Корпус савременог српског језика (СрпКор): стање и перспективе

Милош Утвић

Семинар за језичке технологије
и ресурсе, 29. I 2015. године

План излагања

- Кратак увод у корпусну лингвистику
- Параметри текуће верзије СрпКор-а
  - Структура
  - Анотација
  - Могућности претраге
  - Неки резултати и примене
- Даљи рад

Увод

Семинар за језичке технологије
и ресурсе, 29. I 2015. године

Дефиниције корпуса

- „колекција аутентичних машински читљивих текстова који представљају репрезентативни узорак појединачног језика или језичког варијетета” ([McEnery et al., 2006])
- „*колекција текстова*. Ако се то чини прешироко... колекција текстова који су предмет истраживања језика или књижевности” ([Kilgarriff & Grefenstette, 2003: 334])
- „емпиријски материјал намењен истраживању језика”

Корпусна лингвистика

- Термин је ушао у употребу крајем осамдесетих година прошлог века.
- Дисциплина или методологија?
- Доживљава процват упоредо са статистичким методама за обраду природног језика
- Примене у лингвистици (лексикографија, морфологија, синтакса, семантика, дискурс, прагматика) и рачунарској лингвистици (језички модели као подаци за обучавање програма који обележавају/класификују)

Параметри и класификација корпуса (1)

- Носач:
  - пределектронски
  - електронски
- Домен и намена:
  - општи
  - специјализовани (лексикографски, граматички, дијалекатски, регионални, нестандардни, корпуси језика као нематерњег...)
- Величина:
  - статички
  - динамички (енг. monitor-corpora)
  - опортунистички
- Период:
  - синхрони
  - дијахрони

Параметри и класификација корпуса (2)

- Извор/медијум:
  - говорни
  - писани
  - мултимодални
- Анотација
  - неанотирани
  - анотирани (библиографска, структурна, морфолошка, синтаксичка, семантичка, прагматичка, стилистичка анотација, анотација кореференције)
- Вишејезичност:
  - Једнојезични
  - Вишејезични (паралел(изова)ни и упоредни)

Текућа верзија СрпКор-а

– Кратак историјски преглед
– Параметри
– Фазе изградње
– Могућности претраге
– Анализа и примене

Пројекти (1)

- *Математичка и рачунарска лингвистика*, СР Србија, 1981-1985. године;
- *Аутоматска обрада текста*, ОЗН Београда, СР Србија, 1983-1985;
- *Рачунарство са применама*, РЗН СРС, 1986-1990;
- *Језичке индустрије* (енг. *Language Industries*), Европска унија, 1989-1991. године;
- *Транс-европска инфраструктура језичких ресурса I-II* (енг. *Trans European Language Resources Infrastructure I-II*, скр. *TELRI I-II*), 1995-2001. године;

Пројекти (2)

- 1743 *Интеракција текста и речника*, Министарство за науку Републике Србије, 2002-2004. године;
- 148021 *Теоријско-методолошки оквир за модернизацију описа српског језика*, Министарство за науку Републике Србије и САНУ, 2005-2010. године;
- *Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages*, SEE-ERA.NET (ICT 10503 RP), 2007-2008;

Пројекти (3)

- III 178006 *Српски језик и његови ресурси: теорија, опис и примене*, Министарство за образовање и науку Републике Србије, 2011-2015. године;
- III 47003 *Инфраструктура за електронски подржано учење у Србији*, Министарство за образовање и науку Републике Србије 2011-2015. године;
- *Ресурси Средње и Југоисточне Европе* (енг. *Central and South-East European Resources*, skr. *CESAR*), 2011-2013. године (ICT Policy Support Programme, Grant agreement no.: 271022).

Учесници на изградњи корпуса (2003-2014)

- Група за језичке технологије Универзитета у Београду (Душко Витас, Цветана Крстев, Гордана Павловић Лажетић, Иван Обрадовић, Ранка Станковић, Небојша Васиљевић, Милош Утвић)
- Спољни сарадници:
  - Сандра Гуцул, Вања Радуловић, Катарина Тодоровић
  - Жељко Пајкић, Саша Стевановић, Милан Вукосављевић, Душко Вишић
  - Горан Ракић, Јелена Андоновски, Биљана Ђорђевић, Катарина Станишић, Тијана Стојковић, Биљана Лазић
  - Зоран Ристовић, Бојана Ђорђевић
  - Студенти Филолошког, Математичког, Рударско-геолошког факултета у Београду (семинарски радови).
- Уз свесрдну подршку осталих колега који су учествовали као истраживачи на наведеним пројектима:
  - Универзитет у Београду (Катедра за српски језик и књижевност Филолошког факултета у Београду; Математички факултет; Рударско-геолошки факултет; Пољопривредни факултет; Факултет организационих наука)
  - Универзитет у Новом Саду (Катедра за српски језик и лингвистику Филозофског факултета)

