Slovenský závislostný korpus

Mária Šimková – Katarína Gajdošová
Slovenský národný korpus Jazykovedného ústavu Ľ. Štúra SAV, Bratislava
{marias, katarinag}@korpus.juls.savba.sk

In this article, we introduce a subproject of the Slovak National Corpus: The Slovak Dependency Treebank (SDT), in particular its analytical layer. SDT has a wide coverage of styles and genres, including novels, fairy tales, journal texts, scientific articles and historical texts. These texts were independently annotated by two different annotators.

In 2007, the analytical annotation entered its 2nd phase. Our annotation team began correcting the manually annotated data. In this article, we describe some of the corrections they had to perform. The corrections can be divided into three groups: technical, morphological and syntactic. They are based on rules that are applied to SDT automatically and the non-conformant data is then corrected manually.

S tvorbou Slovenského závislostného korpusu (ďalej SZK) sa v oddelení Slovenského národného korpusu Jazykovedného ústavu Ľ. Štúra SAV v Bratislave (ďalej SNK) začalo v roku 2005. Práca na projekte prvej, analytickej roviny syntaktickej anotácie je rozdelená do niekoľkých fáz.

1. Prvou fázou je ručná anotácia, ktorá prebieha od júna 2005. Temer všetky texty predikované na syntaktickú anotáciu boli vopred ručne morfologicky anotované. Tým sa zabezpečila možnosť využitia morfologickej roviny pri poloautomatizovanej-poloručnej oprave anotovaných textov. Súčinnosť morfológie a syntaxe a ich vzájomné dopĺňanie pri opravách ručne anotovaných textov SNK uvedieme nižšie.

Databáza pracovnej verzie SZK sa v súčasnosti (stav k 20. 9. 2007) skladá z 39 483 viet, ktoré sú anotované dvojmo, teda v skutočnosti je zanotovaných 78 966 viet. Pracovná verzia SZK obsahuje momentálne aj texty syntakticky anotované len jedenkrát vzhľadom na to, že druhá nezávislá anotácia iným anotátorom práve prebieha. Táto časť textov predstavuje 2 198 viet. Projekt SZK je v prvej fáze plánovaný v rozsahu 50 000 jednotlivých viet, čo sa môže reálne naplniť do konca roka 2007, keďže súbežne s prvými opravami prebieha anotovanie ďalších textov, ktoré sa po fáze kontroly zaradia do celku SZK.

Primárnym východiskom slovenskej anotácie je funkčno-generatívny prístup aplikovaný v

Ústave formálnej a aplikovanej lingvistiky Matematicko-fyzikálnej fakulty UK v Prahe pri anotácii analytickej roviny Pražského závislostného korpusu (ďalej PZK). Z len minimálne modifikovaných princípov Anotace na analytické rovině. Návod pro anotátory (http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/cz/a-layer/pdf/a-man-cz.pdf) vychádzal slovenský anotátorský kolektív pri vytváraní databázy SZK za pomoci softvéru poskytnutého partnerským pracoviskom ÚFAL MFF UK v Prahe. Hlbšie zásady a teoretické východiská anotácie pracovnej verzie SZK sú zhrnuté v príspevku Syntaktická anotácia vybraných textov Slovenského národného korpusu (Gajdošová, v tlači).

Na rozdiel od Pražského závislostného korpusu, ktorý obsahuje výlučne publicistické texty, je štýlovo-žánrová štruktúra SZK pestrejšia:

- **beletria** (román Georgea Orwella *1984*, román Ladislava Balleka *Pomocník*, preklady rozprávok ako napr. *Aladin, Ferdo mravec, Herkules, Kniha džunglí, Včielka Maja*; mládežnícky román *Prvá láska*)
- **odborné texty** (historické monografie, lingvisticko-spoločenské štúdie)
- **populárno-náučný štýl** (texty z internetovej stránky <http://sk.wikipedia.org/wiki/>
- **publicistika** (texty z denníka SME a internetového časopisu InZine)

2. Oprava pracovnej verzie SZK je v súčasnosti v začiatočnom štádiu vzhľadom na dosiaľ absentujúcu programátorskú zložku. Nižšie uvedené postupy a pravidlá opráv SZK sú teda len úvodnou sondou do druhej fázy tvorby SZK.

Fáza opráv SZK je primárne založená na poloautomatizovanej-poloručnej oprave textov. Zvolený postup bol čiastočne inšpirovaný postupom pri oprave textov PZK anotovaných na analytickej rovine (Hladká – Pajas, 2001). Oprava analytickej roviny PZK sa však uskutočnila plne automatizovane a za iných podmienok. Nevýhodou anotácie analytickej roviny PZK je anotácia textov iba jedným anotátorom na rozdiel od slovenských textov, ktoré sú anotované dvoma anotátormi nezávisle od seba. Opravy analytickej roviny PZK sa robili s časovým odstupom paralelne s opravami ďalšej – tektogramatickej roviny PZK, teda nie bezprostredne po skončení ručnej anotácie analytickej roviny, resp. v jej priebehu. Východiská pre opravy analytickej roviny sa utvárali akoby druhoplánovo pri realizácii opráv tektogramatickej roviny PZK.

