Morfologická anotácia korpusu

Martina Forróová – Alexander Horák

Príspevok odznel na konferencii Slovenčina na začiatku 21. storočia v Prešove (7. - 9. 3. 2003) a bude publikovaný v zborníku z tejto konferencie.

Abstract: *In this paper, we discuss the general annotation of texts in corpora, with detailed insight into morphological analysis of texts in corpora. In the first part of our paper, we deal with general annotation of corpora, in the second part we open the question of morphological analysis and in the third part we present several different Slavic languages and existing systems for their morphological annotation and analysis.*

0. Úvod

Koncom 20. a začiatkom 21. storočia sa výrazne aktualizuje počítačové spracovanie prirodzeného jazyka (NLP) ako v rovine počítačovej podpory lexikografických projektov, tak aj v rozličných interdisciplinárnych aplikáciách (umelá inteligencia, automatizované systémy spracovania dát, neurolingvistika, analýza a syntéza reči, jazykové a terminologické databázy atď.).

Jednou z aplikácií NLP, ktorú priblížime v tomto príspevku, je automatická morfologická analýza. Zameriame sa na jej uplatnenie pri anotácii korpusu.

V prvej časti upriamime pozornosť na anotáciu korpusu vo všeobecnosti. V druhej časti načrtneme morfologickú analýzu a jej aplikáciu pri značkovaní korpusu v teoretickej rovine. V tretej časti si predstavíme niekoľko odlišných prístupov uplatnených v korpusoch niektorých slovanských jazykov (čeština a poľština). Na záver by sme chceli, na pozadí vopred spracovaných informácií, otvoriť otázku možnosti automatického morfologického spracovania pre slovenčinu.

1. Anotácia korpusu

1.1 Pojem anotácie

V štúdiách z oblasti korpusovej lingvistiky je termín anotácia (angl. *annotation*) frekventovaný, no používa sa často v rozličných významoch. Vo všeobecnosti ho možno chápať ako pridávanie informácií k textom tvoriacim korpus. V teórii korpusovej lingvistiky sa objavujú v zásade dva protichodné názorové prúdy o zmysle anotácie korpusov.

Hlavným predstaviteľom odporcov anotácie korpusov je John Sinclair, tvorca známeho anglického výkladového slovníka Collins Cobuild, vychádzajúceho z materiálu korpusu *Bank of English*. Základnou myšlienkou, z ktorej tento prúd vychádza, je „čistota“ vstupných dát, ktoré samy o sebe najlepšie reprezentujú prirodzený jazyk. To znamená, že každá vnesená informácia do korpusu by v konečnom dôsledku znamenala dezinterpretáciu jazyka. Týmto korpus stráca svoj pôvodný význam (reprezentácia reálneho jazyka).

Na druhej strane zástancovia anotácie (napr. Geoffrey Leech) argumentujú širšou využiteľnosťou anotovaných korpusov v lexikografii, automatizovanom preklade, pri verifikácii lingvistických teórií, tvorbe štatistických modelov jazyka a návrhu tagerov (programov na značkovanie korpusov). Táto skupina si tiež uvedomuje možnosť lingvistickej dezinterpretácie informáciami vnášanými do korpusu. G. Leech preto navrhuje tieto zásady anotácie korpusov:

1. **eliminovateľnosť** – možnosť návratu k neanotovanému korpusu;
2. **extrahovateľnosť** – možnosť extrakcie anotácie z neanotovaného korpusu;
3. **prístupnosť** – anotácia je založená na zásadách prístupných koncovému používateľovi (aj nelingvistovi);
4. anotácia má charakter **pomôcky**, nie predpisu;
5. **konsenzus** vedeckých teórií, teoretická „neutralita“;
6. **neautoritatívnosť**, zlučiteľnosť s inými anotačnými štandardami;
7. **transparentnosť** údajov o spôsobe a autoroch anotácie.

