**За полезноста на електронските јазични корпуси**

**(врз примерот на еден тип на именската фраза во чешкиот јазик)**

*Јармила Паневова, Кирил Рибаров*

Прага

Како компјутерски лингвисти кои работат во условите што можат да се наречат лабораториски, ѝ ја оддаваме нашата длабока почит на славеничката која во скромни услови, неспоредиви со нашите, се вклучува во собирање и обработка на материјалот од македонските дијалекти, вградувајќи ја во овие истражувања - како впрочем и оние од областа на граматиката на современиот полски јазик - својата длабока лингвистичка ерудиција, учествувајќи на тој начин и во создавањето на нормата на современиот македонски јазик. Од позиција на истражувачи од компјутерска лабораторија, во нашиот јубилеен прилог сакаме да се осврнеме на аспектите што во современата лингвистика ги внесува користењето на современите компјутерски средства.

Новите компјутерски генерации, карактеризирани со параметри на брзини на обработка мерени во гигахерци и капацитети на меморија што наскоро ќе бидат изразувани во терабајти, отвораат нови перспективи за лингвистичка обработка на јазикот. Мерено преку капацитет изразен во бројот на страници, на постарите сметачи им беа потребни часови за обработка на само неколку стотини страници, додека денешното време нуди речиси неверојатен контраст со претходно опишаната состојба: обработка на податоци од милиони страници на текст во делови од микросекунди, експоненцијално растечки капацитет, присутен во целосветските компјутерски мрежи, со едноставен пристап. И не само тоа: дури од поголемо значење е тоа што бележиме и присутност на голем број софтверски алатки со кои му се овозможува не само на

информатичарот пристап кон електронските димензии и со кои пребарувањето на милиони единици нема да трае подолго од неколку секунди.

Нашите денешни лингвистички погодности, кога ограничувањата на компјутерското време и простор се предимно од релативен карактер, ги согледуваме пред сè во можноста на *постојано* и *повеќекратно* пребарување на јазичен материјал достапен во електронска форма. Ако на споменатиот јазичен материјал му ги доделиме и атрибутите на внатрешна структурираност, унифицираност (често и индексираност), на неговата зборовната исцрпност, вообичаено присутна во текстуална форма, и кога се организира така да служи за одредена цел во однос на која се поставува како репрезентативен, зборуваме за јазичен корпус (в. на пр. Čermák, 1998). Вака карактеризираната структура ни овозможува, меѓу другото, постојано зголемување на електронскиот фонд на зборови кој сакаме да го пребаруваме, и (алтернативно) неговата лексичка и граматичка структура да ја збогатуваме со додавање на информации и релации поврзани со секоја зборовна и/или реченичка и/или текстуална појава.

Патот до една голема јазична колекција во внатрешно подредена и унифицирана електронска форма не е директен; подразбира постоење на текстови во електронска форма (во спротивен случај претходи нивно дигитализирање), усогласување на нивното кодирање, чистење на истите од најразлични форматирачки карактеристики (поврзани со нивното визуално презентирање а не со нивната содржина) при едновременото задржување на компактност, нивното меморирање, како и нивното (вклучувајќи автоматско) сегментирање на текстуално релевантни сегменти како, на пример, зборови, реченици, пасуси, документи.

Пристапот кон ваквиот материјал (корпус) е овозможен со помош на т.н. корпусни менаџери, одн. софтвер кој овозможува задавање на пребарувачки критериуми и структурирано презентирање на пребаруваната информација.

Гореспоменатата структура е карактеристична за основен јазичен корпус во кој може да се пребарува по зборформи и по нивната контекстуална околина[1]. Ако сакаме, на пример, да пребаруваме зборови членувани со -та, немаме друг избор туку да ги бараме сите зборови што завршуваат на „та“ (вклучувајќи ги и оние што очигледно не се членувани, но завршуваат со „та“). Немаме ни можност да правиме разлика меѓу именки, броеви или придавки (или други зборовни групи), сè додека информацијата за зборовните групи не ја внесеме и асоцираме со секој збор во јазичниот корпус (во овој случај станува збор за морфолошка анотација на јазичниот корпус која подразбира еднозначно определување на зборовната група и нејзините категории)[2]. Друг тип на збогатување на корпусот претставува, на пример, доделување на лема кон секоја зборформа - неопходно за успешно пребарување во јазици со богата флексија каков што е, на пример, чешкиот.

