ЕЗИКОВИ РЕСУРСИ И ТЕХНОЛОГИИ ЗА БЪЛГАРСКИ ЕЗИК

АКАДЕМИЧНО ИЗДАТЕЛСТВО
„Проф. МАРИН ДРИНОВ“

BULGARIAN ACADEMY OF SCIENCES

INSTITUTE FOR BULGARIAN LANGUAGE "PROF. LYUBOMIR ANDREYCHIN"

LANGUAGE RESOURCES AND TECHNOLOGIES FOR BULGARIAN LANGUAGE

Editor *Svetla Koeva*

Sofia • 2014

Prof. Marin Drinov Academic Publishing House

БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ
ИНСТИТУТ ЗА БЪЛГАРСКИ ЕЗИК „ПРОФ. ЛЮБОМИР АНДРЕЙЧИН“

ЕЗИКОВИ РЕСУРСИ И ТЕХНОЛОГИИ ЗА БЪЛГАРСКИ ЕЗИК

Съставител *Светла Коева*

СОФИЯ • 2014

АКАДЕМИЧНО ИЗДАТЕЛСТВО
„Проф. МАРИН ДРИНОВ“

Сборникът „Езикови ресурси и технологии за български език“ съдържа студии и статии, които представят научни и научноприложни резултати, получени в рамките на участието на Секцията по компютърна лингвистика към Института за български език при БАН в международните проекти *CESAR: Централно- и южноевропейски езикови ресурси* и *ATLAS: Система за създаване и поддържане на съдържание, базирана на езикови технологии*.

Представени са разнообразни езикови ресурси (едно- и многоезикови корпуси, анотирани и паралелни корпуси, речници, лексикално-семантични мрежи и др.) и програми за обработка на езика. Езиковите ресурси и технологии са приложими в разнообразни социално ориентирани софтуерни решения и технологични продукти – говорещи устройства за незрящи; програми за автоматично резюмиране на множество документи, написани на различни езици; приложения за гласово търсене; програми за автоматичен превод в специализирани тематични области и за разнообразни двойки езици; интелигентни асистенти за извличане на важна информация от интернет, съобразена с различни потребителски интереси, и много други.

Научни редактори: *Светла Коева, Диана Благоева*

ISBN 978-954-322-797-6

Съдържание

Българско-английски паралелен корпус със съотнесени (прости) изречения

Екатерина Търпоманова, Цветана Димитрова***

Софийски университет „Св. Климент Охридски“*,
Институт за български език – БАН**

Abstact: This paper presents the Bulgarian-English Sentence- and Clause-Aligned Corpus (BulEnAC) that is an excerpt from the Bulgarian-English Parallel Corpus – a part of the Bulgarian National Corpus (BulNC). Both the Bulgarian and the English parts of the corpus were automatically sentence-split and sentence-aligned, with splitting and alignment manually verified by experts. The clause boundaries in English were determined by an OpenNLP parser, and the annotation was manually post-edited. The Bulgarian sentences were split into clauses manually. The clause-introducing conjunctions were identified, with the type of relation, the type of the clauses involved in the relation, and the direction of the relation. Finally, the parallel clauses within corresponding sentence pairs have been manually aligned. The paper discusses the conventions followed for sentence and clause splitting and annotation. The annotation is in compliance with the specific syntactic rules and the grammar tradition and annotation practices for the respective languages.

1. ВЪВЕДЕНИЕ

Паралелният корпус представлява колекция от текстове с еднакво съдържание на два или повече езика. Текстовете може да са създадени на единия език и да са преведени на един или повече езици, но най-често всички са преводни. Паралелният корпус може да включва текстове само на два езика (*Чешко-английският паралелен корпус*[1], *Hunglish*[2], *Полско-руският паралелен корпус*[3]), но много корпуси включват текстове на повече от два езика (*OPUS*[4]; *Europarl*[5]; *JRC-Acquis Multilingual Parallel Corpus*[6]; *MultiUN*[7]). Посоката на превода не е постоянна и може и да не е известна (например текстовете да са компилация от преводни и оригинални текстове). Много от паралелните корпуси са събирани чрез автоматично обхождане на уебстраници (кроулиране) и сваляне на текстове от тях или чрез ръчно сваляне на публично достъпни материали от интернет. Желателно е текстовете да покриват различни стилове, тематика и жанр, но много от познатите ни паралелни корпуси се ограничават до няколко тематични области заради ограничения достъп до паралелни данни (Коева и др. 2012а).

[1] http://ufal.mff.cuni.cz/czeng/czeng10/
[2] http://mokk.bme.hu/resources/hunglishcorpus/
[3] http://www.pol-ros.polon.uw.edu.pl/
[4] http://opus.lingfil.uu.se/
[5] http://www.statmt.org/europarl/
[6] http://ipsc.jrc.ec.europa.eu/index.php?id=198
[7] http://www.euromatrixplus.net/multi-un/

Паралелните текстове може да бъдат анотирани, сегментирани и съотнесени на ниво дума, фраза и изречение (Коева и др. 2012в; Коева и др. 2012д).

Корпусите с паралелни текстове намират широко приложение в компютърната лингвистика като източник на тренировъчни и тестови данни при създаването на редица приложения в сферата на езиковите технологии. В същото време анотацията, сегментирането и съотнасянето на езиковите данни увеличават приложимостта им.

В статията се разглежда един от частично анотираните подкорпуси на *Българския паралелен корпус* (част от *Българския национален корпус* – по-нататък БНК) – *Българско-английският паралелен корпус със съотнесени (прости) изречения (БулЕнАК)*[8], като се представят принципите на частичната синтактична анотация и сегментация, които са следвани при анотирането му (Коева и др. 2012в; Коева и др. 2012г; Коева и др. 2012д).

2. ПАРАЛЕЛНИ РЕСУРСИ ЗА БЪЛГАРСКИ ЕЗИК

2.1. Българският паралелен корпус

Паралелният корпус в състава на БНК включва 47 паралелни корпуса на различни езици, като задължително единият от двойката паралелни корпуси е с текстове на български. Паралелните корпуси се различават по големина и по състав (присъствие на текстове от различни типове), а разнообразието на текстовете се определя от достъпните в интернет материали. Всеки паралелен корпус се състои от текстове, които имат българско съответствие, като българският текст може да бъде оригинал, превод от другия език или превод от трети език. Паралелните корпуси са неделима част от БНК и следват неговия модел. Текстовете са снабдени с подробни метаданни, които най-често са извлечени автоматично и при необходимост са обработени ръчно. Структурата на всеки паралелен корпус отразява структурата на ядрото на БНК – едноезиковия български корпус, като повтаря класификацията по стил, тематична област и жанр. Паралелните корпуси непрекъснато се увеличават и обогатяват, като общата големина на включените в тях текстове е над 4,2 милиарда токъна. Най-големият паралелен корпус в БНК е *Българско-английският паралелен корпус*, който съдържа около 280 милиона токъна за език. Шест корпуса са с големина от 200 до 250 милиона токъна, 14 корпуса са по 150–200 милиона, три са между 100 и 150 милиона, 11 са между 1 и 15 милиона и 15 корпуса са под един милион токъна.