НЕТК/СрпКор2003

- Доступан од децембра 2003. године на адреси http://www.korpus.matf.bg.ac**.yu (**касније **.rs)**;
- НЕТК = Колекција неанотираних текстова величине 22,2 милиона речи;
- Веб-сучеље омогућава претрагу засновану на CQP-регуларним изразима
- Накнадна анотација корпусних текстова одговарајућим библиографским описима
- СрпКор2003 = НЕТК + библиографске информације о корпусним текстовима;
- У периоду када је био једина верзија корпуса регистровало се око 300 корисника.

Учесници (2011-2013)

- Душко Витас, Цветана Крстев, Иван Обрадовић, Ранка Станковић, Милош Утвић (део Групе за језичке технологије Универзитета у Београду);
- Горан Ракић, Јелена Андоновски, Биљана Ђорђевић, Катарина Станишић, Тијана Стојковић, Биљана Лазић (спољни сарадници);
- Студенти Филолошког, Математичког, Рударско-геолошког факултета у Београду (семинарски радови).

Детаљна анализа СрпКор2003

- Cvetana Krstev, Duško Vitas, “Corpus and Lexicon - Mutual Incompletness”, in *Proceedings of the Corpus Linguistics Conference*, 14-17 July 2005, Birmingham, eds. Pernilla Danielsson and Martijn Wagenmakers, ISSN 1747-9398, http://www.corpus.bham.ac.uk/PCLC/, 2005.
- На основу СрпКор2003 су генерисане листе учестаности секвенци *n* узастопних корпусних речи, где је $1 \leq n \leq 3$ (униграми, биграми и триграми).

СрпКор2013

- СрпКор2013 је текућа верзија СрпКор-а, настала у периоду од јула 2009. до јануара 2013. године (уз индиректну помоћ неколико пројеката: 148021, III 178006, III 47003, CESAR)
- СрпКор2013 је доступан уз претходну бесплатну регистрацију на истој адреси као досадашње верзије: http://www.korpus.matf.bg.ac.rs
- Контакт (захтев за креирање корисничког налога): korpus@matf.bg.ac.rs