Postup zvolený pri opravách pracovnej verzie SZK je založený na lingvisticko-programátorskej spolupráci a kombinácii poloautomatizovaných-poloručných opráv. Celý komplex

opráv, ktoré v súčasnosti začínajú prebiehať v SNK, môžeme rozčleniť na niekoľko stupňov: technicky nevyhnutné, morfologické, syntaktické opravy. Všetky typy opráv musia prebiehať vo vzájomnej súčinnosti, niektoré z nich je nevyhnutné vykonať ihneď na začiatku, iné nie sú závislé od času, v ktorom sa uskutočnia.

Ako základná technicky nevyhnutná oprava je potrebná kontrola segmentácie viet. Problém so zdvojenými, príp. viacerými vetami na mieste, kde mala stáť regulárne len jedna veta, sa ukázal hneď pri prvých textoch, v ktorých vetnú segmentáciu vykonal nástroj automaticky. Po zistení uvedenej skutočnosti sme prešli na ručnú segmentáciu, resp. kontrolu automatickej segmentácie viet pri príprave textov na ručnú anotáciu. Ukázalo sa, že tento krok odstránil mnohé chybné rozdelenia viet spôsobené najmä iným zakončením vety, ako je štandardná koncová interpunkcia (., ?, !) – problematická bola predovšetkým priama reč, teda kombinácie s úvodzovkami, ale aj neukončené výpovede, kde boli, naopak, koncové tri bodky rozdelené nástrojom do troch samostatných formálnych štruktúr zaradených do systému v súbore ako samostatné vety. Napriek tomu však, prihliadajúc na ľudský faktor, nemôžeme zaručiť, že všetky texty, ktoré prešli ručnou kontrolou, sú absolútne správne rozdelené. Preto je vhodné aplikovať ešte opätovnú kontrolu segmentácie v už zanotovaných vetách. Na základe zoznamu potenciálnych zakončení viet a začiatkov po nich nasledujúcich viet v slovenčine (zoznam v prílohe) určíme v opravovaných dátach všetky možné chybné segmentácie viet a po manuálnej kontrole výpisu potenciálne chybne rozdelených viet skutočne chybné vety automaticky rozdelíme.

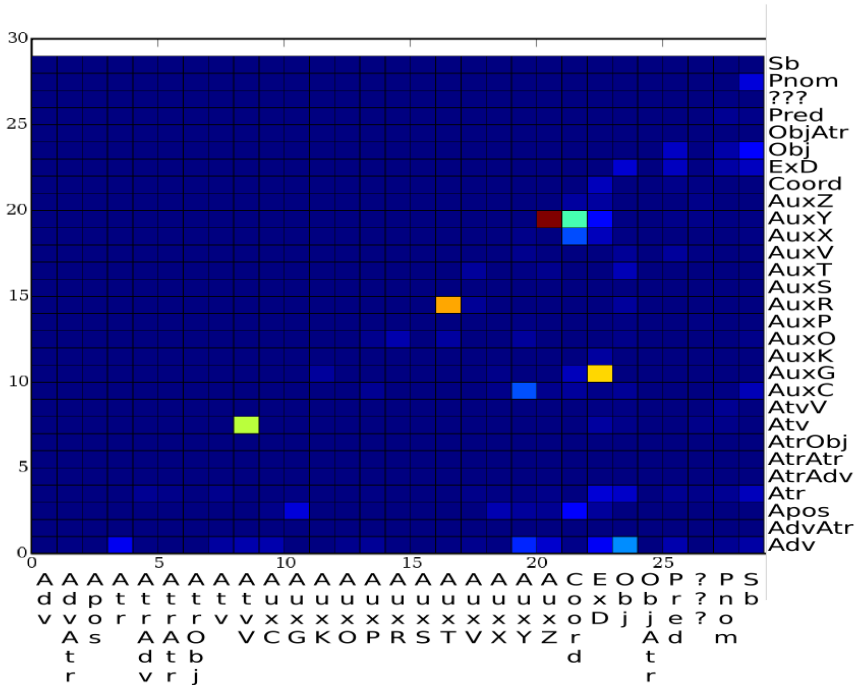
Ďalšou do istej miery technickou opravou je zistenie a ručné opravenie všetkých znakov ??? pri uzloch, ktorým z rôznych príčin – najčastejšie z dôvodu nepozornosti anotátora či chybného preklopenia viet z pôvodných fs formátov do pml formátov – neboli priradené syntaktické funktory. Nezanotované vrcholy predstavovali 0,74 % všetkých vrcholov, čo je v skutočnosti 7922 neoznačených vrcholov v celom rozsahu doteraz zanotovaných viet.