В *Българско-английския паралелен корпус* една част от текстовете са токънизирани, сегментирани и частично анотирани, като някои влизат в състава на *Българско-английския паралелен корпус със съотнесени (прости) изречения (БулЕнАК)*.

---

[8] Статията представя резултата от работата по *Българско-английския паралелен корпус със съотнесени (прости) изречения (БулЕнАК)* на следния екип: Светла Коева, Ивелина Стоянова, Светлозара Лесева, Борислав Ризов, Ангел Генов, Екатерина Търпоманова, Росица Декова, Цветана Димитрова, Христина Кукова.

2.2. Анотиране на паралелните ресурси

Паралелните текстове могат да бъдат сегментирани и съотнесени на ниво дума и фраза, но най-честото равнище на сегментация е изреченското, тъй като именно на това равнище се реализират отношенията между думите, словосъчетанията и простите изречения в състава на сложното, а явления като междуезиковите различия в словореда и във фразовата структура могат да бъдат по-добре описани и формализирани на равнището на простото, отколкото на сложното изречение.

При обработка на паралелните текстове на изреченско равнище трябва да се имат предвид фактори като дължина и сложност на изреченията и брой и относителен ред на простите изречения в състава на сложното. За наблюдение на тези фактори изреченските единици в паралелните текстове трябва да се сегментират (в рамките на едноезиковите корпуси) и да се съотнесат (като част от паралелен корпус). Допълнителната анотация на едноезиковите и на паралелните корпуси на равнището на простото изречение (например анотация на вида на съюзните връзки) подпомага работата по изграждане на приложения за автоматичен лингвистичен анализ, автоматичен превод и други подобни задачи.

Паралелните корпуси със съотнесени прости изречения се използват за трениране на модели за автоматичен превод, които превеждат изречение към изречение, както и за модели, основани върху метод на преподреждане на изреченията. Паралелни корпуси със съотнесени единици намират приложение за тестване на различни подходи, например при превод от немски към английски (Коуан, Кучерова, Колинс 2006) и др.

Корпусите със съотнесени (прости) изречения обикновено съдържат краен брой изречения в текстове, които принадлежат към определен стил, тематична област или жанр, като например биомедицински, юридически и др. По-нататъшното категоризиране на стиловете може да включва разпределение на текстовете в тематични области (например административният може да включва текстове от икономическата, юридическата и други области) и жанрове (например проза и поезия при художествената литература).

Тук ще разгледаме структурата, състава и формата на *Българско-английския паралелен корпус със съотнесени (прости) изречения*. Накратко ще опишем избрания подход за разпознаване на границите на простите изречения, включително в състава на сложните, и съпоставянето им, както и принципите на анотация на изреченското деление с оглед на вида на простите изречения и връзката между тях.

3. БЪЛГАРСКО-АНГЛИЙСКИЯТ ПАРАЛЕЛЕН КОРПУС СЪС СЪПОСТАВЕНИ (ПРОСТИ) ИЗРЕЧЕНИЯ (БулЕнАК)

3.1. Състав на БулЕнАК

Както беше посочено, *Българско-английският паралелен корпус със съотнесени (прости) изречения* е подкорпус на *Българско-английския паралелен корпус*, който е част от БНК. *Българско-английският паралелен корпус* съдържа близо 280,8 милиона токъна в 8,2 милиона изречения в българската част и 283,1

милиона токъна в 8,9 милиона изречения в английската част. Текстовете са токънизирани и лематизирани, а изреченията са сегментирани. За обработка на българския подкорпус е използвана Българската многокомпонентна система за първична обработка и лингвистична анотация на текстове[9], а за английската са използвани Apache OpenNLP и Stanford CoreNLP (Коева и др. 2012г).

БулЕнАК включва общо 366 865 токъна, от които 176 397 са в 14 667 изречения на български, със средна дължина на изречението 12,02 токъна, и 190 468 токъна – в 15 718 изречения на английски, със средна дължина на изречението 12,11 токъна. Средният брой на простите изречения в състава на сложното е 1,67 за българския подкорпус и 1,85 за английския.

Текстовете в корпуса принадлежат към различни стилове, тематични области и жанрове. Под *стил* тук разбираме комплексна категория, която се основава на взаимодействието на вътрешноприсъщите характеристики на текстовете с екстралингвистични фактори (Коева и др. 2012г). Разнообразието на езиковите данни позволява да се правят прогнози и изводи за представянето на различни методи за обработка на данни при различни текстове. По стил текстовете в БулЕнАК се отнасят към следните пет категории: административни текстове (20,5 %), художествена литература (21,35 %), публицистика (37,13 %), научни текстове (11,16 %) и разговорни/художествени текстове (9,84 %).

3.2. Формат на корпуса

Текстовите данни се съхраняват под формата на файлове в плитък xml формат, който е подходящ за представяне на дистантно разположени конституенти на фразата / простото изречение. Форматът е ефективен за представяне на анотацията и на синтактичната йерархия между двойките прости изречения чрез посочване на вида на съюзната връзка между простите изречения в състава на сложното.

Текстът е представен като поредица от xml елементи от типа <word> – вж. в (1а) представянето на изречението *Джуканович подаде оставка в понеделник, за да заеме премиерския пост* и в (1б) представянето на преводния еквивалент на английски *Djukanovic resigned as president Monday to assume the post of prime minister.*

(1а) <word cl="146091856" l="PUNCT" w="====/><word cl="146091856" l="Джуканович" w="Джуканович"/><word cl="146091856" cl_al="c7a15f60ee58459ab20072c1bcee5785" l="подам" w="подаде"/><word cl="146091856" l="оставка" w="оставка"/><word cl="146091856" l="в" w="в"/><word cl="146091856" l="понеделник" w="понеделник,"/><word cl="146091600" l="PUNCT" w="====/><word cl= "146091600" cl2="146091856" l="за" m="N_S" p="146091216:0" w="за"/><word cl="146091600" l="да" p="146091216:1" w="да"/><word cl="146091600" l="заема" w="заеме"/><word cl="146091600" l="премиерски" w="премиерския"/><word cl="146091600" cl_al="03a77b5721da4101aaa2b4bea751c3da" e="True" l="пост" sen="dfae9869f23c4a8083903940b90ad9ea" w="пост."/>

(1б) <word cl="170450380" l="djukanovic" w="Djukanovic"/><word cl="170450380" l="resign" w="resigned"/><word cl="170450380" l="as" w="as"/><word cl="170450380" l="president" w="president"/><word cl="170450380" cl_al="c7a15f60ee58459ab200

[9] http://dcl.bas.bg/DCLservices-bg.html

72c1bce e5785" l="monday" w="Monday"/><word cl="170449004" cl2="170450380" cl_al="03a77b5721da4101aaa2b4bea751c3da" l="to" m="N_S" w="to"/><word cl="170449004" l="assume" w="assume"/><word cl="170449004" l="the" w="the" /><word cl="170449004" l="post" w="post"/><word cl="170449004" l="of" w= "of" /><word cl="170449004" l="prime" w="prime"/><word cl="170449004" e="True" l="minister" sen="dfae9869f23c4a8083903940b90ad9ea" w="minister."/>

На всеки <word> елемент са приписани атрибути, съдържащи следната информация за словоформата:

1. Лексикална (равнище на лематизация) – чрез атрибутите ‘l’ и ‘m’, които посочват информация съответно за лема и словоформа, например l="подам", w="подаде" в (1а).

