ПЛОВДИВСКИ УНИВЕРСИТЕТ „ПАИСИЙ ХИЛЕНДАРСКИ“ – БЪЛГАРИЯ
НАУЧНИ ТРУДОВЕ, ТОМ 44, КН. 1, СБ. А, 2006 – ФИЛОЛОГИЯ
PLOVDIV UNIVERSITY „PAISII HILENDARSKI“ – BULGARIA
SCIENTIFIC PAPERS, VOL. 44, BOOK 1, PART A, 2006 – PHYLOLOGY

# ПРОБЛЕМИ ПРИ СЪЗДАВАНЕТО НА ЕЗИКОВИ КОРПУСИ С ТРАНСКРИБИРАНА БЪЛГАРСКА РАЗГОВОРНА РЕЧ

***Атанас Атанасов***
***Софийски университет „Св. Климент Охридски“***

Един от основните проблеми при създаването на езикови корпуси е свързан с тяхната стандартизация, т.е. с превръщането на необработения текстов архив в анотиран корпус. През последните години в областта на компютърната лингвистика и т. нар. „компютърна обработка на естествения език“ (Natural Language Processing – NLP) при работата с големи текстови архиви все повече се налагат езиците за маркиране (markup languages). XML (Extensible Markup Language) е един от тези езици. Той е базиран на езика SGML (Standard Generalized Markup Language), чието основно предназначение е да се използва като универсален начин за форматиране на произволен тип данни. Тъй като SGML е прекалено усложнен и тежък език, XML е създаден като негово подмножество, което е значително опростено, но също така изключително мощно средство при описването на данни. Повечето компютърни лингвисти се насочват именно към XML заради голямата му гъвкавост и възможността да бъде описана езиковата структура на абсолютно всички нива (в зависимост от конкретната задача).

Целта в настоящата статия е да бъдат представени някои от трудностите при обработването на транскрибирана разговорна реч и да бъде направен опит за тяхното адекватно разрешаване. Както беше посочено по-горе, за да бъде превърнат един текстов архив в анотиран корпус, е необходимо да бъде следван определен стандарт. Спряхме се на спецификацията, предложена от консорциума TEI (Text Encoding Initiative), тъй като тя съдържа специален модул за

обработка и анотиране на транскрибирана реч и през последните години се е утвърдила като един от водещите световни стандарти. За съставянето на дефиницията на документния тип (DTD) беше използван разработеният от TEI софтуер „Pizza Chef“ (http://www.tei-c.org/pizza.html).

Първият важен въпрос е какъв тип информация е необходимо да бъде включен в анотацията. В нашия корпус (http://bgspeech.net) тя е разделена на две основни части – „метаданни“ (съдържащи общи сведения за текста, описание на участниците, продължителност на записа и т.н.) и информация, описваща лингвистичните явления (на различни равнища – фонологично, морфосинтактично, прагматично и дискурсно).

Основната структура на документа съдържа следните елементи:

```
- <TEI.2>
  - <teiHeader>
    + <fileDesc></fileDesc>
    + <profileDesc></profileDesc>
    + <revisionDesc></revisionDesc>
    </teiHeader>
  - <text>
    + <body></body>
    </text>
  </TEI.2>
```

В елемента <teiHeader>, където се намират метаданните, има три главни елемента: <fileDesc>, <profileDesc> и <revisionDesc>. Първият съдържа информация за файла (заглавие, автор, продължителност на записа, място и дата на публикуване, информация за източника и т.н.). Вторият дава информация за текста (език, дата и място на осъществяване на записа, описание на участниците, характеристика на текста и т.н.). Третият съдържа информация за верификацията на транскрипцията.

