Български език 63 (2016), 4, 54-74

СВЕТЛОЗАРА ЛЕСЕВА, ИВЕЛИНА СТОЯНОВА

КЪМ АВТОМАТИЧНОТО РАЗРЕШАВАНЕ НА РЕФЕРЕНЦИИ В БЪЛГАРСКИЯ ЕЗИК

SVETLOZARA LESEVA, IVELINA STOYANOVA

TOWARDS A SYSTEM FOR ANAPHORA RESOLUTION IN BULGARIAN

(Abstract)

The paper presents the first steps towards the implementation of a system for automatic anaphora resolution for Bulgarian. The present study is focused on the following types of anaphora: (a) pronouns: personal pronouns in the nominative, accusative and dative, possessive, reflexive/reciprocal personal and possessive pronouns, as well as relative pronouns; and (b) elliptic subject. The system is based on a set of rules which reflect specific language phenomena and dependencies in Bulgarian, such as free word order, ellipsis, impersonal sentences, and morphosyntactic agreement. The rules fall into two types – filtering rules which pose restrictions and thus dispose of invalid candidates, and ranking rules which are used to sort the candidates and select the most reliable one. Different anaphoras are resolved by a different combination of the proposed rules related to the scope of the reference, as well as the syntactic and morphosyntactic features. The method is applied on the Bulgarian Semantically-Annotated Corpus (BulSemCor) and the results are compared with the results from a baseline which relies only on agreement and assigns the reference to the closest candidate preceding the anaphora. The results are preliminary and the detailed analysis of the rules and the preferences is among the tasks for future work.

Keywords: anaphora, anaphora resolution, automatic methods, Bulgarian

1. Въведение

В книгата си „Кохезията в английския език“ (Halliday 1976) авторите описват механизмите за изграждане на повърхнинната структура на текста, или *кохезията* (cohesion), при която интерпретацията на едни елементи от дискурса зависи от интерпретацията на други елементи. Различават няколко типа кохезия: референция, субституция, елипса, съюз и лексикална кохезия (колокации). Кохезията на текста, включително и разрешаването на анафората, има важно значение за автоматичната обработка

на езика. Разработването на ефективни автоматични методи намира приложение в области като машинния превод, извличането на информация, автоматичното резюмиране на документи и много други.

В настоящата разработка се приема следната дефиниция за понятието *анафора*: дума (или съставна единица), чиято интерпретация зависи от друга дума (или израз), наричан *антецедент*. В изследването обхватът на разглежданите явления се ограничава, като се поставят следните условия: разглеждат се само случаите на местоименна анафора, които има антецедент в предходния контекст.

В литературата са известни различни видове анафора. Хърст (Hirst 1981) описва 14 типа анафора в зависимост от начина на изразяване (местоимения, именни фрази с местоименна функция, елипса и др.) или към какво се реферира (към обект, глагол, прилагателно, изречение и др.). Обект на тази разработка са следните типове (по Hirst 1981):

1) местоимения (Пример 1а, 1б);
2) референции към изречение или просто изречение (Пример 1в, 1г);
3) нереферентни местоимения (Пример 1д);
4) неизразен подлог (Пример 1е).

Пример 1. Различни типове анафора (изразите с еднакъв референт ще бъдат означавани с еднакъв индекс; с ***pro*** се бележи неизразеният подлог при личните изречения).
(а) ***[Надя]i*** *искаше пръстен, а Иван* ***[ѝ]i*** *купи гривна.* ***[Тя]i*** *беше разочарована.*
(б) ***[Иван]i*** *искаше* ***[pro]i*** *да си почине, а* ***[жена [му]i]j*** *настояваше* ***[pro]i*** *да* ***[ѝ]j*** *помогне.* ***[[Неговото]i*** *разочарование]****k*** *беше голямо.*
(в) ***[Между депутатите избухна голям скандал]i, [което]i*** *сложи край на преговорите.* (***което*** реферира към цялото събитие: *„между депутатите избухна голям скандал“*)
(г) *[Писателят]****i*** *не само подозираше, че* ***[книгата ще стане толкова популярна]j****. [Той]****i*** *[го]****i*** *знаеше. (****го*** реферира към цялото събитие: *„книгата ще стане толкова популярна“*)
(д) *Излизам аз, а* ***то*** *вали дъжд!* (няма референт)
(е) ***[Иван]i*** *се обади и* ***[pro]i*** *каза, че* ***[pro]i*** *няма да дойде.*

В настоящото изследване са разгледани следните типове местоимения: **лични** – в именителен, винителен и дателен падеж, включително кратките им форми (например *той, него, го, я, ѝ, им*), **притежателни**, включително кратките им форми (например *негов, нейно, техни, им, ѝ*), **възвратни лични** и **възвратни притежателни** (например *се, си, свой*) и част от **относителните местоимения** (*който* и формите му). Разглеждани са и случаите на неизразен подлог, които са типични за българския език.

Разрешаването на анафората е процесът на (автоматично) определяне на антецедента, към който тя реферира. Единици в текста (прости и със-

тавни съществителни имена, местоимения и т.н.), които се отнасят към един и същ референт, се наричат кореферентни и формират така наречените кореферентни вериги. Кореференцията обаче остава извън обхвата на тази разработка.

Представеното тук изследване е насочено към създаването на система за разрешаване на референцията на анафорите. Разработката се фокусира върху лингвистичните особености на местоименните анафори и формулирането на правила за откриване на обсега (контекста), в който следва да бъдат търсени техните антецеденти и съответно – установяване на кандидатите за антецеденти в съответния контекст. Изведени са и правила за разпознаване на възможните позиции, в които има неизразен подлог. Описани са фактори за ранкиране на потенциалните кандидати за антецеденти като процедура за избор на най-вероятния кандидат.

Тъй като в представените в статията правила обикновено се разглежда не сказуемото като цялост и функция, а определени негови компоненти (основният глагол или спомагателен глагол), за улеснение и краткост при формулирането на правилата и зависимостите ще използваме „глагол“ .

2. Преглед на изследванията, посветени на автоматичното разрешаване на анафора

Основният интерес към разрешаването на анафората е през 90-те години на 20. век. Достиженията на съществуващите подходи и системи за разрешаване на анафора през този период са представени в обзора на Митков (Mitkov 1999), в който са обобщени и основните принципи и използвани стратегии за имплементация. Актуално изследване върху разрешаването на референцията на анафорите е представено у Шмолц (Schmolz 2015).

