Български език/Bulgarian Language 65 (2018), Приложение/Supplement, 144–154

Интердисциплинарен подход за идентифициране на изречения с участието на адюнкти (върху корпус с XML описания)

Елисавета Балабанова
Университет по библиотекознание и информационни технологии – УниБИТ
e.balabanova@unibit.bg

An Interdisciplinary Approach to Identifying Sentences with Adjuncts (Based on a Corpus of XML Representations)

Elisaveta Balabanova
University of Library Studies and Information Technologies – UniBIT
e.balabanova@unibit.bg

The paper presents an interdisciplinary approach (combination of linguistics and computer linguistics) for identification of sentences with adjuncts in contemporary Bulgarian. The investigation is made within the corpus Bultreebank (www.bultreebank.org). The texts in the corpus are in XML format. The identification and extraction of sentences is made by use of the Clark system – a system for corpora development (http://bultreebank.org/en/clark/). All positions of adjuncts in the sentence are defined (four positions). An informal word order model with all positions, in which adjuncts appear, is defined. XPath expressions, identifying the adjuncts in the particular positions in the XML documents, are written. Defined are four groups of sentences, according to the position of the adjunct. Statistics is made, evaluating the frequency of each adjunct, appearing in each particular position in the sentence.

Keywords: word order models, adjuncts, XML, XPath

1. Увод

Словоредът в българското изречение не е бил обект на изследователски интерес в продължение на десетилетия. След станалите христоматийни трудове на Ел. Георгиева отпреди приблизително 40 години (Георгиева/Georgieva 1974; Георгиева/Georgieva 1987) за дълъг период от време в българското езикознание се появяваха само спорадични изследвания, които засягаха (най-често като съпътстващи към осветляването на други езиковедски проблеми) предимно частни словоредни въпроси.

В последните няколко години се забелязва нараснал интерес към словоредната тематика, която за българския език представлява съчетание от въпроси, свързани с разпределението на информацията в изречението, с

проблематиката на дискурса, а също и с пресечната точка между синтаксис и семантика.

Тук ще цитираме някои от най-новите изследвания по словоредни въпроси, като в никакъв случай не претендираме за изчерпателност (вж. Тишева/Tisheva 2011; Тишева/Tisheva 2013; Теофилова/Teofilova 2017, Вилимовска/Vilimovska 2017). Нараства и броят на словоредните изследвания, свързани с особеностите на разговорната реч като част от тенденцията за разкриване на закономерностите на разговорната реч – тендеция, започнала преди 15-ина години в нашето езикознание.

По-голямата част от учените отчитат факта, че в словореда се отразяват различни аспекти на езика – разпределението на информацията в изречението, семантичните особености на единиците, участващи в словоредните модели, както и това, че словоредът е отражение на дискурсни фактори. Разбира се, при оформянето на словоредните модели в разговорната и в неразговорната реч се наблюдава доминация на различни типове фактори, но не това е обект на настоящия текст.

Настоящото изследване се вписва в поредицата изследвания върху словоредни въпроси, като се занимава с представянето на всички словоредни позиции, в които се реализират адюнкти[1] в простото и в сложното изречение в българския книжовен език. То представя първия етап от едно по-голямо изследване, замислено с цел да опише всички словоредни модели в българското изречение, в които участват адюнкти.

2. Изследването е направено върху корпуса със синтактични описания на българския език BulTreeBank (http://www.bultreebank.org/BTBDescriptionTreebank.html).

BulTreeBank е корпус със синтактични описания на български изречения в XML формат. Корпусът е балансиран – съдържа текстове от всички стилови регистри с изключение на разговорния. Всички текстове в корпуса са маркирани съгласно указанията на TEI[2] и са в XML формат, което позволява извличането на информация на съответните езикови нива (морфологично и синтактично). Вариантът на корпуса, върху който е извършено изследването, е маркиран морфологично и синтактично. За да посочим как извличаме информация за словоредните модели с адюнкти, е необходимо накратко да представим езиците XML и XPath.