2. Синтактична (равнище на самостоятелното изречение) – чрез комбинацията от два атрибута e=True и sen=senID, която се появява в края на всяко изречение, като приписва уникален идентификационен номер на изречението в корпуса, например e="True" sen="dfae9869f23c4a8083903940b90ad9ea" в (1а).

3. Синтактична (равнище на простото изречение в състава на сложното) – атрибутът cl съдържа идентификационния номер на простото изречение, в което се появява дадена словоформа – в (1а) има две прости изречения: *Джуканович подаде оставка в понеделник* – cl="146091856", и *за да заеме премиерския пост* – cl="146091600".

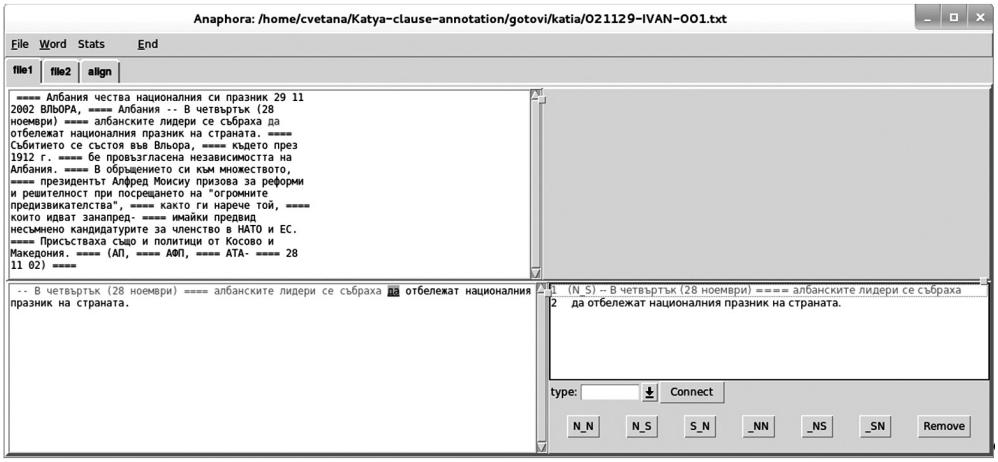
4. Синтактична (свързващи елементи – подчинителни и съчинителни връзки) – атрибутът cl2 се включва в елемента <word> на съюзни думи и фрази, които свързват две прости изречения в състава на сложното. Стойността на атрибута е идентификационният номер на простото изречение, с което е свързано изречението, в което се намира съответната съюзна дума – в примера в (1а) *за да* съдържа атрибути cl="146091600" и cl2="146091856". Атрибутът m отбелязва вида на връзката между двете прости изречения cl и cl2 (съответно – съчинено свързване или координация и подчинено свързване или субординация), посоката на връзката и позицията на свързващия елемент спрямо други прости изречения – в (1а) *за да* има атрибут m="N_S". Допълнителен атрибут p се използва за сложни съюзи, той получава еднаква стойност за двата елемента, като към първия се добавя 0, а към втория – 1, например в (1а) *за* – p="146091216:0", *да* – p="146091216:1".

5. Съотнасяне – атрибутите sen и cl_al отбелязват съответно съотнесени самостоятелни изречения и съотнесени прости изречения. Съотнесените изречения в два паралелни текста имат един и същ идентификационен номер, както се вижда в (1а) и (1б) – sen="dfae9869f23c4a8083903940b90ad9ea". Съотнесените прости изречения в състава на сложното също получават една и съща стойност – cl_al="c7a15f60ee58459ab20072c1bcee5785" за първите съотнесени прости изречения (*Джуканович подаде оставка в понеделник* се съотнася с *Djukanovic resigned as president Monday*) и cl_al="03a77b5721da4101aaa2b4bea751c3da" – за вторите (*за да заеме премиерския пост* се съотнася с *to assume the post of prime minister*).

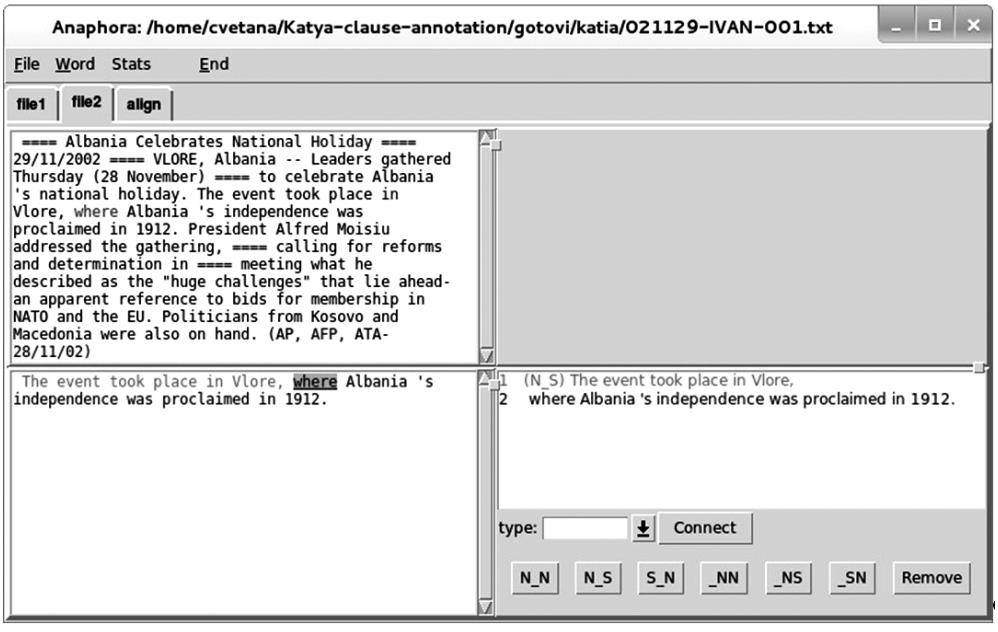
Празните думи (w="====") са формални елементи, които се появяват в началото на простото изречение, ако свързващият елемент не е експлицитно изразен или ако връзката е изразена чрез пунктуационен знак. Пунктуационните знаци не са отделени като самостоятелни токъни, а са анотирани заедно с предхождащия ги токън.

