Grosjean, F.: Life with Two Languages. An introduction to bilingualism. Cambridge Massachusetts, Harvard University Press, 1982.

Hammerová, L. B. – Ripka, I.: Jazykové prejavy amerických Slovákov. Bratislava: VEDA, 1994. ISBN 80-224-0151-X

Michalčák, L.: Kolonizácia Slovákov v Rumunsku. In: Slováci v Rumunsku. Zborník materiálov z vedeckej konferencie 7. júna 1995. Bratislava: Dom zahraničných Slovákov, 1995, s. 22-30.

Štefanko, O.: Slovenské osídlenie v Rumunsku. In: *Atlas ľudovej kultúry Slovákov v Rumunsku.* Zodp. red. Ondrej Štefanko. Nadlak: Vydavateľstvo kultúrnej a vedeckej spoločnosti Ivana Krasku, 1998, s. 15-21. ISBN 973-9292-35-6

Weinreich, U.: Languages in Contact. The Hague: Mouton, 1953.

Winford, D.: An introduction to contact linguistics. Blackwell, 2003.

Syntaktická anotácia vybraných textov Slovenského národného korpusu

Katarína Gajdošová

Jazykovedný ústav Ľ. Štúra SAV, Bratislava

Syntax, jedna z najdôležitejších disciplín lingvistiky, má svoje zastúpenie aj v počítačovom spracovaní prirodzeného jazyka. S počítačovým spracovaním slovenského jazyka sa začalo intenzívne až v roku 2002, keď bolo v Jazykovednom ústave Ľudovíta Štúra SAV v Bratislave utvorené oddelenie Slovenského národného korpusu (ďalej SNK), ktoré sa systematicky venuje počítačovému spracovávaniu prirodzeného jazyka (http://korpus.juls.savba.sk).

Syntaktickú anotáciu zahŕňame spolu s morfologickou anotáciou do internej anotácie textov v korpuse. Štýlovo-žánrová a bibliografická anotácia textov sú externými anotáciami SNK (bližšie Forróová – Horák, 2004).

Práca na ručnej syntaktickej anotácii vybraných textov Slovenského národného korpusu sa začala v polovici roka 2005. Začiatku prác predchádzali teoretické prednášky i praktické kurzy pracovníkov z partnerského pracoviska Ústavu formální a aplikované lingvistiky Matematicko-fyzikální fakulty Univerzity Karlovy v Prahe (ďalej ÚFaL). Na syntaktickej anotácii vybraných textov SNK začal pracovať 6-členný tím študentov, ktorého členovia sa priebežne obmieňali. Teoretické východiská, resp. vedomosti anotátorov zodpovedajú jednotlivým lingvistickým „školám“ na Slovensku (Trnava – prof. Kačala, Prešov – prof. Nižníková). Na konci roku 2006 mal anotátorský kolektív 4 stále členky, jednu externú spolupracovníčku a za sebou 34 613 viet rôznych štýlov a žánrov, pričom každý text je syntakticky anotovaný dvakrát, dvoma rôznymi anotátormi (bližšie Šimková – Garabík, 2006). Celkový počet zanotovaných

viet je teda dvojnásobný – takmer 70 tisíc viet. Jadro anotovaných dát tvoria nasledujúce texty:

- **beletria** (román Georgea Orwella 1984, román Ladislava Balleka Pomocník, preklady rozprávok, mládežnícky román)
- **odborné texty** (historická monografia, lingvisticko-spoločenské štúdie)
- **populárno-náučný štýl** (texty z internetového portálu <http://sk.wikipedia.org/wiki/>
- **publicistika** (texty z denníka SME a internetového časopisu InZine)

Teoretické východiská syntaktickej anotácie SNK predstavujú práce:

Hajič, Jan – Panevová, Jarmila – Buráňová, Eva – Urešová, Zdeňka – Bémová, Alla – Kárník, Jiří – Štěpánek, Jan – Pajas, Petr: Anotace na analytické rovině. Návod pro anotátory. Praha: ÚFaL MFF UK 1999. 214 s.

Nižníková, Jolana – Sokolová, Miloslava: Valenčný slovník slovenčiny. Prešov: Filozofická fakulta v Prešove, Prešovská univerzita 1998. 270 s.

Oravec, Ján – Bajzíková, Eugénia: Súčasný slovenský spisovný jazyk. Syntax. Bratislava: Slovenské pedagogické nakladateľstvo 1982. 261 s.