XML[3] (eXtensible Markup Language) е ново поколение език за структурно описание, създаден първоначално за описание и обмен на данни по интернет. XML произлиза от езика SGML (Standard Generalized Markup Language), но за разлика от него е много по-гъвкав, по-лесен за имплементация и с много големи възможности за приложение.

XML представя структурата на документа като поредица от елементи. Целият документ е елемент, който съдържа останалите елементи. Структурните елементи на документа са маркирани посредством тагове[4]. Таговете могат да ограждат съдържанието на даден елемент или могат да маркират отделни негови части. Таговете биват два типа: отварящи (<text>) –

маркират началото на елемента, и затварящи </text> – маркират края на елемента.

В компютърната лингвистика езикът XML се използва широко поради необятните му възможностите за представяне на езиковата структура на всички нива. Това е възможно, тъй като елементите, чрез които се описва езиковата структура, могат да се дефинират от самия лингвист с оглед на целта. Ето защо за всяка конкретна цел може да бъде създадена съответна XML структура (дефинирането на елементите, които участват в дадена структура, както и на алгоритмите на тяхното взаимодействие става в документ, наречен *дефиниция на документен тип* (DTD – Document Type Definition).

Всички файлове от корпуса BulTreeBank са под формата на XML документи. За да може да се направи извличане на информация от тях, използваме езика XPath, който е мощен език за селектиране на елементи от XML документи. XPath разглежда всеки XML документ като дърво, в което възлите представят елементите на документа. Най-високият възел е коренът на дървото, а децата на един възел са съдържанието на съответния елемент.

За нашата цел беше нужно да започнем с написването на XPath изрази, които да откриват изречения с адюнкт в съответната словоредна позиция.

3. XPath изрази, разпознаващи адюнктите в съответните словоредни позиции в XML документи.

3.1. Разпознаване на изречения с адюнкт между подлога и опората се извършва чрез XPath израза //VPA[child::nid]/descendant::VPS/child::DiscA.

Изразът гласи: Търси фраза от вид опора адюнкт, която има дете елемент от вид ненепосредствено доминиране (nid) и наследник фраза от вид опора подлог с дете елемент, който причинява дистантна фразова реализация (дистантната фразова реализация е термин, който посочва, че елементите на дадена фраза не се реализират контактно един до друг). Елементът DiscA показва, че адюнктът не принадлежи към фразата от вида VPS, но словоредно се реализира вътре в нея.

Пример с изречение от корпуса (под изречението е даден XML документът, изваден от корпуса):

България по никакъв начин не би могла да бъде арбитър в отношенията между Русия и Украйна.

```
<LC>
<CoIndex>
<identifier ="id12"/>
<CoIndex/>
<N sort="NE-Loc">
<name cat="lex" sort="NE-Loc">Българ**ия**</name>
</N>
```

```
</LC>
<I>
<DiscAidref="id12">
<PP>
<Prep>по</Prep>
<NPA>
<Pron>никакъв</Pron>
<N>начин</N>
</NPA>
</PP>
</DiscA>
</I>
<RC>
<VPC>
<V>
<T>не</T>
<V>би</V>
<Participle>могла</Participle>
</V>
<CLDA>
<VPC>
<V>
<T>да</T>
<V>бъде</V>
</V><NPA>
<N>арбитър</N>
<PP>
<Prep>в</Prep>
<NPA>
<N>отношения</N>
<PP>
<Prep>между</Prep>
<CoordP>
<ConjArg>
<N sort="NE-Loc">
<name amb="on" cat="lex" sort="NE-Loc">Русия</name>
</N>
</ConjArg>
<Conj>
<C>и</C>
</Conj>
<ConjArg>
<N sort="NE-Loc">
<name cat="lex" sort="NE-Loc">Украйна</name>
```

```
</N>
</ConjArg>
</CoordP>
</PP>
</NPA>
</PP>
</NPA>
</VPC>
</CLDA>
</VPC>
<pt>.</pt>
</RC>
</L>
</textBTBV4>
```

3.2. Разпознаване на изречения с адюнкти между опората и комплемента на глаголната фраза се осъществява чрез XPath израза //VPA[child::nid]/descendant::VPC/child::DiscA.