3.3. Програма за анотация

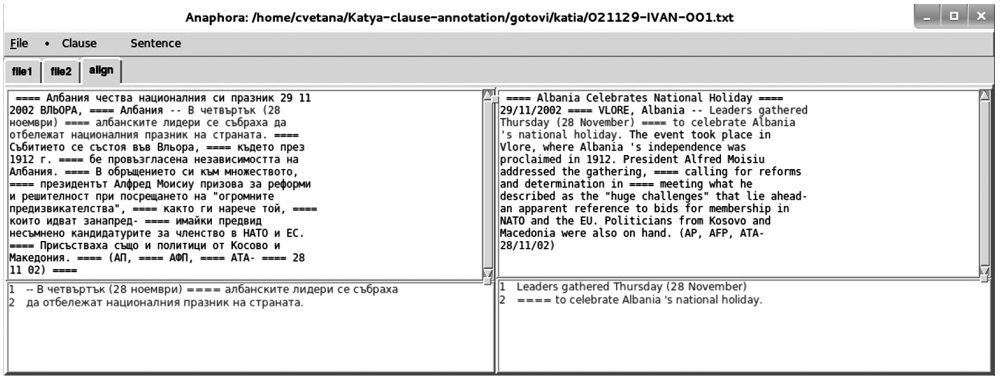
Ръчното съотнасяне и анотация на изреченските единици, както и проверката и редактирането на резултатите от автоматичното съотнасяне се извършват чрез специално приложение – ClauseChooser, разновидност на програмата за анотация Chooser (Ризов 2010). То поддържа два оперативни режима: едноезиков – за ръчно редактиране и анотация на текстовете от паралелния корпус (на български и на английски), и многоезиков – за съотнасяне на паралелните единици.



Фиг. 1. Едноезиков режим на ClauseChooser (български текст)



Фиг. 2. Едноезиков режим на ClauseChooser (английски текст)



Фиг. 3. Режим на ClauseChooser за съотнасяне на паралелните изреченски единици

Едноезиковият режим позволява: разделяне на изреченските единици (включително на простите изречения в състава на сложното); корекция на грешно разделени изреченски единици (чрез сливане или разделяне); анотация на съюзни връзки (включително на вида на съюзната връзка между изреченията).

Фигура 1 и фигура 2 представят едноезиковия режим на ClauseChooser за сегментация на изреченските единици и за анотация на съюзните връзки между простите изречения в състава на сложното. Режимът във фигура 1 е с български текст, а във фигура 2 – с английски текст. В левия долен панел се разделят изреченските единици, а резултатът от сегментацията се отразява в десния долен панел. Видът на съюзната връзка се посочва чрез сивите бутони N_N, N_S и др.

Многоезиковият режим, илюстриран във фигура 3, използва резултата от едноезиковата сегментация на изреченските единици, като позволява ръчно съотнасяне на изреченските единици и ръчно съотнасяне на простите изречения в състава на сложното.

4. АНОТАЦИЯ НА БулЕнАК

4.1. Общи принципи на анотацията на синтактично равнище

Частичната синтактична анотация на БулЕнАК включва:

а) определяне на границите на самостоятелното изречение и на простите изречения в състава на сложното;

б) определяне на вида на съюзната връзка (съчинителна или подчинителна) между простите изречения в състава на сложното;

в) идентификация на лингвистичните маркери, които въвеждат простите изречения в състава на сложното – съчинителни и подчинителни съюзи, съюзни наречия, местоименни съюзи, пунктуационни знаци и др.

Изреченските единици в корпуса са сегментирани и съотнесени според предварително уговорени конвенции (които ще разгледаме по-подробно в 4.2.), като всяко просто изречение в състава на сложното е свързано с друго просто изречение. Анотират се съюзната връзка и видът на изреченията едно спрямо друго. В примерите от анотацията с N (nuclear) се отбелязват самостоятелните (прости)

изречения или главните изречения в състава на сложното, а със S (satellite) – подчинените изречения (спрямо главно N). Така конструкция от типа N_N означава две съчинено свързани прости изречения в състава на сложното (отбелязани с удебелен шрифт са съюзните думи; при безсъюзно свързване връзката между изреченията се маркира с празна дума в анотацията и във файла, а в примерите за илюстрация липсва подчертан елемент). N_S отбелязва конструкция с първо главно (N) и второ подчинено (на N) изречение (S), като подчинителният съюз е отбелязан с удебелен шрифт (например *докато* и *която* в (2). Дадено просто изречение може да е подчинено (S) спрямо главно просто изречение (N), но да е главно (N) спрямо друго просто изречение (S). В (2) *докато тази страна не получи правна система* е подчинено (S1) на (N1) *Не можем да почиваме*, но е главно (N2) спрямо *която вкарва престъпниците в затвора* (S2).

В (2) е приведен пример за съюзно и безсъюзно свързване в рамките на едно изречение: (б) е подчинено на главното (а), въведено от съюза *докато*, а (в), (г), (д) и (е) са съподчинени на (б), като само първото от тях е въведено от относителното местоимение *която*, а следващите са свързани безсъюзно.

(2) (а) [N1 Не можем да почиваме, N1_S1
(б) [S1/N2 докато тази страна не получи правна система, N2_S2
(в) [S2/N3 която вкарва престъпниците в затвора,] N3_N4
(г) [N4 защитава невинните,] N4_N5
(д) [N5 привлича чуждестранни инвестиции] N5_N6
(е) [N6 и стимулира местния бизнес.]]]

4.2. Анотационни конвенции

При анотацията на изреченията в български и в английски[10] се спазват възприетите синтактични правила за всеки от езиците според избрани официални граматики (за български – ГСБКЕ 1983а, ГСБКЕ 1983б; за английски – Даунинг, Лок 2002). Анотационни конвенции се прилагат, за да бъде постигнат единен модел на анотация, включително при спорни случаи, които не са еднозначно разрешени в официалните граматики. Основният подход, който възприемаме в такива случаи, е възстановяването на смисловите отношения в изречението, когато те не са експлицитни.

4.2.1. Безглаголни изречения

Основният признак на изречението е предикативността и според редица синтактични теории глаголът е ядрото на всяко изречение. Безглаголни са изреченията с елипса на глагол. Приемането им за изречения е въз основа на експлицитното изразяване на дадена същина, която предполага наличие на предикативен признак. В безглаголни изречения, които се състоят от (опорна дума) съществително, прилагателно, наречие, числително, местоимение,

[10] В статията повечето примери са на български, примери от английските текстове се дават при разглеждането на междуезиковата асиметрия и принципите, следвани при съотнасянето. За повече примери от английските текстове вж. Коева и др. 2012д.

липсващият глагол, най-често глагол за екзистенция, може да бъде възстановен от контекста.

(3) а. [N1 Ръка, N1_S1
[S1 която се протяга с отчаянието на удавник над струпаните бели престилки.]] < *Това е ръка, която ...*[11]
б. [N Долу!] < *Долу е.*

Безглаголните изречения се анотират като изречения независимо от това дали са самостоятелни, или са прости изречения (главни или подчинени) в състава на сложно изречение.

(4) а. [N1 Най-добре N1_S1
[S1 да проверим.]] < *Най-добре е да проверим.*
б. [N1 Каква полза N1_S1
[S1 да я плаши още повече.]] < *Каква полза има да я плаши още повече.*
в. [N1 Съществуващите квоти ще бъдат увеличени с 410 000 литра, N1_S1
[S1 от които 200 000 за Испания.]] < *от които 200 000 са за Испания.*

Самостойно изречение може да бъде образувано и от междуметие или частица. Анотира се като отделно изречение, ако се състои само от една дума (междуметие или частица) с пунктуационен маркер за край на изречение. Статутът на междуметията и частиците в състава на изречението е описан по-нататък.