1.2 Typy anotácie

Informácie, vnášané do korpusu, sú dvojakého druhu:

1. Informácie o reprezentácii zaznamenávajú štruktúrne vlastnosti textu ako nadpis, kapitola, odsek, veta, interpunkcia a pod. Zaznamenávajú sa pomocou štandardizovaných kódovacích metajazykov ako SGML a XML.

2. Interpretatívne informácie tvorí navrhnutý súbor značiek, zachytávajúci lingvistické vlastnosti textu. Zvyčajne je to množina znakov, ktorá je výsledkom formálneho opisu jazyka, alebo formalizácia už existujúcich lingvistických deskripcií.

Vzhľadom na povahu priraďovaných údajov sa anotácia zvyčajne rozdeľuje na:

1. externú, ktorú tvoria:
   - bibliografické údaje (pôvod, autor, zdroj, typ, jazyk textu) a
   - štruktúrne údaje (kapitola, odsek, veta ...)[1]
2. internú, ktorá zachytáva všetky jazykové roviny:
   - fonologickú (fonetickú, prozodickú) – pri hovorených korpusoch
   - morfologickú (častejšie označovanú ako morfosyntaktická)
   - syntaktickú
   - sémantickú (dezambiguácia lexikálnych významov, pragmatická, diskurzová)

2. Morfologická rovina spracovania korpusu

2.1 Morfologická analýza (MA)

V písaných korpusoch predstavuje rovina morfologickej analýzy vstup do hlbších analytických rovín. MA sa všeobecne chápe ako proces klasifikácie slov prirodzeného jazyka do gramaticko-sémantických tried a priradenie gramatických kategórií (v tradičnom ponímaní) týmto slovám. To, čo sa javí ako relatívne nenáročná úloha pre jazykovú kompetenciu človeka, môže byť pomerne náročnou úlohou pre počítač, resp. automatické spracovanie jazyka.

2.2 Automatická morfologická analýza (AMA)

Pre spracovanie morfológie prirodzeného jazyka automatickým spôsobom je nevyhnutné popísať proces MA formálne. MA môžeme vyjadriť ako zobrazenie, ktoré každému slovu (slovnému tvaru) priradí dvojicu lema – značka alebo množinu takýchto dvojíc:

$MA(f) \rightarrow \{<l,t>; l \in L, t \in T\}$, kde

$f \in A^+$ je slovný tvar zložený z písmen abecedy A,

L je množina lem,

T je množina značiek (tagset) pre daný jazyk.

1 Podrobnosti v príspevku R. Garabíka v tomto zborníku.

Možno teda konštatovať, že v priebehu AMA dochádza k dvom úzko prepojeným procesom:

1. lematizácia – priradenie základného (slovníkového) tvaru slovným tvarom v texte,
2. tagovanie (značkovanie) – priradenie súboru lingvistických značiek slovným tvarom v texte.

Výsledky AMA môžu byť v dôsledku tvarovej homonymie slov[2] nejednoznačné, a preto sa následne korigujú morfologickou dezambiguáciou (zjednoznačnením výstupu morfologickej analýzy).

2.2.1 Tokenizácia

Predpokladom každej automatickej morfologickej analýzy je prevedenie textov do jednotného formátu (SGML, XML) a tokenizácia, čiže identifikácia „slov“ v texte – rozdelenie textu na tokeny (reťazce znakov medzi dvomi medzerami).Takéto chápanie slova sa dostáva do rozporu s lingvistickým, kde je slovo definované podľa jednotlivých rovín ako:

- súhrn všetkých slovných tvarov (morfológia);
- komponent syntagmy (syntax);
- nositeľ vecného významu (lexikológia).

