БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ
„ИЗВЕСТИЯ“ НА ИНСТИТУТА ЗА БЪЛГАРСКИ ЕЗИК „Проф. ЛЮБОМИР АНДРЕЙЧИН“
Книга XXVII

ЛИНГВИСТИЧНА АНОТАЦИЯ

Светла Коева

Резюме. Еднозначното разграничаване и лингвистичната анотация на езиковите единици са важни за правилното им групиране в надредни езикови единици, за установяването на граматични и семантични връзки и за правилното анализиране на различните видове значение. В студията представяме накратко основните принципи, които са приети при разделянето на български писмен текст на думи и изречения, при определянето на граматичните характеристики и основната форма на лексикалните единици, при определянето на границите и опората на някои типове именни словосъчетания и при определянето на някои анафорични зависимости. Докато определянето на границите на езиковите единици има общ характер – не е съобразено с конкретно приложение при компютърната обработка на езика, анализирането на именните фрази и анафоричните зависимости е свързано с конкретна цел, а именно – идентифицирането на относително малък брой лингвистични единици (ключови думи или фрази), които могат да се асоциират със съдържанието на документа.

Abstract. The unambiguous splitting of language units and the assignment of relevant linguistic annotation are important for proper analysis of bigger linguistic units, identification of existing grammatical and semantic relations and adequate interpretation of different meanings. In this study we briefly present the basic principles adopted in automatic splitting of Bulgarian texts in words and sentences, automatic tagging for part of speech, lemma and grammatical features of the words, and automatic recognition of certain types of noun phrases, named entities and anaphora resolution. While the delimitation of linguistic units does not comply with a specific application in natural language processing, the recognition of noun phrases, named entities and anaphora resolution presented here is related with a specific purpose, namely the identification of relatively small number of linguistic units (keywords or phrases) that are relevant for the semantic interpretation of the content of the document.

Keywords: linguistic annotation, natural language processing.

1. Въведение

Автоматичното разделяне на текста на единиците, които го съставят (думи, словосъчетания, прости изречения в състава на сложното изречение, изречения, параграфи), е относително проста, но важна задача. По-сложна и също важна задача е автоматичното приписване на лингвистична информация към езиковите единици (**лингвистичната анотация**): информация за синтактичните отношения между простите изречения в състава на сложното и техния вид; информация за синтактичните отношения между лексикалните и синтактичните категории, които образуват словосъчетания; информация за основната форма и граматичните характеристики на думите и съставните думи и т.н.

Еднозначното разграничаване и характеризиране на езиковите единици е важно както за правилното им групиране в надредни езикови единици, така и за правилното анализиране на значението на дадена езикова единица в обхвата на по-голямата езикова структура, в която е включена.

Лексикалната единица означава уникално понятие, затова под лексикална единица разбираме абстрактната двойка: лема – лексикално значение. Лексикалната единица може да бъде както дума, така и съставна дума[1] (две или повече думи, не задължително последователни, означаващи уникално и константно понятие, в релация на еквивалентност с дума от същия или друг език) (Коева 2006: 201).

При автоматичен анализ различни езикови единици могат да се използват като характеристики, ключови за идентификацията на съдържанието на документа – словоформи, основни форми на думите, именни фрази, собствени имена, граматични характеристики, n-грами[2] от думи, n-грами от фрази, n-грами от граматични характеристики, смесени n-грами и т.н. Често използвана и относително проста техника за претегляне на корелацията между конкретна характеристика и характеристиките на целия документ е честотата на срещане tf_{ij} – където tf е броят срещания на характеристиката i в документа j (Манинг, Рагаван и Шютце 2008). Модификация на този вид претегляне е нормализираната честота tf_{ij}/n, където n е броят на всички характеристики в документа. Връзката на характеристиката спрямо колекция от документи се претегля с $\log (N/df_i)$ – където df_i показва броя на документите, в които i се среща, а N е броят на всички документи. Tf*idf отразява отношението между локалната и глобалната информация за характеристиката в множество от документи – $tf_{ij}*\log(N/df_i)$, и дава повече тежест на единици с голяма честота в даден документ, ако не се срещат в много документи. Ясно е обаче, че честотата може да е ниска за целия документ, но висока за дадена част от документа, особено при относително дълги документи. Това налага по-комплексни методи за изчисление, за да се осигури по-голяма достоверност на резулта-

---

[1] В настоящата студия употребяваме като синоними термините *дума – лексикална единица*, *съставна дума – съставна лексикална единица*.

[2] Последователност от n на брой елемента.

та. Например идентифицирането на ключови за съдържанието на документа думи може да се базира не само на честотата на срещане, но и на относителното място (например на първото срещане) на ключовата дума в документа.

Характеристиките, които се приписват на даден документ, се базират на свойствата на лингвистичните единици, които го съставят. Затова е важен изборът: (а) на лингвистичните единици, които се приемат за важни за изразяване на съдържанието на даден документ, (б) на характеристиките, които експлицитно се асоциират с тези лингвистични единици. Добре избраните отношения между лингвистични единици и характеристики могат да подобрят тяхното групиране и ефективното извличане на ключовите за съдържанието на даден документ думи. Някои от обичайните филтри са: отстраняване на рядко срещани думи под определен праг, отстраняване на най-често срещаните думи – служебни думи и други, обичайно наричани стоп думи, и автоматичното определяне на основните форми (лематизация). Разбира се, всеки избор на праг, под/над който дадена честота се смята за ниска или висока, може да повлияе на качеството на резултата.

В студията представяме накратко основните принципи, които са приети при разделянето на писмен български текст на думи и изречения, при определянето на граматичните характеристики и основната форма на лексикалните единици, на границите и главната част на някои типове именни словосъчетания и на някои анафорични зависимости. Докато определянето на границите на езиковите единици, представено в студията, има общ характер – не е съобразено с конкретно приложение при компютърната обработка на езика, разпознаването на именни фрази и анафорични зависимости е свързано с определена цел, а именно – с идентифицирането на относително малък брой лингвистични единици (ключови думи или фрази), които могат да се асоциират със съдържанието на документа.

2. Определяне на границите на изречение

За определянето на границите на изречение с относителна степен на достоверност за много езици е достатъчно следното правило:

Ако бъде намерена последователност от пунктуационен знак за край на изречение[3], празен символ[4], главна буква[5], маркер за край на изречение се поставя след пунктуационния знак.

---

[3] Пунктуационните знаци за край на изречение за български са точка, многоточие, двоеточие, въпросителен знак, удивителен знак, комбинация от въпросителен и удивителен знак. След знака за край на изречение може да има затваряща скоба или затваряща кавичка или комбинация от тях.

[4] Празният символ може да е интервал, включително табулация или нов ред.

[5] Българско изречение може да започне с главна буква на кирилица или латиница или с цифра. Отпред може да има отваряща скоба, отваряща кавичка, комбинация от тях или знак за пряка реч.

Това правило обаче не е прецизно във всички случаи, защото думата, която започва с главна буква след точка (не само от българската азбука), може да е собствено име след съкращение (*ас. Иванов*) или начало на ново изречение (включително собствено име в началото на ново изречение). За да се илюстрират накратко трудностите, ще бъдат изброени възможностите за употреба на точка в нормативно правилни български текстове:

• пунктуационна употреба на точката – край на изречение, част от многоточие;

• непунктуационна употреба на точката – част от дроб, част от номерация при изброяване, част от дата, част от графично съкращение, част от лексикално съкращение;

• смесена употреба – част от съкращение и край на изречение едновременно (по-рядко).

