КОРПУСНИ ДАННИ ЗА АНАЛИЗ НА СИНТАКТИЧНАТА РЕАЛИЗАЦИЯ НА СЕМАНТИЧНИТЕ ФРЕЙМОВЕ[1]

Светлозара Лесева[а], Ивелина Стоянова[б]

Институт за български език „Проф. Любомир Андрейчин“, Българска академия на науките[а,б]

zarka@dcl.bas.bg[а], iva@dcl.bas.bg[б]

Резюме. Статията представя работата по разработването на двуезичен корпус, съдържащ семантична и синтактична анотация на глаголи за комуникация, движение и промяна и техните задължителни обкръжения в английски и български език. Компилирането му се основава на универсалните аспекти на семантичното описание и синтактичната реализация и структурираната организация на познанието в двата основни ресурса, използвани в изследването (Уърднет и Фреймнет), както и на възможността за пренос на семантичното (и отчасти на синтактичното) описание на семантичните фреймове и прилежащите им фреймови елементи от Фреймнет към глаголите в Уърднет. Семантичната анотация е съотнесена със синтактичното ниво чрез маркирането на думите и фразите, изразяващи фреймовите елементи, синтактичната им категория и в редица случаи – граматичната им функция в изречението, като по този начин се извършва семантико-синтактично съотнасяне и се очертават основните модели на изразяването на аргументите на изследваните глаголи. Корпусът дава емпирична основа за теоретични и практически наблюдения, базирани на реални езикови данни.

Ключови думи: *семантично описание; български език; английски език; семантична анотация; синтактична анотация*

1. Мотивация на изследването

Съчетаването на семантичната информация за глаголите от Уърднет (WordNet) с лексикални ресурси като Фреймнет (FrameNet) дава възможност за тяхното по-пълно и многопластово описание, като по този начин се разширяват възможните приложения на ресурсите за целите на синтактичния и семантичния анализ (Baker, Fellbaum 2009; Schneider et al. 2012; Das et al. 2014).

В настоящата статия се разглежда методология за създаване на двуезиков паралелен анотиран корпус, основаваща се на относително универсалните аспекти на семантичното описание на глаголите, които правят възможно съотнасянето на ресурсите за различни езици и преноса на компоненти от описанието от един език към друг (в случая – от английски към български). Създаденият и анотиран корпус от примери на български и английски език има за цел да предостави илюстративен материал за наблюдения върху реализацията на семантичните фреймове и фреймовите елементи за глаголите от няколко големи семантични класа – глаголи за комуникация, глаголи за движение и глаголи за промяна. Междуезиковата съпоставка позволява и изучаването на езиково специфичните особености при синтактичната реализация на обкръженията на глаголите и може да се използва както за теоретични изследвания, така и за проверка на семантичната информация, приписана на синонимните множества в Уърднет.

Използваните в изследването лексикално-семантични и корпусни ресурси са представени в точка 2. В точка 3 е описан процесът на компилиране на корпуса – подбор на глаголи, асоциирани с определени семантични фреймове, извличане на подходящи примери от корпусните ресурси, идентифициране на синтактичните компоненти, реализиращи отделните фреймови елементи. В точка 4 са изложени резултатите от работата върху корпуса и броят анотирани примери за български и английски език. Статията завършва с представяне на възможните приложения на корпуса с анотирани примери, както и насоки за бъдеща работа.

[1] **Благодарности:** Разработката е осъществена по проекта „Обогатяване на семантичната мрежа УърдНет с концептуални фреймове“, подкрепен от Фонд „Научни изследвания“, Договор КП-06-Н50/1 от 2020 г.

2. Използвани ресурси

По-долу са описани лексикално-семантичните и корпусните ресурси, въз основа на които е съставен разгледаният корпус.

2.1. Лексикално-семантични ресурси

Принстънският уърднет, или само **Уърднет**[2] (Miller 1995; Fellbaum 1998), е лексикално-семантична база от данни, която представя изчерпателно лексикалния състав на езика в структуриран вид под формата на мрежа. Възлите ѝ представляват синонимни множества (синсети) от един или повече синоними, свързани помежду си с редица концептуално-семантични, лексикални и други релации, като хиперонимия, меронимия, антонимия и др. Значението на всяко синонимно множество е определено еднозначно чрез тълковна дефиниция, (в общия случай) илюстрирано с примери за употреба и при необходимост – допълнително пояснено с помощта на бележки за употребата, граматичните специфики и др. Съществителните и глаголите са определени от гледна точка и на принадлежността им към определен езиково независим семантичен (онтологичен) примитив като noun.person (съществителни за лица), noun.artifact (съществителни, означаващи произведения на човешката дейност), noun.plant (съществителни, означаващи растения), verb.communication (глаголи за комуникация), verb.emotion (глаголи, изразяващи емоции и чувства) и т.н. Тези примитиви, общо 25 за съществителните и 15 за глаголите[3], поделят синонимните множества от двете части на речта на относително малък брой обобщени семантични класове (Miller et al. 1993a).