Z definície tokenu ako reťazca znakov medzi dvomi medzerami vyplýva, že ním okrem textového slova môžu byť aj čísla, interpunkčné znamienka, skratky atď. Základným problémom z lingvistického hľadiska je tu teda nemožnosť identifikácie hraníc slova s hranicami tokenu, čo môže viesť na ďalších rovinách automatického spracovania k odlišnej lingvistickej interpretácii. Ide napríklad o javy, keď:

- je význam jazykovej jednotky vyjadrený viacerými slovnými tvarmi: analytické slovesné tvary (*bol by som pracoval*), propriá (*Nové Mesto nad Váhom*), termíny (*oxid uhličitý*) ...
- sú významy viacerých jazykových jednotiek vyjadrené jednoslovne: napr. *zaň, preň, tys'*

2.3 Morfologické značkovanie

Morfologické značkovanie zahŕňa priebeh a výsledok morfologickej analýzy po dezambiguácii. Pre morfologické značkovanie sú nevyhnutné morfologický slovník, tagset, tager, trénovací korpus a testovacie dáta.

2 Tvarová homonymia sa v počítačovom spracovaní jazyka chápe zvyčajne ako akákoľvek formálna identita znakových reťazcov, teda značne širšie než tradične v lingvistike.

2.3.1 Morfologický slovník

Súbor informácií o kmeňoch a koncovkách slovných tvarov a o možných značkách (kombináciách morfologických hodnôt), ktoré koncovkám prislúchajú, napr.:

„ „ = NNIS1-----A---- (podst. m., všeobecné, mužské neživotné, j. č., 1. p., afirmatív) alebo
NNIS4-----A---- (podst. m., všeobecné, mužské neživotné, j. č., 4. p., afirmatív)

„u“= NNIS2-----A---- (podst. m., všeobecné, mužské neživotné, j. č., 2. p., afirmatív) alebo
NNIS3-----A---- (podst. m., všeobecné, mužské neživotné, j. č., 3. p., afirmatív) alebo
NNIS6-----A---- (podst. m., všeobecné, mužské neživotné, j. č., 6. p., afirmatív)

Množina týchto koncoviek a ich značiek tvorí vzor, ktorý sa v priebehu morfologickej analýzy priraďuje danému kmeňu.

2.4 Tagset

Množina značiek reprezentujúcich lingvistické kategórie. Formálne ho možno vyjadriť ako karteziánsky súčin morfologických kategórií, ktorých hodnoty sa reprezentujú značkami.

$T \subseteq K_1 \times \ldots \times K_n$

K_i … i-tá morfologická kategória

kategória je množina hodnôt, napr. $K_{rod} = \{M, I, F, N\}$

značka $t = (k_1, \ldots, k_n) \in T$, k_i je hodnota i-tej kategórie

Morfologické značky sú pri slovných tvaroch zapisované dvomi spôsobmi:

1. **pozičný** (Hajič): každej pozícii zodpovedá jedna značka – hodnota gramatickej kategórie; hodnoty irelevantných kategórií sú vyznačené pomlčkou
   politikou NNFS7-----A---- (podst. m., všeobecné, ž. r., j. č., 7. p., afirmatív)
2. **skrátený/atribútový** (Multext-East): vyznačujú sa iba relevantné hodnoty pre daný slovný tvar
   budeme Vcif1pan (sloveso, kopula, indikatív., budúci čas, 1.os., mn.č., činný rod, nenegatívne)

Pri voľbe tagsetu je dôležitý práve typ notácie. Spôsob pozičného zapisovania je jednoduchšie počítačovo spracovateľný, kým atribútový (skrátený) systém je čitateľnejší pre používateľa.

2.5 Reprezentácia gramatických kategórií v tagsete

Pri voľbe tagsetu dochádza ku konfrontácii medzi reprezentáciou tradične chápaných gramatických kategórií a možnosťami počítača a automatického spracovania jazyka.

V lingvistickom prístupe sú kritériá morfologickej klasifikácie nejasné, často prichádza k prelínaniu viacerých jazykových rovín. Napríklad pre slovenský jazyk sa v doteraz platnej vysokoškolskej učebnici morfológie (Oravec – Bajzíková – Furdík, 1984, s. 13 a n.) uvádza ako klasifikačné kritérium pre slovné druhy komplex morfologických, syntaktických, lexikálnych a sémantických vlastností slov. Aj jednotlivé slovné druhy sa klasifikujú kombinovanými formálno-sémantických kritériami: substantíva lexikálno-gramatickými, adjektíva sémanticko-slovotvornými, zámená a slovesá sémanticko-gramatickými. Medzi morfologické kategórie sa zaraďujú aj lexikálno-gramatické kategórie intencie, vidu a stupňa. Za striktne formálne (gramatické) sa považujú rod, číslo, pád, čas, osoba a spôsob.