Pri pohľade na syntaktickú anotáciu SNK je potrebné uvedomiť si, že anotácia SNK je inšpirovaná pražskou závislostnou syntaxou a projektom PDT 1.0 – analytickou rovinou českého závislostného korpusu. Teoretickým východiskom PDT 1.0 bola Šmilauerova Novočeská skladba, z ktorej boli niektoré časti prispôsobené počítačovému spracovaniu, teda spôsobu vešania jednotlivých syntaktických závislostí vo vete. Rovnako sa autori opierali o valenčnú teóriu profesorky Jarmily Panevovej.

Základnou príručkou pri anotovaní SNK bol manuál *Anotace na analytické rovině. Návod pro anotátory.* Dôvodov na zvolenie smerovania v línii pražskej závislostnej syntaxe bolo viacero:

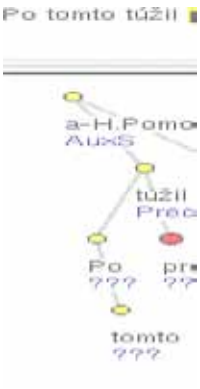
- blízkosť syntaktickej štruktúry oboch jazykov;
- blízkosť teoretických východísk na analytickej rovine;
- bohaté skúsenosti českých kolegov z niekoľkoročného anotovania;
- vynikajúco zvládnutá technická stránka spracovania syntaktickej roviny jazyka.

Rovnako však i ústretovosť a ochota pomôcť zo strany lingvistického i programátorského kolektívu ÚFaL. V čase, keď sa syntaktická anotácia SNK začínala, prebiehala v ÚFaLe už tretia, nadstavbová anotácia lingvistických dát – anotácia tzv. tektogramatickej roviny. Z toho vyplýva, že mnohé cenné skúsenosti, ale i omyly pražských lingvistov pri anotácii na analytickej rovine mohli byť naplno využité pri syntaktickom anotovaní SNK.

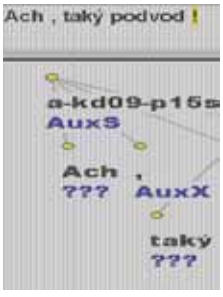
Hoci východiskom syntaktickej anotácie SNK bola aj súčasná syntaktická teória slovenčiny, v niektorých javoch a spôsoboch zachytenia jednotlivých syntaktických vzťahov vo vete museli zaužívané konvencie v jazyku ustúpiť počítačovému spracovaniu vetných štruktúr. Z konkrétnych prípadov osobitos-

tí spracovania niektorých syntagiem pri anotovaní uvedieme vzhľadom na rozsah príspevku a šírku teoretického výkladu len niektoré, sumár odlišností bude zachytený v pripravovanom slovenskom syntaktickom manuáli. (Zoznam používaných funktorov v syntaktickej anotácii vybraných textov SNK uvádzame v prílohe.)

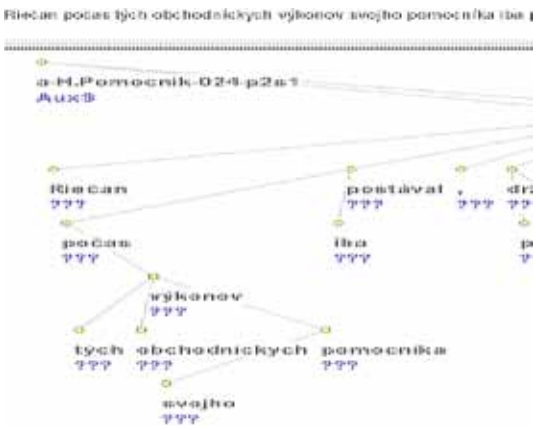
Vrcholom každej syntaktickej štruktúry (ďalej strom) je hlavný syntaktický uzol vety, v pozadí označený symbolom # – ide o technické označenie vety. V ňom sa uvádzajú informácie o názve súboru/textu, z ktorého veta pochádza, a o poradí vetnej konštrukcie v súbore i v celom texte. Technický vrchol vety má automaticky priradenú imaginárnu funkciu AuxS (koreň stromu).
V ukážke vidíme nezanotovanú vetu, ako sa zobrazí anotátorovi. V hornej časti vidí vetu reálneho textu lineárne usporiadanú, v dolnej časti sú jednotlivé uzly zavesené na seba jeden po druhom bez špecifikácie vetnočlenskej funkcie. Anotátor musí vetu „rozvešať“ a jednotlivým vetným členom priradiť zodpovedajúce funktory.