Приема се, че до голяма степен концептите, включени в Уърднет, отразяват относително езиково независими фрагменти от човешкото познание, за разлика от лексикалното им изразяване, което представлява езиково зависим параметър. Квазиуниверсалният характер на концептите позволява създаването на аналогични на Принстънския уърднет ресурси и за множество други, разнообразни в типологично отношение езици[4]. В настоящото изследване наред с Принстънския уърднет се използва и разработеният за български **Булнет** (Koeva 2021), като еквивалентните синонимни множества в двата ресурса са съотнесени помежду си посредством еднозначен междуезиков идентификатор. Глаголите в Булнет са организирани в 14 103 синонимни множества, включващи и специфични за българския език глаголни значения.

В Уърднет липсва системно кодиране на участниците в ситуациите, описвани от предикатите, както и на тяхната предикатно-аргументна структура, а информацията за синтактично поведение е относително ограничена.

Фреймнет[5] (Baker et al. 1998; Baker 2008) е лексикално-семантичен ресурс, който представя знанието за съчетаемостта на лексемите чрез система от семантични фреймове в теоретичния апарат на фреймовата семантика. Фреймовете представляват схематични описания на концептуалната структура на ситуациите посредством участниците, обстоятелствата и други концептуални роли, наречени фреймови елементи (Baker et al. 1998; Ruppenhofer et al. 2016). В зависимост от статуса си те се поделят на ядрени, периферни и извънтематични. Ядрените фреймови елементи съответстват на първостепенните участници в ситуацията, които в специфичната си конфигурация представят най-съществената информация за съчетаемостта на лексемите, а в случая с глаголите – за техните семантични аргументи. Такива са например фреймовите елементи **Тема** (**Theme**) и **Цел** (**Goal**) във фрейма **Използване на превозно средство** (**Ride vehicle**) в пример 1. Периферните елементи характеризират ситуациите с оглед на обстоятелствата, при които протичат, например фреймовият елемент **Време** (**Time**) (пример 1). Извънтематичните фреймови елементи поставят фрейма в контекста на друг фрейм или на по-обща ситуация: такъв е фреймовият елемент **Ко-тема** (**Co-theme**) в пример 1[6]:

---

[2] https://wordnet.princeton.edu/

[3] Класове, определени от примитивите: https://wordnet.princeton.edu/documentation/lexnames5wn.

[4] http://globalwordnet.org/resources/wordnets-in-the-world/

[5] https://framenet.icsi.berkeley.edu/fndrupal/

[6] Посоченият извънтематичен фреймов елемент реферира към фрейма **Движение** (**Motion**), където запълва позицията на движещия се обект (**Темата**), чиято траектория съвпада с тази на возещия се (**Темата** във фрейма **Използване на превозно средство**) (Ruppenhofer et al. 2016: 98).

1. [*Утре*]$_{\text{Време}}$ [*майка ми*]$_{\text{Тема}}$ ***ЩЕ ЛЕТИ*** [*до Лондон*]$_{\text{Цел}}$ [*с приятели*]$_{\text{Ко-тема}}$.

По-долу фокусът ще бъде поставен върху ядрените фреймови елементи, доколкото те съответстват на глаголните аргументи (Baker et al. 1998), а където е релевантно, ще бъдат споменати и някои от периферните.

Концепция за представянето на семантичните и синтактичните свойства на глаголите в българския език, основаващо се на Фреймнет, е изложена у Коева (Коева / Koeva 2010). Въз основа на това описание и създадения единен теоретичен модел за формално представяне на езиковите единици е разработена уеббазираната система за пълно и непротиворечиво описание на семантиката и синтактичните свойства на глаголите с помощта на семантични фреймове Булфрейм (Koeva, Doychev 2022). В настоящото изследване стъпваме както на разработките за английски, така и на тези за български език.

Една от най-важните характеристики на Фреймнет като лингвистичен ресурс е корпусът от изречения, илюстриращи фреймовете, в които синтактичните групи, чрез които са реализирани фреймовите елементи, техният тип – например именна фраза (NP), предложна фраза (PP) и т. н. – и в част от случаите, функцията им в изречението (подлог, допълнение), са анотирани еднозначно от експерти. Корпусът онагледява синтактичната реализация на семантичните фреймове и не само дава възможност за извършването на лингвистични обобщения за английския език, но може да служи и като отправна точка за съпоставителни изследвания върху други езици.