Двете основни тенденции при методите за разрешаване на тази задача могат да се обобщят като подходи, основаващи се на правила, и подходи, основаващи се на данни.

Базираните на правила подходи разчитат на синтактична, морфосинтактична и друга лингвистична информация, въз основа на която се формулират закономерности за възможните антецеденти на анафорите. Хобс (Hobbs 1978) използва различни синтактични ограничения върху прономинализацията, които се прилагат върху синтактичната репрезентация (синтактични дървета) на изреченията. Един от най-влиятелните базирани на правила алгоритми – RAP (Resolution of Anaphora Procedure – Процедура за разрешаване на анафори), разработен у Лапин (Lappin 1994), използва за откриване на антецедентите информацията от синтактично парсиран текст, като впоследствие кандидатите се филтрират с помощта на морфологични и синтактични филтри. Друг много популярен метод е предложен у Митков (Mitkov 1998). Методът използва морфологична информация, ограничена синтактична информация, извлечена от повърхнинната структура, и набор от процедури за ранкиране на кандидатите. Подобрена негова версия е имплементирана в автоматизираната система MARS (Mitkov 2002). У

Хагиги (Haghighi 2009) е предложен метод, който включва не само разрешаване на референцията на местоименните анафори, а и разпознаване на кореференцията при именни фрази. При този метод се използват синтактичните пътища между анафорите и потенциалните антецеденти, извлечени от парсиран текст, които след това се филтрират с помощта на синтактични ограничения. Съществен принос на метода е семантичният модул, чрез който се оценява семантичната съвместимост между лични имена и опори на именни фрази, като кандидатите се филтрират по семантичен критерий. Семантичната информация се извлича от неанотирани данни.

Методите, базиращи се на данни, обикновено използват анотирани с анафори корпуси, а в някои случаи и неанотирани данни (Schmolz 2015), които служат за трениране на алгоритми за машинно самообучение. Един от най-популярните подходи е предложеният от Сун (Soon 2001) метод за разрешаване както на референцията на местоименни анафори, така и на кореферентността на именни фрази. Лингвистичната обработка, която той използва, включва: токънизация, сегментация на текста на изречения, тагиране на думите по част на речта, частично синтактично парсиране, включително разпознаване на именни фрази и именувани обекти, семантични класове. Машинното самообучение използва алгоритми, основаващи се на дърво на решенията. Системата BART, предложена от Вързли (Versley 2008), използва разширен вариант на същия алгоритъм и по-обогатен набор от характеристики за машинно самообучение. Системата Reconcile, представена у Стоянов (Stoyanov 2010), която ползва същия метод, прилага и клъстериране на кандидатите, получени от класификатора, за определяне на кореферентните вериги. Урюпина (Uryupina 2010) предлага системата Corry, в която машиното самообучение съчетава метода на Сун (Soon 2001) с други модели, като използва богат набор от езиково мотивирани характеристики, включително семантични класове, извлечени от Уърднет (Miller 1995).

Толдова и колеги (Toldova 2016) представят системите за разрешаване на анафора за руски език, участвали в състезанието RU-EVAL, заедно с анализ на грешките. Повечето от изследваните типове грешки и предизвикателства са валидни и за българския като морфологично богат език – граматическа многозначност, свободен словоред, специални случаи на местоимения без референт, особености на възвратните и реципрочните местоимения и др.

За български език са известни малко разработки върху разрешаването на референцията на анафорите. Ограничен брой са и имплементираните системи, предлагащи тази функционалност.

Тук ще се спрем по-подробно на описания от Митков (Mitkov 1998) метод за разрешаване на референцията на третолични местоимения, тъй като той се базира на ограниченото използване на езиково познание и ресурси за автоматична обработка на езика. Това, от една страна, го прави

приложим към различни езици, за които няма разработени системи за семантична, синтактична и дискурсна анотация, а от друга – позволява имплементирането му в приложения, работещи в реално време с големи по обем данни.

Методът използва тагер за определяне на частите на речта и проста фразова граматика за разпознаване на именните фрази и работи с помощта на набор от индикатори на антецеденти. Като използва резултата от работата на тагера (определените по част на речта и граматични характеристики форми), системата извършва разпознаване на именните фрази на разстояние от 2 изречения вляво от анафората, извършва проверка за съвместимост на съгласувателните морфологични признаци род и число и прилага набор от индикатори за ранкиране на възможните антедецеденти на неразрешените анафори, като на всеки индикатор се приписва определено тегло, установено по емпиричен път. Антецедентът се определя като кандидата с най-голямо общо тегло от индикаторите.

Включени са следните индикатори: **определеност на именната фраза** (определените фрази са по-вероятни кандидати за антецедент); **именната фраза представя вече известна информация** (представлява информационната тема); **принадлежност на глагола, предшестващ фразата, към определена група глаголи** (именните фрази, непосредствено следващи глаголи като *дискутирам, обобщавам, представям* и под., са вероятни кандидати за антецедент); **лексикална повторителност** (дума или фраза, която се повтаря в рамките на даден абзац, е по-вероятен кандидат за антецедент); **поява на фразата в заглавие на раздел** (именните фрази в заглавия на части от текст са по-вероятни кандидати за антецеденти от останалите); **самостойност на именната фраза** (именните фрази, които не са част от предложни фрази, са по-вероятни кандидати за антецедент от фрази, които са комплементи на предлог); **колокативност** (именните фрази в колокация с глагол (глагол + NP), които имат съответствие колокат глагол + местоимение, са по-вероятни кандидати за антецедент); **непосредствено реферирane** (именни фрази в непосредствено предхождащо просто изречение с аналогична синтактична (линейна) структура като простото изречение, в което се среща анафората, са по-вероятни кандидати за антецедент); **разстояние** (именна фраза, намираща се в предходно просто изречение, е по-вероятен кандидат за антецедент от по-отдалечена фраза); **терминология от съответната област** (в терминологични текстове – термините от съответната област са по-вероятни кандидати за антецедент).

Методът е адаптиран от Танев и Митков (Tanev 2000) за разрешаване на референцията на анафорите в текстове на български език. Предложената от авторите система LINGUA включва няколко модула – за разпознаване на частта на речта, за разделяне на текста на параграфи, изречения и прости изречения, за повърхностен синтактичен анализ и за разрешаване на референцията на анафорични третолични местоимения. Освен индика-

торите, посочени по-горе, са формулирани и още няколко, включително: **именната фраза да е именуван обект** (собствените имена са по-вероятни кандидати за антецедент от нарицателните съществителни), **именната фраза да е разширена с прилагателно име** (фразите с адективни модификатори са по-вероятни кандидати от останалите).