Изразът гласи: Търси фраза от вид опора адюнкт, която има дете елемент от вид ненепосредствено доминиране (nid) и наследник фраза от вид опора комплемент с дете елемент, който причинява дистантна фразова реализация.

Пример с изречение от корпуса:

Като университетски преподавател аз мога да твърдя отговорно, че българката е много високо образована, интелигентна и можеща.

```
<textBTBV4>
<L source="random3.1.ttt">
<LC>
<CoIndex>
<identifier id="id220">
</identifier>
<identifier id="id219">
</identifier>
</CoIndex>
<PP>
<Prep>Като</Prep>
<NPA idref="id219">
<A>университетски</A>
<N>преподавател</N>
</NPA>
</PP>
<Pron idref="id219">аз</Pron>
<V>мога</V>
```

<nid idref="id220"/>
<V>
<T>**да**</T>
<V>
<pro-ss idref="id219"/>**твърдя**</V>
</V>
</LC>
<I>
<DiscA idref="id220">
<Adv>**отговорно**</Adv>
</DiscA>
</I>
<RC>
<CLCHE>
<pt>,</pt>
<C>**че**</C>
<VPS>
<N>**българката**</N>
<VPC>
<V>**е**</V>
<CoordP>
<ConjArg>
<APA>
<AdvPA>
<Adv>**много**</Adv>
<Adv>**високо**</Adv>
</AdvPA>
<Participle>**образована**</Participle></APA></ConjArg><Conj><pt>**,**</pt></Conj><ConjArg><A>**интелигентна**</A></ConjArg><Conj><C>**и**</C></Conj><ConjArg><Participle>**можеща**</Participle></ConjArg>
</CoordP>
</VPC>
</VPS>
</CLCHE>
<pt>.</pt>
<source>
</source>
<source>ba0005:p0868</source>
</RC>
</L>
</textBTBV4>

3.3. Разпознаване на изречения с адюнкти в абсолютна препозиция и абсолютна постпозиция

След като чрез първите два XPath израза от корпуса са извадени всички изречения, в които адюнктът е разположен между подлога и опората и между опората и комплемента, **върху останалата част от корпуса** е приложен следващият XPath израз, който разпознава адюнктите в началото и в края на изречението. Този израз е: //S/VPA. Изразът гласи: Търси изречение, в което има фраза от тип опора адюнкт.

Пример от корпуса за изречение с адюнкт в абсолютна препозиция:
Военна академия работи по всички стандарти.

<!DOCTYPE Extract SYSTEM "BTB.dtd">
<S h="6" source="[Root] random2.ttt">
<Discourse>
<InDiscourse></InDiscourse>
<OutDiscourse></OutDiscourse></Discourse>
<Index></Index>
<VPA h="5">
<VPS h="2">
<NPA>
<Pron>
<A>
<w >**Военна**</w></A>
<N h="2">
<>**академия**</w></N></NPA></NPA>
<w >**работи**</w></V></VPS>
<PP>
<Prep h="2">
<w>**по**</w></Prep>
<NPA>
<Pron>
<w>**всички**</w></Pron>
<N h="2">
<w>**стандарти**</w></N></NPA></PP></VPA>
<pt n="n6997">.</pt></S>

3.4. Допълнителни обработки, необходими за разпознаването на словоредните позиции на адюнктите и на вида на адюнктите

3.4.1. Маркиране на словоредната позиция на адюнкта

След извличането на изреченията с адюнкти чрез израза //S/VPA беше неоходимо да се разграничи групата на изреченията, в които адюнктът се реализира в началото на изречението, от групата на изреченията, в които той се реализира в абсолютна постпозиция. Това беше постигнато чрез допълнителна обработка на езиковия материал. Ето защо беше извършено ръчно маркиране на адюнктите със следните четири маркера:

- маркер за абсолютна препозиция
- маркер за абсолютна постпозиция
- маркер за абсолютна препозиция и за абсолютна постпозиция едновременно
- маркер за адюнкт, реализиран вътре в изречението (в случай че са останали неразпознати изречения след пускането на предишните 2 израза върху корпуса).