Тъй като информацията, включена в Уърднет и Фреймнет, отразява взаимно допълващи се аспекти на семантичното (и синтактичното) описание на езиковите единици, обогатяването на синонимните множества с концептуална информация от Фреймнет, както и обратното – попълването на фреймовете във Фреймнет с подходящи единици от по-богатия по отношение на лексикалния състав Уърднет, води до по-изчерпателно семантично и синтактично описание на думите в езика, в частност – на глаголите. Взаимното обогатяване на двата ресурса е осъществено чрез автоматичното им съотнасяне (Tonelli, Pighin 2009; Palmer 2009; Palmer et al. 2014; Leseva, Stoyanova 2020), при което на синонимни множества в Уърднет са приписани семантични фреймове от Фреймнет. Глаголите в едно синонимно множество имат общо лексикално значение, което предполага и принадлежността им към един семантичен фрейм, при все че от практическа гледна точка се наблюдават и разминавания. Относителната универсалност на системата от концепти в Уърднет и на фреймовете във Фреймнет позволява междуезиковия пренос на тази информация.

2.2. Корпусни ресурси

За компилиране на корпуса с анотирани илюстративни примери за синтактичното изразяване на глаголните фреймове и техните елементи в настоящото изследване са използвани няколко изходни ресурса. На първо място това са семантично анотираните корпуси – Семкор (SemCor) за английски и Булсемкор (BulSemCor) за български език, като и двата са анотирани със значения от Уърднет.

Семкор (SemCor) (текуща версия 3.0) (Miller et al. 1993b; Miller et al. 1994; Landes et al. 1998) е компилиран от екипа на Принстънския уърднет. На единиците в корпуса са приписани части на речта и граматични характеристики, а пълнозначните думи са семантично анотирани, като всяка дума е отнесена еднозначно към синсет в Уърднет. Семкор е най-големият ръчно анотиран корпус от този вид с общ обем от 226 040 анотирани единици.

Булсемкор (BulSemCor) (Koeva et al. 2006; Koeva et al. 2011) е създаден по модела на оригиналния Семкор, като са добавени критерии за осигуряване на покритие на съвременната обща лексика. В допълнение към пълнозначните думи Булсемкор включва анотация и на служебните думи: предлози, съюзи, частици, местоимения и междуметия. За тази цел Българският уърднет е разширен със синонимни множества от тези части на речта. Размерът на корпуса е близо 100 000 анотирани единици.

Обемът на двата корпуса не е достатъчен, за да се осигурят достатъчно примери за много от изследваните глаголи, затова са използвани и примери, извлечени от други корпуси.

Българо-английският паралелен корпус със съотнесени изречения и прости изречения[7] **(BulEnAC)** (Koeva et al. 2012a) се състои от съотнесени двойки паралелни изречения на двата езика и съдържа 366 865 думи (176 397 за български и 190 468 за английски). Синтактичната анотация

[7] https://dcl.bas.bg/en/resources_list/bulenac/

включва: а) определяне на границите на изреченията и простите изречения в състава на сложните; б) маркиране на вида на синтактичната връзка (съчинителна или подчинителна) между простите изречения; в) маркиране на елементите, които въвеждат изречението: съюзи, комплементизатори и пунктуация. Корпусът е подходящ за извличане на паралелни примери, илюстриращи употребата на изследваните глаголи. Така се улеснява идентифицирането на съответните преводни еквиваленти в рамките на съотнесените изречения, но се изисква определяне на конкретното лексикално значение на глаголите.

Българският национален корпус (БНК) (Koeva et al. 2012b) е най-големият корпус за български език с общ обем от около 5,4 милиарда думи. Състои се от едноезикова част, включваща текстове само на български език, и 47 сателитни паралелни корпуса, в които българските текстове имат съответствие на друг език. Българската част включва около 1,2 милиарда думи. Системата за търсене в БНК позволява извличане на примери въз основа на комбинации от лексикални и граматически критерии и синтактични шаблони.

3. Създаване на английско-български корпус с анотирани примери за синтактичната реализация на семантичните фреймове

При компилирането на корпуса сме се фокусирали върху три големи класа глаголи – глаголи за комуникация, глаголи за движение и глаголи за промяна. По-долу е описан процесът на създаването му – от подбора на подходящи илюстративни примери до автоматичната им обработка и ръчната анотация на реализациите на фреймовите елементи.

3.1. Събиране на илюстративни примери

Като отправна точка са използвани валентните конфигурации и техните синтактични реализации за глаголите, принадлежащи към даден фрейм от Фреймнет. Примерите за английски език включват анотираните изречения от корпуса, разработен в рамките на Фреймнет[8] (Burchardt, Pennacchiotti 2008), като впоследствие наборът от данни е допълнен с примери от Семкор, за да се илюстрира използването на определени глаголни значения, които не са добре застъпени във Фреймнет.

