Български POS анотиран корпус - особености на граматичната анотация

Мария Тодорова и Росица Декова

Секция по компютърна лингвистика
Институт за български език, БАН

Езикови ресурси и технологии за български
София, 30.09.2013 г.

Мотивация

- Статията представя създаването на Българския POS анотиран корпус с оглед на особеностите на граматичната анотация.
- Незаменим лингвистичен ресурс за български език с широки приложения:
  - компютърната лингвистика
  - други научни области

Морфологична анотация

- Обработката на думите, така че за всяка една дума от текста, която представлява низ от символи (или фонеми) на входа на морфологичния анализатор, да бъде извършен коректен морфологичен анализ и в резултат на това да ѝ бъдат прикрепени тагове (етикети) с информация относно граматичните ѝ характеристики.
- Основен етап в обработката на лингвистичните ресурси и предпоставка за следващи етапи.

Структура на БулПосКор

- Представителна част от Българския “Браун” корпус (ББК);
- Извадка от минимум 300 думи от всеки файл в ББК, разширена до край на изречение;
- Текстовете са разделени в 15 категории от 2 типа – художествени и информативни;
- 217 210 токъна, вкл. 172 482 думи, 42 058 знака и 2 670 числа

Анотация на корпуса

- Токънизация.
- Предварителен етап.
- Същинско автоматичното анотиране.
- Ръчно разрешаване на случаите на многозначност.
- Валидация.

Предварителен етап на анотиране

- Въз основа на Българския граматичен речник (Коева, 1998);
- Въвежда се информация за граматичния клас на дадена лексикална единица и съответните стойности на граматичните категории, характеризиращи единицата.
- Приписването на тагове следва формализма
  <атрибут:стойност>
- Таговете са построени според стандартната **DELAF структура,** състояща се от:

<лема, граматични категории на лемата; граматични характеристики на словоформата>

Автоматично граматично анотиране

- Използвана е програмата **BgTagger**, разработена в СКЛ, ИБЕ-БАН
  - SVM тагер за еднозначно автоматично определяне на частите на речта в произволни текстове;
  - предсказва частта на речта въз основа на множество от характеристики, които описват думата и нейния контекст.
- На 51,9% от токъните е приписано едно значение;
- 46,7% получават повече от едно значение;
- 1,4% от токъните са оставени нетагирани
  - вкл. редки и/ или чужди думи или лични имена, които не присъстват в речника

Ръчно разрешаване на случаите на многозначност

- Граматическите характеристики на всяка словоформа са определени еднозначно от лингвисти1.
- Системата за анотиране на текстове **Chooser,** създадена в СКЛ (Ризов 2010).
- Корпус, който е коректно и еднозначно анотиран за частите на речта на отделните словоформи.

1. Ивелина Стоянова; Петя Плачкова; Албена Копринарова

Многозначност

- Наличие на асиметрия между формата и съдържанието на даден езиков знак, когато на една форма съответства повече от едно значение (Петрова 2009)
- Отстраняването на многозначността е необходимо условие при създаването на различни видове програмни приложения за обработване на естествения език, както и за целите на теоретичните и теоретично-приложните лингвистични изследвания.

Видове многозначност в БулПосКор

В зависимост от граматичните признаци, които се застъпват в обща словоформа (Коева и др. 2006):

Класове многозначност

- Проявите на системна многозначност са групираниспоред **пресичането на атрибутивните стойности**, зададени при тагирането:

  [+/–лема; +/–грам. категории на лемата;

  +/–грам. х-ки на словоформата]

- Плюс “+” – съвпадение на признаци при дадена словоформа, а минус “–” – наличие на два или повече граматични признака при една словоформа.
- Мястото, където се налага да бъде направен изборът на коректното морфологично значение определя класа на многозначност.

Класификация

Морфо-граматична многозначност

Граматична многозначност

[+лема; +грам. категории на лемата; –**грам. характеристики на словоформата**]

Категориална многозначност

[+лема; –**грам. категории на лемата**; +грам. характеристики на словоформата]

Категориално-граматична многозначност

[+лема; –**грам. категории на лемата**; **-грам. характеристики на словоформата**]

Морфо-лексикална многозначност

Лексикално-граматична многозначност

[**–лема; +грам. категории на лемата;** –грам. характеристики на словоформата],

Комбинирана

Лексикално-категориална многозначност

[**–лема; –грам. категории на лемата**]

Граматична многозначност

Многозначността на дадена словоформа се дължи на общ израз на две или повече граматични характеристики

словоформа	лема	граматични категории на лемата	граматични характеристики на словоформата
човека	+ човек	+ същ, м.р.	– бройна форма/ кратка членна ф-ма
катереше	+ катеря	+ глагол, минало несв. вр.	– 2 л. / 3 л. ед. ч

Категориална многозначност

Многозначността се дължи на различие на граматичните категории на дадена словоформа

- глаголи от II спрежение, при които формите за 3 л. ед. ч., сегашно и минало свършено време, съвпадат;
- несвършени преходни глаголи, при които формите за минало свършено и минало несвършено деятелно причастие, в ж. р., ед. ч., нечл. съвпадат (*казвала*)

словоформа	лема	граматични категории на лемата	граматични характеристики на словоформата
гради	+ градя	– глагол, сег вр/ мин вр.	+ 3 л. ед. ч

Категориално-граматична многозначност

При една словоформа, под обща лема се появяват различни граматични категории, като в рамките на една или няколко категории се появява граматична многозначност.