След маркирането на адюнктите според словоредната им позиция беше извършена екстракция от корпуса на изреченията с адюнкти в абсолютна препозиция, абсолютна постпозиция, в позиция между подлог и опора и в позиция между опора и комплемент. Като резултат се обособиха **4 групи изречения** в зависимост от местоположението на адюнкта в тях.

Оформи се следният неформален словореден модел:

{адюнкт} подлог {адюнкт} опора {адюнкт} комплемент {адюнкт}.

3.4.2. Маркиране на вида на адюнкта

След маркирането на словоредната позиция на адюнкта в оформилите се 4 групи изречения беше извършено ръчно маркиране на вида на адюнктите. Използваната от нас класификация включва следните видове: *адюнкти за време*, *за място*, *за начин*, *за количество и степен, адюнкти за вторична предикация*[5], *адюнкти за условие*, *адюнкти за причина* и *адюнкти за цел*.

4. Групи изречения според позицията на адюнкта в тях

Изречения с адюнкти в четирите словоредни позиции: 2 200 изречения.

Разпределението на изречения в четирите словоредни позиции е следното:

препозиция	*между подлог и опора*	*между опора и комплемент*	*постпозиция*
399 изр.	770 изр.	563 изр.	487 изр.

От изложените данни се вижда, че най-голям е броят на адюнктите, разположени между подлога и опората, докато в другите три позиции броят на адюнктите е приблизително еднакъв.

4.1. Изречения с разполагане на адюнкта в абсолютна препозиция

Разпределението на адюнкти в тази изреченска позиция е следното:

време – 161 изр.
начин – 96 изр.
място – 70 изр.
количество и степен – 30 изр.
причина – 18 изр.
условие – 8 изр.
цел – 8
отстъпка – 8

4.2. Изречения с разполагане на адюнкт между подлога и опората на VP

Разпределението на адюнкти в тази изреченска позиция е следното:

време – 311 изр.

начин – 258 изр.
количество и степен – 102 изр.
място – 42 изр.
вторична предикация – 21 изр.
условие – 12 изр.
причина – 10 изр.
цел – 2 изр.

4.3. Изречения с разполагане на адюнкт между опората и комплемента на VP

Разпределението на адюнкти в тази изреченска позиция е следното:
начин – 202 изр.
време – 162 изр.
количество и степен – 79 изр.
място – 70 изр.
вторична предикация – 23 изр.
цел – 11 изр.
условие – 4 изр.
причина – 1 изр.

4.4. Изречения с адюнкт в абсолютна постпозиция

Разпределението на адюнкти в тази изреченска позиция е следното:
начин – 141 изр.
време – 110 изр.
вторична предикация – 77 изр.
място – 72 изр.
цел – 31 изр.
причина – 27 изр.
количество и степен – 17 изр.
условие – 8 изр.
отстъпка – 4 изр.

5. Заключение. Изводи

В това изследване показахме един интердисциплинарен подход (комбинация между лингвистика и компютърна лингвистика) за разпознаване и извличане на изречения с адюнкти от корпус с XML описания.

Използването на този подход има несъмнени предимства пред класическите начини за извличане на информация, тъй като позволява обработката на голямо количество езикови данни по бърз и прецизен начин, като също така дава реална представа за разпределението на явлението в съвременния български книжовен език.

От извлечените данни могат да се направят следните *изводи*:

- Най-предпочитаната позиция за адюнктите е между подлога и опора на VP.
- Във всички словоредни позиции най-често (на първо и второ място

по честота на срещане) се реализират адюнктите за време и адюнктите за начин, а на 3-о и 4-о място съответно адюнктите за количество и степен и за място.
- Адюнктите за време предпочитат началото и вътрешността на изречението.
- Адюнктите за вторична предикация винаги се реализират в края на изречението.

Предимства на интердисциплинарния подход:
- Обработване на голямо количество реални примери
- Възможност за допълнителна ръчна лингвистична обработка, която позволява максимално прецизиране на резултатите
- Възможност за моментална статистическа оценка на разпределението на явлението.

БЕЛЕЖКИ / NOTES

[1] За целта е използвана най-обстойната съществуваща класификация на обстоятелствените пояснения в българския език (ГСБКЕ/GSBKE 1983), допълнена с още един вид адюнкти – за условие (Пашов/Pashov 1994). В работата си приемаме за равностойни термините *адюнкт* и *обстоятелствено пояснение*, т.е. за нас адюнктите не са аргументи на опората на глаголната фраза.