(5) а. [N Ох!]
б. [N – Имаш ли часовник?]
[N – Да.]

4.2.2. Пряка реч

Основният проблем тук е типът на свързване между простите изречения в пряката и авторовата реч. При свързването се следва концепцията за смисловите връзки между простите изречения, макар че те не са експлицитни. Изхожда се от трансформацията на пряката реч в непряка. При такива трансформации авторовите думи формират главното изречение, а цитираната реч – подчиненото.

(6) [[S1/N2 Партиите, _S1N1
[S2 които спечелиха настоящите избори в Босна и Херцеговина] N2_S2
S1/N2 трябва да реформират централното правителство и системата на митниците] N2_S3
[S3/N3 ако не искат N3_S4
[S4 да загубят чуждестранната помощ и потенциалните инвеститори]]
N1 каза Върховният представител Пади Ашдаун.]

[11] В курсив предаваме възстановени конструкции.

При повече от едно изречение в цитираната реч думите на автора се свързват с изречението от пряката реч, което е в най-близко съседство, обикновено това, с което са свързани и пунктуационно. Предходните и следващите изречения от пряката реч се маркират самостоятелно, без връзка с авторовите думи.

(7) *Проклети идиоти! До един продават хероин и се тъпчат с него – заяви шофьорът. – Остави се! Не е за разправяне!>*

[N – Проклети идиоти!]
[[S1/N1 До един продават хероин] _S1N1
[N2 и се тъпчат с него] – N1_N2
N1 заяви шофьорът.] –
[N Остави се!]
[N Не е за разправяне!]

Ако в пряката реч липсва глагол, тя се смята за просто безглаголно изречение и се свързва като подчинено изречение с авторовата реч:

(8) [[S – Ооо!] – N възкликна Анди смутен.] _SN

Най-често авторовите думи се въвеждат от глагол от семантичното поле за речева дейност или мисловен процес, което улеснява трансформацията от пряка в непряка реч и смисловата връзка между главното и подчиненото изречение. В художествените текстове обаче се срещат и авторови думи, които представляват отделно изказване и при които липсва директна логическа връзка с пряката реч. В тези случаи се следва правилото за свързване на пряка и авторова реч в отношения на подчинено и главно изречение, като мотивът е възстановимостта на глагола за речева или ментална дейност в смисловата структура на изречението.

(9) [[[S1 – Ти] – _S1N1
N1 мисълта му се изгуби] N1_N2
[N2 и остана само болката.]]
< – Ти – (каза той) и мисълта му се изгуби и остана само болката.

4.2.3. Обособени части и вметнати изрази

Обособени части или вметнати изрази, които не съдържат глагол, се анотират като част от изречението, към което се отнасят.

(10) [N Детето, червено като домат, се разплака.]

Вметнати изрази, които съдържат глагол, се обособяват като прости изречения и се свързват с останалите прости изречения според логическата връзка между тях. Вътрешната синтактична структура на вметнатия израз също се отразява в анотацията според това колко глагола съдържа.

(11) [N1 Книгата, N1_S1
[[S1/N2 щеш] N2_N3
[N3 не щеш]],
N1 ще я вземеш.]

Вметнати изречения се анотират отделно, без да се маркира връзка с изречението, в което са вмъкнати, ако представляват цялостни изказвания, които могат да функционират самостоятелно, и ако не може да се установи пряка логическа и синтактична връзка между двете изречения.

(12) *И ето го тук, предал документа, изчакал цигароубиеца Уонлес да си иде (вярно, че напомня на доктора от оня циклопски филм), отговаря на въпроси за ...*
> [N1 Вярно, N1_S1
[S1 че напомня на доктора от оня циклопски филм.]]

4.2.4. Модални и фазови глаголи

Модалните и фазовите глаголи са част от съставното сказуемо и не се анотират отделно. Модалните изрази от типа *необходимо е*, *нужно е*, *възможно е* и др., които са синонимни на модалните глаголи в българския език, се анотират отделно като главно изречение с подчинено, въведено от съюза *да*.

(13) а. [N Трябва да вървим.]
б. [N1 Възможно е N1_S1
[S1 да дойдат.]]

4.2.5. Безсъюзно свързване

При безсъюзно свързване на простите изречения в сложното основният проблем е да се определи типът връзка – съчинителна или подчинителна. Свързването се определя според контекста и възможните трансформации в съюзно въведени прости изречения.

(14) а. [[[S1 Ако е така,] _S1N1
N1 то всичко им е пределно ясно:] N1_S2
[[S3 щом Анди се сдобие с някакви пари,] N2_S3
N2/S2 за известно време странните неща престават.]]
< *им е пределно ясно, (че) ...*

б. [N1 Събуждай се, N1_S1
[S1 пристигнахме.]
< *защото пристигнахме*

в. [[N1 Хвърли поглед към Вики] – N1_N2
[N2 тя се бе вторачила в него с широко отворени и изплашени очи.]]
< *... а тя ...*

4.2.6. Частици

Като служебни думи частиците се характеризират с липса на самостоятелност, отнасят се към дума или фраза в изречението или към цяло просто изречение (ГСБКЕ 1983а: 476). Анотират се като част от изречението, в чийто състав влизат.

(15) [N1 – И да не си мисли, N1_S1
[S1 че е имало туй-онуй,]
а?]

4.2.7. Междуметия

Употребата на междуметията в изречението е различна в сравнение с останалите части на речта. Междуметията имат отношение към съдържанието на изречението, но нямат синтактична връзка с останалите му конституенти (ГСБКЕ 1983а: 468). Анотират се като част от простото изречение, с което са смислово свързани:

(16) [[N1 О, морската трева е готина,] N1_N2
[N2 морската трева е купон.]]

Глаголите, употребени като междуметия, се анотират като отделно просто изречение. Основание е запазената синтактична връзка с останалите прости изречения в сложното. Типът свързване се избира според синтактичните отношения между глагола междуметие и другото просто изречение. При *моля*, *заповядай*, *кажи* и под. връзката е подчинителна, тъй като простото изречение, което се отнася към глагола междуметие, е във функция на комплемент (*моля да*, *заповядай да*, *кажи дали*). Глаголи като *виж*, *слушай*, *чуй*, употребени като междуметия, се свързват съчинително със съседното просто изречение.

(17) а. [[S1 – Стиснете си юмрука, _SN
N1 моля.]

б. [[N1 – Слушай, Сали, N1_N2
[N2 разбирам N2_S1
[S1 как се чувстваш,]]
[N3 но всичко ще ти обясня.]] N2_N3

4.2.8. Граница на просто изречение

Проблеми с поставянето на граница на простото изречение възникват в случаите, когато след преходен глагол в главното изречение е употребена именна група, за която има колебание дали е пряко допълнение в главното, или подлог в подчиненото изречение. Синтактичната двузначност се разрешава чрез тест със заместване с местоимение. Ако именната група може да се замести с кратка винителна форма на личното местоимение, тя е допълнение в главното изречение – като в (18).

(18) [N1 Папата призова босненците N1_S1

[S1 да обърнат гръб на насилието...]]

