

Институт за български език „Проф. Л. Андрейчин“  
Българска академия на науките  
София, България

## КОМПЮТЪРНАТА ЛИНГВИСТИКА В СПИСАНИЕ „БЪЛГАРСКИ ЕЗИК“

За втори път от 2009 г. насам сп. „Български език“ подготвя специален брой, посветен на компютърната лингвистика. В дигитална ера, в условията на коренно променящи се начини на комуникация и организиране на информацията се извършва стремителен преход от традиционните методи на събиране и обработка на езикови данни към компютърно базирани методи, които предоставят повече възможности за точно и пълно описание на различни явления от езиковата действителност. Компютърната лингвистика е сравнително ново, но много бързо развиващо се научно направление, свързано с прилагането на математически модели за описание на естествения език, със създаването на електронни езикови ресурси и на различни програми за тяхната обработка, които не само оптимизират работата на лингвистите, но и намират широко приложение в редица области.

Броят се открива с преведената на български език статия на Гюнеш Еркан и Драгомир Радев „LexRank: лексикална централност на графи като мярка за значимост при резюмирането на текст“. Тя представя стохастичен метод, базиран на графи, за изчисляване на относителната важност на отделните текстови единици за целите на автоматичната обработка на естествения език. Представените техники са тествани в областта на резюмирането на текст.

Иван Держански разглежда думите за кратки периоди от време в български и украински език, като използва материал от паралелни текстове, включващи оригинална българска и украинска художествена проза и преводи от други езици. По същество това е корпусно изследване на употребата на съществителни имена, назоваващи кратки периоди и единици време (*минута, миг, момент, секунда*), в български и в украински език.

Статията на Светлозара Лесева и Ивелина Стоянова представя първите стъпки към разработването на система за автоматично разрешаване на анафората за български език. Системата се основава на правила, отразяващи редица особености на българския език като свободния словоред, елип-

---

сата при подлога, безличните изречения и съгласуването. Формулират се два вида правила: филтриращи и ранкиращи. За различните анафори се използват различни комбинации от така формулираните правила, свързани с обсега за откриване на референции, както и със синтактичните и морфосинтактичните особености. Методът е приложен върху семантично аотирания корпус на български език БулСемКор и резултатите са съпоставени с базовия метод, който използва само съгласуване и приписва референция на най-близкия възможен кандидат.

Валентина Стефанова и Цветана Димитрова разглеждат въпроса за класификацията на прилагателните имена в българската езиковедска литература в контекста на подхода при класифицирането им в лексикално-семантичната мрежа Уърднет. Представени са първоначални наблюдения върху проблематиката, осъществени в рамките на проект за семантична класификация на прилагателните в Българския уърднет, за да се мотивира нуждата от семантичен подход.

Росица Декова, Светлозара Лесева, Мария Тодорова и Екатерина Търпоманова представят някои структурни особености на фрази с опора прилагателно име в българския език. Резултатите са получени в рамките на задачата за разработване на демонстрационен софтуер за тагери и лематизатори по проекта „Интегриране на нови практики и знания в обучението по компютърна лингвистика“. В резултат от работата по проекта е съставен набор от синтактични правила, обобщаващи структурата на фраза с опора прилагателно име и различни разширения – числително име, местоимение, наречие и предложна фраза. Правилата се основават на езикови хипотези, засвидетелствани чрез формални заявки в Българския национален корпус.

В статията на Тодор Лазаров се анализират ресурсите за създаването на статистически модел за превод на глаголни форми от български на английски. Представени са основните трудности, породени от граматичните особености на двата езика, които изникват при формалното описание на глаголните форми за целите на създаването на трансферни правила за машинен превод. Разглеждат се количествените и качествените характеристики на възможните ресурси, както и техните недостатъци и трудностите, свързани с бъдещата работа върху статистическия модел.

Включените в този брой на сп. „Български език“ статии представят само една част от научните резултати на Секцията по компютърна лингвистика към Института за български език при БАН, както и на нейни партньори от различни наши и международни проекти. В същото време те очертават огромните перспективи, които се разкриват пред това научно направление.

✉ сп. „Български език“  
*balgarskiezik@mail.bg*