Непунктуационната употреба на точка се открива лесно с регулярни изрази (последователност от символи, синтактично организирани като шаблон, който описва множество от последователности от символи). Например следният регулярен израз открива комбинации от арабски цифри, латински цифри и точка, които означават дати:

((|0)[1-9])|([1|2][0-9])|(30)|(31)\.((|0)[1-9])|(10)|(11)|(12)|((|I)V)|V|(I|II|III)|((|I)X)|X|(I|II)\.((19)[0-9][0-9])|([0-9][0-9])|([20][0-9][0-9])|([0-9][0-9])

Въпросът за определяне на границите на изречение би бил частично решен, ако има информация дали думата в началото на изречението е (част от) собствено име, или не е. Ако думата с главна буква не е собствено име, тогава горното правило със сигурност идентифицира край на изречение. Обратно, ако точката не е край на изречение, тогава думата с главна буква е собствено име. Ако думата пред точка не е съкращение, тогава горното правило идентифицира край на изречение. Обратно, ако думата пред точка е съкращение, което изисква собствено име, то точката не маркира край на изречение. Формулирани са следните правила (Микеев 2002):

Ако точката следва дума, която не е съкращение, точката означава край на изречение.

Ако точката следва съкращение, което е последната дума в параграф, точката означава край на изречение (и е част от съкращението).

Ако точката следва съкращение и след нея не следва дума с главна буква, точката не означава край на изречение (и е част от съкращението).

Ако точката следва съкращение и след нея следва дума с главна буква, която не е собствено име, точката означава край на изречение (и е част от съкращението).

Проблемите, свързани с използването на горните правила, са, че те се основават на разпознаването на собствените имена и съкращенията, а няма и

не може да има пълен списък на съкращенията, нито пълен списък на собствените имена. Също така, както някои съкращения, така и някои собствени имена съвпадат с думи, които не са съкращения или собствени имена (*с, Маргарита*). Част от съкращенията са многозначни: *с.* може да означава както *село*, така и *страница*, докато *стр.* означава само *страница*. В зависимост от начина, по който е прието да се означават пунктуационно, графичните съкращения могат да завършват с точка (*стр., с., гр.*), или не (*л, м, км*), но това разграничение няма отношение към показателите за определяне на края на изречението.

Може да се използва проста евристика за автоматично извличане на кандидатите за съкращения – ако точката не се следва от дума с малка буква, цифра или друга пунктуация, то най-вероятно точката означава графично съкращение (Графенстет и Тапанайнен 1994). За определяне на границите на българско изречение се използват регулярни правила, асоциирани с речници на българските графични съкращения, в които съкращенията са класифицирани в зависимост от позицията в изречението, в която се срещат, и в зависимост от това дали задължително изискват собствено име, число, или и двете. В български можем да отделим следните групи графични съкращения спрямо позицията им в изречението:

• графични съкращения, които се срещат само в средата на изречението: например *инж., проф., ас., полк., гр., р.*;

• графични съкращения, които се срещат както в средата, така и в края на изречението: например *стр., г.*

В зависимост от думата, която следва след графичните съкращения в български, можем да отделим следните групи (имащи значение при прилагането на правила за определяне на границите на изречение):

• графични съкращения, след които винаги има собствено име с главна буква: например *полк., инж., р.*;

• графични съкращения, след които винаги има число: например *тел., вътр.*;

• графични съкращения, след които винаги има число или следват число или числително име: например *стр., ал.*;

• графични съкращения, след които има или собствено име, или число: например *ул.*, *бл.*

Точката след графичните съкращения, които се срещат само в средата на изречението и след които следва или собствено име, или число, не означава край на изречение (и е част от съкращението). Честотата на употреба на някои от тези съкращения е голяма. Например съкращението *чл.* (член) се среща 36 001 пъти в извадка от около 300 милиона думи от Българския национален корпус, *ал.* (алинея) – 21 312 пъти, *стр.* (страница) – 12 345 пъти, *проф.* (професор) – 3680 пъти, и т.н.

Графичните съкращения на лични и бащини имена не са описани в речник, тъй като се подчиняват на просто правило – съкращават се след първата

буква или след първата буква и една или повече съгласни. Начините за съкращаване на лични и бащини имена са описани с регулярни изрази, за да се избегне неправилното поставяне на разделител на изречения в случаи като *С. Коева, Св. Коева* и др. В някои науки се използва главна буква или комбинация от главни букви за различни типове означения, например *витамин С, ъгъл А* и др. Ако след тях следват точка и собствено име, то тази точка е знак за край на изречение, а не за съкращение – възможните случаи също са описани с регулярни изрази.

Един от основните проблеми за определянето на граница на изречение в български текстове е (не)пряката реч: в кои случаи думите на автора и пряката реч се приемат за едно изречение. Разграничаваме следните случаи (Коева 2001):

Ако въвеждащото авторово изречение е пред пряката реч, след него се поставя двоеточие, а пряката реч започва с тире на нов ред.

Ако авторовите думи са след пряката реч, те започват с малка буква. Отделят се от нея с тире, като след пряката реч се поставя такъв препинателен знак (но не и точка), какъвто изисква смисълът на изречението. В този случай авторовата и пряката реч също се разглеждат като главно и подчинено изречение, т.е. като едно изречение. Ако пряката реч продължава, тя се отделя като ново изречение.

Когато авторовите думи се намират между частите на пряката реч, те се отделят от нея с тирета от двете страни и започват с малка буква. Ако след пряката реч, където се вмъкват думите на автора, има запетая, тя се пренася след авторовите думи. В този случай авторовата и пряката реч също се разглеждат като главно и подчинено изречение, т.е. като едно изречение.

Вместо с тире и на нов ред пряката реч може да бъде отделена и с кавички. Няма специални условия, които да налагат избора на кавички или на тире. Правописните правила при пряката реч, въведена с кавички, са аналогични на правописните правила при пряката реч, въведена с тире. Разликата е, че когато се огражда с кавички, пряката реч не се изнася на нов ред.

В обобщение, за автоматично определяне на границата на изречение се използват относително прости правила, които са формулирани като регулярни изрази и които отчитат употребата на пунктуация за край на изречение, непунктуационната употреба на точка и правилата за употреба на пряка реч.

3. Определяне на графичните думи в текста[6]

Писменият текст представлява последователност от символи, включително празните символи, – при автоматична обработка на текста трябва да е ясно кои последователности от символи съответстват на думи от даден език. Думите в писмен български текст (като думите в останалите индоевропейски езици и за разлика от някои други езици, например японски и китайски) представляват последователност от букви между празни символи. Това, разбира се, не се отнася за всички думи, защото думите освен букви могат да съдържат пунктуационни знаци в непунктуационна употреба (например тире при полуслято писане на сложни думи или точка при графични съкращения); комбинация от букви, цифри и тире (*4-и*), само цифри (*1961*) или цифри и тире или точка (*10.10.2010*). Има и много случаи, при които думите съдържат празни символи (така наречените съставни лексикални единици – *смея се, Стара Загора, Обединена българска банка*).

В произволен текст е възможно да се срещнат последователности от букви, цифри, пунктуационни знаци, специални символи, празни символи и комбинации от изброените. Просто правило идентифицира поредици от символи между два празни символа и отделя с други празни символи пунктуационните знаци в началото и края на поредиците (например запетая, двоеточие, кавички, скоби) – така се обособяват графичните думи, които могат да са или не (част от) лексикална единица (например *добре, алфа-лъчи, Стара (планина), *ааба*). Точката при графичните съкращения е част от самото съкращение и не се отделя. На това равнище задачата е да се определят графичните думи и посредством дефиниране на „типа“ им да се определи кои графични думи са лексикални единици от българския език (например н*ов, ?вно,* но **нов-1, *нов-*) и кои последователности от графични думи еднозначно могат да се класифицират като съставни лексикални единици (*благодарение на*), собствени имена (*Иван Петров*), изрази за време (*10.X.)*, числа, парични стойности и др. (*10%, http://dcl.bas.bg*).

Приписва се следната анотация за комбинациите от символи[7]: малки кирилски букви (и малко тире); първа главна и следващи малки кирилски букви (и малко тире); главни кирилски букви (и малко тире); главни и малки кирилски букви (и малко тире); малки латински букви (и малко тире); първа главна и следващи малки латински букви (и малко тире); главни латински букви (и малко тире); главни и малки латински букви (и малко тире); цифри; цифри, малко тире и главни и малки кирилски букви; главни и малки кирилски и латински букви и тире. Някои последователности от символи съответстват на

---

[6] Познато е под термина *токънизация*, тук ще използваме термина *графична дума* като синоним на термина *токън*.