Българската част на корпуса включва изречения, извлечени от Булсемкор, допълнени с примери от паралелния корпус със съотнесени изречения (добавени са двойките паралелни изречения и за двата езика) и БНК. Глаголите в примерите, които не са семантично анотирани (т. е. всички без изреченията от Булсемкор), са ръчно съотнесени с уникален синсет в Уърднет в хода на работата.

В процеса на подбор на примери на английски и български език са взети предвид както синтактичните модели, засвидетелствани във Фреймнет, които са валидни и за двата езика, така и езиково специфичните структури.

3.2. Набори от валентни конфигурации и синтактични модели

Синтактичното описание включва две нива. Първото се състои от валентните конфигурации от Фреймнет, които определят комбинациите от фреймови елементи (ядрени и периферни), реализиращи се в речта. Това ниво е в голяма степен езиково независимо, тъй като се отнася към сферата на човешкото познание и концептуализацията на ситуациите, затова информацията за английски език може да бъде пренесена и върху съответните глаголи със същото лексикално значение в български език.

За даден фрейм са обобщени всички валентни конфигурации, характеризиращи глаголните лексикални единици, включени в него, заедно с общата им честота на срещане. Конфигурации с по-малко от 3 примера са изключени, тъй като показват специфична употреба на отделни глаголи и не са показателни за фрейма. Таблица 1 илюстрира извлечените валентни модели за фрейма **Задаване на въпрос** (**Questioning**).

---

[8] http://framenet.icsi.berkeley.edu/

Таблица 1. Валентни конфигурации за фрейма **Задаване на въпрос**

Честота	Комбинации от фреймови елементи			
117	Говорещ	Съобщение	Адресат	
41	Говорещ		Адресат	Тема на съобщението
33	Говорещ	Съобщение	Адресат	Начин
20	Говорещ		Адресат	
11	Говорещ	Съобщение		

Валентните конфигурации с най-висока честота включват ядрените фреймови елементи **Говорещ** (**Speaker**), **Съобщение** (**Message**), **Адресат** (**Addressee**), **Тема на съобщението** (**Topic**) и неядрения фрейм **Начин** (**Manner**). Тези обобщени модели дават нагледна представа за срещащите се в езика комбинации. В конкретния случай се вижда, че за фрейма е типично едновременното (първият и третият модел в Таблица 1, вж. също пример 2.а.) или алтернативното (вторият и четвъртият модел в Таблица 1) синтактично изразяване на **Съобщението** и **Адресата**. Наред с това, **Темата на съобщението** много често няма израз, когато **Съобщението** е експлицирано (първият, третият и петият модел в Таблица 1), и се изразява по-свободно в контексти, когато **Съобщението** е имплицитно (вторият модел в Таблица 1, вж. също пример 2.б.):

2.а. [*They*]$_{\text{Говорещ}}$ ***ASKED*** [*Stephen*]$_{\text{Адресат}}$ [*if he can spare them a few minutes*]$_{\text{Съобщение}}$.
[*Те*]$_{\text{Говорещ}}$ ***ПОПИТАХА*** [*Стивън*]$_{\text{Адресат}}$ [*дали ще им отдели две минути*]$_{\text{Съобщение}}$.
2.б. [*He*]$_{\text{Говорещ}}$ ***QUESTIONED*** [*us*]$_{\text{Адресат}}$ [*about our journey*]$_{\text{Тема на съобщението}}$.
[*Той*]$_{\text{Говорещ}}$ [*ни*]$_{\text{Адресат}}$ ***РАЗПИТВА*** [*за пътуването*]$_{\text{Тема на съобщението}}$.

Второто ниво на синтактично описание се отнася до конкретната синтактична реализация на фреймовите елементи: синтактичната им категория и граматичната им функция в изречението. Въпреки наличието на езикови специфики в редица случаи при (сравнително) близки езици се наблюдава сходна синтактична реализация на основните участници в ситуацията. Това, разбира се, не означава пълно съвпадение по отношение на синтактичните категории или граматичните функции в междуезиков план и трябва да се разглежда преди всичко като отправна точка за изследване.

Таблица 2 показва възможни синтактични реализации на най-често срещаната валентна конфигурация за фрейма **Задаване на въпрос**.

Таблица 2. Синтактични модели, съответстващи на валентната конфигурация **Говорещ** – **Съобщение** – **Адресат** за фрейма **Задаване на въпрос**

Говорещ	Съобщение	Адресат
NP.Ext	Sinterrog	DNI
NP.Ext	Sinterrog	NP.Obj
NP.Ext	QUOTE	DNI

Говорещият се изразява като именна група в подложна позиция (в деятелен залог); **Съобщението** – като подчинено допълнително изречение, въведено с въпросителна дума (пример 3.а.), или като пряка реч (пример 3.б.); **Адресатът** най-често остава имплицитен, ако има конкретен, известен от непосредствения или по-широкия контекст референт (пример 3.в.)[9], а при изразяване се реализира като именна фраза в двойнообектна конструкция (double object construction) (пример 3.а., 3.б.).