Pri použití takejto klasifikácie je vzhľadom na jej formálnu nejednoznačnosť značná pravdepodobnosť zlyhania automatického spracovania. Možnosť uplatnenia tradičného lingvistického prístupu v automatickej morfologickej analýze je zložitá. Pri návrhu tagsetu je nutné zvoliť čo najvyhovujúcejší z možných a doteraz aplikovaných prístupov:

- lingvisticky optimistický: uplatnenie čo najväčšieho počtu gramatických kategórií v tagsete → dezambiguácia pomocou pravidiel (Petkevič – Oliva)
- lingvisticky pesimistický: kompromis medzi lingvistickým a „inžinierskym“ prístupom[3] (Hajič, Multext-East, väčšina existujúcich)
- minimalistický: „načo nám je anotácia?“ (Sinclair, Belica)

2.6 Tager

Počítačový program, ktorý vykonáva morfologickú analýzu (každému slovnému tvaru v texte priraďuje možné morfologické interpretácie) a morfologickú dezambiguáciu (z množiny možných značiek vyberá správnu). Základom funkcie tagerov je algoritmus, ktorý môže pracovať podľa viacerých modelov:

1. pravdepodobnostné, založené na:
   - skrytých markovovských modeloch (*Hidden Markov Modesl*), napr. Hajičov tager
   - maximálnej entropii (*Maximum Entropy Tagger*), napr. tager Adwaita Ratnaparkhiho
   - trigramoch (*Trigrams ´n´ Tags*), napr. tager Thorstena Brantsa
2. nepravdepodobnostné, založené na:
   - pamäti (*Memory Based Tagger*), napr tager Waltera Daelemansa a Jakuba Zavrela
   - pravidlách (*Rule Based Tagger*), napr. tager Erica Brilla, tager V. Petkeviča a K. Olivu

3 Pod inžinierskym prístupom rozumieme morfologickú klasifikáciu, ktorá vyhovuje matematickému modelu tagera; porov. príklad z Českého národného korpusu, kap. 3.

Častou metódou na zefektívnenie morfologickej analýzy a dezambiguácie je kombinácia viacerých metód tagovania, napr. hybridný systém Hajiča a kol. (2001).

2.7 Trénovacie a testovacie dáta

Trénovacie dáta tvorí vopred ručne označkovaný a dezambiguovaný korpus, na ktorom sa tager „učí“ pomery medzi slovnými tvarmi a ich značkami. Testovacie dáta predstavuje korpus, na ktorom prebieha testovanie tagera. Pri značkovaní platí zásada, že testovacie dáta sa musia odlišovať od trénovacích (nemôže na nich prebiehať trénovanie). Typicky sa na testovanie používa 10 % dát, zvyšok tvoria dáta trénovacie.

3. Tagsety v slovanských jazykoch

Na príkladoch tagsetov Českého národného korpusu (ČNK), Pražského závislostného korpusu (PZK), Multext-East a Ústavu základov informatiky Poľskej akadémie vied (IPI PAN) si predstavíme niekoľko odlišných prístupov pri rozdeľovaní atribútov a priraďovaní ich morfologických kategórií. Konkrétne sa zameriame na atribút SUBPOS (TYPE, klasa gramatyczna) a na názorných príkladoch pre zámená porovnáme odlišnosti v menovaných tagsetoch.