Квалитативно различни анотации добиваме ако започнеме со анотирање на контекстите, т.е. означување на релации меѓу зборовите во и/или надвор од речениците. Додавањето на површинската синтаксичка информација е само еден таков пример. Реченици збогатени со синтаксички информации формираат т.н. синтаксички корпуси[3], а во зависност од имплементираниот граматички формализам ни се отвораат

[1] Критериумите за пребарување можат да бидат најразчлични, но секојпат во рамките на регуларен израз со кој се дефинираат.

[2] Морфолошката анотација може да биде рачна или автоматска. Рачната се дефинира како единствена исправна анотација и истата служи за тренирање на алгоритми кои се во состојба релативно успешно морфолошки да анотираат. На морфолошката анотација ѝ претходи чекор на т.н. морфолошка анализа која има за цел кон секој збор да ги асоцира сите негови возможни (контекстуално независни) морфолошки особености. Во морфолошката анализа на чешкиот јазик вршиме околу 3500 морфолошки дистинкции.

[3] Англ. Treebank; соодветно на морфолошката анотација, и во овој случај постои рачна и автоматска синтаксичка анотација (анализа).

сосема нови можности на автоматски пребарувања и структурирања на бараната информација.

Како што понатака ќе биде покажано врз примерот на еден тип на именски фрази во чешкиот јазик, корпусите и нивните софтверски алатки имаат и свои недостатоци, но пред сè предности; компјутерските средства за обработка на јазикот започнуваат непосредно да бидат корисни, како за современи, така и за класични лингвистички цели.

3. Тополињска проучувала бројни прашања од областа на синтаксата и семантиката на словенските јазици. Но со нејзиното име неразделно се поврзува нејзиниот исцрпен и сестран поглед врз именските фрази - понатака NP според англ. Noun Phrase; сп. Тополињска (1981). Уверени сме дека славеничката нема да ни замери дека во овој текст нема да стане збор за суптилните разлики поврзани со референцијата, квантификацијата и со останатите аспекти што во описот на NP ги има воведено. Во овој текст ќе се усредоточиме само на синтаксата на еден тип на именски фрази кои Тополињска (1981) ги анализира во одделот *Determiners expressing predicates whose primary exponents are verbs* (стр. 108 и сл.), каде изнесува мислење дека *„expanded principal constructions **usually** stand to the right of the CM (constitutive member) of the NP“* (стр. 111, болдирано од страна на авторите на статијава). Мораме да констатираме дека во современиот чешки јазик овој „вообичаен“ редослед релативно често се нарушува што, следователно, претставува извор на синтаксички повеќезначности, в. ја разликата меѓу примерите (1) и (2). Покрај тоа, овој тип на синтаксичка хомонимија, може да се појави и кај препозитивното проширување на придавката која не спаѓа меѓу продуктивните одглаголски деривати од чешките партиципи.

(1) Dívka rovná **ve výloze vystavený** kabát.

(2) Dívka si koupila **ve výloze vystavený** kabát.

(3) Ladronka je hotelem, ve kterém **bydlí** (**v Čechách** zatím téměř **neznámá** skupina squaterů).

Од синтактичка гледна точка реченициte (1) - (3) се подеднакво повеќезначни: предлошкиот падеж *ve výloze* во (1) и (2) е или (адвербална) прилошка определба, или е во состав на NP и го развива адјективизираниот партицип. Во (1) присутноста на две различни синтактички структури тешко може да влиае врз условите на вистинитоста на реченицата (највероjатно станува збор за аранжерка која го аранжира палтото во излогот, и истото истовремено се наоѓа во излогот). Во примерот (2) една од структурите, од прагматична гледна точка малку веројатна, ја елиминираме само со преместување на целиот детерминатор во постпозиција. Иако во (3), синтаксичката интерпретација *bydlí v Čechách* не е ни формално ни содржински блокирана, сепак од контекстот е јасно дека *v Čechách* е препозитивно проширување на (NP) *skupina squaterů*. Во овој, и во многу други случаи, со промената на збороредот се постигнува синтаксичка еднозначност, а уверени сме, и подобрување на стилските квалитети на реченицата.

Реченицата (4) е пример за проблематична разбирливост на една реченица, што може барем делумно да се елиминира со преместување на проширениот дел од NP од препозиција во постпозиција.

(4) Když přičteme **delegáty posvěcené paragrafy** ve stanovách umožňující soudu vyloučit oponenty, ... → Když přičteme **paragrafy ve stanovách posvěcené delegáty** umožňující soudu vyloučit oponenty, ...

