ЕЗИКОВИ РЕСУРСИ И ТЕХНОЛОГИИ ЗА БЪЛГАРСКИ ЕЗИК

АКАДЕМИЧНО ИЗДАТЕЛСТВО
„Проф. МАРИН ДРИНОВ“

BULGARIAN ACADEMY OF SCIENCES

INSTITUTE FOR BULGARIAN LANGUAGE "PROF. LYUBOMIR ANDREYCHIN"

LANGUAGE RESOURCES AND TECHNOLOGIES FOR BULGARIAN LANGUAGE

Editor *Svetla Koeva*

Sofia • 2014

Prof. Marin Drinov Academic Publishing House

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

ЕЗИКОВИ РЕСУРСИ И ТЕХНОЛОГИИ ЗА БЪЛГАРСКИ ЕЗИК

Съставител *Светла Коева*

СОФИЯ • 2014

АКАДЕМИЧНО ИЗДАТЕЛСТВО
„Проф. МАРИН ДРИНОВ“

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

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

Научни редактори: *Светла Коева, Диана Благоева*

ISBN 978-954-322-797-6

Съдържание

Предговор / **7**

Tamás Váradi. Serving Multilingual Europe: The CESAR Project / **9**

Светла Коева. Българският национален корпус в контекста на световната теория и практика / **29**

Мария Тодорова, Росица Декова. Български POS анотиран корпус – особености на граматичната анотация / **53**

Мария Тодорова, Христина Кукова, Светлозара Лесева. Семантично анотирани ресурси за българския език – БулСемКор / **80**

Екатерина Търпоманова, Цветана Димитрова. Българско-английски паралелен корпус със съотнесени (прости) изречения / **105**

Атанас Атанасов, Марина Джонова. Мултимедиен корпус на българската устна реч – структура и приложение / **127**

Хетил Ро Хауге, Йовка Тишева. Паралелен корпус с данни за българската разговорна реч – структура и приложение / **142**

Светла Коева. WordNet и БулНет / **154**

Борислав Ризов. Софтуерна система за работа с WordNet – Hydra / **174**

Ивелина Стоянова, Мария Тодорова. Разработване на речници на съставните лексикални единици в българския език за целите на компютърната лингвистика / **185**

Диана Благоева, Сия Колковска. „Инфолекс“ – лексикални ресурси за българския език / **202**

Ивелина Стоянова, Светлозара Лесева. Уикипедия като източник на езикови ресурси – корпуси, речници и езикови модели / **216**

Руси Николов. Генериране и управление на езикови ресурси с многофункционалната програма *TREFL* / **231**

Диман Карагьозов. Проектиране и реализация на система за откриване на плагиатство на български език / **248**

Диман Карагьозов, Анелия Белогай, Ангел Генов. Извличане на семантична информация в системата за управление на съдържание *АТЛАС* / **258**

Max Silberztein. Various Computational Devices for Various Linguistic Phenomena with NooJ / **298**

Предговор

Ако под „компютърна лингвистика“ разбираме приложението на формални методи за описание на езиковите данни и подобряване на точността и бързината на анализ с помощта на специализирани компютърни програми, то тогава съвременната лингвистика до голяма степен е компютърна лингвистика. Компютърна лингвистика в нашето разбиране е по-широко понятие. Освен формалното описание на естествения език това понятие включва и т.нар. компютърна обработка на естествения език, което означава както създаване на езикови технологии и изчислителни модели, така и тяхното приложение в системи, които влияят на качеството и ефективността на общуването между хората (и машините) за: проверка на правописа, автоматичен превод, търсене и извличане на информация, трансформиране на писмен текст в реч и обратно, разпознаване на текст от изображения и много други.

Езиковите технологии са една от най-бързо развиващите се области и това е предизвикателство към специалистите, ангажирани в научните и научноприложните разработки, свързани със създаването на езикови ресурси и технологии.

Сборникът „Езикови ресурси и технологии за български език“ съдържа студии и статии, които представят част от резултатите, получени в рамките на два международни проекта – *CESAR: Централно- и южноевропейски езикови ресурси*, и *ATLAS: Система за създаване и поддържане на съдържание, базирана на езикови технологии*. Основният резултат от проекта CESAR е разширяването и разпространението на широк кръг от езикови ресурси и програми за обработка на езика. Постижение на проекта ATLAS е интеграцията на езикови технологии в уеббазирана система за управление на съдържание.

В сборника е представена малка част от целенасочените усилия на широк кръг от специалисти в целия свят за развитие на езиковите технологии, които ще осигурят по-големи възможности за междуезикова комуникация и сътрудничество, както и еквивалентен достъп на носителите на различни езици до информация и познание.

Изданието е предназначено за изследователи (лингвисти, компютърни лингвисти, математици и информатици), частни и бизнес създатели и потребители на езикови ресурси и технологии, както и за всички, които проявяват интерес към съвременните езикови технологии.

Проф. Светла Коева

ЕЗИКОВИ РЕСУРСИ
И ТЕХНОЛОГИИ
ЗА БЪЛГАРСКИ ЕЗИК

Българска
Първо издание

Редактор *Мила Вълкова*
Художник *Константин Жеков*
Дизайн *Десислава Георгиева*

Формат 167×237 mm
Печатни коли 19,50

Печатница на Академично издателство „Проф. Марин Дринов“
1113 София, ул. „Акад. Г. Бончев“, бл. 5

www.baspress.com

ISBN 978-954-322-797-6