3.1 Tagset ČNK, PZK

V tagsete pre Český národný korpus[4] a Pražský závislostný korpus[5] je použitá pozičná notácia. Obsahuje 13 atribútov, z ktorých atribút *POS* vyjadruje slovný druh (má 12 hodnôt), atribút *SUBPOS* vyjadruje bližšie určenie slovného druhu (má 74 hodnôt!). Ďalších 11 atribútoreprezentuje morfologické kategórie: rod, číslo, pád, privlastňovací rod, privlastňovacie číslo, osoba, čas, stupeň, negácia, slovesný rod, štýl, 2 pozície rezerva.

agset ČNK, PZK je vhodným príkladom „inžinierskeho“ prístupu. Možno ho ilustrovať atribútom *SUBPOS* (bližšie určenie slovného druhu) pre slovný druh zámeno, ktorý obsahuje až 19 možných hodnôt:

- vzťažné privlastňovacie zámeno *jehož, jejíž*
- vzťažné zámeno s adjektívnym skloňovaním (*jaký, který, ...*)

4 http://ucnk.ff.cuni.cz/

5 http://shadow.ms.mff.cuni.cz/pdt/index.html

- zámeno *on* v tvaroch po predložke (*něj, něho*, ...)
- reflexívne zámeno *se* v dlhých tvaroch (*sebe, sobě, sebou*, ...)
- reflexívne zámeno *se, si* iba v týchto tvaroch a tvaroch *ses, sis*
- privlastňovacie zámeno *svůj*
- vzťažné zámeno *jenž, již* po predložke (*něhož, níž*, ...)
- ukazovacie zámeno *ten, onen*
- vzťažné zámeno *což*
- vzťažné zámeno *jenž, jíž* bez predložky
- opytovacie, alebo vzťažné zámeno *kdo* vrátane tvarov s *-ž* a *-s*
- neurčité zámeno *všechen, sám*
- samostatne stojace zámená *svůj, nesvůj, tentam*
- osobné zámena (aj tvar *tys*)
- opytovacie/vzťažné zámená *co, copak, cožpak*
- privlastňovacie zámená *můj, tvůj, jeho* (aj plurál)
- záporné zámená (*nic, nikdo, nijaký, žádný*)
- zámeno *co* spojené s predložkou (*oč, nač, zač*)
- neurčité zámená (*nějaký, některý, číkoli, čosi*)

Z uvedeného je zrejmé, že klasifikácia zámen podlieha výlučne formálnym kritériám siahajúcim až na úroveň jedinečného slovného tvaru (*tys, což, svůj*).

3.2 Tagset Multext -East

Medzinárodný projekt Multext-East[6] bol zameraný na formálny opis siedmich európskych jazykov (anglický, rumunský, slovinský, český, bulharský a maďarský). Jeho cieľom bolo zjednotiť a štandardne opísať spomínané jazyky a následne vytvoriť anotovaný viacjazyčný korpus.

Tagset pre Multext-East bol vytvorený na princípe skrátenej notácie. Pre češtinu sa vymedzuje 12 slovných druhov (gramatických tried). Každá trieda je potom bližšie špecifikovaná 25 atribútmi, z ktorých *Type* je bližšie určenie príslušného slovného druhu. Ďalšie atribúty reprezentujú (podľa relevantnosti vo vzťahu k slovným druhom) nasledujúce morfologické kategórie: rod (Gend), číslo (Numb), pád (Case), určitosť (Def), klitika (Clitic), klitika *s* (Clitic_s), životnosť (Anim), privlastňovacie číslo (OwnN), privlastňovacia osoba (OwnP), privlastňovací rod (OwnG), slovesný tvar (VForm), čas (Tense), osoba (Pers), slovesný

6 http://nl.ijs.si/ME/

rod (Voice), negácia (Neg), zvratnosť (RefT), syntaktický typ (SynT), opytovacia forma (WhT), tvar zámena (PrFr), trieda (Class), typ spojky (Coord_Type), stupeň (Degree), tvar (Form), útvar (Formation). Každý z týchto atribútov potom môže nadobudnúť 2 až 14 hodnôt.