В елемента <body> се намира самият корпус. Той ще бъде разгледан по-подробно, тъй като основните трудности са именно тук. Текстът е конструиран от изказвания, като всяко от тях е включено в елемент <u>. Той има атрибут „*who*“, който осигурява идентификатор на говорещия, като неговата стойност съвпада с тази на атрибута „*id*“ на елемента <person> (намиращ се в частта, описваща участниците). Понякога изказванията могат да бъдат прекъсвани от паузи. Елементът <pause> е празен и има атрибут „*dur*“, показващ приблизителната продължителност на паузата – кратка, средна или дълга (символът „\“ отбелязва ударението):

```
- <u who="AA">
    \ти та\ка ше разбе\реш \кой \е
    <pause dur="short"/>
    разб\ра ли бе\бона ?
  </u>
```

По-надолу ще бъдат разгледани някои екстралингвистични явления и ще бъде показан начинът за разрешаването на някои трудности при тяхната анотация.

## НЕПЪЛНИ ИЗКАЗВАНИЯ

Често говорещият е прекъсван от друг участник в разговора, като възможностите тук са три:

```
- <u who="BA">
    \ох
    <pause dur="short"/>
    \ужас
    <pause dur="short"/>
    <seg part="I">о\тиваме \е \са в \петък пре\ди \да </seg>
  </u>
  <u who="AA"> на \следващия \ден ? </u>
```

Този пример демонстрира изказване, което е непълно (незавършено). Говорещият започва репликата си, но е прекъснат, т.е. само началната част на неговото изказване е налице. Елементът <seg> обозначава част от изказването, а неговият атрибут „*part*“ показва коя точно част е изпусната и коя е налична. Например <seg part="I">...</seg> означава, че даден фрагмент е „начален“ (initial), т.е. не е завършен; <seg part="F">...</seg> означава, че даден фраг-

мент е „краен“ (final), т.е. изпусната е началната част, както в този случай:

```
<u who="AA"> е как\во \нищо ? </u>
- <u who="BA">
    <vocal desc="ъ"/>
    <seg part="F">и я зап\лашвал с то\ва </seg>
    <pause dur="short"/>
    \тя \да си о\тиде на \костинброд и нак\рая \тя си о\тиде на \костинброд
  </u>
```

Следващият пример илюстрира по-сложен случай на частично изказване – първият говорещ започва изказването си, вторият го прекъсва, след което първият довършва изречението си:

```
    <pause dur="medium"/>
    <seg part="I"> \снощи </seg>
  </u>
- <u who="BA">
    <seg part="M"> в \осем </seg>
  </u>
- <u who="AA">
    <seg part="F">в \девет си си \легнала </seg>
    <pause dur="medium"/>
```

Изказването на втория говорещ е маркирано като фрагмент, чийто атрибут има стойност „M“ – междинна част от непълно изказване, т.е. не са налични нито началната, нито крайната част. Всъщност в конкретния случай те се намират в репликите на първия говорещ, а репликата на втория е вмъкната между тях, така че трите фрагмента заедно образуват една синтактична единица.

### ФАЛСТАРТ

С този термин са означени един вид непълни изказвания – когато някой от участниците започва да говори, но решава да смени темата, преди да е завършил изказването си; фрагментът в елемента <seg> е незавършен (затова стойността на неговия атрибут винаги е „I“ – initial):





```
  </u>
- <u who="AA">
    сак\сията \е
    <seg part="I">със </seg>
    стра\хотна
    <pause dur="medium"/>
    \имам \нарове
  </u>
```

### КОРЕКЦИИ (SPEECH REPAIR)

Това явление е подобно на фалстарта. Разликата е, че тук темата на разговора не е променена; говорещият се връща към предходна част от собственото си изказване, за да го прецизира или коригира:

```
<seg part="I"> \чак нак\рая се запоз\нава </seg>
<pause dur="short"/>
\деси \най - нак\рая се запоз\нава със \нея
<pause dur="short"/>
```

### „ПРЕКЪСНАТИ“ ДУМИ

Корекциите, фалстартът или просто повреден запис могат да бъдат причина и за наличието на непълни думи в корпуса. Тъй като те не образуват синтактични единици, елементът <seg> не е подходящ при тяхното анотиране; вместо това е използван елементът <w>. Когато причината за прекъсването е прагматична (а не повреда в записа), единствената възможна стойност за атрибута „*part*“ е „I“ (initial):

```
- <seg part="I">
    на \другия \ден теле\фона му е \цял
    <w part="I">д </w>
  </seg>
  \те не из\лизат вечер\та на \другия \ден му \е \цял \ден теле\фона изк\лючен
  <pause dur="short"/>
```