[9] Даден фреймов елемент може да се предполага на концептуално равнище, но да е синтактично имплицитен, т.е. да има нулева реализация. Тя се подразделя на няколко вида: определена (DNI – definite null instantiation), при която референтът е известен от близкия контекст; неопределена (INI – indefinite null instantiation), при която

3.а. [*We*]$_{\text{Говорещ}}$ ***ASKED*** [*Ruby*]$_{\text{Адресат}}$ [*what kind of food they ate*]$_{\text{Съобщение}}$.
[]$_{\text{Говорещ-DNI}}$ ***ПОПИТАХМЕ*** *[Руби]*$_{\text{Адресат}}$ [*с какво са се хранили*]$_{\text{Съобщение}}$.
3.б. [*"Which is heavier?"*]$_{\text{Съобщение}}$ [*he*]$_{\text{Говорещ}}$ ***ASKED*** [*the child*]$_{\text{Адресат}}$.
[*– Кой тежи повече?*]$_{\text{Съобщение}}$ – ***ПОПИТА*** [*той*]$_{\text{Говорещ}}$ [*детето*]$_{\text{Адресат}}$.
3.в. [*"And how is Syl?"*]$_{\text{Съобщение}}$ ***ASKED*** [*Lily*]$_{\text{Говорещ}}$ [ ]$_{\text{Адресат-DNI}}$.
[*– Как е Сил?*]$_{\text{Съобщение}}$ – ***ПОПИТА*** [*Лили*]$_{\text{Говорещ}}$ [ ]$_{\text{Адресат-DNI}}$.

Междуезикови различия се наблюдават и по отношение на инвентара от лексикални елементи, въвеждащи синтактичните групи (предлози, съюзи и т. н.), езиково специфичните граматични особености и конструкции. Например, за разлика от английски, в българския език няма нефинитни изречения, поради което пропозиционалните фреймови елементи се реализират чрез финитни изречения. Налице са различия и по отношение на задължителността на синтактичното изразяване (включително на подлога), специфичните за езика диатези, конструкции, словоред, морфосинтактични особености и т. н.

За да се осигури съотносимостта на синтактичните модели за българските данни, оригиналните синтактични модели за английските глаголи във Фреймнет в някои случаи са обобщени. Например моделите, включващи въпросително подчинено изречение (Sinterrog), финитно съюзно подчинено изречение (Sfin) и предложна фраза с *-ing* форма (VPing), са групирани заедно и се разглеждат като подкласове на категорията Clause (просто изречение в състава на сложното), така че да се позволи съпоставка с различната им реализация на български, където някои от тези подкласове не присъстват (напр. VPing). Аналогично, предложните фрази с различна опора, реализиращи един и същ фреймов елемент, например PP[of], PP[from], които служат за изразяване на фреймовия елемент **Компоненти** във фрейма **Построяване** (**Building**), също се групират заедно в обобщен вид: предложна фраза (PP). Насоките за бъдеща работа включват пълното описание на синтактичната реализация във всеки от езиците и анализа на езиково специфичните синтактични особености.