Na porovnanie si uvedieme atribút *Type* pre češtinu, opäť sa sústredíme na zámeno:

- personal (osobné): *já*
- demonstrative (ukazovacie): *ten*
- indefinite (neurčité): *někdo*
- possesive (privlastňovacie): *její*
- interrogative (opytovacie): *kdo*
- relative (vzťažné): *jenž*
- reflexive (zvratné): *se*
- negative (záporné): *nikdo*
- general („totálne“): *každý*

3.3 Tagset IPI PAN

V tagsete vytvorenom v Oddelení lingvistického inžinierstva Ústavu základov informatiky Poľskej akadémie vied[7] sa dôsledne uplatňuje kritérium formálnej morfológie vo výbere atribútov a ich hodnôt. Rozoznáva až 29 gramatických tried slov („slovných druhov“): substantívum (subst), depreciatívna forma substantíva (depr), adjektívum (adj), preadjektívne adjektívum (adja), popredložkové adjektívum (adjp), príslovka (adv), číslovka (num), zámeno v 1. a 2. osobe (ppron12), zámeno v 3. osobe (ppron3), zámeno *siebie* (siebie), sloveso v neminulom čase (fin), sloveso *być* v budúcom čase (bedzie), aglutinant slovesa *być* (aglt), *l*-particípium/pseudoparticípium (praet), sloveso v imperatíve (impt), sloveso v neosobnej forme *-no/-to* (imps), sloveso v infinitíve (inf), prechodník minulý (pant), prechodník prítomný (pcon), gerundium (ger), činné príčastie (pact), trpné príčastie (ppas), sloveso typu *winien*, predikatív (pred), predložka (prep), spojka (conj), častico-príslovka (part), nominálny cudzí element (xxs), iný cudzí element (xxx). Týmto triedam slov môžu byť priradené nasledujúce gramatické kategórie: číslo, pád, rod, osoba, stupeň, vid, negácia, akcentovanosť, popredložkovosť, akomodačnosť, aglutinačnosť, vokalizovanosť.

7 http://dach.ipipan.waw.pl/

4. Súbor morfologických značiek pre slovenčinu

Ktorý prístup zvolíme pre slovenčinu? Uprednostníme konvenčný lingvistický, ktorý rešpektuje komplex sémantických, syntaktických a gramatických kritérií v morfológii, aj napriek nižšej úspešnosti tagera? Zameriame sa na „inžiniersky“ a prispôsobíme gramatickú klasifikáciu slovnej zásoby matematickému modelu tagera? Zdá sa, že prístupom, pri ktorom sa dá predpokladať vyššia úspešnosť tagera a ktorý by zároveň rešpektoval lingvistické klasifikačné kritériá, je formálno-gramatický. Pri tejto voľbe však treba mať na pamäti lingvistické znalosti používateľa korpusu, ktoré na Slovensku vychádzajú z tradičných školských preskriptívnych gramatík. Na druhej strane predpokladáme, že viacerí budúci používatelia už majú isté znalosti o korpuse a možnostiach jeho použitia a sú schopní lingvistickej abstrakcie.

LITERATÚRA

Bański, Piotr (2003): *Anotacja zewnętrzna: wpływ architektury korpusu IPI PAN na efektywność jego tworzenia oraz wykorzystania.*
<http://dach.ipipan.waw.pl//CORPUS/banski_polonica.pdf>

Bański, Piotr (2001): *The proposed encoding scheme for the IPI PAN corpus.* Raport techniczny. Warszawa: Instytut Podstaw Informatyki PAN.
<http://dach.ipipan.waw.pl/CORPUS/banski_raport.pdf>

Dębowski, Łukasz (2001): *Tagowanie i dezambiguacja*. Prace IPI PAN 934. Instytut Podstaw Informatyki PAN. <http://www.ipipan.waw.pl/~ldebowsk/raporty/kropka934.pdf>

Džeroski, Sašo – Erjavec, Tomaž – Zavrel, Jakub (2000): *Morphosyntactic Tagging of Slovene: Evaluating Taggers and Tagset.* In: Proceedings of the Second International Conference on Language Resources and Evaluation, s. 1099–1044.
<http://www.coli.uni-sb.de/~thorsten/tnt/papers/lrec2000-dzeroski-ea.pdf>