СрпКор2013 – параметри

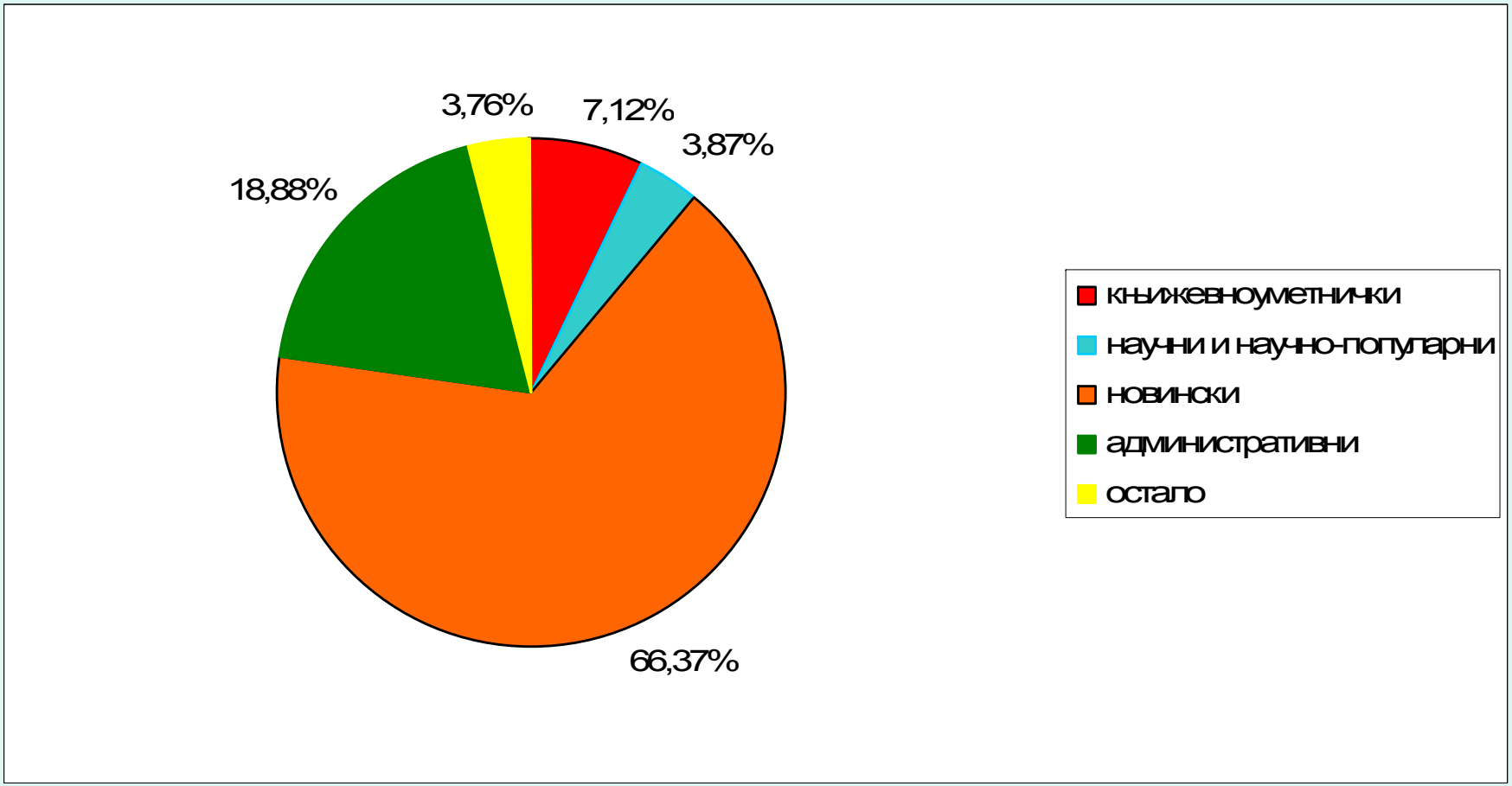
- Носач: електронски
- Извор/медијум: корпус углавном писаних текстова
- Вишејезичност: једнојезични
- Величина: динамички корпус
  - Порекло текстова: веб, донације аутора, семинарски радови (прекуцавање и сканирање)

СрпКор2013 (структура)