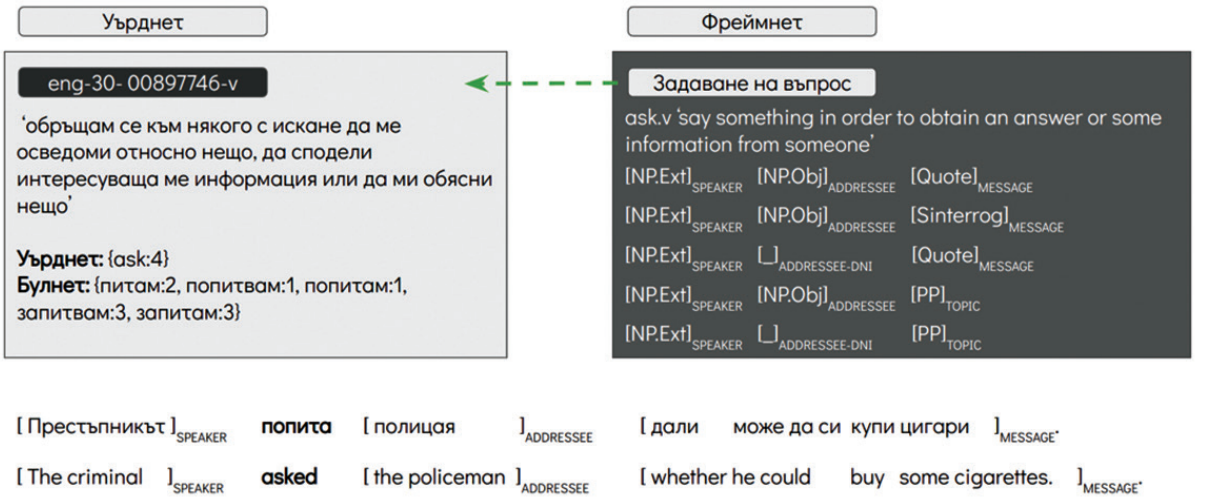
3.3. Анотация на фреймовите елементи

Съотнасянето на семантичното и синтактичното ниво се постига чрез: а) анотиране на фреймовите елементи в изречението; и б) определяне на синтактичната категория и в редица случаи граматичната им функция, например подлог (NP.Ext), пряко допълнение (NP.Obj), предложна фраза (PP), адвербиална фраза (AdvP), просто изречение (Clause). Фигура 1 показва анотирана двойка примери за глаголите *питам* и *ask*, в която е включена още информация за: а) съответстващите си синонимни множества, към които принадлежат глаголите в двата езика: eng-30-00897746-v[10] {питам; попитвам; попитам; запитвам; запитам} и {ask}, 'обръщам се към някого с искане да ме осведоми относно нещо, да сподели интересуваща ме информация или да ми обясни нещо'; б) приписания на синсета фрейм **Задаване на въпрос;** и в) синтактичните модели във Фреймнет, отнасящи се към фрейма.

Примерите, включени в българската и английската част на корпуса, са предварително обработени с инструменти за автоматична лингвистична анотация, в резултат от което за всяка словоформа са определени частта на речта и граматичните ѝ характеристики и основната ѝ форма (Koeva et al. 2020). Въз основа на така извършената морфосинтактична анотация и лематизация в хода на настоящото изследване са определени синтактичните групи, чрез които се реализират основните изреченски елементи: (а) глаголът, изпълняващ ролята на сказуемо; (б) фразите, чрез които се изразяват фреймовите елементи: именни групи, изпълняващи функцията на подлог и пряко допълнение; предложни фрази с различна функция (PP); подчинени изречения. За всяко подчинено изречение е маркиран лексикалният елемент или фраза, които го въвеждат (въз основа на което изреченията са

липсва конкретен референт, но наличието и спецификата на възможния референт се подразбират по силата на обичайна интерпретация или конвенция, например *Той ядеше [ ] с апетит*; конструктивна (CNI – constructional null instantiation), при която липсата на изразен референт е обусловена от синтактичната конструкция, например неизразен **Агент** в страдателен залог (Petruck 2019).

[10] Уникален идентификатор, чрез който се осигурява съответствието между синонимните множества с еквивалентно значение в Уърднет и Булнет, както и в лексикално-семантичните мрежи за всички други езици, разработени по сходна методология.



Фигура 1. Анотиран пример

разделени на съюзни, въпросителни и пряка реч), а в примерите от Фреймнет е определено и дали изреченията са финитни или нефинитни (и съответно: инфинитивни или герундивни).

Частите на изречението, чрез които са изразени основните фреймови елементи, са ръчно анотирани и маркирани както с името на фреймовия елемент (напр. **Агент**, **Тема**, **Инструмент** и т. н.), така и с вида на синтактичната категория, чрез която той се реализира – NP.Ext (подлог), NP.Obj (пряко допълнение), PP (предложна фраза), AdvP (адвербиална фраза), Clause (просто изречение в състава на сложното) и др.

4. Резултати

Английската част от корпуса с илюстративни примери обхваща 211 глагола (лексикални единици във Фреймнет), съотнесени към 135 синсета в Уърднет. Обемът ѝ възлиза на 13 295 примера, онагледяващи 3577 различни синтактични модела. Във всяко изречение са анотирани глаголното сказуемо (съотнесено с конкретно синонимно множество) и компонентите на изречението, отбелязващи реализацията на ядрените фреймови елементи.

Към момента българската част на корпуса съдържа данни за 146 глагола, принадлежащи към 125 синсета в Булнет, и включва 2184 анотирани примера, представящи 290 различни модела. Анотацията на българските примери е извършена по аналогичен начин като в английски (Фиг. 1).

Таблица 3 онагледява данните в корпуса, разпределени по език. Посочени са броят разгледани фреймове и броят синсети, един или повече синоними от които фигурират в корпуса. За всеки от езиците е представен и броят примери и общият брой анотирани фреймови елементи.