[2] http://www.tei-c.org/index.xml

[3] Повече информация за XML – на http://www.w3c.org

[4] Тагът представлява текст, който описва даден елемент в XML. Таговете се различават от данните, които описват, по това, че са оградени със скоби от типа <> (пример: <text>).

[5] В работата си използваме термина *адюнкт за вторична предикация* за означаване на сказуемните определения. Основанието сказуемните определения да бъдат определени като адюнкти е фактът, че в общия случай те не са аргументи на предиката опора и следователно синтактичният им характер е по-близък с този на адюнктите.

ЛИТЕРАТУРА / REFERENCES

Вилимовска 2017: *Вилимовска, Д.* Словоред и информационна структура на изречението (с оглед на адвербиалните модификатори, изразяващи епистемична модалност). Дисертация, ръкопис. [Vilimovska 2017: *Vilimovska, D.* Slovored i informatsionna struktura na izrechenieto (s ogled na adverbialnite modifikatori, izrazyavashti epistemichna modalnost). Disertatsia, rakopis.]

Георгиева 1974: *Георгиева, Е.* Словоред на простото изречение в българския книжовен език. София, Изд. на БАН. [Georgieva 1974: *Georgieva, E.* Slovored na prostoto izrechenie v balgarskiya knizhoven ezik. Sofia, Izd. na BAN.]

Георгиева 1987: *Георгиева, Е.* Словоред на усложненото просто изречение. София, Изд. на БАН. [Georgieva 1987: *Georgieva, E.* Slovored na uslozhnenoto prosto izrechenie. Sofia, Izd. na BAN.]

ГСБКЕ 1983: *Граматика на съвременния български книжовен език*. Том 3. Синтаксис. Под ред. на Константин Попов. Изд. на БАН, София. [GSBKE 1983:

Gramatika na savremenniya balgarski knizhoven ezik. Sintaksis. Vol. 3. Ed. Konstantin Popov. Izd. na BAN.]

Пашов 1994: *Пашов, П.* Практическа българска граматика. София, Просвета. [Pashov 1994: *Pashov, P.* Prakticheska balgarska gramatika. Sofia, Prosveta.]

Теофилова 2017: *Теофилова, С.* Словоред на обстоятелствените пояснения за време и за място и посока в разговорната реч. Филологически форум, бр. 1 (5). [Teofilova 2017: *Teofilova, S.* Slovored na obstoyatelstvenite poyasneniya za vreme i za myasto i posoka v razgovornata rech. Filologicheski forum, br. 1 (5).]

Тишева 2011: *Тишева, Й.* Устната комуникация – стилове, стандарти и регистри. Пловдив. Научни трудове, том 49, кн. 1, сб. А, 2011 – Филология. Пловдив, Университетско издателство, 86–96. [Tisheva 2011: *Tisheva, Y.* Ustnata komunikatsiya – stilove, standarti i registri. Plovdiv. Nauchni trudove, tom 49, kn. 1, sb. A, 2011 – Filologia. Plovdiv, Universitetsko izdatelstvo, 86–96.]

Тишева 2013: *Тишева, Й.* Прагматични аспекти на устната реч – Littera et Lingua Series Dissertations 3. Изд. на Факултет по славянски филологии, Софийски университет „Св, Климент Охридски“. София. [Tisheva 2013: *Tisheva, Y.* Pragmatichni aspekti na ustnata rech – Littera et Lingua Series Dissertations 3. Izd. na Fakultet po slavyanski filologii, Sofiyski universitet „Sv, Kliment Ohridski“. Sofia.]

✉ *Гл. ас. д-р Елисавета Балабанова*
Университет по библиотекознание и информационни технологии – УниБИТ
бул. „Шипченски проход“ № 69А, 1574 София, България

✉ *Chief Assist. Prof. Elisaveta Balabanova, PhD*
University of Library Studies and Information Technologies – UniBIT
bul. Shipka Pass № 69A, 1574 Sofia, Bulgaria