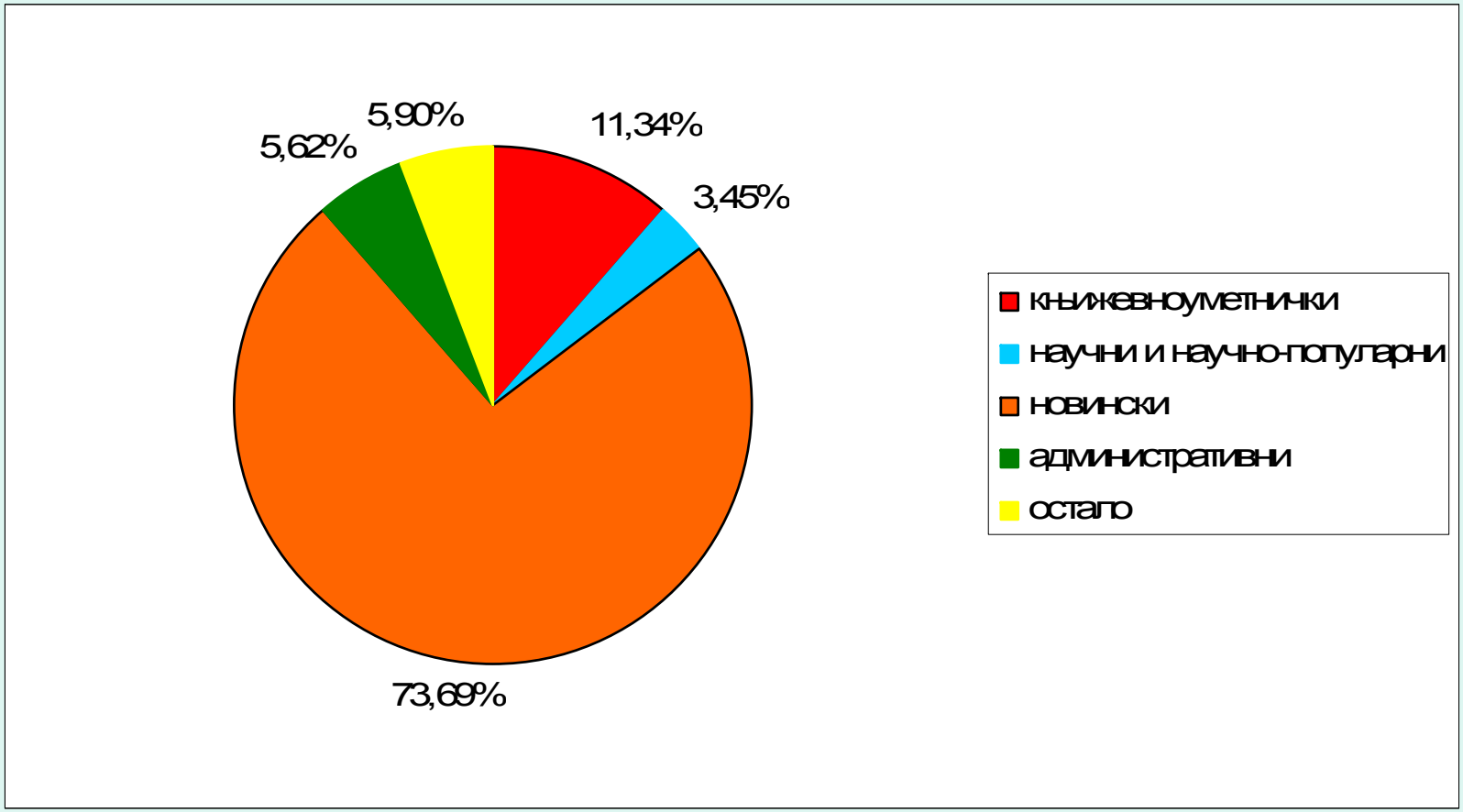
текстови	токени	типови	корпусне речи	корпусни типови
4889	152.540.721	1.424.899	122.255.064	1.402.664

статус текста у односу на изворну верзију текста	**текстови (%)**	**токени (%)**	**типови (%)**	**корпусне речи (%)**	**корпусни типови (%)**
написан на српском	92,74%	90,47%	93,50%	90,32%	93,44%
превод на српски	7,26%	9,53%	6,50%	9,68%	6,56%

Расподела текстова по функционалним стиловима



Расподела корпусних речи по функционалним стиловима



СрпКор2013 - параметри (2)

- Домен и намена: претендује на општост, још увек није достигао жељену балансираност (удео новинских текстова се повећао у односу на СрпКор2003)
- Период: синхрони корпус (доминирају текстови објављени после 2000. године, ~87%)

СрпКор2013 - анотација

- Сваком токену корпуса су придружене одговарајуће вредности следећих седам позиционих атрибута:
  - идентификатор корпусног текста,
  - идентификатор аутора корпусног текста,
  - година издања изворног текста,
  - идентификатор функционалног стила,
  - идентификатор статуса текста у односу на језик оригиналне верзије изворног текста,
  - врста речи и
  - лема.
- На основу вредности идентификатора корпусног текста атрибута се приступа одговарајућем библиографском опису корпусног текста у табели релационе базе података

Морфолошка анотација (1)

- Приликом одабира алата за аутоматску морфолошку анотацију текста, тестиране су могућности три алата:
  - Unitex (Sébastien Paumier, http://www-igm.univmlv.fr/unitex/)
  - TnT (Thorsten Brants, http://www.coli.uni-saarland.de/~thorsten/tnt/)
  - TreeTagger (Helmut Schmid, http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/)
- TreeTagger је изабран методом елиминације
- TnT и TreeTagger су дали најбоље резултате приликом евалуације (прецизност / удео непознатих речи у грешакама је 93,86% / 58,36% за TnT, односно 91,78% / 36,71% за TreeTagger)

Морфолошка анотација (2)