Докато разгледаният метод се фокусира върху третоличните анафори, други автори се насочват към задачата за откриване и разрешаване на така наречената нулева анафора, при която местоименният подлог на дадено просто изречение не е изразен, но се подразбира от контекста. Тази задача изисква формулирането на две подзадачи: а) да се открие дали в дадено просто изречение има неизразен подлог; б) ако има – да се открие антецедентът на съответното лично местоимение. Този тип анафори и разрешаването им са обект на изследване от Григорова (Grigorova 2011, Григорова 2014). Авторката разрешава задачата с помощта на метод, основаващ се на евристичен алгоритъм и машинно обучение.

Коева (Коева 2014) дефинира границите на областта, в която се търси антецедентът, до предходното просто изречение, независимо дали то се намира в същото или в предходно сложно изречение. Представените правила се основават на съвместимостта на съгласувателните характеристики и близостта на антецедента, като се избира най-близкият подходящ антецедент, а при равни условия се предпочита собствено или членувано име. Ако анафората остане неразрешена, се прилага правило, което съотнася граматичната функция на анафората със съответна линейна позиция в предходното просто изречение (при неутрален словоред) – за лично именително местоимение се избира антецедент вляво от глагола, за винително местоимение – антецедент вдясно от глагола, за дателно местоимение – антецедент вдясно от глагола и след предлог.

3. Основни особености при разрешаването на анфората в българския език

При разрешаването на анафората в българския език следва да се вземат предвид и редица езиково специфични явления.

3.1. Особености на анафората в подложна позиция

При поява на анафора има няколко характеристики, които подсказват синтактичната ѝ позиция в изречението. В подложна позиция може да има следните типове анафора: а) лично местоимение в именителен падеж; б) относителното местоимение *който* (и формите му за род и число) в именителен падеж, което едновременно с това не участва в предложна фраза; в) неизразен подлог.

За българския език е типично изпускането на подлога. При липса на изразен подлог съществуват два варианта: а) глаголът да е безличен; б) подлогът да е елиптиран. Във втория случай на мястото на подлога непосредствено преди глаголната форма в простото изречение се поставя празна

подложна позиция и системата се опитва да идентифицира неговия антецедент в окръжаващия контекст.

За българския език са характерни и изречения, чийто подлог е изразен с подчинено подложно изречение. В този случай системата за откриване и разрешаване на анафори не трябва да търси лично местоимение или позиция за неизразено лично местоимение. За разпознаването на подобни случаи може да се приложи допълнителна процедура, която взема предвид лексикалните, морфосинтактичните и синтактичните особености на този тип подложни изречения. За момента те остават извън настоящата разработка.

3.2. Безлични изречения

За българския език са характерни безличните изречения, при които синтактичната позиция на подлога остава винаги свободна и не се подразбира от контекста за разлика от изреченията с изпуснат местоименен подлог (Коева 2001). Сказуемите в безличните изречения могат да са изразени: а) от безлични глаголи като *няма*, *има* и др.; б) от лични глаголи, претърпели определени синтактични трансформации (какъвто е случаят с оптативната конструкция в *Танцува ми се* (вж. Коева 2005); в) от именни или други неглаголни сказуеми: *Яд ме е, че не дойде*. Допълнителен фактор, който затруднява разпознаването, е, че някои глаголи и техни диатези могат да се появяват както в безлични, така и в лични изречения: срв. *Вали* и *Вали дъжд*; *Танцува ми се* и *Танцува ми се хоро*.

3.3. Свободен словоред

За българския език е характерен относително свободният словоред. Тази особеност предпоставя различни промени в неутралния словоред, като например подлогът да се намира вдясно, а допълнението вляво от сказуемото. Това от своя страна предполага по-малката точност на правилата, които се основават на словоредни фактори, за разлика от езици с по-строго фиксиран словоред.

3.4. Съгласуване между анафората и антецедента

В българския език личните местоимения дават информация за лицето, рода, числото и синтактичната функция на референциалния израз, който заместват, в съответното просто изречение.

Тъй като в българския език родът на съществителните имена е лексикално-граматична категория, водещ при съгласуването между антецедента и анафората по категорията род е именно родът на именната опора на антецедента. Същевременно в анотираните текстове са открити примери, в които се отдава предпочитание на семантичното пред граматичното съгласуване в случаи, когато граматичният и естественият род не съвпадат (Пример 2а).

Особеностите, които се срещат при съгласуването по число между анафората и антецедента, са свързани преди всичко със случаите на съчинително свързани антецеденти в единствено число, при което анафората е в

множествено число (Пример 2б). Тъй като в тези случаи има привидно несъответствие, за да се разпознаят съответните фрази в единствено число като кандидати за антецедент на анафората (по-конкретно – за да удовлетворят морфосинтактичните филтриращи правила, представени по-долу) е необходимо да се разпознаят координираните структури. Разрешаването на анафората в тези специфични случаи изисква по-задълбочен семантичен и/или синтактичен анализ и остава като задача за бъдеща работа.

Пример 2.
(а) *[**Момичето**]**i** се сепна и [**pro**]**i** разтри очи. Мъжът [**ù**]**i** подаде чашата.*
(б) *[**Иван** и **Мария**]**i** избягаха от час. Учителят каза [на майка [**им**]**i**]*.

4. Използвани ресурси за целите на изследването

4.1. Анотиран корпус с анафори на български език

За целите на изследването е използван Българският семантично анотиран корпус – БулСемКор (Коева 2010; Тодорова 2014). При създаването на корпуса е следвана методологията на семантично анотирания корпус СемКор (Landes 1998) в съчетание с някои специфични принципи (Коева 2010). БулСемКор е анотиран ръчно със значения от Българския уърднет (Булнет) (Коева 2007), а обемът му е съпоставим с този на много от съществуващите семантично анотирани корпуси: 101 791 токъна, от които 99 480 – анотирани езикови единици.

Важна особеност на БулСемКор в съпоставка с много от съществуващите семантично анотирани корпуси е, че се приписва значение на всички думи в корпуса, докато в традиционната практика се е наложило да се анотират преди всичко пълнозначни несъставни думи или част от тях (основно съществителни имена и глаголи). В това отношение анотацията в БулСемКор предлага по-голяма пълнота, а оттам – и по-големи възможности за езикови наблюдения и приложения в областта на компютърната лингвистика.