Na hlavnom uzle visí predikát vety. Pri počítačovom spracovaní syntaktických štruktúr je potrebný jeden hlavný uzol a predikát vety považovaný za nadradený vetný člen, ktorému sú podradené na ňom závislé vetné členy, teda aj subjekt vety. Toto poňatie primárnosti predikátu vo vete je jedným z rozdielov medzi tradičnou syntaxou a počítačovým spracovaním syntaktickej roviny jazyka. Rovnako visí na hlavnom uzle vety i koncová interpunkcia, ktorá má podobne ako tento uzol automaticky priradenú imaginárnu syntaktickú funkciu AuxK (koniec vety).

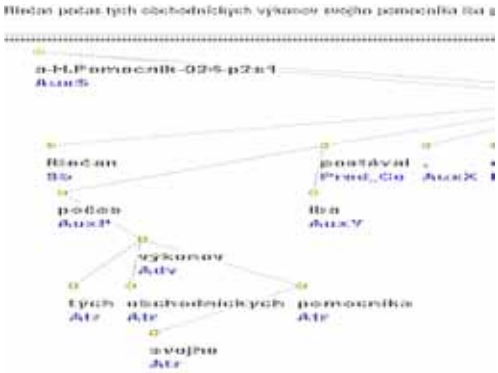


V prípade vety bez predikátu visia všetky nadradené vetné členy na hlavnom uzle.



Ak je anotovanou syntaktickou štruktúrou súvetie, visí na koreni vety koordinačná spojka alebo čiarka celého súvetia. V prípade spojenia troch a viacerých viet – uvedený príklad – visí na koreni vety posledná koordinačná spojka alebo čiarka.

Po vyhľadaní a zavesení predikátu vety dochádza k identifikovaniu jeho obligatórnych a fakultatívnych doplnení, ako aj k priradeniu ostatných rozvíjacích vetných členov k nadradeným vetným členom.



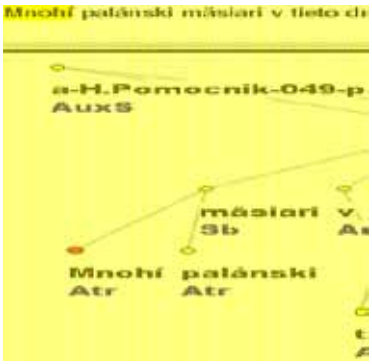
Po analýze syntaktickej konštrukcie nasleduje priradenie adekvátnych funkcií jednotlivým uzlom – označeniu vetných členov vo vete. Na analytickej rovine sa funktor priraďuje každému slovu vo vete, aj synsyntaktikám a interpunkcii.

Ako sme už vyššie spomenuli, nie vždy je funkcia či metóda počítačového spracovania vetných členov vo vete totožná so spôsobom analýzy tých istých vetných členov z pohľadu klasickej štruktúrnej syntaxe. Niektoré rozdiely, skôr technického charakteru, sme už uviedli, v nasledujúcej časti poukážeme na základné rozdiely v hodnotení jednotlivých vetných členov. Ku každému rozdielu uvádzame príklad.

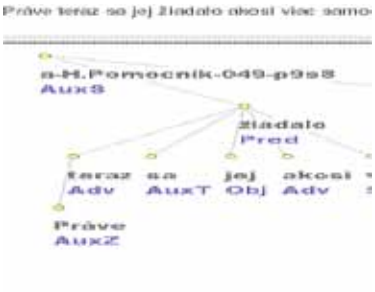
Klasická slovenská syntax vyčleňuje viacero neplnovýznamových, pomocných slovies, ktoré sú súčasťou slovesno-menného prísudku. Počítačové spracovanie hodnotí ako jediné pomocné sloveso sponové sloveso *byť*. Ostatné pomocné slovesá sú chápané ako plnohodnotné predikáty vety. Slovesá, ktoré stoja v tradičnej

syntaxi za pomocným slovesom a hodnotia sa ako časť zloženého prísudku, sú v počítačovom spracovaní považované za objekty.