Za technickú opravu pokladáme aj odstránenie chýb (_ERR), ktoré vznikli pri nesprávnom zavesení funktora na vetný člen, ktorý nie je garantom koordinačného alebo apozičného vzťahu, resp. garant koordinačného alebo apozičného vzťahu je nesprávne označený, keďže na ňom nie sú závislé žiadne vetné členy, ktoré by boli vo vzťahu koordinácie či apozície.

Po týchto viac-menej technických opravách bude z dvoch textov anotovaných dvoma anotátormi nezávisle od seba utvorený jeden výsledný súbor, v ktorom sa zobrazia odlišnosti v anotáciách. Nástrojom btred bude možné všetky odlišnosti anotátorov porovnať, príp. odstrániť. Odstraňovanie odlišností by sme však radi efektívne skoordinovali s ďalšími typmi opráv, napr. na

báze morfologicko-syntaktických pravidiel.

Chybovosť anotácie, teda zhodu či nezhodu dvoch anotátorov dokumentuje nasledujúca ilustratívna tabuľka. Čím je farba štvorčeka na priesečníku označenia dvoch uzlov svetlejšia, tým je vyššia percentuálna nezhoda anotátorov: v najbledších miestach, ktoré predstavujú predovšetkým priesečníky rôznych druhov častíc, interpunkcie označenej ako AuxG alebo ExD, presného určenia doplnku či tradične problémového rozlíšenia niektorých typov predmetov a príslovkových určení, sú anotátori najviac náchylní zamieňať konkrétne vetné funktory. Chybovosť znázornená v tabuľke je však do istej miery skreslená automatickým preklopením pôvodných súborov fs do formátu pml, takže momentálne nie je možné presne vyčísliť percento chybovosti jednotlivých funktorov.



Oprava pracovnej verzie SZK je neodlučiteľne spätá s prípravou špecifických zoznamov taxatívne spracovaných slovných druhov, resp. tried, na základe ktorých bude možné jednotne a prakticky poloautomatizovane zachytiť niektoré členy závislostného vzťahu, najmä v prípade nesyntagmatických slovných druhov. Uvedené zoznamy vznikajú ako pracovný materiál z aktuálnych slovenských kodifikačných príručiek (Krátky slovník slovenského jazyka, 2003; Pravidlá slovenského pravopisu, 2000) a prvého zväzku Slovníka súčasného slovenského jazyka (A – G, 2006).

3. V nasledujúcej časti príspevku uvedieme výberovo niekoľko základných lingvistických pravidiel, ktoré sú podkladom pre poloautomatizované opravy a budú súčasťou komplexnej dokumentácie opráv analytickej roviny SZK. Takisto jednotlivé požiadavky, tzv. makrá, ktoré sa budú používať pri opravách pracovnej verzie SZK, sú aktuálne v procese dopĺňania a skvalitňovania tak, aby boli efektívne vzhľadom na množstvo chybne anotovaných vzťahov a potrebu zvládnutia ich ručných opráv v reálnom čase.

Jednou z požiadaviek na texty, ktoré boli predikované na zaradenie do SZK, bola predchádzajúca ručná morfologická anotácia textov na m-rovine. Táto požiadavka zabezpečila možnosť formulovať pravidlá na opravu syntaktickej roviny anotácie v spolupráci s ručným morfologickým označkovaním textov. Formulovanie morfologickej časti pravidiel na poloautomatizované počítačové opravy pracovnej verzie SZK vychádza zo zásad ručnej morfologickej anotácie vybraných textov SNK (doplnená verzia Garabík – Gianitsová – Karčová – Šimková, 2006). Na základe týchto pravidiel predpokladáme, že vzájomnou kontrolou obidvoch druhov anotácie dosiahneme odstránenie chýb nielen v syntaktických dátach, ale aj v tých istých ručne morfologicky anotovaných dátach. Hoci morfologická anotácia vybraných textov SNK prešla niekoľkými fázami opráv, stále prebieha ich sporadické opravovanie na základe príležitostne zistených chýb používateľmi ručne morfologicky anotovaného korpusu r-mak.

Na ilustráciu uvedieme dva príklady pripravovaných lingvistických pravidiel pre poloautomatizované opravy, resp. kontroly dát pracovnej verzie SZK.
(Pozn.: AuxV a Sb sú funktory analytickej roviny syntaktickej anotácie, AuxV sú pomocné časti zloženého slovesného tvaru, Sb je subjekt; značky tagov uvádzame podľa slovenského tagsetu, porov. Garabík – Gianitsová – Karčová – Šimková, 2006.)