- СрпКор2013 је аутоматски анотиран уз помоћ прерађеног подскупа електронског морфолошког речника српског језика (СрпМД) у формату LADL/DELA. За разлику од полазног речника који садржи вредности свих морфолошких категорија, прерађени подскуп за сваку одредницу бележи само информацију о леми и врсти речи.
- Поред речника, као скуп података за обучавање алата TreeTagger је коришћена српска верзија паралелног корпуса SELFEH (INTERA) коју су ручно анотирали Ц. Крстев, Д. Витас, Г. Павловић Лажетић, Иван Обрадовић и Сандра Гуцул (сваком токену су придружене одговарајућа лема и врста речи).
- Приликом евалуације TreeTagger-а десетоструким унакрсним тестом над SELFEH-ом, просечна прецизност анотације корпусних речи из речника је 96,57%, а удео непознатих речи у односу на број грешака приликом анотације је 11,93% (Утвић, М. 2011. “Анотација Корпуса савременог српског језика”. *ИНФОтека* 12(2), 39-51)

Скуп етикета (1)

- ознаке десет врста речи у српском језику
  - N (именице),
  - A (придеви),
  - V (глаголи),
  - PRO (заменице),
  - NUM (бројеви),
  - PREP (предлози),
  - ADV (прилози),
  - CONJ (везници),
  - INT (узвици),
  - PAR (речце или партикуле).

Скуп етикета (2)

- шест специјалних ознака:
  - ABB (скраћеница)
  - PREF (префикс)
  - RN (римски број),
  - PUNCT (знак интерпункције),
  - SENT (ознака краја реченице) и
  - остало (?)

Претрага

- Софтвер којим је индексиран корпус омогућава претрагу помоћу језика CQP.
- Уколико корпус није анотиран, претрага се своди на примену регуларних израза:

`zima`	{zima}
`zim[a-z]`	{zima, zime, zimu, zimi,…}
`zim[a-z]+`	{zima, zimom, zimski,…}
`[a-z]+ski`	{zimski, rimski, srpski,…}
`z[a-z]+ski`	{zanatski, zetski, zimski,…}

Џокери

- Токен је
  - низ алфабетских карактера (‘реч’, корпусна реч),
  - низ цифара (‘број’),
  - сваки неалфанумерички карактер се третира као појединачни токен.
- Произвољан токен
  `je []* rekao` {je rekao, je brzo rekao, je mnogo brzo rekao, …}
- Произвољан карактер
  `zim.+` {zima, zimom, zimski,…}. Али и zimski,
- У случају да се регуларни израз односи на корпусну реч, тј. садржи бар једно слово, онда се џокер ‘.’ своди на `[a-z] zim.` {zima, zime, zimu, zimi, zim…}

Претрага - примери

- `"star"`
  корпусна реч *стар*
  `[word="star"]`
- `[lemma="star"] [lemma="slava"]`
  све варијације облика лема стар и слава
- `[pos="A" & lemma=".*ski"]{2} [pos="N"]`
  два узастопна облика релационих придева на -*ски*, за којима следи произвољна именица

Претрага - слагање

- `"s" a:[pos="N"] "na" b:[pos="N"] :: a.word=b.word;`
  {s čoveka na čoveka, s mladića na mladića,...}
- `"s" a:[pos="N"] "na" b:[pos="N"] :: a.lemma=b.lemma;`
  {s brda na brdo, s kolena na koleno,...}

Претрага - савет

- Избегавати 'свеобухватне' упите. На пример, тражимо облике повратних глагола (*радујем се*, *се радујем*).
- Уместо једног упита
  `([pos="V"] "se") | ("se" [pos="V"]);`

боље је рашчланити упит:

- `[pos="V"] "se";`
- `"se" [pos="V"] ;`

СрпКор2013 – прве анализе

- генерисане су листе учестаности речи у корпусу, сортиране алфабетски и по учестаностима (фреквенцијски речник корпуса);
- генерисане су листе учестаности низова *n* узастопних корпусних речи, $1 \leq n \leq 5$ (*n*-рами)
- генерисана је укупна листа учестаности лема, као и листе учестаности лема које припадају истим променљивим врстама речи (именице, глаголи, придеви)

Даљи рад

- Око 200 нових текстова је припремљено за следећу верзију СрпКор-а, а далеко више чека на обраду
- Истраживање могућности да се корпусне речи додатно анотирају преосталим морфолошким информацијама (ЛАДЛ/ДЕЛА, Unitex, генерисани n-грами)
- Делимична семантичка анотација СрпКор-а коришћењем српске верзије лексичке семантичке базе WordNet

Хвала на пажњи

Семинар за језичке технологије
и ресурсе, 29. I 2015. године