Таблица 3. Анотирани примери по езици и класове глаголи

Клас глаголи	Брой фреймове	Брой синсети	Български		Английски	
			Брой примери	Брой анотации	Брой примери	Брой анотации
Глаголи за комуникация	7	51	1025	2921	5065	14 822
Глаголи за движение	8	48	744	2016	4748	13 304
Глаголи за промяна	6	26	415	866	3482	9048

5. Насоки за бъдеща работа

Работата по описанието на семантичните и синтактичните свойства на българските глаголи показва, че семантичното описание, кодирано в рамките на Фреймнет, е до голяма степен езиково независимо и поради това – преносимо от един език към друг. Езиковите специфики се наблюдават предимно при синтактичната реализация на фреймовите елементи.

Представеният корпус от илюстративни примери използва потенциала на Уърднет и Фреймнет за междуезиков анализ и трансфер на информация. Извлечените от Фреймнет валентни модели и синтактични шаблони се разглеждат за български език въз основа на събраните примери за употреба на глаголите.

Бъдещата работа върху корпуса ще се фокусира върху разширяването му с по-голям по обем и по-разнообразен по съдържание материал на български език. За целта ще бъдат извлечени допълнителни примери от БНК. Разработват се и методи за частична автоматична анотация – автоматично определяне на значението на глагола, идентифициране на фразите в изречението въз основа на синтактичен анализ, съотнасяне на компоненти от синтактичната структура на изречението към фреймови елементи и т.н. Автоматичните процедури имат за цел улесняването на анотацията и по-ефективното увеличаване на броя примери.

Методите за пренос на информация между ресурсите и езиците, основаващи се на универсалността на семантичното описание, са обещаващ подход за разработване на мащабни семантични ресурси за езици с по-слаба обезпеченост с ресурси и технологии, като се използват знанието и данните, налични за езици като английски. Ръчно анотираните ресурси могат да бъдат използвани за трениране на приложения за автоматичен семантичен анализ, анотиране на семантични роли и др.

Цитирана литература / References

Коева 2010: Коева, Св. *Българският ФреймНет*. София: Институт за български език „Проф. Любомир Андрейчин“.

Baker, Fellbaum 2009: Baker, C. F., C. Fellbaum. WordNet and FrameNet as Complementary Resources for Annotation. – In: *Proceedings of the Third Linguistic Annotation Workshop (ACL-IJCNLP ’09)*. Association for Computational Linguistics, pp. 125 – 129.

Baker 2008: Baker, C. F. FrameNet: Present and Future. – In: Webster, N. Ide, A. Chengyu Fang (Eds.), *The First International Conference on Global Interoperability for Language Resources*. Hong Kong: City University.

Baker et al. 1998: Baker, C. F., C. J. Fillmore, J. B. Lowe. The Berkeley FrameNet Project. – In: *COLING-ACL ’98: Proceedings of the Conference*. Montreal, Canada, pp. 86 – 90.

Burchardt, Pennacchiotti 2008: Burchardt, A., M. Pennacchiotti. FATE: a FrameNet-Annotated Corpus for Textual Entailment. – In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*. Marrakech, Morocco: European Language Resources Association (ELRA). <http://www.lrec-conf.org/ proceedings/lrec2008/pdf/143_paper.pdf> [18.01.2024].

Das et al. 2014: Das, D., D. Chen, A. F. T. Martins, N. Schneider, N. A. Smith. Frame-Semantic Parsing. – *Computational Linguistics*, vol. 40(1), pp. 9 – 56.

Fellbaum 1998: Fellbaum, C. (Ed.). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Koeva et al. 2006: Koeva, S., S. Leseva, M. Todorova. Bulgarian Sense Tagged Corpus. – In: Proceedings of LREC 2006, 79 – 86.

Koeva et al. 2011: Koeva, S., S. Leseva, B. Rizov, E. Tarpomanova, T. Dimitrova, H. Kukova, M. Todorova. Design and Development of the Bulgarian Sense-annotated Corpus. – In: Carrio Pastor, M. L., M. A. C. Mora (Eds.), *Information and Communication Technologies: Present and Future in Corpus Analysis*: *Proceedings of the III International Congress of Corpus Linguistics*, pp. 143 – 150.

Koeva et al. 2012a: 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)*. Lisboa: Colibri: pp. 51 – 62.

Koeva et al. 2012b: Koeva, S., I.,Stoyanova, S., Leseva, R., Dekova, T., Dimitrova, E. Tarpomanova. The Bulgarian National Corpus: Theory and Practice in Corpus Design. – *Journal of Language Modelling*, vol. 1, pp. 65 – 110.

Koeva et al. 2020: Koeva, S., N. Obreshkov, and M. Yalamov. Natural Language Processing Pipeline to Annotate Bulgarian Legislative Documents. – In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association.