Во електронските курпуси на чешкиот јазик ќе се обидеме да побараме конструкции со наведениот, не многу вообичаен, збороред каде самиот именски детерминатор е препозитивно проширен со именска фраза. Потоа ќе пристапиме кон

проверка на нашата хипотеза, имено дека со преместување на детерминаторот во повообичаената постпозитивна позиција во пишуваната форма на јазикот добиваме поеднозначна и стилски поиздржана реченица. Доколку ова преместување воопшто има некакво влијание врз реченичната перспектива, станува збор за ситни разлики во чии подетални анализи не навлегуваме.

Најпрво, служејќи се со Чешкиот национален корпус (ЧНК), ќе формулираме пребарување чиј резултат ќе биде во вид низа на збороформи. За локализирање на конструкциите кои се во нашиот случај од интерес, прашањето го задаваме во наредниот формат:

```
[tag="V.*"] [tag="R.*"] []{0,2} [tag="N.*"] [tag="A.*"] [tag="N.*"]
```

Ова прашање[4] ги содржи следните определби: непосредно зад глаголот (V) се наоѓа предлог (R) по кој, во оддалеченост од највеќе 2 позиции (збора) десно, се наоѓа именка (N) проследена од придавка (A) и именка (се претпоставува дека последниот член одговара на СМ). Со тоа е дефиниран критериумот според кој ќе се пребарува насекаде во корпусот. Контекстот за бараната низа е дефиниран со должина од 1 реченица пред бараната појава и 1 реченица по истата - конкорданса (I). Корпусниот манаџер не овозможува едноставно, во рамките на само едно прашање, да се тестира конгруенцијата меѓу последната именка N (СМ) и придавката која директно претходи, нити тестирање на конгруенцијата меѓу барањата на рекцијата на предлогот и именката која се наоѓа дури за две позиции десно од неа. Значи, мора да земе предвид дека во резултатите од пребарувањето нема да постојат само примери што се од наш интерес. Резултатите ги добиваме на мониторот како стрингови од определена должина каде

[4] За помошта при формулирањето на критериумите за пребарување во ЧНК а понатака во текстот и во Прашкиот синтаксички корпус, и за пребарување во истите, им се заблагодаруваме на м-р Вероника Ржезничкова и на д-р Роман Ондрушка. Едновремено ги користиме работните услови создадени благодарение на Центарот за компјутерска лингвистика при Факултетот за математика и физика при Карловиот Универзитет во Прага.

меѓу заградите „<“ и „>“ е оној дел од примерот кој одговара на критериумот за пребарувањето. Секој пример е обележан со реден број од конкордансата.

На овој начин од ЧНК (околу 100 мил. зборформи) добивме конкорданса од 212 264 појави. Подолу наведените реченици (5) и (6) се примери каде пребарувачкиот критериум е важечки, но излезните примери не се од наш интерес.

(5) [12] Pak <vytáhl z kapsy hubertusu kožený řemínek> a přivázal jej…

(6) [19] Když dojeli k dělnicím, které již, všechno sbalené, auto netrpělivě vyhlížely, <udělal s nimi Vítězslav krátkou poradu>

Реченицата (5) при поинаков подбор на лексемите, ќе одговараше на бараниот критериум, но не и реченицата (6).

За исклучување на реченициte од типот (6), критериумите на пребарување треба да се дополнат со специфични информации за зборовите и нивните форми што можат, одн. не можат да стојат меѓу предлошкиот израз и придавката. Ако сакаме да ги редуцираме редундантните примери, ја користиме најновата верзија на корпусот (SYN2000B) во кој анотацијата овозможува пребарување само на девербативни придавки. Во овој случај прашањето е во следниот формат:

```
[tag="V.*"] [tag="R.*"] []{0,2} [tag="N.*"] [(tag="A.*") &
                (lemma=".*t[)]")] [tag="N.*"]
```

Во оваа конкорданса (II) ги губиме примерите од типот *Slunce* ***viděli na okraji temně rudé****, uprostřed zářivě červené,* но едновремено се враќаме кон типот NP со партиципското проширување на СМ, анализирано во 3.2.2 кај Тополињска (1981). Бројот на откриените примери овој пат е намален на 21 316, што изнесува 10% од претходно пронајдените низи. Но и во овој случај, во конкордансата (II) која претставува подмножество на конкордансата (I), се наоѓаат многу конструкции што не се од наш интерес, како на пример (7) и (8).