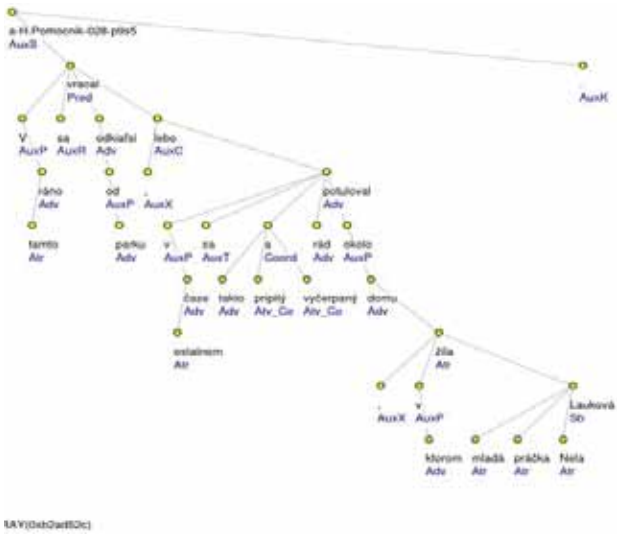
Na uvedenom príklade možno vidieť, že počítačové spracovanie syntaktických štruktúr sa opiera i o kontextové začlenenie viet do celku. Ak by sme mali prístup k celému textu, zistili by sme, že veta nadväzuje na predchádzajúcu výpoveď, a preto je zaznačená cez koordinačný uzol ako pokračovanie predchádzajúcej výpovede, teda súčasť súvetia. Tento fakt sa odráža na slovese vety – je mu priradená funkcia Pred_Co, čiže predikát v koordinácii.



Ďalším rozdielom medzi klasickou slovenskou syntaxou a počítačovým spracovaním syntaktických štruktúr sú postupne rozvité prívlastky. Sémantické vzťahy medzi jednotlivými adjektívnymi tvarmi sú vyjadrené nadradenosťou, resp. podradenosťou – čiže umiestnením (zavesením) podradených atribútov jedného na druhý. Počítačové spracovanie, vzhľadom na náročnosť segmentácie, volí jednoduchú možnosť samostatného zavesenia každého z postupne rozvitých atribútov na rozvíjané substantívum.



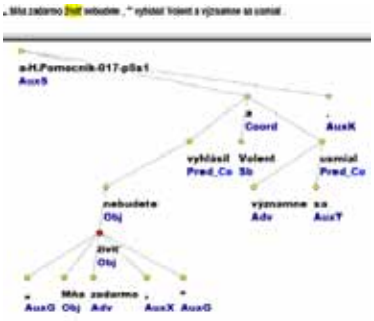
Rovnakú funkciu atribútu má člen recipročnej syntagmy. Tu možno najzreteľnejšie vidieť rozdiel v spôsobe zavesenia vetného člena. Ak by nebolo počítačového pravidla, ktorému je potrebné vyjsť v tomto spôsobe spracúvania dát v ústrety, bolo by znázornené umiestnenie presne opačné.



Klasická slovenská, a nielen slovenská syntax má rozpracovanú teóriu o doplnku ako o vetnom člene viažucom sa naraz na dva vetné členy vo vete. Buď na subjekt a predikát vety, vtedy hovoríme o subjektovom doplnku, alebo na objekt a predikát a vtedy hovoríme o objektovom doplnku. Oba vzťahy (subjekt – doplnok a predikát – doplnok, resp. objekt – doplnok a predikát – doplnok) sú rovnocenné, teda doplnok má plnohodnotný vzťah k predikátu i subjektu či objektu.

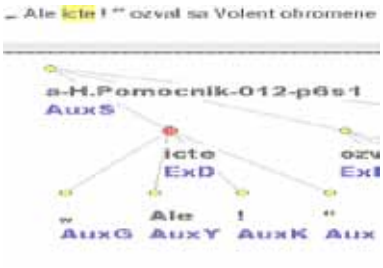
Pri syntaktickej anotácii je zachytenie tohto vzťahu doplnku odlišné. Ak je vo vete vyjadrený subjekt, na ktorý sa doplnok viaže, je doplnok priradený k nemu a dostáva označenie Atv. Pokiaľ subjekt nie je vyjadrený, doplnok je priradený k predikátu a dostáva funkciu AtvV. V prípade objektového doplnku je doplnok priradený k objektu, ku ktorému sa vzťahuje, s funkciou Atv.

Na uvedenom príklade možno vidieť aj špecifický spôsob značenia vedľajších viet v súvetí. Predikát vedľajšej vety je označený funkciou vetného člena, ktorý táto veta vyjadruje. V uvedenom prípade je predikát vedľajšej vety prívlastkovej označený funktorom Atr, teda funkciou prívlastku.