Корпусът съдържа подробна лингвистична анотация на различни езикови равнища, част от която е извършена в хода на създаването на семантичния корпус, а друга част е допълнена за целите на настоящото изследване:

(а) автоматична основна анотация, която включва токънизиране, сегментиране на изречения, тагиране с части на речта, лематизация и приписване на граматически характеристики, извършени с Българската многокомпонентна система за първична обработка и лингвистична анотация на текстове (Koeva 2011);

(б) ръчна семантична анотация – всяка контекстуална употреба на всяка лексикална единица е съотнесена еднозначно със семантично множество в Булнет. Подборът на най-правилното измежду възможните значения се основава на множество от процедури, при които се вземат предвид

другите членове на синонимните множества, тълковната дефиниция, мястото на синонимното множество в структурата на Булнет (Тодорова 2014);

(в) автоматично разделяне на сложните изречения на прости изречения с помощта на специален модул за определяне на границите на простите изречения в рамките на сложното (Stoyanova 2016). Разпознаването на границите на простите изречения е съществено за определяне на областта, в която се осъществява референцията при местоименията (вж. точка 6);

(г) автоматична анотация на именни фрази с помощта на модул от контекстносвободни правила за разпознаване на структурата на именните фрази и определяне на опорната им дума;

(д) маркиране на именните фрази, които представляват наименования на хора, географски обекти, организации и други;

(е) автоматично разпознаване на глаголните форми с помощта на метод, базиран на шаблони (Leseva 2016). Правилното разпознаване на глаголните форми е предусловие за прецизното прилагане на по-долу описаните процедури, при които се използва информация за залога и словореда на спомагателния и главния глагол и/или на глагола и местоименните частици, които са част от глаголната лексема;

(ж) полуавтоматична анотация на анафори чрез автоматично прилагане на предварително дефинирани правила и последвала ръчна редакция и верификация.

4.2. Процедура за анотация на анафорите

За анотация на анафорите в корпуса е използвана следната трифазова процедура:

(1) Автоматично идентифициране на всички именни фрази, тъй като са потенциални антецеденти. Стремежът е, където е възможно, съставните лексеми, включително имена, да се разпознават като една единица и един антецедент.

(2) Автоматично идентифициране на всички местоименни и нулево-местоименни анафори. Нулевите анафори са означени с **pro** и са приписани автоматично в случаите, когато в дадено просто изречение глаголът не се предхожда от именна фраза, в простото изречение има съществително име от мъжки род, членувано с кратък член, или местоимение, което не е в именителен падеж.

(3) Автоматично съотнасяне на всяка анафора с антецедент. За целта е използван прост метод за съотнасяне с най-близкия предхождащ антецедент с единственото ограничение за съгласуване по род и число между анафората и антецедента (вж. точка 5 по-долу). Този метод служи като база за сравнителен анализ на резултатите от същинския метод.

(4) Ръчна верификация и редакция на анафорите – анафорите и приписаните им антецеденти са ръчно верифицирани от експерт. Редакцията включва:

(а) отстраняване на грешно приписана нулева анафора (т.е. изречението има изразен подлог или е безлично);

(б) отстраняване на грешно тагирани анафори, ако означеният като анафора елемент представлява друга част на речта или не функционира като анафора (*то* като съюз и др.);

(в) корекция на грешно приписан антецедент;

(г) отстраняване на грешно приписан антецедент на анафора, чийто антецедент не е наличен в текста.

5. Процедури за идентифициране на анафорите

Процедура 1. Идентифициране на изразен или неизразен анафоричен подлог в рамките на простото изречение.

(1) Проверява се дали глаголът (или главният глагол при сложните форми) е личен, третоличен или безличен.

(2) Ако е разпознат от тагера като безличен, не се търси подлог.

(3) Ако е разпознат като личен, съществуват следните възможности:

(а) Да има подлог, изразен чрез именна фраза, лично местоимение, подложно изречение;

(б) Ако не може да се идентифицира подлогът, предполагаме наличието на неизразен подлог;

(4) За установяване на подлога се проверява в кое лице е глаголът:

(а) Ако е в първо или второ лице, се търси местоименен подлог (именително лично местоимение в съответното лице и число). Ако не бъде открит местоименен подлог, се приписва нулевоместоименен подлог в съответното лице и число;

(б) Ако е в трето лице, се търси потенциален подлог:

– търси се NP с пълен член или именително лично местоимение вляво и вдясно от глагола в рамките на същото просто изречение, което да заема подложната позиция;

– търси се NP (което не е комплемент на PP) между началото на простото изречение и глаголната форма;

– търси се NP (което не е комплемент на PP) вдясно от глаголната форма, ако преди нея има винително местоимение или NP, членувано с непълен член;

– търси се NP (което не е комплемент на PP) вдясно от глаголната форма, ако преди нея има въпросителна дума;

– търси се относително местоимение *който*, което е в именителен падеж (и не е част от PP), предхождащо глаголната форма;

– ако спомагателен глагол, *се* или *си* следва (главния) глагол, се търси подлог вдясно от него; ако не се намери подходящ кандидат за подлог, се приписва *pro* вдясно от сказуемото.

Процедура 2. Идентифициране на анафора във винителен или дателен падеж. Маркират се тагираните пълни форми на винителните лични местоимения; комбинациите от предлог и пълна форма на винително лично местоимение; кратките форми на винителните и дателните лични местоимения; пълните и кратките форми на възвратноличното местоимение, когато глаголът не е маркиран като рефлексивен или реципрочен; формите на винителните относителни местоимения, срещащи се самостоятелно или като комплемент на предлог и формите на именителните относителни местоимения, срещащи се като комплемент на предлог. Пълната форма на дателното лично местоимение и дателната форма на относителното местоимение *който* също се разпознават.

Процедура 3. Идентифициране на притежателни анафори. Маркират се тагираните пълни и кратки форми на притежателните и възвратните притежателни местоимения.

6. Метод за автоматично разрешаване на анафора в българския език

6.1. Базов метод за разрешаване на анафора

Базовият метод използва ограничен брой прости правила за разрешаване на анафорите. Правилата са общи за всички типове анафора. Антецедентът се търси вляво от местоимението в рамките на текущото и две предходни изречения. Налага се изискване за съгласуване по род и число между анафората и именната фраза, представляваща антецедента. Базовият метод не използва информация за границите на простите изречения в рамките на сложното, за граматическите характеристики на глаголните форми, а също така не отчита семантични особености при глагола и кандидатите за антецеденти.