[7] Изброени са само част от тях за илюстрация, например не са изброени комбинациите от главни и малки латински букви и апостроф.

лексикални единици, числа, изрази за време и т.н., а други – не. Например, последователност от една гръцка буква, тире и кирилски букви може да е приемлива българска дума, но последователност от кирилска буква, запетая и гръцки букви – не. Пунктуационните знаци и символи като *@, °, &, |, %, +, =, №* и др. също получават анотация, която ги определя еднозначно. С помощта на прости средства – регулярни правила, се разпознават някои изрази, които означават дати, време, мерни единици, дроби, номерация, телефонни номера, имейли, интернет адреси и др. Например следният регулярен израз маркира последователност от символи, които са имейл адреси:

[A-Za-z0-9._%+-]+@[a-z0-9.-]+\.[a-z]{2,4}

Правилна стратегия е на това равнище да се анотират собствени имена, съкращения и съставни лексикални единици, които могат да се разпознаят еднозначно винаги или при определени обстоятелства (*Светла* – прилагателно или собствено име в началото на изречение, но *Светла* – винаги собствено име в средата на изречение; *под.* – или съкращение, или съществително в края на изречение, но *под.* – винаги съкращение в средата на изречение). Това се отнася и до съставните лексикални единици, които могат да се идентифицират еднозначно винаги (*Атлантически океан*) или при определени условия (*Стара планина* в началото на изречение). Известна трудност могат да представляват различните формати за записване, които не винаги отговарят на установените правописни правила, например за дата: *08.08.2013, 08-08-2013, 8.8.13, 8/8/2013* и т.н.

При автоматично определяне на графичните думи също се използват относително прости правила, които са формулирани като регулярни изрази. Пунктуационните знаци (когато са употребени в същинската си роля) се разделят от останалите символи. Приписва се анотация, която описва поредиците от символи. Някои лексикални единици, които могат да бъдат еднозначно определени без допълнителна лингвистична информация (съкращения, собствени имена, числови изрази, интернет адреси, граматично еднозначни съставни лексикални единици) също получават анотация.

4. Определяне на части на речта[8], основни форми и граматични характеристики

Известно е, че една дума може да се използва с различно граматично значение в различните контексти – граматичното значение е отношение, изразено в структурата на езика. При морфологичните категории граматичните

[8] Терминът, който е прието да се използва за автоматично определяне на частта на речта, е тагиране, съответно за автоматично определяне на основна форма и граматични характеристики на словоформата - *лематизиране*.

значения се изразяват на морфологично равнище, а при лексикално-граматичните категории – на лексикално равнище. Българският език притежава следните морфологични и лексикално-граматични категории: число, род, лице, време, наклонение, преизказване, залог, определеност, степенуване, падеж и вид на глагола. Проявата на отделните морфологични и лексикално-граматични категории е свързана с дадени класове думи и в зависимост от това лексикалните единици са групирани в различни групи – наречени граматични класове, подкласове и типове, отразяващи парадигмите от синтетични словоформи (Коева 1998; Коева 2010). Граматичното значение може да се раздели на:

• категориално – характеризиращо основните форми (и словоформите, свързани с дадена основна форма) и показващо групирането на думите по класове – части на речта (например съществително, глагол);

• парадигматично – характеризиращо основните форми (и словоформите, свързани с дадена основна форма) и показващо групирането на думите по граматични подкласове (например мъжки, женски и среден род за съществителните имена);

• граматично – характеризиращо образуването на словоформите и показващо групирането на думите по граматични (флективни) типове (всички думи от един и същи тип имат еднакво формообразуване).

Следователно граматичният клас се състои от три компонента: класове думи, граматични подкласове и граматични типове, и специфицира множеството от езиково специфични морфо-синтактични свойства на основната форма.

Класовете (по части на речта) обхващат думите, които имат обща парадигма, разбирана като потенциална възможност за реализация (Коева 2010). Традиционните части на речта – съществително, прилагателно, глагол, числително, местоимение, наречие, предлог, съюз, частица и междуметие, са обединени в следните класове: съществително, прилагателно, глагол, числително, лично местоимение, притежателно местоимение, местоимение, неизменяема част. Например граматичният клас съществително име обединява всички словоформи, които принадлежат на класа, и задава възможностите за граматична реализация в рамките на отделните подкласове: единствено число, нечленувано; единствено число членувано; единствено число членувано с кратък член; звателна форма; множествено число нечленувано; множествено число членувано; бройна форма.

Граматичният подклас характеризира част от лексикално-семантичните и синтактичните свойства на лексикалните единици, които обединява, при които се реализират точно определени членове на парадигмата, зададена от граматичния клас. Граматичните подкласове при съществителните се определят от лексикално-граматичната категория род на съществителните (парадигмите на граматичните подкласове съществително мъжки род, съществително женски род и съществително среден род са подмножества на общата парадиг-

ма за класа) и от следните характеристики[9]: нарицателни и собствени, конкретни и абстрактни, броими и неброими, в рамките на неброимите – само с форма за единствено или за множествено число, имена за лица (имат звателна форма) и нелица, в рамките на нелицата – мъжколични съществителни и немъжколични съществителни (имат бройна форма).

За описанието на граматичните подкласове при глаголите се използват стойностите на категориите парадигматичност, инхерентна транзитивност и вид на глагола, тъй като пардигматичността детерминира дистрибуцията на категориите лице и число; преходността – образуването на пасивно причастие; а видът – образуването на деепричастие и сегашно деятелно причастие (както и на отглаголни съществителни, които обаче не се разглеждат като част от глаголната парадигма). Всеки глагол се определя като личен, безличен (неутрализация на категориите лице и число) и третоличен (неутрализация на категорията лице) (Коева 1998). Изборът на стойност рефлектира върху следващите възможности за избор при дефиниране на глаголния подклас – личните глаголи могат да бъдат както преходни, така и непреходни, но безличните и третоличните са само непреходни с изключение на акузатива тантум. Наблюдават се и други ограничения в дистрибуцията по лице и число, които се отнасят за ограничена група от глаголи: такива, които се употребяват само в множествено число – *да насядаме*; такива, които се употребяват само в трето лице множествено число – *елементите оформят структурата*; такива, които се употребяват само в трето лице единствено число, но не са безлични – *завива ми се (свят)*. Стойностите на категорията вид са свършен вид (тук спадат първичният свършен вид – *родя*; свършен вид, образуван с префиксация от първичен свършен вид – *изродя*; свършен вид, образуван с префиксация от първичен несвършен вид – *проходя*), несвършен вид (тук спадат първичният несвършен вид – *ходя*; несвършен вид, получен от първичен свършен вид чрез суфиксация – *раждам*; несвършен вид, получен от непрефигиран несвършен вид чрез префиксация – *израждам*) и вторичен несвършен вид (несвършен вид, получен от префигиран свършен вид чрез суфиксация – *прохождам*). Така наречените двувидови глаголи – *анализирам*, се класифицират като глаголи от несвършен вид. Известно е, че преходността (респективно непреходността) не е лексикално свойство на глаголите, а резултат от определена синтактична реализация на даден аргумент – именна група, която е в прякообектна позиция. Всеки глагол се дефинира като транзитивен и нетранзитивен (аналогично, както няма двувидови глаголи в контекста, по същия начин няма и двутранзитивни глаголи). В рамките на двете големи групи са отделени класове глаголи в зависимост от лексикалните им свойства – да образуват съставна лексикална единица с „рефлексивна“ частица и местоименна „винителна“ или „дателна“ клитика (Коева 2004).

[9] По-подробна класификация е представена в Буров (2004).

Подкласовете при прилагателните имена се образуват в зависимост от това дали са изменяеми, или не, а в рамките на изменяемите – дали са относителни, или качествени.