(7) [19] Prozradil jsem jim, že se <chystám vypravit do Země splněných přání>

(8) [34] …začal <přecházet po místnosti přihrbeným způsobem> šelmy lapené v kleci

Од 430 случајно одбрани примери (односно од приближно 20%) од конкордансата (II), 15 конструкции се покажаа како двозначни и по употребата на семантички и прагматички критериуми (примери (9) и (10) - група (А)), 35 примери го потврдија бараното препозитивно проширување на СМ, додека пак во 16 од нив постои несогласување меѓу падежот управуван од предлогот (група (Б), примери (11), (12)). При понатамошното прецизирање на критериумот појавите од типот (Б) би можеле да се пребаруваат посебно како најсомнителни примери во даденото множество на појави.

(9) [6] … zaplesal jesenický hajný, když <uviděl v regálech naskládané hrnce>

(10) [85] … nevelkou místnost <dělil na dvě poloviny rozedraný závěs>

(11) [7] … postavy v uniformách <zapadaly do sluncem rozžhaveného města>

(12) [93] Úzká pěšina nás <zavedla na štěrkem vysypané prostranství>

(13) [130] A pak, jako by ze země vyrostl, <objevil se stářím shrbený pavián>

Реченицте од типот како во (13) не можат да бидат елиминирани ниту со тестирање на конгруенцијата меѓу предлогот и именката (општо, формата *se* се анотира и како рефлексивна заменка и како предлог - дезамбигуацијата досега не е извршена во користената верзија на корпусот). Ако извршиме преместување на адјективното проширување од препозицијата во постпозиција (зад СМ), во групите (А) и (Б) во 43 случаи од 50 добиваме стилски подобри реченици, без структурна повеќезначност. Во 6 случаи станува збор за фразеологизирана употреба, што ја оправдува вообичаеноста на препозицијата (сп. (14), (15), (16)). Во ретки случаи (в. (17)) испаѓа дека постпозицијата е „подобро решение“, но со зборoредот N (СМ) - (N - Adj).

(14) [69]… vrah <je do nebe volající packal>

(15) [72] Muž <měl do hněda opálený hrudník>

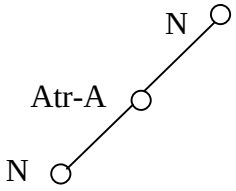
(16) [398] … před poštou <stála po zuby ozbrojená četa>

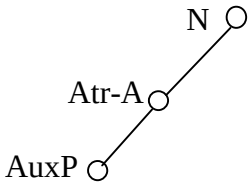
(17)[94] Ve Wolfově tónu se <ozvala s obtížemi zachovávaná trpělivost> → … se ozvala trpělivost s obtížemi zachovávaná

Прашкиот синтаксички корпус[5] (ПСК) во кој реченицитe се анотирани на т.н. аналитичко ниво[6] (в. на пр. Hajič, 1998), ни овозможува да пребаруваме поконкретно, одн. само на пр. NP, а дури само NP од одреден тип. Софтверската алатка NETGRAPH овозможува внесување на пребарувачки критериум според кој можеме да локализираме поддрво со јазол N (во произволна функција) на кој зависи придавка во функција на атрибут (Atr), каде атрибутот е развиен. Недостаток на овој софтвер е дека тој, во сегашниот стадиум од неговиот развој, нема можност за задавање на негативни услови (поради што не можеме однапред да ги исклучиме конструкциите како *silně znečištěné ovzduší*), но позитивниот услов можеме да го ограничиме на конструкции каде придавката е развиена со именка (со тоа се елиминира конструкцијата како *pět Oscarů získavší snímek*). По извршување на пребарувањето, на мониторот се појавува ситаксичко дрво на целата реченица со обоен јазол на бараната потструктура.

Потструктурите, што сакаме да ги побараме во овој корпус, ги дефинираме со пребарувачките критериуми (III) и (IV), т.е. бараме NP чиј корен, изразен со именка N (СМ) е проширен со Atr(ibut) изразен со придавка (А), на која зависи именка (N) (N во прост - III, или предлошки падеж - IV). Како што кажавме погоре, алатката NETGRAPH ни овозможува откривање на проширувањата на СМ, како од лево така и од десно. Со критериумот III добиваме 2077 единици (безпредлошки проширувања на придавката - група (В)), додека со критериумот IV добиваме 2701 единици (N е зависно на А преку предлогот AuxP - група (Г)).

[5] Составен дел од ЧНК.
[6] Површинска синтакса од зависносен тип.