< *Папата **ги** призова да обърнат гръб на насилието...*

Ако заместването е с именителна форма на личното местоимение, а цялото подчинено изречение се замества с местоимението *това*, именната група след преходния глагол е подлог в подчиненото изречение и се анотира като част от него – като в (19).

(19) [N1 ЕС иска N1_S1

[S1 Швеция да гласува „за“.]] –

< **ЕС **я** иска да гласува „за“.*

< *ЕС иска **тя** да гласува „за“.*

< *ЕС иска **това**.*

Някои глаголи, например перцептивните, допускат и двете интерпретации, когато именната група не е изразена чрез лично местоимение. Например изречението в (20) може да бъде трансформирано по два начина:

(20) Видях студентите как влизат. > *Видях ги как влизат.*

> *Видях те как влизат.*

Съществителното *студентите* след перцептивния глагол може да бъде анализирано и като пряко допълнение в главното изречение, и като подлог в подчиненото.

4.3. Сегментиране и съотнасяне

Самостоятелните изречения и простите изречения в състава на сложните са сегментирани и съотнесени с помощта на приложения за автоматична обработка. Следват се официалните граматики на двата езика и анотационните конвенции. При превода понякога се наблюдава преразпределение на съдържанието между различни изречения, като на едно изречение в българския текст може да съответстват две или повече изречения в английския и обратно или съдържанието да се преразпредели между простите изречения в рамките на сложното. Практиката на съотнасяне в подобни случаи следва съдържанието, т.е. съотнасят се изреченията, които съдържат съответстващи си предикати и по-голямата част от информацията.

4.3.1. Сегментиране на самостоятелни изречения

Изреченията в българските текстове са сегментирани посредством съответното приложение от Българската многокомпонентна система за първична обработка и лингвистична анотация на текстове, което използва регулярни правила и лексикон. Изреченията в английските текстове са разделени чрез приложението OpenNLP с предварително тренирани модели, като сегментацията е проверена от експерти. Заглавия на текстове, както и други изрази от този тип (наименования на илюстрации, фигури, таблици; дати и източници на информация в новинарски

текстове и документи; рубрики в документи и др.) са анотирани като отделни изречения и са съотнесени със съответните изрази в паралелния текст.

4.3.2. Съотнасяне на самостоятелни изречения

Самостоятелните изречения са съотнесени автоматично чрез HunAligh (Варга и др. 2005; за повече подробности вж. Коева и др. 2012г), като получените двойки изречения са проверени и редактирани от експерти.

Преобладаващият модел при съотнесените изречения е 1:1, т.е. изреченията са напълно паралелни като в (21).

(21) а. [[S1 Мечтите не стават за ядене], N1 каза жената.] _S1N1
b. [[S1 You can't eat hope], N1 the woman said.] _S1N1

Съотнасяния 0:1 и 1:0 се появяват в случаи, в които на изречение в единия език не отговаря паралелно изречение от другия език (например ако изречението не е преведено).

Съотнасяния1:2 и 2:1 обхващат случаи, в които една и съща информация е предадена в едно изречение в единия език и в две – в другия, като в (22), където английското безсъюзно свързано изречение *a change that will not lead to reduced combative operability* се предава на български със самостоятелното изречение *Тази замяна няма да доведе до намалена бойна оперативност*.

(22) а. [N1 Съществуват планове N1_S1
[[S1/N2 Македония да извади от употреба четирите си изтребителя Сухой Су-25] N2_N3
[N3 и да ги замени с транспортни и бойни хеликоптери.]]]
[N *Тази замяна няма да доведе до намалена бойна оперативност.*]

б. [[N1 There are also plans for Macedonia N1_S1
[[S1/N2 **to** discard its four Suhoi Su-25s] N2_N3
[N3 **and** replace them with a transport and combat helicopters]] N3_S2
[S2/N4 – *a change* N4_S3
[S3 *that will not lead to reduced combative operability.*]]]]

В таблица 1 е показано разпределението на моделите на съотнасяне (български към английски), като в категорията *Други* са модели с ниска честота (1:3, 3:1, 2:2 и пр.).

Таблица 1
Модели на съотнесени изречения

BG:EN	Брой в корпуса	Процент
0:1	1187	7,60
1:0	225	1,44
1:1	13 697	87,74
1:2	294	1,69
2:1	187	1,20
Други	15	0,33

4.3.3. Сегментиране на простите изречения в състава на сложното

Границите на простите изречения в състава на сложното в английските текстове са определени чрез предварително трениран OpenNLP парсер, след което сегментацията е проверена от експерти. В българския подкорпус простите изречения в състава на сложното са определени ръчно от експерти. Сегментацията зависи от езика и следва синтактичните правила, както са описани в традиционните граматики и анотационните практики за съответните езици.

4.3.4. Съотнасяне на простите изречения в състава на сложното

Простите изречения в състава на сложното са съотнесени ръчно в процеса на анотация, като съотнасянето се извършва в рамките на двойка паралелни самостоятелни изречения. Преобладаващият модел при съотнесените прости изречения в състава на сложното също е 1:1.

(23) [[S1 Мечтите не стават за ядене,] _S1N1
[[S1 You can't eat hope,] _S1N1
[N1 каза жената.] [N1 the woman said.]

Наблюдават се различни случаи на асиметрия, тъй като информацията може да се разпредели в различни прости изречения. Съотнасяния 1:0 и 0:1 се наблюдават в случаи, в които просто изречение в единия език няма съответствие в просто изречение в другия език, като например в (24), където се съотнасят, както следва: първото изречение [S] от (24а) с второто [S] от (24б), второто [N] от (24а) с първото от (24б), а последното от (24а) остава без паралел (съотнасяне 1:0), защото не е преведено.

(24) а. [[S1 Докато лапаше поредната лъжица качамак,] _S1N1
[[N1 пресметна] N1_N2
[N2 и рече:]
б. [[N1 He made some calculations,] N1_S1
[S1 while he sipped a spoonful of mush.]

Съотнасяния от типа 1:N, N:1 (N>1) отразяват случаи, в които на едно просто изречение в единия език отговарят две или повече прости изречения в другия език, като в (25), където две прости изречения в състава на едно самостоятелно сложно изречение в български съответстват само на едно изречение в английски (2:1).

(25) а. [N1 Ще ни даде възможност N1_S1
[S1 да се храним в продължение на три години.]]
б. [N He'll feed us for three years.]

Модели на съотнасяне с участие на повече от едно просто изречение – N:M (N, M>1), отразяват сложно разпределение на информацията. В (26) моделът на съотнасяне е 2:2 поради невъзможност простите изречения в състава на сложното да бъдат съотнесени едно към друго. В български авторовата реч представлява главно изречение, докато същата информация в английското изречение е въведена чрез предлог и без глагол, следователно не представлява отделно изречение. Освен това обособената част с причастие в български се анотира

като част от изречението, към което се отнася, а в английски – като отделно подчинено изречение.

(26) а. [[S1Партиите, представени в новия парламент на БиХ, постигнаха _S1N1
единодушие по въпроса за планираната реформа,]
N1 уведомиха от Службата на Върховния представител.]
б. [N1The parties N1_S1
[S1 represented in the new BiH Parliament]
have reached a general consensus about the planned reform,
according to the Office of the High Representative.]