Основните подкласове при числителните имена са числителни редни и числителни бройни. Могат да се отделят и мъжколични числителни, числителни за изразяване на приблизително количество и числителни за изразяване на дроби.

Подкласовете при класа местоимение включват обобщително, неопределително, отрицателно, показателно и относително местоимение. Личното местоимение има два подкласа – лично и възвратно лично местоимение. Притежателното местоимение има два подкласа – притежателно и възвратно притежателно местоимение.

Подкласовете при неизменяемите думи включват наречие, изменяемо наречие (което се степенува), местоименно наречие, предлог, съчинителен съюз, подчинителен съюз, частица, междуметие.

Под граматичен тип (флективен тип) се разбира съвкупността от думи с еднаква парадигма, разбирана не само като еквивалентност на редовете на парадигмата, но и като еквивалентност при звуковите и акцентните редувания. Тоест в рамките на един подклас може да има няколко граматични типа в зависимост от различните начини за образуване на словоформите. Думите *дете* и *куче*, макар че са от един подклас – съществително нарицателно среден род, принадлежат към различни граматични типове, защото формите им за множествено число се образуват по различен начин.

Всички думи в езика, които имат еднакви граматични характеристики (граматичен клас и подклас) и еднакво множество от окончания и редувания, принадлежат към един граматичен тип. В Граматичния речник на българския език (Коева 1998) лексикалните единици са представени с основна форма, граматично значение и принадлежност към граматичен тип. Граматичните типове са описани в рамките на дефинираните подкласове думи. Даден граматичен тип дефинира за всеки ред от парадигмата окончанието, звуковите и акцентните редувания и стойностите на граматичните характеристики, които характеризират словоформата. Например качествените прилагателни *хубав, нов, стар* и др. принадлежат към един граматичен тип, при който няма акцентни и звукови редувания, а окончанията, определителният член и стойностите на граматичните характеристики на словоформите са, както следва:

–	мъжки род, единствено число, нечленувано
– ия	мъжки род, единствено число, членувано с кратък член
– ият	мъжки род, единствено число, членувано с пълен член
– а	женски род, единствено число, нечленувано
– ата	женски род, единствено число, членувано
– о	среден род, единствено число, нечленувано
– ото	женски род, единствено число, членувано
– и	множествено число, нечленувано
– ите	множествено число, членувано

Категориите, които характеризират словоформите при съществителното, са: число (единствено и множествено), бройна форма, звателна форма и определителен член (определено, определено с пълен член, определено с кратък член, неопределено), при прилагателното – род (мъжки, женски и среден), число (единствено и множествено), определителен член (определено, определено с пълен член, определено с кратък член, неопределено), при глагола – време (сегашно, минало свършено, минало несвършено), лице (първо, второ и трето), число (единствено и множествено), наклонение (изявително и повелително), форми (лични, сегашно деятелно причастие, минало свършено деятелно причастие, минало несвършено деятелно причастие, минало свършено страдателно причастие, деепричастие), за причастията – род (мъжки, женски и среден), число (единствено и множествено), определителен член (определено, определено с пълен член, определено с кратък член, неопределено).

Граматичните типове са представени като крайни преобразуватели, така че при компютърна обработка на всяка дума, разпозната от крайния преобразувател, може да се припише съответната основна форма, граматични характеристики на основната форма (клас и подклас) и граматични характеристики на словоформата. Граматичните типове на думите и съставните думи се разглеждат в рамките на унифициран подход. Парадигмата на съставната дума включва всички форми от парадигмата на главната дума, които се употребяват в езика. При комбинация с несъгласувани модификатори, които също се употребяват с повече от една форма, редовете от парадигмата на главната дума се умножават по редовете от парадигмите на тези модификатори.

При компютърен морфологичен анализ на текст част от думите получават повече от едно граматично значение, част – само едно, а друга част – остават неразпознати. Анотациите имат формат, съдържащ лема, граматични характеристики на лемата, граматични характеристики на словоформата (ако има словоформи). Например при автоматична анотация на 217 210 единици, от които 172 482 думи, 42 058 пунктуационни знака и 2670 цифри, 51,9% от единиците получават едно граматично значение, 46,7% – повече от едно, и 1,4% – нито едно (Коева и др. 2006). Неанотираните думи обикновено са редки думи, чужди думи или собствени имена, които не са включени в речника, но могат да бъдат и грешно изписани думи. Степента на многозначност (броят на различните граматични значения) варира при различните класове думи. Най-висока е граматичната многозначност при пунктуацията и затворените класове думи, които имат множество граматични значения за разлика от отворените класове, – например запетаята има 25 различни значения според функцията си в изречението.

Най-общо казано, анализирането на думите по части на речта и съответните стойности на граматичните категории включва въвеждането на многозначни граматични характеристики и отстраняването на граматичната многозначност. При автоматичното определяне на частта на речта е необходимо да се открият тези характиристика на думата и зависимости на непосредствения

контекст (с лингвистични правила или посредством статистически методи), които позволяват думата да се асоциира с правилната част на речта и частични граматични характеристики на словоформата.

Системата за автоматично определяне на частите на речта, която се използва в момента, е базирана на Метода на опорните вектори (Support Vector Machines) и предсказва частта на речта въз основа на множество от характеристики, които описват думата и нейния контекст. Използва се SVMTool[10] (Гименес и Маркес 2004), отворен код за трениране на езикови модели и прилагането им върху текст. SVMTool се нуждае от корпус за трениране – в случая това е ръчно анотираният за част на речта и граматични характеристики корпус на българския език (Коева и др. 2006). Изборът на стратегия за трениране на модела за автоматично граматично анотиране е от особена важност: посока – от ляво на дясно, от дясно на ляво или и двете; дължината на контекста – приема се, че колкото е по-голям контекстът, толкова са по-добри резултатите; дефиницията на параметрите – n-грами на думи или граматични анотации или класове на многозначност, лексикализирани параметри като префикси, суфикси, краища на изречения, анотация към графична дума, главни букви и др. Стратегията за трениране (Коева и Генов 2011) има следните параметри: две обхождания вляво и вдясно; контекстов прозорец от пет графични думи, като анотираната дума е на втора позиция; дву- и триграми от думи или части на речта, лексикални параметри като префикси, суфикси, граница на изречение, главни букви и др. За граматичното анотиране се използва редуцирано множество от анотации (75 класа в съпоставка с 1029 уникални граматични анотации в речника), което осигурява минималната необходима информация за еднозначно съотнасяне със съответната лема. Пример за резултата след действието на системата е представен по-долу, като описанието на тага и коментарът са дадени за улеснение:

Словоформа	Анотация	Описание на анотацията	Коментар
Писмо	Ns	съществително, единствено число	
до	R	предлог	
Иван	NHs	съществително собствено, единствено число	
Иоцов	NHs	съществително собствено, единствено число	
от	R	предлог	
Враца	Ns	съществително, единствено число	непълно
.	U	пунктуация	

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

То	Ps	местоимение, единствено число
е	Vs	глагол, единствено число
последното	As	прилагателно, единствено число
Ботево	As	прилагателно, единствено число
писмо	Ns	съществително, единствено число
.	U	пунктуация

След автоматичното приписване на част на речта и частични граматични характеристики на словоформата се определя основната форма на думата и се преписва подробна граматична информация от Граматичния речник. Търси се съвпадение между словоформата в текста и словоформа в Граматичния речник и се използва дефинираното съответствие между автоматично приписаните частични граматични характеристики и съответните им пълни граматични характеристики в речника. За разрешаване на многозначността се прилагат малък брой правила, част от които са представени по-долу:

Ако думата няма съответствие в речника, използва се само редуцираната анотация.

Ако думата започва с главна буква и не е в началото на изречението, тя се асоциира с лема за съществително собствено име от речника (ако такава е налична), независимо че автоматичната анотация може да е избрала друга анотация като по-вероятна.