Базовият метод е приложен за първоначалното автоматично анотиране на анафорите в изследвания корпус БулСемКор (вж. 4), като върху резултата от действието му е извършена ръчна верификация и редакция. На основата на анализа на грешките, допускани от базовия метод, са разработени принципите и конкретните правила за разрешаване на различните типове анафора.

6.2. Област на действие за разрешаване на различните видове анафори

Личните именителни местоимения изпълняват функцията преди всичко на подлог в простото изречение, по-рядко на предикатив (*Аз съм ти.*). Първо- и второличните именителни местоимения остават извън обхвата на задачата за разрешаването на анафорите, тъй като реферират към участниците в речевия акт и изискват дискурсен анализ, при който се вземат предвид различни контекстови фактори, включително смяната на ролите между говорещия/пишещия и възприемателя. Основен обект на формулирана-

та задача са третоличните форми, които означават кореферентност с друг именен или местоименен израз в окръжаващия контекст (Ницолова 2008: 149).

Пълните винителни форми, употребени самостойно или в съчетание с предлог, могат да бъдат пряко допълнение, непряко допълнение и обстоятелствено пояснение, докато кратките винителни форми субституират само пряко допълнение. Кратките дателни форми може да изпълняват функцията на непряко допълнение, а в съчетание с някои локативни предлози – на обстоятелствено пояснение за място (Ницолова 2008: 155).

Обикновено антецедентът на винителните и дателните местоимения се намира в предходно (просто) изречение. Едно от основните изключения представляват случаите на удвояване на допълнението, което се проявява системно в българския език, а в редица случаи е задължително. При това явление единият елемент винаги е кратко винително или дателно местоимение, а другият е именна фраза (при удвояване на прякото допълнение) или предложна фраза (при удвояването на непрякото допълнение).

От гледна точка на разрешаването на анфорите се наблюдават следните два варианта: 1) антецедентът е изразен: с именна фраза с опора съществително (или субстантивирана част) или местоимение от следните видове местоимения съществителни: въпросително, неопределително, отрицателно, обобщително (при удвояване на прякото допъление) или с предложна фраза с комплемент именна фраза или местоимение от посочените видове; 2) удвояването е реализирано чрез две анафори – пълна и кратка форма на личното местоимение или относително местоимение и кратка форма на личното местоимение.

В първия случай, след като бъдат разпознати, антецедентът и анафората следва да се маркират като кореферентни. Във втория случай следва да се маркира кореферентността между двете анафори, а антецедентът на относителното местоимение или на пълната форма на личното местоимение да се търси в предходния контекст.

Притежателните местоимения означават отношение между притежател и обект на притежанието. Пълните форми на тези местоимения носят информация за притежателя и за притежавания обект, а кратките форми – информация само за притежателя (Ницолова 2008: 165).

Антецедентът на притежателните местоимения е именно притежателят. Той може да бъде както в настоящото просто изречение, така и в предходно (просто) изречение. Антецедент в същото просто изречение е възможен, когато: а) антецедентът не е опора на фразата в подложна позиция (Пример 3а–г); б) когато даден обект е притежание на подлога и подлогът включва говорещия или пишещия (т.е. подлогът е в 1 или 2 лице) (Пример 3д–е); в) при удвояване на определението с кратко притежателно местоимение (Пример 3ж–з) (Савова 2004). Удвояването на определението се

наблюдава в разговорната реч и няма задължителен характер. За момента не са формулирани правила за разпознаването му.

Пример 3.
(а) ***[pro]i*** *Видях* ***[директора]j*** *[в колата* ***[му]j]k****.* (антецедент пряко допълнение)
(б) ***[pro]i*** *Казах* ***[на директора]j*** *[за колата* ***[му]j]k****.* (антецедент непряко допълнение)
(в) ***[pro]i*** *Влязох* ***[в магазина]j*** *[през страничния* ***[му]j*** *вход]****k****.* (антецедент обстоятелствено пояснение)
(г) *[Мнението ми* ***[за Иван]i]j*** *не променя [постъпката* ***[му]i]k****.* (антецедент разширение на подлога)
(д) ***[pro]i*** *Купих подаръка [на майка* ***[си/ми]i]j****.*
(е) ***[pro]i*** *Дай играчката [на брат* ***[си/ти]i****].*
(ж) *[Брат* ***[му]i]j*** *[****на Иван]i*** *замина за чужбина.*
(з) ***[На нея]i*** *[брат* ***[ѝ]i]j*** *замина за чужбина.*

Както пълните, така и кратките форми на личните и притежателните местоимения показват лицето и числото на антецедента, а в трето лице единствено число – и рода. Възвратното местоимение означава кореферентност с антецедент, който е главен аргумент в състава на просто изречение или конструкция с второстепенен предикат. Антецедентът е: а) най-често подлог (Пример 4а); б) пряко или непряко допълнение, когато допълнението е експериенцер (субект на психическо или физиологическо състояние) (Пример 4б, в); в) антецедентът може да е субект на второстепенен предикат: причастие (Пример 4г), деепричастие, прилагателно (Пример 4д), отглаголно съществително (Пример 4е); субектът на второстепенния предикат може да бъде (Пример 4г, д) или да не бъде (Пример 4е) кореферентен с подлога (Ницолова 2008: 158–159).

Пример 4.
(а) *В този момент* ***[тя]i*** *мразеше* ***[себе си]i.***
(б) *Болно* ***[му]i*** *беше преди всичко* ***[за себе си]i.***
(в) ***[Него]i*** ***[го]i*** *беше страх* ***[за себе си]i.***
(г) *В помещението нахълтват* ***[неколцина пожарникари]i,*** ***[pro]i*** *мъкнещи* ***[със себе си]i*** *маркуч и пожарогасители.*
(д) ***[pro]i*** *Доволен* ***[от себе си]i,*** ***[той]i*** *застана пред огледалото.*
(е) *Често* ***[го]j*** ***[pro]i*** *порицаваха за доволството* ***[от себе си]j,*** *което* ***[pro]j*** *демонстрираше.*

Кратките форми изпълняват функция на допълнение, което е кореферентно с подлога или със субект на вторичен предикат (причастие).

Пример 5.
[Тя]i *стоеше там,* ***[pro]i*** *разглеждайки* ***[се]i*** *с удоволствие в огледалото.*

Основното затруднение при разпознаването на кратките възвратни местоимения идва от това, че прономиналните частици *се* и *си* участват в състава на глаголни лексеми. Ако глаголите, с които формите *се* и *си* се срещат, са тагирани като рефлексивни или реципрочни глаголи, ги смятаме за частици, в противен случай – за местоимения.