Koeva, Doychev 2022: Koeva, S., E. Doychev. Ontology Supported Frame Classification. – In: *Proceedings of the Fifth International Conference Computational Linguistics in Bulgaria*. Sofia: Institute for Bulgarian Language, pp. 203 – 214. https://aclanthology.org/2022.clib-1.23 [18.01.2024]

Landes et al. 1998: Landes, S., C. Leacock, R. Tengi. Building Semantic Concordances. – In: C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database*.

Litkowski 2014: Litkowski, K. The FrameNet Frame Element Taxonomy. <https://www.clres.com/online-papers/FETaxonomy.pdf> [18.01.2024]

Miller et al. 1993a: Miller, G., R. Beckwith, C. Fellbaum, D. Gross, G. A. Miller. *Introduction to WordNet: an Online Lexical Database. Five Papers on WordNet Princeton*. NJ: Princeton University.

Miller et al. 1993b: Miller, G. A., C. Leacock, R. Tengi, R. T. Bunker. A Semantic Concordance. – In: *Human Language Technology: Proceedings of a Workshop Held at Plainsboro*, New Jersey, March 21 – 24, 1993, pp. 303 – 308. <https://aclanthology. org/H93-1061> [18.01.2024]

Miller et al. 1994: Miller, G. A., M. Chodorow, S. Landes, C. Leacock, R. G. Thomas. Using a Semantic Concordance for Sense Identification. – In: *Human Language Technology: Proceedings of a Workshop held at Plainsboro*, New Jersey, March 8 – 11, 1994, pp. 340 – 343. <https://aclanthology.org/H94-1046> [18.01.2024]

Miller 1995: Miller, G. A. WordNet: A Lexical Database for English. – *Communication of the ACM*, 38 (11), pp. 39 – 41.

Palmer 2009: Palmer, M. Semlink: Linking PropBank, VerbNet and FrameNet. – In: *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon. Sept. 2009, Pisa, Italy: GenLex-09*, pp. 9 – 15.

Palmer et al. 2014: Palmer, M., C. Bonial, D. McCarthy. SemLink+: FrameNet, VerbNet and Event Ontologies. – In: *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929 – 2014)*, Baltimore, Maryland USA, June 27, 2014, Stroudsburg, PA: Association for Computational Linguistics, pp. 13 – 17.

Petruck 2019: Petruck, M. Meaning Representation of Null Instantiated Semantic Roles in FrameNet. – In: *Proceedings of the First International Workshop on Designing Meaning Representations*, Stroudsburg, PA: Association for Computational Linguistics, pp. 121 – 127.

Ruppenhofer et al. 2016: Ruppenhofer, J., M. Ellsworth, M. R. L. Petruck, C. R. Johnson, C. F. Baker, J. Scheffczyk. *FrameNet II: Extended Theory and Practice*. Berkeley, California: International Computer Science Institute.

Schneider et al. 2012: Schneider, N., B. Mohit, K. Oflazer, N. A. Smith. Coarse Lexical Semantic Annotation with Supersenses: An Arabic Case Study. – In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Stroudsburg, PA: Association for Computational Linguistics, pp. 253 – 258.

Tonelli, Pighin 2009: Tonelli, S.a, D. Pighin. New Features for Framenet – Wordnet Mapping. – In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL'09).* Boulder, USA.

Corpus Data for the Validation of the Syntactic Realisation of Semantic Frames

Svetlozara Leseva[a], Ivelina Stoyanova[b]

Institute for Bulgarian Language Prof. Lyubomir Andreychin, Bulgarian Academy of Sciences[a,b]

zarka@dcl.bas.bg[a], iva@dcl.bas.bg[b]

Abstract. This paper presents the work on compiling a bilingual corpus that demonstrates the syntactic realisation of the conceptual description of verbs in English and Bulgarian and in a broader context, facilitates cross-linguistic studies by providing the foundation for theoretical and practical observations based on the two languages. The compilation of the corpus relies on the universal aspects of conceptual description and syntactic realisation by harnessing the main structural principles of the two main resources employed (WordNet and FrameNet), allowing for cross-linguistic alignment and transfer of information from one language to another (in this case from English to Bulgarian) and between resources, in particular transferring the semantic description from FrameNet onto the verb synsets in WordNet. The resulting resource links the semantic level of the conceptual frames and the frame elements with the syntactic level (couched in the form of patterns representing the syntactic realisation of the frame elements in terms of their syntactic categories and grammatical function).

Keywords: *conceptual description; Bulgarian; English; semantic annotation; syntactic annotation*

Svetlozara Leseva
Institute for Bulgarian Language
52 Shipchenski prohod Blvd., Bl. 17
Sofia 1113, Bulgaria
Ivelina Stoyanova
Institute for Bulgarian Language
52 Shipchenski prohod Blvd., Bl. 17
Sofia 1113, Bulgaria

https://doi.org/10.7546/ConfIBL2024.39