Ако думата е разпозната като сегашно деятелно причастие, членувана форма на минало свършено деятелно причастие или членувана форма на минало страдателно причастие, приписва ѝ се част на речта прилагателно, за основна форма – нечленуваната форма за мъжки род единствено число на причастието и съответните граматични характеристики на формата на причастието. Ако думата е разпозната като минало несвършено деятелно причастие, приписва ѝ се част на речта глагол, съответната основна форма на глагола и характеристиките на формата на причастието.

Пример за резултата след действието на системата за приписване на лема и разширени граматични характеристики е представен по-долу, като описанието на анотацията е дадено за улеснение:

Словоформа	Лема	Граматична анотация	Описание на анотацията
Писмо	писмо	NCNson	съществително нарицателно, среден род, единствено число, нечленувано
до	до	R	предлог
Иван	Иван	NHMsom	съществително собствено, мъжки род, единствено число, нечленувано
Иоцов	Иоцов	NHs	съществително собствено, нечленувано
от	от	R	предлог

Враца	Враца	NHso	съществително собствено, единствено число, нечленувано
.	.	U	пунктуация
То	аз	PHi3sn	лично местоимение, трето лице, единствено число, среден род
е	съм	VLINr3s	глагол, личен, несвършен вид, непреходен, сегашно време, трето лице, единствено число
последното	последен	Asnd	прилагателно, среден род, единствено число, членувано
Ботево	Ботев	Asno	прилагателно, среден род, единствено число, нечленувано
писмо	писмо	NCNson	съществително нарицателно, среден род, единствено число, нечленувано
.	.	U	пунктуация

5. Определяне на именни фрази

Българският език се характеризира с няколко особености, които представляват трудност при автоматичното определяне на синтактичната му структура: относително свободен словоред – докато мястото на прилагателните като предпоставени модификатори на съществителни, както и на наречията като предпоставени модификатори на съществителни, прилагателни или други наречия може да се приеме за фиксиран, то редът на субекта и комплементите към глагола е относително свободен – така в изречение с три члена (субект, глагол, обект) всички шест словоредни комбинации са възможни; в изречение с четири члена (субект, глагол, пряк обект, непряк обект) са възможни двадесет и четири комбинации; с пет члена – сто комбинации и т.н. На равнище конституенти (субект, предикат, комплемент, адюнкт) словоредните разместванията са възможни, като, разбира се, всяка промяна от т.нар. основен словоред (субект глагол обект) носи допълнително значение за фокуса, който се поставя при предаването на съобщението. Например в изречението: *Мария търси Иван цяла сутрин* – не е ясно кой конституент е подлогът извън по-широк контекст: *Мария* или *Иван*. За разлика от други езици с относително свободен словоред (например останалите славянски езици) в български не може да се съди за синтактичните отношения между думите по падежни окончания. Отличителна черта на българския език е загубата на морфологичното изразяване на падежните отношения при съществителните, а оттам и при прилагателните, които ги модифицират. Още една особеност на българския, която затруднява автоматичното определяне на синтактичната му структура, е свободното изпускане на подлога, което в комбинация с възможността за разместване на позицията на субекта и обекта прави задачата още по-трудна. Сложността на анализа се определя и от трудностите при дефиниране

на границите на фразите основно при предложните съчетания – дали дадено предложно словосъчетание е модификатор на непосредствено предхождащия го конституент, или на някой друг.

Задачата за автоматичен анализ на именни групи може да бъде формулирана по следния начин: след като се определят с достатъчна степен на достоверност съществителните имена, тяхната основна форма и граматични характеристики, трябва да се идентифицират границите на именните словосъчетания в текста (дали дадено съществително име е опора на именно словосъчетание), конституентите, които ги съставят, и видът на синтактичните отношения между тях. За определяне на структурата на всяка именна фраза е необходим синтактичен анализ, който в много случаи не може да бъде коректен без подходящ семантичен анализ. Тъй като става въпрос за произволен текст, автоматичното решение на задачата е недостатъчно точно, независимо от избора на средства: лингвистични правила, статистически методи, машинно обучение или комбинация на подходите. По тази причина предлаганият подход за анализ на именните фрази се основава на правила, при формулирането на които са спазени следните принципи:

• Структурата на именните фрази се дефинира независимо от конкретна лингвистична теория.

• Анализират се фразово-структурни, а не дependентни отношения.

• Словосъчетания със субстантивирана употреба на други части на речта не се разглеждат.

• Словосъчетания, в които опорното съществително име е елиптично, не се разглеждат.

• Координативните именни словосъчетания не се разглеждат.

• Анализират се структури от съществително и предпоставени и/или задпоставени модификатори – думи или фрази.

• Идентифицират се границите на именното словосъчетание и се анотират именната фраза и именната опора.

• Категориалната и словоредната съчетаемост се дефинират в правила, ако са засвидетелствани (относително голям брой) примери, които ги илюстрират. Например в Българския национален корпус не се наблюдава нито един случай на удвояване на предложна фраза с притежателна клитика в рамките на словосъчетание с главна част съществително: *Чантата ѝ на Мария*.

В Българския национален корпус (над един милиард и двеста милиона думи) е идентифициран следният брой последователности от категории (с уговорката, че в някои случаи те не образуват именни словосъчетания): съществително – 26 859 749 пъти; прилагателно, съществително – 12 714 675 пъти; съществително, съществително – 9 745 197 пъти (грешката е по-голяма, защото в много случаи може да принадлежат на две различни словосъчетания); съществително, предлог, съществително – 9 148 758 пъти; съществително, предлог, прилагателно, съществително – 3 335 800 пъти; прилагателно, съществително, предлог, съществително – 2 869 101 пъти; прилагателно, прилагателно, съ-

ществително – 1 909 770 пъти; съществително, прилагателно, съществително – 1 166 428 пъти (грешката е по-голяма, защото в много случаи първото съществително може да принадлежи към друго словосъчетание); прилагателно, притежателно местоимение (клитика), съществително – 293 602 пъти; притежателно местоимение, прилагателно, съществително – 158 720 пъти; прилагателно, предлог, съществително, съществително – 129 659 пъти и т.н.

Формализмът за дефиниране на лингвистични правила, който е използван, накратко може да бъде описан по следния начин (Карагьозов и др. 2012). Всяко правило се състои от елементи и може да има десен и/или ляв контекст (елементите и контекстът имат еднакъв синтаксис и семантика). Елементите могат да бъдат дума, лема, част на речта и граматични характеристики, лексикон, не-дума[11], не-лема, не-част на речта и не-граматични характеристики, регулярен израз, описващ частта на речта и граматичните характеристики. Между елементите може да се приложи някой от следните оператори: или (или единият, или другият елемент), въпросителна (нула или едно срещане на предходния елемент), плюс (едно или повече срещания на предходния елемент) и звезда (нула или повече срещания на предходния елемент) на Клини, група (показва групата, върху която работи оператор). Формализмът за дефиниране на правила поддържа унификация на елементите по част на речта и граматични характеристики в рамките на дадено правило, но към момента унификацията не се използва, като се предполага, че именните групи в даден текст са употребени без граматични грешки при съгласуването. Под лексикон се разбира списък от думи, принадлежащи на един и същ клас: например географски понятия, собствени имена и т.н. – при правилата за маркиране на именни групи лексикони не се използват. Формализмът поддържа каскадно прилагане на правилата чрез групиране на правила в групи с предварително зададен приоритет.

Категориите, които могат да разширяват съществително име, са дефинирани (Пенчев 1993; Бъркалова 1997) и до известна степен са подробно описани (Петрова 2009), както следва:

• в препозиция – прилагателни, словосъчетания с главна част прилагателно с определена структура; наречия, словосъчетания с главна част наречие с определена структура и класове наречия; класове съществителни и словосъчетания с тях, всички видове местоимения (без личните и рефлексивните лични местоимения) и числителни имена;

• в постпозиция – предложни групи без ограничение за структурата; наречия, словосъчетания с главна част наречие с определена структура и класове наречия; подчинени изречения.