Антецедент на възвратното притежателно местоимение може да е подлогът в изречението или по-рядко субектът на второстепенен предикат (нелична глаголна форма, прилагателно) (Ницолова 2008: 171). Приемаме, че употребата му е задължителна, когато подлогът на изречението (или субектът на второстепенния предикат) не е нито говорещият, нито слушащият (ОПРБЕ: 28) (иначе се конкурират с притежателните местоимения). В действителност ситуацията е доста по-комплексна (вж. Ницолова 2008: 171–174).

6.3. Основни правила, използвани за разрешаване на анафора

За разрешаване на анафората в настоящото изследване се използват синтактични и морфосинтактични правила.

Синтактичните правила се определят от функцията на анафората като част от простото изречение, в което се намира, и произтичащите от това ограничения, свързани с обсега, в който се търси референтът. В Пример 6в *му* не може да се отнася към подлога на изречението.

Морфосинтактичните правила засягат, от една страна, съгласуването на анафората в подложна позиция със сказуемото по лице и число и когато е приложимо – по род (6а, 6б), а от друга страна – съгласуването на анафората и нейния антецедент по граматичните им характеристики род и число (6в).

Пример 6.
(а) *[**Мария** и **Иван]i** се скарали.*
(б) *[**Тя]i** много често отсъствала.*
(в) *Иван върна [**на Тодор]i** [книгата [**му]i]j.*

Правилата за разрешаване на анафора са от два типа – филтриращи и ранкиращи. Филтриращите отстраняват неподходящите кандидати, докато ранкиращите дават количествена оценка на всеки кандидат, така че да се позволи подбиране на най-подходящия, без другите да бъдат изключени. Ранкиращите правила се прилагат последователно, при което стойността на оценката се обновява.

По-долу са представени обобщени правила, които се прилагат в различни конфигурации за отделните типове анафора.

Правило 1. (филтриране) Определяне на обсега, в който да се търси антецедентът.
Правило 1А. Антецедентът се търси само в текущото просто изречение.
Правило 1Б. Антецедентът се търси само в предходни прости изречения.

Правило 1В. Антецедентът се търси както в текущото просто изречение, така и в предходни.

Правило 2. (филтриране) Съгласуване по морфосинтактични характеристики.

Правило 2А. Антецедентът се съгласува с анафората по род и число.

Правило 2Б. Анафората (а от там и антецедентът) се съгласува с глагола по лице, число и ако е приложимо – по род.

Правило 2В. Съгласуване с координирани части: анафора в множествено число може да се съгласува с цяла фраза, съдържаща координирани части в единствено или множествено число.

Правило 3. (ранкиране) Идентифициране на антецедента на анафоричен подлог.

Правило 3А. Ако при последователни прости изречения в едно сложно изречение е изразен само един подлог, като се запазва лицето и числото на глагола, този подлог определяме като антецедент на *pro* в другите прости изречения (Пример 7а и 7б).

Правило 3Б. Ако във второто от две поредни прости изречения има изразен местоименен подлог, най-вероятно има смяна на подлога (Пример 7в и 7г).

Пример 7.

(а) *След като **[pro]i** влезе, **[той]i** затвори вратата и **[pro]i** заключи.*

(б) *След като **[pro]i** влезе, [родителите [му]**i**]**j** затвориха вратата и **[pro]j** заключиха.*

(в) ***[Иван]i** опита **[pro]i** да убеди **[Мария]j,** но **[тя]j** все пак не се отказа от искането си.*

(г) ***[Иван]i** опита **[pro]i** да убеди **[Мария]j,** но **[pro]i** накрая се отказа.*

Правило 4. (ранкиране) Семантично съответствие на анафората като аргумент на глагола.

Правило 4А. При глагол в деятелен залог, изискващ одушевен субект, изразен с местоимение, се търси одушевен антецедент.

Правило 4Б. При глагол в деятелен залог, изискващ неодушевен субект, изразен с местоимение, се търси неодушевен антецедент.

Правило 5. (филтриране) Допълнителни условия за обсега при възвратните анафори.

Правило 5А. Антецедентът на възвратното местоимение е подлог в рамките на простото изречение (с изключение на условията в 5Б).

Правило 5Б. Ако местоимението е в (предикативна) конструкция с опора причастие, деепричастие или прилагателно, антецедентът е (неизразеният) субект на този предикат, който най-често е кореферентен с най-близкото изразено местоимение или съществително вляво от второстепенния предикат.

Правило 6. (филтриране) Допълнителни правила за всички невъзвратни личноместоименни анафори.
Правило 6А. Ако местоимението е в (предикативна) конструкция с опора причастие, деепричастие или прилагателно, антецедентът не може да бъде (неизразеният) субект на този предикат и не може да бъде кореферентен с най-близкото изразено местоимение или съществително вляво от второстепенния предикат.

Правило 7. (филтриране) Допълнителни правила за анафора притежателно местоимение.
Правило 7А. Антецедентът на притежателното местоимение не може да е подлог в текущото просто изречение или субект на предикативна $_{\text{конструкция}}$ (неговият антецедент трябва да се търси в предходни изречения – вж. Пример 8).

Пример 8.
*[**Той]i** може да си изгради [[**образ]j** на механизма], [**pro]j** отговарящ на всички [[**негови]i** наблюдения]**k.***
Правило 7Б. Ако притежателното местоимение е в границите на пряко допълнение на просто изречение, неговият антецедент вероятно трябва да се търси в предходни прости изречения.
Правило 7В. Ако притежателното местоимение е в границите на непряко допълнение или обстоятелствено пояснение в просто изречение, неговият антецедент може да бъде пряко допълнение или непряко допълнение в същото просто изречение.

В Таблица 1 са представени комбинациите от правила за разрешаване на референции при отделните типове анафора.