Модели на съотнасяне от типа N:M (N, M>1) са рядко срещани:

Таблица 2
Модели съотнесени прости изречения в състава на сложното

BG:EN	Брой в корпуса	Процент
0:1	1745	7,05
1:0	482	1,95
1:1	18 997	76,80
1:2	2256	9,12
1:3	239	1,33
1:4	99	0,40
2:1	621	2,51
2:2	87	0,32
Други	128	0,52

Преразпределянето на информацията между простите изречения е причина за сериозен процент от съотнасянията 0:1 (7,05 %) и 1:2 (9,12 %) от български към английски, а противоположните модели са доста по-малък процент – 1,95 % за 1:0 и 2,51 % за 2:1. Тези резултати показват силна тенденция моделът 1:N (N>1) да се среща по-често при съотнасяне от български към английски, отколкото при съотнасяне от английски към български.

5. МЕЖДУЕЗИКОВА АСИМЕТРИЯ

Междуезикова асиметрия се наблюдава в случаите, когато между двата езика има несъответствие в превода на лексикално ниво или в граматичната структура. Условно обособяваме следните типове междуезикова асиметрия: 1) различни езикови явления; 2) различен превод в двата езика; 3) към едни и същи езикови явления се прилага различен подход във всеки от езиците.

Първият тип е илюстриран с възможността да се пропусне съюзът *that* в английски при трансформиране на пряка реч в непряка, когато в български на него съответства съюзът *че* – вж. първите подчинени изречения [S] в (28). Английското *that* във функция на относително местоимение, което

е допълнение в подчиненото изречение, също се изпуска свободно, докато в български относителното местоимение е задължително – вж. (27). Асиметрията се изразява в различен тип свързване – безсъюзно в английски и съюзно в български.

(27) а. [N1 Големите промени в конституцията, N1_S1
[S1/N2 **които** правителството планира N2_S2
[S2 да извърши,]]
имат за цел N1_S3
[S3 да изпълнят критериите на ЕС.]]

б. [N1 The broad constitutional reforms N1_S1
[S1/N2 the government is planning N2_S2
[S2 to carry out]]
are aimed N1_S3
[S3 at meeting the EU criteria.]]

Друга проява на този тип асиметрия е различната лексикализация в двата езика – на израз, съдържащ глагол в единия език, в другия съответства дума, различна от глагол, при което се получава несъответствие в броя на простите изречения. В (28) на глаголната фраза *did you* в български съответства частица.

(28) а. [[N1 … не мислеше, N1_S1
[S1 че ще получиш целия океан,]
нали?]
б. [[N1 … you didn't think N1_S1
[S1 you'd get the whole ocean,]]
[N2 *did you?*]] N1_N2

Вторият тип различия са свързани с избора на различни лексикални или граматични средства от преводача, въпреки че съществува възможност за идентичен превод. С висока честота е превеждането на глагол със съществително, като в английски се предпочита глагол, а в български – съществително, макар че се срещат и обратни примери. Резултат от това несъответствие е асиметрията при съотнасянето – различен брой прости изречения в състава на сложното.

(29) а. [[N1 … каза тя] N1_N2
[N2 и добави, N2_S1
[S1 че ЕС е отпуснал 1 милион евро
за *финансирането* на процеса...]]]
б. [[N1 … she said] N1_S1
[S1/N2 adding N2_S2
[S2/N3 that the EU has allocated 1m euros N3_S3
[S3 to *finance* the process...]]]]

Към втория тип се отнасят и различията в подреждането на простите изречения в състава на сложното, което води до различно свързване в сложното изречение: N_S в английски и _SN в български.

(30) а. [[S1 За да стигне до печката,] _S1N1
[N1 тя трябваше да мине покрай тях.]
б. [[N1 She had to make a detour N1_S1
[S1 to get to the stove.]]

Към този тип спада и изборът на глаголи с различна структура на комплементите, така че границата на простите изречения в състава на сложното в двата езика се различава. В (30) английският преходен глагол *urge* е преведен на български с непреходния *настоявам*, което променя границата между простите изречения – в английски *Croatia* е пряко допълнение в главното изречение (*urges it*), следователно влиза в състава му, а в българския превод *Хърватия* е подлог в подчиненото изречение (*тя да сътрудничи*) и се включва в неговия състав.

(31) а. [N1 Европейският парламент настоява N1_S1
[S1 Хърватия да сътрудничи напълно на трибунала.]]
б. [N1 European Parliament urges Croatia N1_S1
[S1 to fully cooperate with the Tribunal.]]

Най-много несъответствия се наблюдават при различен анализ на едни и същи явления. Разликата е най-забележима при причастията с функция на прилагателни, които въвеждат обособени части, и при деепричастията. В английски в тези случаи се смята, че причастията и деепричастието въвеждат отделни (подчинени) изречения в рамките на сложното, докато в български се следва концепцията, че не образуват отделни изречения. Различният подход в двата езика дава отражение в голям брой несъответствия при анотирането – съотнасяне на две или повече прости изречения в английски към едно в български.

(32) а. [N Изразявайки повторно подкрепата на ООН …, Щайнер подчерта необходимостта от свободно придвижване на хората ...]
б. [[S1 Reiterating UN support...], _S1N1
[N1 Steiner stressed the need for free movement of people...]

(33) а. [N1 Компенсацията, N1_S1
[S1 която се изплаща независимо от финансирането,
предоставяно на Революционна народна република Гвинея
съгласно Конвенцията от Ломе,]
се мобилизира в съответствие със специалната процедура …]
б. [N1 This compensation, N1_S1
[S1/N2 which shall be paid without prejudice to financing, N2_S2
[S2 accorded to the Revolutionary People's Republic of
Guinea under the Lome Convention,]]
shall be mobilised in accordance with the special procedure …]

5.1. Структура на сложното изречение – по данни от корпуса

Съотношението на сложните съчинени (N_N) към сложните съставни (N_S и _SN) изречения в корпуса е около 1:4 (вж. таблица 3). Разликата се дължи до голяма степен на факта, че подчинените прости изречения често се появяват в аргументна позиция и се изискват от субкатегоризационната рамка на глагола в главното изречение.

В сложните съчинени изречения съюзната връзка е между двете прости изречения, като се анотира към второто изречение (в български няма конструкции от типа _NN; сложните съчинени изречения със съотносителни съюзи (и – и, или – или, хем – хем и т.н.) са анотирани по модела N_N).

С най-ниска честота са подчинените конструкции от типа _SN, в които подчиненото изречение е в първа позиция, въведено е с подчинителна връзка, която се маркира като част от него, а главното изречение е на втора позиция. Това са конструкции, в които подчиненото изречение е преди главното и се въвежда със съюзи като *за да*, *докато*, *преди да*, *въпреки че* и други, както и конструкции с пряка реч (вж. 3.3.2), при които авторовите думи следват репликата. Това обяснява и най-високата им честота в подкорпусите с художествени текстове (ниската им честота в подкорпусите със субтитри, които се състоят предимно от пряка реч (реплики), се дължи на липсата на авторова реч – поради особеностите на контекста, в който се появяват репликите).