---

[11] Това означава, че даден елемент може да бъде лема, част на речта и граматични характеристики, лексикон, но не и дума. По аналогичен начин се интерпретират не-лема, не-част на речта и не-граматични характеристики.

Тъй като целта ни не е да правим пълен синтактичен анализ, а да извлечем именните групи, които са най-важни за интерпретацията на съдържанието на даден текст, някои от тези разширения не представляват интерес, а именно: наречия, словосъчетания с главна част наречие с определена структура и класове наречия, подчинени изречения. Подходът е да се маркира последователност от категории, които съответстват на именно словосъчетание, без да се определя структурата на словосъчетанието. Както беше посочено, правилата не се грижат за проверката на съгласуването, както и за други прояви на граматична употреба, например членуването или словореда – правилата откриват редици от категории, които с голяма вероятност са именни групи. Например правилото[12]:

(((обобщително или отрицателно или неопределително или въпросително или показателно местоимение)? числително редно (притежателно или рефлексивно местоимение)?)? или ((притежателно или рефлексивно местоимение) числително редно)?)? или ((числително редно или прилагателно) (притежателна или рефлексивна клитика)?)? или (числително бройно? (притежателно или рефлексивно местоимение))? или (квантифициращо наречие? (притежателно или рефлексивно местоимение))?)?)? (прилагателно (предлог прилагателно съществително)?) (прилагателно* съществително)? съществително (предлог (((обобщително или отрицателно или неопределително или въпросително или показателно местоимение)? числително редно (притежателно или рефлексивно местоимение)?)? или ((притежателно или рефлексивно местоимение) числително редно)?)? или ((числително редно или прилагателно) (притежателна или рефлексивна клитика)?)? или (числително бройно? (притежателно или рефлексивно местоимение))? или (квантифициращо наречие? (притежателно или рефлексивно местоимение))?)?)? (прилагателно (предлог прилагателно* съществително)?) (прилагателно* съществително)? съществително)?*

маркира словосъчетания като: *всяка втора нова книга, никоя втора нова книга, някоя втора нова книга, коя втора нова книга, тази втора нова книга, тази моя нова книга, тази своя нова книга, втората ми нова книга, много мои нови книги, интересната ми нова книга, една своя нова книга, интересна книга, интересна нова книга, интересната за ранобудни студенти книга, нова група интересни книги, Иван Петров, втората ми нова книга за деца* и т.н.

Разбира се, правилото може да бъде още по-комплексно, но всъщност целите на изследването изискват да бъде опростено. Местоименията, най-общо

---

[12] Посоченото правило е илюстративно и показва комбинаториката, задължителността и повторяемостта. Правилото, което се използва на практика, работи с регулярни изрази върху символите за части на речта и граматични характеристики и в него допустимите комбинации са изброени експлицитно.

казано, допринасят за текстовата свързаност, а числителните изразяват количество и поредност. За определянето на именните групи, които са важни за съдържанието на даден текст, всъщност трябва да се идентифицира опорното съществително и стесняващите понятието модификатори – други съществителни и прилагателни. Така правилото може да бъде сведено до:

прилагателно (предлог прилагателно съществително)? (прилагателно* съществително)? съществително (предлог прилагателно* съществително)?*

Основният проблем, който остава, е свързан с предложната група като задпоставен модификатор – с подобен анализ не може да се установи дали предложната група се отнася към съществителното, или към друга дума в изречението, например глагола. Относително сигурни критерии са следните:

• Много е вероятно предложната група да се отнася към съществителното, ако и двете са пред глагола на простото изречение, в което се намират.

• Много е вероятно предложната група да се отнася към съществителното, ако самото съществително е част от предложна група.

6. Идентифициране на именувани същности

Терминът **именувани същности** (именувани обекти, назовани обекти, Named Entity) се използва за означаване на собствени имена и тяхната класификация като имена на лица, организации и места (като именувани същности обикновено се означават и някои изрази за време, числа, парични стойности и др.), например *Стара планина, ул. „Ален мак“, Иван Петров, ас. Тодоров, фирма „Континентал“, „Тетраком“ ООД*. Известни са и по-подробни класификации, в рамките на повече от двеста категории (Секине и Нобата 2004) – например имената на места могат да се класифицират като имена на континенти, държави, области, градове, села, улици, площади, реки, планини и т.н.

Съществуват различни начини за идентифициране на именуваните същности (Надю и Секине 2007) – основно посредством правила и статистически методи. Възприетият в нашата разработка подход се основава на правила, комбинирани с лексикони, които съдържат списъци от именувани същности, класифицирани в зависимост от това дали са лични имена – собствени (около 13 000 уникални лични имена, които не се срещат в речника, използван за автоматично определяне на основната форма и граматичните характеристики), или фамилни (около 7000 уникални фамилни имена, които не се срещат в Граматичния речник), географски имена (около 13 000 уникални имена, които не се срещат в Граматичния речник), имена на организации (около 3000 уникални названия, които не се срещат в Граматичния речник). Правилата задават модел, характерен за отделните видове именувани същности: например географските имена (в зависимост от вида си) могат да бъдат предхождани от думите *град, гр., село, с.* и т.н.; имената на лица могат да бъдат предхождани от приложение за тяхната професия, например думите *инженер, инж., профе-*

сор, проф. и т.н. Обичаен подход при разпознаването на именувани същности е да се разчита на автоматично определяне на частите на речта, което идентифицира и собствените имена. Приложенията към именуваните същности (*професор, инженер* и т.н.) също са изброени като списъци и класифицирани в отделни лексикони. Правилото за идентифициране на именувана същност – географско название, най-общо има следната структура:

приложение от лексикон? (географско име от лексикон или съществително собствено)[13].

Правилото за идентифициране на именувана същност – собствено име, най-общо има следната структура:

*приложение от лексикон? (собствено име от лексикон или съществително собствено) (фамилно име от лексикон или съществително собствено)**

Правилото за идентифициране на именувана същност – организация, най-общо има следната структура:

(приложение от лексикон? (собствено име от лексикон или съществително собствено)+) или ((собствено име от лексикон или съществително собствено)+ приложение от лексикон?).

7. Откриване на анафорите

Анафора се нарича лингвистична единица, чиято интерпретация зависи от значението на друга лингвистична единица (антецедент). Значението на част от българските местоимения (лични в трето лице, притежателни в трето лице, рефлексивни, относителни, въпросителни, показателни) и местоименните наречия се интерпретира спрямо значението на дадена лингвистична единица (съществително, именна фраза, предложна фраза, просто изречение и др.) от непосредствения контекст: обикновено предходен ([Иван]$_i$ подари на [Мария]$_j$ новата [си]$_i$ книга [тази сутрин]$_k$, [когато]$_k$ разбра, че [тя]$_j$ има рожден ден), но може да е следходен – тогава свързаните лингвистични единици се наричат катафора и постцедент (В [своето]$_i$ легло [Иван]$_i$ не яде шоколад). Анафорите и класовете антецеденти, които могат да се асоциират с тях, са различни. Тези, които представляват интерес за разработката, са:

• лични местоимения в трето лице и антецедент съществително или именна фраза;

• рефлексивни местоимения и антецедент съществително или именна фраза;

---

[13] Посочените правила са илюстративни и показват комбинаториката, задължителността и повторяемостта. Правилата, които се използват на практика, работят с регулярни изрази върху символите за части на речта и граматични характеристики и в тях допустимите комбинации са изброени експлицитно.

• относителното местоимение *който* и антецедент съществително или именна фраза;

• относителните местоименни наречия *когато* и *където* и антецедент съществително или именна фраза.

Мотивите за това са следните:

• антецедентите се идентифицират относително лесно и с висока коректност на резултата;

• анафорите са конституенти в изречението, които, най-общо казано, се съотнасят със субекта и комплементите.

Подходите за автоматично определяне на анафоричните връзки (Митков 1997) се базират на (комбинация от) различни фактори, които се използват, за да се отсеят възможните кандидати, например съгласуване по род, число и лице (за местоименията), одушевеност, включително човек, позиционна близост, структурни отношения като конституентно командване, селективни ограничения, семантична съвместимост и др.