Erjavec, Tomaž (1999): *Tagging Slavic Corpora.*
<http://nl.ijs.si/et/talks/SFB441/tue-slides2.ps.gz>

Hajič, Jan: *Popis morfologických značek – poziční systém.*
<http://ucnk.ff.cuni.cz/manual/znacky.html>

Hajič, Jan – Krbec, Pavel – Květoň, Pavel – Oliva, Karel – Petkevič, Vladimír (2001): *Serial Combination of Rules and Statistics: A Case Study in Czech Tagging.* Prague

Dependency Treebank. CD ROM. V. 1.0. Praha: Ústav formální a aplikované lingvistiky MFF UK.

Hajič, Jan – Hladká, Barbora: *The Prague Dependency Treebank: Annotation Structure and Support.* < http://ufal.mff.cuni.cz/publications/year2001/IRCSldca.ps />

Hajnicz, Elżbieta – Kupść, Anna: *Przegląd analizatorów morfologicznych dla języka polskiego*. Prace IPI PAN 937. Instytut Podstaw Informatyki PAN.
<http://dach.ipipan.waw.pl/CORPUS/anl.pdf>

Leech, Geoffrey (2000): *Anotační systémy pro značkovaní korpusu*. In: Acta Universitatis Carolinae – Philologica 3–4 – Studie z korpusové lingvistiky. Praha: Univerzita Karlova v Praze, Nakladatelství Karolinum, s. 185–197.

Kol. (1996): *EAGLES Recommendations for the Morphosyntactic Annotation of Corpora*. Eagles Document– CWG–MAC/R.
<http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html/>

Kol. (2001): *Slovenčina a čeština v počítačovom spracovaní.* Bratislava: VEDA Vydavateľstvo Slovenskej akadémie vied.

Kol. (1997): *Specification and Notation for Lexicon Encoding.* COP Project 106 Multext-East Work Package WP1 - Task 1.1 Deliverable D1.1 F Final Report.
<http://nl.ijs.si/ME/CD/docs/mte-d11f/index.html>

Przepiórkowski, Adam (2003): *Składniowe uwarunkowania znakowania morfosyntaktycznego w korpusie IPI PAN.*
<http://dach.ipipan.waw.pl/~adamp/Papers/2003-polonica-dis/polonica-dis.pdf>

Przepiórkowski, Adam – Woliński, Marcin (2003): *The Unbearable Lightness of Tagging. A Case Study in Morphosyntactic Tagging of Polish.*
<http://dach.ipipan.waw.pl/~adamp/Papers/2003-eacl-ws03/ws03.pdf>

Przepiórkowski, Adam – Woliński, Marcin (2003): *A Flexemic Tagset For Polish.*
<http://dach.ipipan.waw.pl//~adamp/Papers/2003-eacl-ws12/ws12.pdf>

Przepiórkowski, Adam – Bański, Piotr – Dębowski, Łukasz – Hajnicz, Elżbieta – Woliński, Marcin (2003): *Konstrukcja korpusu IPI PAN*.
<http://dach.ipipan.waw.pl/~adamp/Papers/2003-polonica-intro/>

Skut, Wojciech – Krenn, Brigitte – Brants, Thorsten – Uszkoreit, Hans (1997): *An Annotation Scheme for Free Word Order Languages*. In: Proceedings of the Fifth Conference on Applied Natural Language Processing, Washington, D.C.
<http://acl.ldc.upenn.edu/A/A97/A97-1014.pdf>

Woliński, Marcin (2003): *System znaczników morfosyntaktycznych w korpusie IPI PAN*.

<http://dach.ipipan.waw.pl/CORPUS/znakowanie.pdf>

Woliński, Marcin – Przepiórkowski, Adam (2001): *Projekt anotacji morfosyntaktycznej korpusu języka polskiego*. Prace IPI PAN 938. Instytut Podstaw Informatyki PAN. <http://dach.ipipan.waw.pl/~adamp/Papers/2001-tagset/ipi938.pdf>