AuxV tvary slovies *som, si, sme, ste:* vždy [lemma="byť"& a tag="VKe[ab].*]

tvary slovies *ne/budem, ne/budeš, ne/bude, ne/budeme, ne/budete, ne/budú:* [lemma="byť" & tag="VB.*"]

tvary slovies *ne/bol(a, o, i):* [lemma="byť" & tag="VLe.c*"]

tvary slovies *ne/býval (a, o, i):* [lemma="bývať & tag="VLe.c*"]

tvar kondicionálovej morfémy *by:* [lemma="by" & tag="Y"]

vždy visí na funktore Pred a súčasne

na ňom už nevisí nič iné

Sb S.*1.*

PS.*1, PP.*1 [lemma="*ja, ty, my, vy*"], PF.*1 PU.*[lemma=".*koľko.*"]

NS.*1, NU.*1 (lemma="*viacero, hodne, .*veľa, viac, .*mnoho, moc, plno, nadostač, .*mál.*, .*menej, tro.*, pár, zopár, dosť, pol, štvrť, trištvrte, štvrte*"]

VI.*

visí na Pred [V.*]

na jednom Pred visí len jeden Sb, inak musí Sb visieť na Pred ako súčasť Coord

môže visieť na Pred, Obj, Atr, Adv, Atv, AtvV, Coord, Pred_Pa, Pred_Ap, ak pred ním visí AuxC

Na základe výpisu viet získa lingvista cielene výber potenciálne chybných viet, ktoré nevyhovujú sformulovaným požiadavkám na správnosť syntagmy, priradenie správnych funktorov jednotlivým vetným členom či správnosť morfologickej informácie v pozadí syntagmatického vzťahu. Tie sa budú ručne opravovať. Po opravení testovacej vzorky dát zistíme, aké sú perspektívy automatickej opravy, resp. správneho umiestnenia konkrétnych vrcholov syntagiem v textoch bez ručnej kontroly. Plne automatizovaný postup však predstavuje priveľké riziko nesprávnych opráv na úkor správnych zachytení jednotlivých syntagiem v prípade, že konkrétne pravidlo neplatí stopercentne. Preto by sme sa mu v našej práci radšej vyhli, príp. ho použili len okrajovo, v pozíciách, kde automatickou opravou nemôže nastať znehodnotenie zanotovaných dát.

4. Fáza poloautomatizovanej-poloručnej opravy pracovnej verzie SZK vyústi do sprístupnenia SZK pre používateľov na Internete. S korpusom SZK-a – analyticky anotovaných dát budú môcť používatelia pracovať prostredníctvom klienta NetGraph, ktorý je známy a úspešne používaný ako vyhľadávací nástroj pre PZK.

Literatúra

Anotace na analytické rovině. Návod pro anotátory. Dostupný z WWW: http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/cz/a-layer/pdf/a-man-cz.pdf

GAJDOŠOVÁ, Katarína: Syntaktická anotácia vybraných textov Slovenského národného korpusu. In: VARIA XVI. Zborník zo XVI. kolokvia mladých jazykovedcov. Bratislava: Slovenská jazykovedná spoločnosť pri SAV(v tlači).

GARABÍK, Radovan – GIANITSOVÁ, Lucia – KARČOVÁ, Agáta – ŠIMKOVÁ, Mária

(2006): Tokenizácia, lematizácia a morfologická anotácia Slovenského národného korpusu. Dostupný z WWW: http://korpus.juls.savba.sk/publications/

HLADKÁ, Barbora – PAJAS, Peter (2001): Post-annotation Checking of the Treebank. Dostupné z WWW: http://ufal.mff.cuni.cz/pdt/Corpora/PDT_1.0/References/AvsM.pdf

Krátky slovník slovenského jazyka. 4., doplnené a upravené vydanie. Red.: J. Kačala – M. Pisárčiková – M. Považaj. Bratislava: Veda 2003.

Pravidlá slovenského pravopisu. 3., upravené a doplnené vydanie. Red. M. Považaj. Bratislava: Veda 2000.

Slovník súčasného slovenského jazyka. A – G. Red. K. Buzássyová – A. Jarošová. Bratislava: Veda 2006.

PRÍLOHA

KOMBINÁCIE KONCOVEJ INTERPUNKCIE NA SEGMENTÁCIU VIET

.
?
!
…
???
!!!

?!
!?
...?
...!
...?!
...!?
...???
...!!!

.“
?“
!“
...“
?!“
!?“
!!!“
???“

...?“
...!“
...?!“
...!?“
...!!!“
...???“

POTENCIÁLNE ZAČIATKY VIET NA SEGMENTÁCIU VIET

… + malé písmeno (číslica)
… + veľké písmeno (číslica)
„ + ... + malé písmeno (číslica)
„ + … + veľké písmeno (číslica)
„ + veľké písmeno (číslica)