Формулирани са различни ограничения за редуциране на кандидатите за антецедент, свързвани с конституентното командване (Чомски 1981):

Именна фраза не може да има същата интерпретация с друга именна фраза (или местоимение), с която е в отношение на конституентно командване (Мария$_i$ му$_j$ разказа всичко за Иван$_k$).

Антецедентът трябва да командва конституентно рефлексивното местоимение (Мария$_i$ обикновено не разпространява слухове за себе си$_i$).

Местоимение не може да има същата интерпретация с именна фраза, с която е в отношение на конституентно командване (Иван$_i$ му$_j$ разказа всичко за него$_k$).

Конституентно командване между възлите на синтактично дърво се дефинира по следния начин:

А конституентно командва Б, ако едновременно са изпълнени:

(а) А не доминира над Б;

(б) Б не доминира над А;

(в) Първият разклонен възел, който доминира над А, доминира и над Б.

Теорията на свързването (Чомски 1981) определя структурните зависимости при свързана интерпретация на именни групи и местоимения по следния начин:

А свързва Б, ако:

(а) А конституентно командва Б;

(б) А и Б са коиндексирани (значението им се интерпретира по един и същи начин).

Или, с други думи, анафоричното свързване на значения се определя от отношението на конституентно командване между коиндексирани елементи. Дефинират се следните принципи:

Принцип А:

Анафората (рефлексивно местоимение) трябва да бъде свързана в своята минимална област за свързване.

Принцип Б:

Местоименията (личните местоимения) трябва да бъдат свободни в своята минимална област за свързване.

Принцип В:

Референциалните изрази (имената) трябва да бъдат свободни навсякъде.

Минималната област се дефинира като областта в изречението, която съдържа анафората, антецедента и субект. Минималната област е или именна фраза, или просто изречение.

За примери като Майката$_i$ на Мария погледна към себе си$_i$ в огледалото, в които изискването за конституентно командване не е спазено, а има анафорична интерпретация на рефлексивното местоимение, се налага дефиниране на така нареченото М-командване (на максимална проекция) (Чомски 1986):

М-командване (на максимална проекция):

(а) А не доминира над Б;

(б) Б не доминира над А;

(в) Първата максимална проекция, която доминира над А, доминира и над Б.

За да се определят отношения на конституентно командване, е необходимо: (а) пълен синтактичен анализ; (б) придържане към определен синтактичен формализъм, в рамките на който пълният синтактичен анализ да е коректен (някои теории въвеждат например функционални възли, които се асоциират с нелексикални и несинтактични категории). Тъй като пълният синтактичен анализ изисква сериозни ресурси, се предпочита подход, при който областта се дефинира достатъчно точно, за да се открие еднозначна интерпретация, без да е задължително минималната област на свързване.

С уговорката, че катафоричната свързаност не се разглежда, приемаме за минимална област, в която се търси антецедент, простото изречение (в рамките на сложното) и преформулираме принципите в следните правила:

(1) Антецедентът на рефлексивно местоимение е максимална именна опора в същото просто изречение, с която анафората се съгласува по род и число (Мария$_i$ не разпространява обикновено слухове за себе си$_i$).

(2) Лично (и относително) местоимение не може да има антецедент – максимална именна опора, в същото просто изречение (Иван$_i$ му$_j$ разказа всичко за него$_k$).

(3) Лично местоимение може да има същата интерпретация с предходна немаксимална именна опора, с която е в същото просто изречение и се съгласува по род и число (Приятелите$_i$ на Иван$_j$ ще го$_{j/k}$ изненадат).

Максимална опора наричаме или съществително без модификатори, или опорното съществително на завършена именна фраза, която може да има модификатори други именни фрази.

Един от алгоритмите, който не се базира на комплексен синтактичен и семантичен анализ, използва множество от експлицитни характеристики за

идентифициране на антецедента (Митков 1998). Изследват се именни групи в най-много две предходни изречения. Кандидатите се редуцират в зависимост от това дали имат същите съгласувателни характеристики по род и число като анафората. Потенциалните кандидати се оценяват на базата на следните характеристики: първа именна група в изречението, повторяемост на именната група в изследваната област, определеност на именната група, именната група да не е част от предложна група, разстояние между антецедента и анафората, най-близка именна група, именната група е собствено име, антецедентът и анафората имат същите синтактични функции.

Алгоритъм (Лозанова и др. 2013), който цели свързване на анафората в български текстове, формулира следните зависимости:

Антецедентът на лично или притежателно местоимение е най-близкото съществително (опора на именна фраза) в даден прозорец отляво на анафората, което удовлетворява съгласувателните свойства на анафората.

Антецедентът на относително местоимение е най-близкото съществително (опора на именна фраза) в предходното просто изречение, което удовлетворява съгласувателните свойства на анафората.

В рамките на всяко просто изречение се определят кандидатите за анафорична свързаност и се елиминират или установяват сигурни или потенциални връзки по горните правила.

За нуждите на изследването предлагаме следните правила:

За местоимения с неразрешена интерпретация се търси подходящ антецедент в предходното просто изречение, независимо дали е в същото сложно изречение, или в предходно (областта на разрешаване е само едно предходно просто изречение).

Местоимение (лично в трето лице) има същата интерпретация с единствена именна опора в предходното просто изречение, с която се съгласува по род и число (Иван$_i$ каза на Мария$_j$, че тя$_j$ трябва да замине, Родителите$_i$ или дават отново [храна на детето$_j$]$_k$, или го$_j$ лишават от нея$_k$.).

Местоимение (относително) има същата интерпретация с последната именна опора от предходното просто изречение, с която се съгласува по род и число.

Имплицитните местоимения не се експлицират и техният антецедент не се търси.

Местоимение (лично в трето лице) може да има същата интерпретация с повече от една именна опора от предходното просто изречение, с която се съгласува по род и число. Тогава се предпочита антецедент, който е собствено или членувано име.

Ако анафората не се разреши – за лично местоимение в именителен падеж се избира антецедент преди глагола на предходното изречение, за лично местоимение във винителен падеж – след глагола на предходното изречение, за лично местоимение в дателен падеж – след глагола на предходното изречение и след предлог.

8. Идентифициране на лексикални вериги

Лексикална верига наричаме последователност от отделни думи (обикновено съществителни) или фрази (обикновено именни словосъчетания), които са семантично свързани. Тъй като тази дефиниция е твърде обща, можем да я разглеждаме в по-тесен и по-широк смисъл. В тесен смисъл под лексикална верига разбираме последователност от близки синоними (форми на думи и словосъчетания с тях) или парафрази, употребени на различно (по-близко или по-далечно) разстояние в текста. Например: *автор, авторът, писател, белетрист; романът на Вазов, Вазовия роман.* В широк смисъл лексикалната верига обединява думи (или словосъчетания с тях), които са свързани със семантична или деривационна релация. Например: *автор, авторът, писател, белетрист, Йордан Йовков, пиша*. Халидей и Хасан (1976) дефинират пет основни типа на свързаност: дума, употребена с идентична референция; дума, употребена с различна референция (*тази ябълка – ябълки*); хипероним – хипоним (*ябълка – плод*); систематично класифицирани семантични отношения – антоними, мероними, наредени множества, ненаредени множества (*хубав – грозен, пръст – ръка, едно – две, червен – зелен*); несистематично класифицирани семантични отношения – колокации (*градина – копая*). Лексикалните вериги са надежден индикатор за определяне на части от текста, които са тематично свързани, както и за определяне на темата на текста и ключовите думи, които го характеризират. Известни са алгоритми за идентифициране на лексикални вериги, които използват уърднет[14] (Хърст и Сентонж 1997). Свързаността между съществителни се дефинира посредством разстоянието между техните срещания и вида на пътя, който ги свързва в уърднет: много силна е свързаността между дума и нейното повторение; силна – между две думи, свързани с релация в уърднет; средно силна – когато пътят между двете думи е повече от една релация (избират се само пътища, които удовлетворяват дадени изисквания). Максималното допустимо разстояние между думите в дадена лексикална верига зависи от силата на връзката между тях: за много силна връзка няма ограничения, за силна – ограничението е прозорец от седем изречения, за средно силна връзка – ограничението е прозорец от три изречения. За да се включи кандидат в дадена лексикална верига, се предпочитат много силните връзки, после силните и така нататък, така че алгоритъмът осигурява и отстраняване на многозначност: думата се включва с подходящото си значение, а значенията на останалите думи във веригата се променят, ако е необходимо, така че всяка дума във веригата, свързана с последната включена дума, да е с най-подходящото значение. Включват се всички възможни връзки, съответно разклонения, и

[14] Лексикално-семантична мрежа, в която думите са организирани в синонимни множества, свързани с различни семантични отношения.