Таблица 1. Приложимост на правилата за отделните типове анафора

Тип анафора	**Правила**
Лични местоимения в именителен падеж и 3 лице	1Б, 2А, 2Б, 2В, 3Б, 4А, 4Б
Лични местоимения във винителен и дателен падеж (дълги и кратки форми)	1В, 2А, 2В, 6А
Притежателни местоимения (дълги и кратки форми)	1В, 2А, 2В, 6А, 7А, 7Б, 7В
Относителни местоимения в именителен падеж, които не са в предложна фраза	1Б, 2А, 2Б, 2В, 3Б, 4А, 4Б

Относителни местоимения във винителен или дателен падеж или в предложна фраза	1Б, 2А, 2В, 6А
Възвратни лични местоимения	1А, 5А, 5Б
Възвратни притежателни местоимения	1А, 2А, 5А,5Б
Неизразен подлог	1Б, 2Б, 2В, 3А, 4А, 4Б

6.3. Ранкиране

С цел да се подбере най-подходящият антецедент, се извършва ранкиране на кандидатите въз основа на множество от характеристики. Част от тях са заимствани от Митков (Mitkov 1998), докато други са обобщени или новоформулирани. Дадени са и възможни тежести на всяка характеристика, които могат да бъдат променяни за оптимизиране на резултата.

(1) Повторение – при срещане на едно съществително име два или повече пъти в рамките на дефинирания обхват на референция съществителното е по-вероятен кандидат за антецедент, тъй като представлява значима за информационната структура единица (+10%).

(2) Участие в колокации – ако анафората и даден потенциален антецедент участват в едни и същи колокации, то този антецедент е по-вероятен кандидат от потенциалните антецеденти, неучастващи в колокации с анафората (+10%).

(3) Синтагматичен паралелизъм – подобната линейна структура на простите изречения, които съдържат съответно анафората и потенциалния антецедент, както и еднаквата им синтактична позиция също покачват достоверността на кандидата (+10%).

***[pro]i** Измийте **[гроздето]j** и**[pro]i** **[го]j** изяжте.*

[pro]i** Измийте **[гроздето]j.** **[pro]i** Изяжте **[го]j.

(4) Разширена именна фраза – дава се преимущество на кандидати, които имат съгласувано определение (по Mitkov 1998) (+5%).

(5) Членувана именна фраза – дава се преимущество на членувани кандидати за антецеденти (+5%).

(6) Имена – имат преимущество подобно на членуваните фрази (+5%).

(7) Предложна фраза – антецеденти в предложна фраза са по-малко вероятни (–5%).

(8) Дистанция между анафората и кандидата за антецедент – може да се измерва както в брой изречения или в брой прости изречения, така и в брой думи. Тук се предпочита измерването в брой прости изречения (–5%* |n–1|, където n е броят прости изречения, разделящи изречението на анафората и изречението на антецедента – по този начин единствено канди-

дати в предходното просто изречение не се „наказват“, тъй като това е най-вероятното местоположение на антецедента).

6.4. Приложение на метода за разрешаване на анафора

Методът се прилага в следната последователност:

(1) Анотация и предварителна обработка на текста, разделяне на прости изречения;

(2) Идентифициране на определени типове фрази, които са от значение за идентифицирането на анафорите и разрешаване на референцията:

(а) именни фрази с техните граматически характеристики като род и число (за проверка на съгласуване), определеност (за някои от дефинираните ранкиращи правила) и др.;

(б) пълните глаголни форми с техните граматически характеристики като лице и число (за съгласуване с подлога), род при някои форми, съдържащи причастия (също за съгласуване с подлога), залог (за откриване на подлога) и др.;

(3) Идентифициране на анафорите с техните граматически (съгласувателни) характеристики;

(4) За всяка анафора – извличане на възможните кандидати в даден обсег (обсегът е фиксиран – обхваща текущото изречение и 2 изречения преди него);

(5) Върху възможните кандидати – прилагане на филтриране;

(6) Ранкиране по зададените преференции;

(7) Идентифициране на най-високо ранкирания кандидат (ако има такъв).

6.5. Оценка на резултатите

Докладваните резултати са предварителни, тъй като работата по формулиране, тестване и оптимизиране на правилата за идентифициране и разрешаване на анафора продължава. Идентифицирането на анафорите става въз основа на тагирането с части на речта, както и на правилата, описани в раздел 5. В корпуса са идентифицирани следните анафори: 668 лични местоимения в именителен падеж, 747 лични винителни и дателни местоимения, 968 относителни местоимения, 729 притежателни местоимения, 1197 възвратнолични и 582 възвратни притежателни местоимения. По описания по-горе метод (точка 5) са идентифицирани и 6286 неизразени подлога, от които 4698 (74.7%) са правилно разпознати.

Базовият метод разрешава референциите с точност 45,1%, докато описаният метод, използващ правилата (точка 6.2), достига точност от 63,6%. Макар и подобрението да е значително, то не е достатъчно за гарантиране на достоверен и надежден резултат. Прилагането на преференциите за ранкиране на кандидатите (точка 6.3) подобрява резултатите до точност 69,7%. Тестовете върху приложението на правилата за ранкиране продължават, с цел да се адаптират към спецификите на българския език и да отразяват реалните количествени съотношения между различните типове кандидати.

7. Насоки за бъдеща работа

Методът ще бъде развиван в посока на включване на повече семантични критерии за осъществяване на сполучливо разрешаване на анафората. Може да бъдат наблюдавани значението на отделните кандидати, тяхната съчетаемост с определени глаголи, контекстът, в който се срещат, и др. По-конкретно една от посоките на работа е формулирането на евристични семантични правила, основаващи се на семантична информация, извлечена от Булнет, които налагат ограничения върху антецедента на анафората в различни синтактични позиции по отношение на категорията одушевеност.

Друга посока на изследване е използването на по-богата синтактична информация, включително йерархизиране на синтактичните позиции от гледна точка на вероятността те да са антецедент на дадена анафора (Lappin 1994). В изследването си авторите предлагат йерархия от вида: подлогът е по-вероятен кандидат от несубектно NP; прякото допълнение е по-вероятен кандидат от останалите комплементи; аргументите на предиката са по-вероятни кандидати от адюнктите.

Резултатите до момента сочат, че използваният метод със съчетаване на филтриращи и ранкиращи правила е приложим за успешното разрешаване на референциите при анафората. Задълбочените емпирични тестове върху ранкиращите правила ще допринесат за по-ефективната параметризация на ранкирането, което може да доведе до значително подобряване на резултатите.

ЛИТЕРАТУРА

Григорова 2014: *Григорова, Д.* Евристичен алгоритъм и машинно обучение за разрешаване на нулево-местоименна анафора в текстове на български език. Дисертация.

Коева 2001: *Коева, Св.* Кратка практическа граматика на българския език. София, Труд.

Коева 2005: *Коева, Св*. Синтактични трансформации. – В: Аргументна структура. Проблеми на простото и сложното изречение, София, СемаРШ.