словоформа	лема	граматични категории на лемата	граматични характеристики на словоформата
гради	+ градя	– глагол, сег. вр/ мин вр. / императив	– 3 л. ед.ч 2 л./ 3 л. ед. ч. 2 л. ед. ч.

Лексикално-категориална многозначност

Характеризира идентични словоформи, принадлежащи към различни леми с различни граматични категории на лемата.

- при неизменяемите думи (*като*, *на*, *че*, *да*)
- изменяема и неизменяема дума или изменяеми думи
- субстантивираните прилагателни (*добро* / *зло*)
- словоформите *ми*, *си*, *се*, и т.н.

словоформа	лема	граматични категории на лемата	граматични характеристики на словоформата
прах	– прах/ пера	– същ., м. р / гл., мин. св.вр.	+ ед. ч., нечл. 1 л. ед. ч

Комбинирана многозначност

Проявява се и при трите параметъра.

словоформа	лема	граматични категории на лемата	граматични характеристики на словоформата
изказа	– изказ / изказвам	– същ. м.р. глагол, мин вр.	– бройна форма/ кратка членна ф-ма 2 /3 л. ед. ч.
катери	– катер / катеря	– съществително / глагол, сег. вр./ глагол,мин. св. вр./ глагол императив	– мн. ч., м. р./ 3 л. ед. ч./ **2 / 3 л. ед. ч/** 2 л. ед. ч.

Критерии за анотиране

- На базата на анализ на възможните класове от многозначни форми се приемат анотационни критерии, които да осигурят единен подход при граматичната анотация.
- В зависимост от реализацията на стойностите на атрибутивни признаци и тяхната комбинация при употребата на многозначната словоформа в конкретния контекст:
  - контекстов анализ;
  - морфологичен анализ;
  - лексикално-семантичен анализ;
  - синтактичен анализ;
  - тестове за замяна;
  - комбиниран анализ.

Контекстов анализ

Според морфологичната или синтактичната функция на многозначната словоформа или според конкретните семантични характеристики, които могат да бъдат изведени от контекста.

- **Например при идентификацията на времето** многозначността <свършено - несвършено време> се разрешава от контекста при комбинацията на причастието със спомагателния глагол *съм*.

1) *В това събрание той [бе избран] с пълно мнозинство.*

Лексикално-семантичен анализ

Многозначността на словоформите се сваля въз основа на техните семантични значения.

- **Многозначност <причастие – прилагателно>** при отделни речникови единици

3) *Най-накрая тя бе **уверена** в правотата му.*

страдателен залог на глагола ***уверя* (*някого да направи нещо*)**

4) *Тя бе напълно уверена в собствените си сили.*

прилагателното ***уверен* (*в нещо*)**

Тестове за заместване

Замяна на многозначни словоформи с немногозначни такава, принадлежащи към същата или към различна лема, в същия контекст, без да причинява неграматичност на изречението или да променя интерпретацията му.

Пример: **различни синтактични значения на *се***

5) *Той **се** прикри от стрелите.*

6) *Той прикри **себе си** от стрелите.*

Комбинирано прилагане на принципи

В процеса на разрешаване комбинираната многозначност, която обхваща 3 или повече стойности на атрибута, може да бъде сведена до чисто категориална (7) или чисто граматична (8).

Примери:

7) *Иван* ***коси*** *ливадата. (5 дни / вече пети ден)*

8) *Вчера* ***гради*** *цял ден.*

Валидация на БулПос Кор

Извършена е вторична анотация с цел валидиране на Българския POS анотиран корпус

- високо качество на създадения езиков ресурс;
- индивидуалният човешки елемент при анотирането е сведен до минимум.

Приложения на БулПосКор

- Тренировъчен или тестови корпус при създаването на програми за автоматично отстраняване на граматична многозначност и лематизация (автоматично определяне на основната форма);
- Усъвършенстване на базата от данни на Българския граматичен речник – допълване и корекция;
- Лесна и бърза употреба, както от специалисти, така и от неспециалисти при ефективно интелигентно търсене на езикови модели и форми чрез системата за търсене на Българския национален корпус.

Благодарим за вниманието!

ПРОЕКТ BG051PO001-*3.3.06-0022 от 19.03.2012 г.*
„Интегриране на нови практики и знания в обучението по компютърна лингвистика“
Проектът се осъществява с финансовата подкрепа на Оперативна програма „Развитие на човешките ресурси”, съфинансирана от Европейския социален фонд на Европейския съюз