се дефинира най-добрата интерпретация. Най-често използваните релации от уърднет за идентифициране на лексикални вериги са: синонимия, хиперонимия, хипонимия, меронимия, холонимия, антонимия, холоними – сестри. Тъй като уърднет е граф, задачата за идентифициране на лексикалните вериги се свежда до: (а) дефиниране на пътища в графа; (б) отстраняване на многозначност. Уърднет не кодира всички семантични връзки между думите, по тази причина се използват дефинициите в уърднет, за да се идентифицират допълнително релации като агент, тема, притежание, локация, начин, инструмент, например *метеоролог* (’предсказвам времето’) и *време* (Молдован и Бланко 2012). По този начин се включват семантични релации между синонимно множество и други синонимни множества, за които се реферира в дефиницията. На лексикалните веригите може да се приписва тегло спрямо различни характеристики, например дължина на веригата, честота на употреба на думите във веригата, йерархия между семантичните релации, свойства на семантичните релации (например транзитивност), композиция на семантичните релации и др.

9. Заключение

Всеки от представените подходи за разграничаване и характеризиране на езикови единици може да бъде разгледан по-подробно от гледна точка на разработването му към момента. От друга страна, всеки от тях подлежи на бъдещо усъвършенстване, което от своя страна ще доведе до въвеждането на разнообразна и богата лингвистична анотация, която да съдържа информация не само за граматичното и лексикалното значение на думите и фразите и за синтактичните и семантичните връзки между тях, но и да експлицира мрежа от разнообразни семантични връзки, които в текстовете нямат явна реализация. Така всеки документ може да се асоциира с абстрактно онтологично представяне на лингвистичната анотация, която е асоциирана с лексикалните и синтактичните единици в него, което ще дава възможност за извличането на разнообразни връзки и изводи.

Литература

Буров 2004: Буров, Ст. *Познанието в езика на българите. Граматично изследване на концептуалната категоризация на предметността.* Велико Търново: „Фабер“.

Бъркалова 1997: Бъркалова, П. *Българският синтаксис познат и непознат*. Пловдив: Пловдивско университетско издателство.

Гименес и Маркес 2004: Giménez, Jesús and Lluís Màrquez. SVMTool: A general POS tagger generator based on Support Vector Machines. – In: *Proceedings of 4th LREC*, pp. 43–46.

Графенстет и Тапанайнен 1994: Gregory Grefenstette and Pasi Tapanainen. What is a word, what is a sentence? Problems of tokenization. – In: *Proceedings of the 3rd Conference on Computational Lexicography and Text Research*. Budapest, Hungary, pp. 79–87.

Карагьозов и др. 2012: Karagiozov, D., A. Belogay, D. Cristea, S. Koeva, M. Ogrodniczuk, P. Raxis, E. Stoyanov and C. Vertan. I–Librarian – Free Onine Library For European Citizens. – In: *INFOtheca – Journal of Librarianship and Informatics*, 2012, May, vol. 13, № 1. Belgrade: BS Print, pp. 27–43.

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

Коева 2001: Коева Св. *Кратка практическа граматика на българския език*. София: Издателска къща „Труд".

Коева 2004: Коева Св. Семантично и синтактично описание на българските диатези. – В: Българско езикознание. Т. 4. София: Издателство на БАН, с. 182–231.

Коева 2006: Koeva Sv. Inflection Morphology of Bulgarian Multiword Expressions. – In: *Computer Applications in Slavic Studies*, Sofia: Boyan Penev Publishing Center, pp. 201–216.

Коева 2010: Коева. Св. *Българският ФреймНет 2010*. София: Институт за български език.

Коева и Генов 2011: Koeva, S. and A. Genov. Bulgarian Language Processing Chain. – In: *Proceeding of the Workshop Integration of Multilingual Resources and Tools in Web Applications*. Hamburg, pp. 29–32.

Коева и др. 2006: Koeva, Sv., Sv. Leseva, M. Todorova. Bulgarian Sense Tagged Corpus. – In: *Proceedings of the 5th SALTMIL Confernece on Minority Languages: Strategies for Developing Machine Translation for Minority Languages*. Genoa, pp.79–87.

Лозанова и др. 2013: Lozanova, S., I. Stoyanova, S. Leseva, S. Koeva and 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.* Sofia, Bulgaria, pp. 39–48.

Манинг, Рагаван и Шютце 2008: Manning, C. D., P. Raghavan and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press.

Микеев 2002: Mikheev, Andrei. Periods, Capitalized Words, etc. – *Computational Linguistics*, 2002, 28(3), pp. 289–318.

Митков 1997: Mitkov, R. Factors in Anaphora Resolution: They Are not the Only Things that Matter. A Case Study Based on Two Different Approaches. – In: *Proceedings of the ACL '07/EACL '97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution*. Madrid, Spain, pp. 14–21.

Митков 1998: Mitkov, R. Robust Pronoun Resolution with Limited Knowledge. – In: *17th International Conference on Computational Linguistics (COLING'98/ ACL'98)*. Montreal, Canada, pp. 969–875.

Молдован и Бланко 2012: Moldovan, D. and E. Blanco. Polaris: Lymba's Semantic parser. In: Chair, N. C. C.; Choukri, K.; Declerck, T.; Dogan, M. U.; Maegaard, B.; Mariani, J.; Odijk, J.; and Piperidis, S., (eds.), *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), pp. 66–72.

Надю и Секине 2007: Nadeau, David and Satoshi Sekine. A survey of named entity recognition and classification. – *Linguisticae Investigationes,* 2007, 30(1), pp. 3–26.
Пенчев 1993: Пенчев, Й. *Българския синтаксис – управление и свързване*. Пловдив: Пловдивско университетско издателство.
Петрова 2009: Петрова, И. *Синтактичен анализ на простото съобщително изречение в българския език* (дисертация).
Секине и Нобата 2004: Sekine, Satoshi, C. Nobata. Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy. – In: *Proc. Conference on Language Resources and Evaluation.*
Халидей и Хасан 1976: Halliday, M. and R. Hasan. *Cohesion in English*. London: Longman Group.
Хърст и Сентонж 1997: Hirst, G. and D. St-Onge. Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms. – In: C. Fellbaum, editor. *WordNet: An electronic lexical database and some of its applications*. Cambrige, MA: The MIT Press, pp. 305–332.
Чомски 1981: Chomsky, N. *Lectures on Government and Binding*: The Pisa Lectures. Dordrecht: Foris.
Чомски 1986: Chomsky, N. *Barriers.* Cambridge, MA: MIT Press.

Проф. д-р Светла Коева
Секция по компютърна лингвистика
Институт за български език
„Проф. Л. Андрейчин“ при БАН
бул. „Шипченски проход“ 52, бл. 17,
1113 София, България
svetla@dcl.bas.bg

Prof. Svetla Koeva, PhD
Department of Computational Linguistics
Institute for Bulgarian Language,
Bulgarian Academy of Scienses
52 Schipchenski prohod, Bl. 17,
1113 Sofia, Bulgaria
svetla@dcl.bas.bg