С най-висока честота в корпуса са подчинените конструкции, в които съюзната връзка е между главно първо и подчинено второ изречение N_S, като самата връзка се маркира към второто изречение. Сериозното разминаване в броя на тези конструкции между английския и българския подкорпус (в научните подкорпуси съотношението е над 2:1) се дължи на различния анализ на обособените причастни конструкции и деепричастието, както беше посочено в 4.1.

Таблица 3

Статистика на видовете изреченски модели в паралелните текстове на английски и на български

	Административни		**Научни**		**Художествени**		**Новини**		**Субтитри**		**Общо**	
	BG	EN	BG	EN	BG	EN	BG	EN	BG	EN	BG	EN
N_N	277	358	143	179	1167	1080	499	533	193	296	2279	2446
N_S	1066	1765	235	557	1561	2043	2186	3234	643	921	5691	8520
_SN	69	66	7	24	622	679	249	313	54	65	1001	1147

Корпусът се използва и за извличане на преводни съответствия на съюзните връзки и наблюдения върху контекста и честотата на срещанията на тези варианти. В таблица 4 е представено разпределението на някои най-често срещани маркери – български и английски съответствия, в различните изреченски модели (съответно N_N, N_S и _SN).

Таблица 4
Разпределение на свързващи маркери в изреченски модели

Маркер в български	Модел в български	Маркер в английски	Модел в английски	Брой
и	N_N	and	N_N	729
и	N_N	PUNCT	N_N	76
и	**N_N**	**to**	**N_S**	**31**
но	N_N	but	N_N	60
или	N_N	or	N_N	49
а	N_N	and	N_N	39
да	N_S	to	N_S	625
да	N_S	PUNCT	N_S	193
да	N_S	and	N_S	30
да	N_S	that	N_S	29
че	N_S	that	N_S	254
че	N_S	PUNCT	N_S	243
че	N_S	to	N_S	32
който	N_S	PUNCT	N_S	136
който	N_S	which	N_S	110
който	N_S	that	N_S	84
който	N_S	who	N_S	68
който	N_S	to	N_S	60

С най-висока честота е двойката съчинителни съюзи *и ~ and* (729 срещания), след тях се нареждат подчинителните *да ~ to* (625), еквивалентност на два пунктуационни знака, изразяващи безсъюзна съчинителна връзка (551), на *че* с *that* (254), на *че* с пунктуационен знак в английски (243), на *да* с пунктуационен знак (193) и т.н. Анотирането на съюзните връзки позволява да се проследят разнообразните преводни еквиваленти, например: съюзът *и* се предава с *and* (729), пунктуационен знак (76), *to* (31), *that* (13), *as* (8), *which* (6), *in* (5), *or* (5) и др.; *който* – пунктуационен знак (136 – заради разликите при причастията, вж. 4.1.), *which* (110), *that* (84), *who* (68), *to* (60), *as* (19), *and* (13) и др.; *ако ~ if* (73), *where* (5), пунктуационен знак (4), *unless* (4), *as* (3), *to* (3) и др. Разнообразието на подчинителни съюзи е много по-голямо, с по-ниска честота на срещанията на всеки един и по-голямо разнообразие на преводните еквиваленти, като има случаи, в които съчинителна връзка в единия език се предава с подчинителна връзка в другия език.

6. ИЗВОДИ

Работата по *Българско-английския корпус със съотнесени (прости) изречения* (БулЕнАК) показва проблемите при изграждане на анотирани паралелни корпуси със съотнесени единици. Целта ни при създаването на този корпус беше да изследваме влиянието на съотнасянето на текстовете по прости изречения, така че останалите равнища на съотнасяне, както и лематизацията бяха само помощни етапи. Анотаторите предварително трябваше да уговорят анотационните конвенции, като всеки от текстовите отрязъци беше преглеждан поне два пъти от всеки анотатор след ревизия на конвенциите. Сегментацията на простите изречения беше допълнително проверена при следващия етап – този на съотнасянето им.

БулЕнАК се използва за разработване на методи за: а) автоматично разделяне на прости изречения в състава на сложното и съотнасянето им; б) пренареждане на простите изречения в състава на сложното с цел изграждане на поподходящи тренировъчни и тестови данни за приложенията за статистически машинен превод; в) съотнасяне на равнище дума и фраза (Коева и др. 2012б). Тези приложения може да улеснят и работата по създаването на по-големи семантично и синтактично анотирани корпуси.

ЛИТЕРАТУРА

Варга и др. 2005: *Varga, D., L. Németh, P. Halácsy, A. Kornai, V. Trón, V. Nagy*. Parallel corpora for medium density languages. – In: Proceedings of the RANLP, 2005, pp. 590–596.

ГСБКЕ 1983а: Граматика на съвременния български книжовен език. Т. 2. Морфология. София: Издателство на БАН, 1983.

ГСБКЕ 1983б: Граматика на съвременния български книжовен език. Т. 3. Синтаксис. София: Издателство на БАН, 1983.

Даунинг, Лок 2002: *Downing, A., Ph. Locke.* English Grammar: A University Course. 2nd edition. New York: Routledge, 2002.

Коева и др. 2012а: *Koeva, S., R. Dekova, I. Stoyanova, B. Rizov, A. Genov*. Bulgarian X-language Parallel Corpus. – In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), 2012, pp. 2480–2486.

Коева и др. 2012б: *Koeva, S., B. Rizov, I. Stoyanova, S. Leseva, R. Dekova, A. Genov, E. Tarpomanova, T. Dimitrova, H. Kukova*. Application of clause alignment for statistical machine translation. – In: Proceedings of SSST-6, Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation. ACL 2012 / SIGMT / SIGLEX Workshop, pp. 102–110.

Коева и др. 2012в: *Koeva, S., B. Rizov, E. Tarpomanova, T. Dimitrova, R. Dekova, I. Stoyanova, S. Leseva, H. Kukova, A. Genov*. Bulgarian-English Sentence- and Clause-Aligned Corpus. – In: Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2), Lisbon, 29 November 2012. Lisbon: Colibri, 2012, pp. 51–62.

Коева и др. 2012г: *Koeva, S., I. Stoyanova, S. Leseva, T. Dimitrova, R. Dekova, E. Tarpomanova*. The Bulgarian National Corpus: Theory and Practice in Corpus Design. – Journal of Language Modelling, 2012, Vol. 0, No 1, pp. 65–110.

Коева и др. 2012д: *Коева, Св., Ив. Стоянова, Цв. Димитрова, Св. Лесева*. Традиции и новаторство в корпусната лингвистика: Българският национален корпус. – Списание на Българската академия на науките, 2012, № 3, с. 34–40.

Коуан, Кучерова, Колинс 2006: *Cowan, B., I. Kucerova, M. Collins*. A discriminative model for tree-to-tree translation. – In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Sydney, 2006, pp. 232–241.

Ризов, Б. Система за анотиране Chooser. – В: *Коева, Св. и др.* Българският семантично анотиран корпус. София: Институт за български език, 2010, с. 43–50.