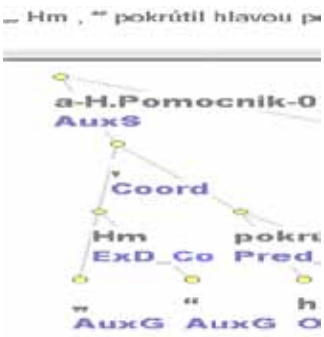


Medzi jeden zo základných rozdielov štruktúrnej a počítačovej syntaktickej analýzy patrí aj pohľad na priamu reč. Priama reč, ktorá nasleduje po slovesách typu *myslieť, hovoriť, povedať, vravieť*..., je označovaná ako objekt.

Na ďalšie typy priamej reči existujú dva rôzne spôsoby zavesenia:



V ilustračnom príklade priamej reči nie je medzi uvádzacou vetou a priamou rečou žiaden formálny vzťah. Predikáty oboch viet dostanú funktor ExD (elipsa) a budú zavesené na technickom hlavnom uzle vety. Všetky ostatné vetné členy budú určené tak, akoby obe slovesá mali funkciu predikátu.



V uvedenom prípade je možné do uvádzacej vety doplniť sloveso „*povedal*“. Preto je možné zavesiť celú koordináciu tak, ako to vidno na príklade. Ak by priama reč obsahovala predikát, bol by členom koordinácie a priradili by sme mu funktor Pred_Co. Vzhľadom na to, že v uvedenom prípade je priama reč vyjadrená prostredníctvom interjekcie, priradíme jej funkciu ExD_Co (účasť v koordinácii).

V súčasnosti prebieha spolu s anotáciou nových textov i prvá fáza opráv ručne syntakticky anotovaných textov. Keďže ide o veľké množstvá textov, nie je možné ich ani efektívne skontrolovať ručne, a to by nakoniec ani neplnilo svoj účel, keďže počítačové spracovanie prirodzeného jazyka, v našom prípade syntaktickej roviny jazyka, je späté s využitím počítačových nástrojov nielen pri samotnej anotácii textov, ale aj pri opravách. Opravy syntaktickej anotácie majú niekoľko fáz, ktoré môžeme pracovne vyčleniť takto:

Fáza odstránenia nezanotovaných vrcholov, označených predanotovaným symbolom ???. Pri opravách syntaktických anotácií sa zistilo, že z celkového počtu 1 070 099 vrcholov je syntakticky neoznačených 7922 vrcholov, teda 0,74% vrcholov. Tieto chyby, ktoré vznikli z nepozornosti, bolo nevyhnutné odstrániť ako prvé. Je pochopiteľné, že tento druh opráv musel byť vykonaný čiastočne ručne, keďže počítačový nástroj je síce schopný automaticky pridelený znak ??? nájsť, ale jeho označenie nedokáže vykonať.

Ďalšou fázou opráv anotovaných textov je **využitie morfologickej anotácie SNK** pri opravách syntaktickej roviny anotácií. Na základe morfologických informácií v pozadí syntaktickej anotácie dokážeme identifikovať chybné funktory v syntagmách. Napr. slovo označené morfologickým tagom T nemôže mať v žiadnom prípade na syntaktickej úrovni žiaden z autosyntaktických funktorov (Pred, Obj, Atr...).

Po tejto úvodnej, zdĺhavej fáze nasleduje porovnanie textu anotovaného dvoma anotátormi, pri ktorom sa zameriavame na rozdielnosti v anotovanom texte a rovnako aj na tzv. **fázu počítačového odstránenia chybných funktorov**, ktorých umiestnenie v syntaktickej štruktúre nie je relevantné z hľadiska syntaktického kontextu funktora. Do tejto skupiny opráv zaraďujeme napríklad:

Skorigovanie funktorov, ktoré majú označenie _ERR, teda chyby pri označovaní vrcholov koordinácie a apozície, ktoré vznikli z nepozornosti alebo pri preklápaní textov zo starého formátu .fs do nového formátu .pml.

Odstránenie funtora Atr, ktorý je zavesený na akomkoľvek inom slovnom druhu ako je nadradené substantívum, pronominum, numerálium.

Odstránenie funktora Obj, ktorý je umiestnený na inom slovnom druhu ako je verbum alebo isté typy adjetív.

Doplnenie, resp. spresnenie umiestnenia interpunkcie. Ide o chybné zavesenie úvodzoviek, pomlčiek, troch bodiek (či už na konci vety alebo v priebehu výpovede), rovnako i prípadné opravenie funktorov na týchto pozíciách, ktoré je spôsobené zmenou umiestnenia v syntaktickej štruktúre.