(III) 

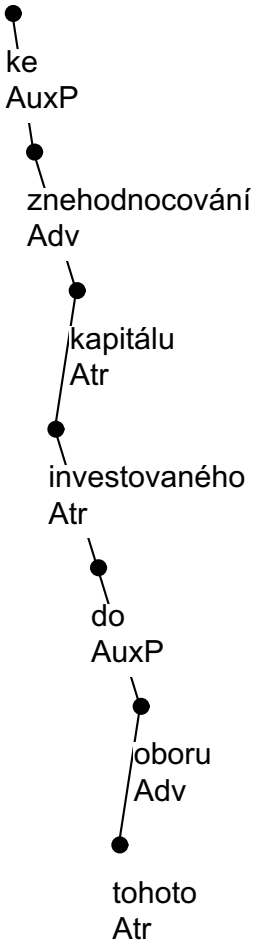

(IV) 

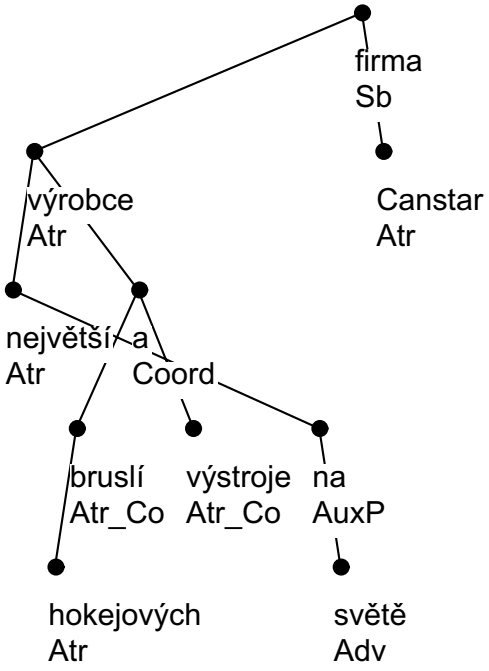

Резултатите од пребарувањето на синтактсичките структури во ПСК се поефективни, па добиваме и конструкции што содржат и т.н. „долги“ (непроективни) зависности (сп. (18), (19)). Во овој случај не моравме да се ограничиме само на пребарување на придавки изведени од партиципи, како што покажува примерот (19), туку имавме можност да ја проучуваме и конкуренцијата меѓу пре- и постпозитивното проширување.

(18) …ceny paliv vedou ke znehodnocování <investovaného kapitálu do tohoto oboru> *(ПСК:ca35am.fs:#32)*

(19) …ve městě sídlí <největší výrobce hokejových bruslí a výstroje na světě> firma Canstar *(ПСК:cb03am.fs:#2)*

Поради ограниченият простор не можеме да пристапиме кон подетална анализа на примерите во групите (А) - (Г) (особено (В) и (Г)). Затоа примерите (18) и (19) ги наведуваме заедно со нивната зависносна структура, и ги оставаме без лингвистички коментар.




Поддрво 1: Кон (18)

Поддрво 2: Кон(19)

Заклучуваме дека нашата примарна цел беше да се демонстрираат можностите што им се отвораат на лингвистите при емпириските истражувања, ако имаат можност да се служат со електронски корпуси и алатки, со помош на кои истите можат најразлично да се пребаруваат. Покрај тоа се обидовме да провериме одредени хипотези што се однесуваат на структурата на проширената номинална фраза во современиот пишуван чешки јазик: (а) проширувањето на придавките образувани од партиципи (но и други видови) со именка (во прост или предлошки падеж) се наоѓа исто така во препозиција; (б) конструкциите според (а) се синтаксички повеќезначни; (в) препозитивното проширување може речиси секогаш (без суштинска промена на

значењето) да биде преместено во постпозиција; (г) речениците што произлегуваат од преместувањето според (в) се поразбирливи, а најчесто, и стилистички поиздржани.

**Литература**

Čermák F. 1998. Český nároní korpus. Во: *Македонско-чешка научна конференција*. Филолошки факултет „Блаже Конески“, Универрзитет „Св. Кирил и Методиј“. Скопје. 41-52.

Český národní korpus. *Úvod a příručka uživatele*. 2000. Ред.: J. Kocek, M. Kopřivová, K. Kučera. ČNK FF UK. Praha.

Hajič J. 1998. Building a syntacticaly annotated corpus: The Prague Dependency Treebank. Во: *Issues of Valency and Meaning - Studies in Honour of Jarmila Panevová*. Ред.: E. Hajičová. Karolinum, Charles University Press. Прага. 106-132.

Topolińska Z. 1981. *Remarks on the Slavic Noun Phrase*. Ossolineum. Wrocław - Warszawa - Kraków - Gdańsk - Łodż.