Коева 2007: *Коева, Св*. БулНет (лексикално-семантична мрежа на българския език) – част от световната лексикално-семантична мрежа. – БЕ, № 1, с. 34–50.

Коева 2010: *Коева, Св*. (ред. и съст.). Българският семантично анотиран корпус. София, ИБЕ–БАН.

Коева 2014: *Коева, Св*. Лингвистична анотация. – В: Известия на Института за български език „Проф. Любомир Андрейчин“, XXVII, с. 7–33.

Ницолова 2008: *Ницолова, Р.* Българска граматика. Морфология. София, Университетско издателство „Св. Климент Охридски“.

Савова 2004: *Савова, И*. Удвояване на определението в българския език. – Български език и литература, № 4. Електронна публикация – списание Liter Net, 22 януари 2004, № 1 (50).

Тодорова 2014: *Тодорова М., Хр. Кукова, Св. Лесева*. Семантично анотирани ресурси за българския език – БулСемКор. – В: Езикови ресурси и технологии за български език. София, Академично издателство „Проф. Марин Дринов“, с. 80–104.

GRIGOROVA D. 2011. Zero Pronoun resolution in Bulgarian. – In: Proceedings of the International Conference on Computer Systems and Technologies – Comp SysTech'11.

HAGHIGHI 2009: *Haghighi, A., D. Klein*. 2009. Simple Coreference Resolution with Rich Syntactic and Semantic Features. – In: Proceedings of the 2009 Conference on Empirical Conference in Natural Language Processing.

HALLIDAY 1976: *Halliday, M. A. K., R. Hasan*. 1976. Cohesion in English. London: Longman.

HIRST G. 1981. Anaphora in Natural Language Understanding: A Survey. Volume 119 of the Series Lecture Notes in Computer Science, Springer-Verlag, 4–32.

HOBBS JERRY R. 1978. Resolving Pronoun References. Lingua, 44, 339–352.

KOEVA 2011: *Koeva, S., A. Genov*. Bulgarian Language Processing Chain. – In: Proceeding of the Workshop Integration of multilingual resources and tools in Web applications, Hamburg.

LANDES 1998: *Landes, S., C. Leacock, C. Fellbaum*. Building Semantic Concordances. – In: WordNet: An Electronic Lexical Database, Chapter 8, 199–216.

LAPPIN 1994: *Lappin, S., H. Leass*. An Algorithm for Pronominal Anaphora Resolution. – Computational Linguistics, 20(4): 535 – 561.

LESEVA 2015: *Leseva, S., I. Stoyanova, S. Koeva*. 2015. Automatic Recognition of Verb Forms in Bulgarian. – In: Paisievi Chetenia, 30–31 October, 2015.

LOZANOVA 2013: *Lozanova, S., I. Stoyanova, S. Leseva, S. Koeva, B. Savtchev.* Text Modification for Bulgarian Sign Language Users. – In: Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations, ACL 2013, Sofia.

MILLER G. A. 1995. WordNet: A Lexical Database for English. – Communications of the ACM Vol. 38, No. 11: 39–41.

MITKOV R. 1998. Robust Pronoun Resolution with Limited Knowledge. – In: Proceedings of the 18th International Conference on Computational Linguistics (COLING'98)/ACL'98 Conference, 869–875, Montreal, Canada.

MITKOV R. 1999. Anaphora Resolution: The State Of The Art. Wolverhampton University Technical Report, UK.

MITKOV R. 2002: *Mitkov, R., R. Evans, C. Orasan.* A New, Fully Automatic Version of Mitkov's Knowledge-poor Pronoun Resolution Method. – In: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, 168–186. Springer-Verlag.

SCHMOLZ H. 2015. Anaphora Resolution and Text Retrieval. A Linguistic Analysis of Hypertexts. Series:Empirische Linguistik / Empirical Linguistics 3. De Gruyter Mouton.

SOON 2001: *Soon, W. M., H. T. Ng, D. C. Y. Lim.* A Machine Learning Approach to Coreference Resolution of Noun Phrases. – Computational Linguistics, 27(4), 521–544.

STOYANOV 2010: *Stoyanov, V., Cardie, C., Gilbert, N., Riloff, E, Buttler, D., Hysom, D*. Coreference Resolution with Reconcile. – In: Proceedings of the ACL 2010 Conference Short Papers.

STOYANOVA 2016: *Stoyanova, I., S. Koeva, S. Leseva*. 2016. Clause Splitting for Bulgarian. – In: Proceedings of the Tenth International Conference on Natural Language Processing (HrTAL2016) (под печат).

TANEV 2000: *Tanev H., R. Mitkov.* 2000. LINGUA – A Robust Architecture for Text Processing and Anaphora Resolution in Bulgarian. – In: Proceedings of the International Conference on Machine Translation and Multilingual Applications in the New Millenium (MT-2000), 20-1-20-8, Exeter, UK.

TOLDOVA 2016: *Toldova, S., I. Azerkovich, A. Ladygina, A. Roitberg, M. Vasilyeva.* Error Analysis for Anaphora Resolution in Russian: New Challenging Issues for Anaphora Resolution Task in a Morphologically Rich Language. – In: Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes, CORBON @NAACL-HLT 2016, June 16, 2016, San Diego, California, USA, 74–83.

URYUPINA O. 2010. Corry: A System for Coreference Resolution. – In: Proceedings of the 5th International Workshop on Semantic Eval uation (Sem Eval'10).

VERSLEY 2008: *Versley, Y., S. P. Ponzetto, M. Poesio, V. Eidelman, A. Jern, J. Smith, X. Yang, A. Moschitti*. BART: A Modular Toolkit for Coreference Resolution. Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).

✉ *Гл. ас. д-р Светлозара Лесева*
✉ *Гл. ас. д-р Ивелина Стоянова*
Секция по компютърна лингвистика
Институт за български език „Проф. Л. Андрейчин" при БАН
бул. „Шипченски проход" 52, бл. 17, 1113 София, България
zarka@dcl.bas.bg
iva@dcl.bas.bg

✉ *Assist. Prof. Svetlozara Leseva, PhD*
✉ *Assist. Prof. Ivelina Stoyanova, PhD*
Department of Computational Linguistics
Institute for Bulgarian Language, Bulgarian Academy of Sciences
52 Shipchenski prohod blvd., bl. 17, 1113 Sofia, Bulgaria
zarka@dcl.bas.bg
iva@dcl.bas.bg