Rozsiahlejšie formy opráv by mali prispieť k skvalitneniu automatizovanej syntaktickej anotácie textov tak, aby boli počítačové nástroje schopné do istej miery sami syntakticky rozanalyzovať text, ktorý im bude predložený. Na základe ručne anotovaných dát bude možné v budúcnosti počítať s poloautomatizovanou-poloručnou syntaktickou anotáciou textov SNK a so sprístupnením Slovenského závislostného korpusu na lingvistické výskumy.

Literatúra

FORRÓOVÁ, Martina – HORÁK, Alexander: Morfologická anotácia korpusu. In: Slovenčina na začiatku 21. storočia. Ed. Mária Imrichová. Prešov: Prešovská univerzita, Fakulta humanitných a prírodných vied 2004, s. 174 – 186. (http://korpus.juls.savba.sk/publications/index.sk.html)

HAJIČ, Jan – PANEVOVÁ, Jarmila – BURÁŇOVÁ, Eva – UREŠOVÁ, Zdeňka – BÉMOVÁ, Alla – KÁRNÍK, Jiří – ŠTĚPÁNEK, Jan – PAJAS, Petr: Anotace na analytické rovině. Návod pro anotátory. Praha: ÚFaL MFF UK 1999. 214 s.

NIŽNÍKOVÁ, Jolana – SOKOLOVÁ, Miloslava: Valenčný slovník slovenčiny. Prešov: Filozofická fakulta v Prešove, Prešovská univerzita 1998. 270 s.

ORAVEC, Ján – BAJZÍKOVÁ, Eugénia: Súčasný slovenský spisovný jazyk. Syntax. Bratislava: Slovenské pedagogické nakladateľstvo 1982. 261 s.

Pracovná verzia Slovenského závislostného korpusu

Slovenský národný korpus (http://korpus.juls.savba.sk)

ŠIMKOVÁ, Mária – GARABÍK, Radovan: Синтаксическая разметка в Словацком национальном корпусе In: Труды международной конференции Корпусная лингвистика – 2006. Sankt-Petersburg: St. Petersburg University Press 2006, s. 389 – 394. (http://korpus.juls.savba.sk/publications/index.sk.html)

Príloha

Zoznam funktorov používaných v syntaktickej anotácii vybraných textov SNK

afun	*Popis*
Pred	Predikát, resp. uzol, ktorý nezávisí od iného uzlu; vešia sa len na # – technický vrchol vety
Sb	Subjekt
Obj	Objekt
Adv	Adverbiále (bez bližšieho rozlíšenia)
Atv	Doplnok technicky zavesený na neslovesnom člene
AtvV	Doplnok visiaci na slovese (chýba druhý riadiaci člen)
Atr	Atribút
Pnom	Predikát nominálny, resp. menná časť prísudku so sponou *byť*
AuxV	Pomocné sloveso *byť* (Auxiliary Verb)
Coord	Koordinačný uzol
Apos	Apozícia
AuxT	Zvratné sa, neoddeliteľné sa – reflexívum
AuxR	Zvratné sa, ktoré nie je Obj ani AuxT (tvorí reflexívne pasívum)
AuxP	Primárna predložka, časť sekundárnej predložky
AuxC	Spojka (podraďovacia)
AuxO	Nadbytočný (odkazovací, emocionálny) element
AuxZ	Zdôrazňovacie slovo
AuxX	Čiarka (nie nositeľka koordinácie)
AuxG	Iné grafické symboly, ktoré neukončujú vetu
AuxY	Príslovky a častice, ktoré nie je možné zaradiť inam
AuxS	Koreň stromu (#), technický vrchol
AuxK	Koncová interpunkcia vety
ExD	Náhradná funkcia pre technické hrany smerujúce miesto od elidovaného člena k „pseudoriadiacemu“ slovu alebo pre hlavný člen vety bez predikátu (Ex-Dependent)
AtrAtr	Riadiacim slovom atribútu môže byť vďaka štruktúrnej viacznačnosti ktorýkoľvek z bezprostredne predchádzajúcich (syntaktických) substantív
AtrAdv	Štruktúrna viacznačnosť medzi adverbiálnou a adnominálnou závislosťou bez sémantických dôsledkov
AdvAtr	Detto, s opačnou preferenciou
AtrObj	Štruktúrna viacznačnosť medzi objektovou a adnominálnou závislosťou bez sémantických dôsledkov
ObjAtr	Detto, s opačnou preferenciou