### НЕЯСЕН ТЕКСТ

Когато в записа има неясни части (повредена лента, шум и т.н.), се използва елементът <gap>. Неговият атрибут „*reason*“ обяснява конкретната причина за неяснотата:

```
- <u who="BA">
    ше му се о\бада и ше му \кажа
    <gap reason="unclear"/>
  </u>
- <u who="AA">
    \даже \тряше \по - \раничко \да му \звъннеш
    <pause dur="short"/>
    спо\ред \мен
  </u>
```

## ФОНЕТИЧНИ ПАРАЛИНГВИСТИЧНИ СРЕДСТВА

В спецификацията на TEI с елемента <vocal> се означават фонационни паралингвистични средства, като значението им и приблизителната им звукова стойност се обясняват в аргумента „*desc*“:

```
- <u who="AA">
    \ти \кво ше \режеш се\га
    <pause dur="short"/>
    <vocal desc="м - въпросително"/>
  </u>
```

## КИНЕМИ

С елемента <kinesic> се обозначават паралингвистични жестови и мимически кинеми, чиято семантика се отбелязва в стойността на атрибута „*desc*“:

```
- <u who="OB">
    \слави \трифонов ха\ресваш ли го ?
    <pause dur="short"/>
  </u>
- <u who="BA">
    <kinesic desc="изкривява устни в израз на отрицание"/>
  </u>
```

## ЗАСТЪПВАНЕ (OVERLAPPING)

Понякога отделни части на изказванията, продуцирани от говорещите, могат да се застъпват. Едновременното говорене на двама (или повече) от участниците представлява една от най-големите





трудности при анотирането на транскрибирани текстове. Тук е избрано много удобно решение – елементът <anchor> дава възможност да се поставят темпорални точки в текста:

```
- <u who="AA">
    \ми \да \де \именно
    <pause dur="short"/>
    <anchor id="t09"/>
    \те от то\ва се полу\чават
    <anchor id="t10"/>
  </u>
- <u who="BA">
    <anchor synch="t09"/>
    не \знам от \кво
    <anchor synch="t10"/>
    се по\лучиха \теа \мравки ?
  </u>
```

По този начин могат да бъдат представени не само най-често срещаните случаи, когато краят на изказването се застъпва с началото на следващото, но и много по-сложни ситуации:

```
- <u who="AA">
    \дай \дай
    <anchor id="t19"/>
    \аз ше си ги
    <anchor id="t20"/>
    на\режа
  </u>
- <u who="BA">
    <anchor synch="t19"/>
    \по - \едро
    <anchor synch="t20"/>
  </u>
```

Тук например цялото изказване на втория говорещ се застъпва с междинна част от изказването на първия говорещ.

## КОМУНИКАТИВЕН СТАТУС НА ИЗКАЗВАНИЯТА

Когато изказването не е индикативно, е маркирана и неговата илокутивна сила (напр. когато е интерогативно или императивно):

```
- <u who="ОВ">
    <seg type="interrogative">\да \не \са със \заешки \пух </seg>
  </u>
```

## ЗАКЛЮЧЕНИЕ

Статията представя основните трудности, които се срещат при опитите да бъде създаден електронен анотиран корпус с българска разговорна реч. Главната цел при бъдещата работа върху корпуса е да бъде направена по-прецизна анотация на текстовете, да се даде по-детайлна информация за морфологичната характеристика на думите и за синтактичните единици (конституентните групи NP, VP, AP и т.н.). Важно е също да бъдат открити основните характеристики, които отличават лингвистичните данни в корпусите с българска разговорна реч от тези в корпусите с писмена (или стандартизирана) реч.

## БИБЛИОГРАФИЯ

http://bgspeech.net
http://bultreebank.org
http://www.tei-c.org