proceedings of the Workshop on Exploitation of multilingual resources and tools for Central and (South) Eastern European Languages, LREC 2010*. pp. 49-53

TIEDEMANN, J. 2009. News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces. – In: *N. Nicolov and K. Bontcheva and G. Angelova and R. Mitkov* (eds.) Recent Advances in Natural Language Processing (vol V), pages 237-248, John Benjamins, Amsterdam/Philadelphia, 2009.

TUFI ET AL. 2009: *Tufiş, D., Koeva, Sv., Erjavec, T., Gavrilidou, M., Krstev., C.* ID 10503 Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages. – In: *Scientific results of the SEE-ERA.NET Pilot Joint Call*, Jana Machaov, Katarina Rohsmann (eds.), Centre for Social Innovation Vienna, Austria, pp. 37-48.

WYNNE, M. 2005. (editor) Developing Linguistic Corpora: a Guide to Good Practice. Oxford: Oxbow Books, 2005. Available online from http://ahds.ac.uk/linguistic-corpora/ [Accessed 02/08/2011].

WALDENFELS, R. 2006. Compiling a parallel corpus of slavic languages. Text strategies, tools and the question of lemmatization in alignment. – In: *Brehmer, B., Zdanova, V., Zimny, R.* (Hrsg.); *Beiträge der Europäischen Slavistischen Linguistik* (POLYSLAV) 9. München, pp. 123-138.

✉ *Проф. д-р Светла Коева*
✉ *Ивелина Стоянова*
✉ *Росица Декова*
Секция по компютърна лингвистика
Институт за български език „Проф. Л. Андрейчин" при БАН
бул. „Шипченски проход" 52, бл. 17, 1113 София, България
svetla@dcl.bas.bg
iva@dcl.bas.bg
rosdek@dcl.bas.bg

✉ *Prof. Svetla Koeva, PhD*
✉ *Ivelina Stoyanova*
✉ *Rositsa Dekova*
Department of Computational Linguistics
Institute for Bulgarian Language, Bulgarian Academy of Sciences
52 Shipchenski prohod, Bl. 17, Sofia 1113, Bulgaria
svetla@dcl.bas.bg
iva@dcl.bas.bg
rosdek@dcl.bas.bg

БЕЛЕЖКИ

[1] Други изследователски центрове, включително Секцията по компютърна лингвистика, също събират тези документи.

[2] http://www.lsi.upc.es/~nlp/SVMTool/