Syntakticky anotovaný korpus českých textů

Milena Hnátková, Petr Jäger, Tomáš Jelínek, Vladimír Petkevič, Alexandr Rosen, Hana Skoumalová
Ústav teoretické a komputační lingvistiky, Filozofická fakulta
Univerzita Karlova

Abstract

We describe a project aiming at a large-size treebank of Czech, with a flexible and customizable interface to represent multiple aspects of sytactic annotation, allowing for underspecification and ambiguities. The treebank is built using a stochastic parser together with a rule-based component, improving the overall error rate.

1. Úvod

Syntaktických teorií je mnoho, a proto je obtížné vytvořit syntakticky anotovaný korpus, který by vyhovoval všem. Při bližším zkoumání se však teorie i přes zdánlivé zásadní rozdíly shodnou na společném jádru podstatných údajů o syntaktických konstrukcích a jevech. A tak si lze představit syntakticky anotovaný korpus (treebank), který by umožňoval takové společné jádro interpretovat více způsoby podle představ uživatele a kromě jiného nabízel třeba různé pohledy na strukturu věty. Jedním z takových pohledů může být třeba závislostní reprezentace syntaktické struktury, jak se jí učí žáci základních a studenti středních škol, další volbou může být složková reprezentace apod.

Takový korpus může být užitečný nejen na akademické půdě a pro všechny odborníky pracující s jazykem profesionálně, ale i pro laické uživatele se zájmem o jazyk. Muselo by se ovšem jednat o korpus dostatečně rozsáhlý, větší než dnes existující treebanky, které obvykle dosahují velikosti nejvýše několika milionů slov. Aby bylo dosaženo velikosti srovnatelné s morfologicky označkovanými korpusy, je třeba použít automatickou proceduru analyzující syntaktickou strukturu vět.

V návaznosti na již vykonanou práci (mj. Hajič 2006, Skut et al. 1997) se chceme přiblížit cíli vybudovat velký korpus s automaticky provedenou syntaktickou anotací, která bude mít rozumnou míru spolehlivosti. Navrhujeme proto schéma morfologické a syntaktické anotace, přístupné uživatelům i aplikacím prostřednictvím více volitelných a nastavitelných uživatelských rozhraní. Automatickou analýzu stochastickou metodou (Holan a Žabokrtský 2006) doplňujeme opravným modulem, vybaveným lingvistickými pravidly, který snižuje chybovost výsledné anotace.

2. Vlastnosti schématu

V této kapitole jsou popsány hlavní vlastnosti anotačního schématu spolu s důvody, proč byly zavedeny.

2.1. Možnost zobrazit syntaktickou strukturu několikerým způsobem

Reprezentaci syntaktické struktury a údaje o gramatických kategoriích uložené v interním formátu lze zobrazit různými způsoby. Interní formát lze interpretovat různě – struktura může být zobrazena jako složkový nebo závislostní strom s volitelným množstvím podrobností a nastavitelnou úrovní abstrakce (zahrnující např. hloubkové nebo povrchové závislosti, interpretaci funkčních slov a identifikaci složených slovesných forem včetně reflexiv). Zobrazení nemusí být nutně ve formě stromu, ale může jít o lineární zobrazení, kde jsou větné členy vyznačeny různými barvami nebo řezy písma. V příkladu (1) mají typografické prvky tento význam: **podmět**, *předmět*, přísudek, <u>členy vykazující shodu.</u>

(1) **Ty** by *se*s byl ušpinil.

Takové zobrazení základních syntaktických vlastností umožňuje zachovat formát konkordancí KWiC i při hledání v syntakticky anotovaném korpusu.

2.2. Víceznačnost a neúplná informace

Anotace korpusu bývá jednoznačná, protože se předpokládá, že ji lze vyřešit na základě širšího kontextu. Někdy je však víceznačnost nerozhodnutelná, protože se netýká významu, ale jde o víceznačnost na úrovni segmentace textu, morfologie nebo syntaxe. Může se jednat např. o valenční pozici vyžadující akuzativ nebo genitiv naplněnou jménem vykazujícím pádový synkretismus, např. ve spojení *využívat zařízení* (Oliva 2001), nebo o strukturu obsahující předložkovou frázi, která může být beze změny významu umístěna v syntaktické stromové struktuře na více místech, např. *uzavřít mír se sousedy* (Hajič et al. 1999). V jiných případech sice víceznačnost v principu rozhodnutelná je, ale kontext nedává dostatek informací pro její vyřešení, a je tedy správnější nerozhodovat arbitrárně, ale víceznačnost zachovat.

Anotační schéma umožňuje zachycení víceznačných nebo nerozhodnutelných jevů pomocí podspecifikace a distributivní disjunkce, a to jak pro hodnoty kategorií, tak pro celé struktury. Chybět může informace libovolného druhu. V extrémním případě je syntaktická struktura zcela triviální – místo stromového grafu reprezentuje větu nebo větný člen pouhý seznam slov. Neúplná analýza může obsahovat informaci o řídícím členu slova, nebo informaci o tom, do jaké složky slovo patří, a ostatní syntaktické vztahy ve větě mohou zůstat nerozhodnuté. Nerozhodnutá víceznačnost není žádoucí, pokud lze dosáhnout úplné desambiguace, ale pro zvláštní případy si tuto možnost ponecháváme.
Aby bylo možné tuto míru volnosti vyjádřit jednotným způsobem, je interní formát pro zachycení syntaktické struktury založen na bezprostředních složkách s kombinací binárního a neomezeného větvení. Formálně nejsou podsložky nedílnou součástí nadřízené složky, ale jsou určeny odkazy do lineárního, nehierarchického seznamu všech složek ve větě. Řešení je patrné na příkladu víceznačné věty (2) – (5).

(2) Zdravotnictví musí zachránit stát.
#1 Stát musí být zachráněn zdravotnictvím.
#2 Zdravotnictví musí být zachráněno státem.

(3) Morfologická analýza s některými neurčenými hodnotami:
[1] zdravotnictví *noun,* case=*X*, num=sg, gend=*n*
[2] musí *verfin,* pers=3, num=sg
[3] zachránit *inf*
[4] stat *noun,* case=*Y,* num=sg, gend=*m*

(4) Složky v jedné z možných struktur:
[5] [[3]zachránit [4]stát]
[6] [[2]musí [5]]
[7] [[1]zdravotnictví [6]]

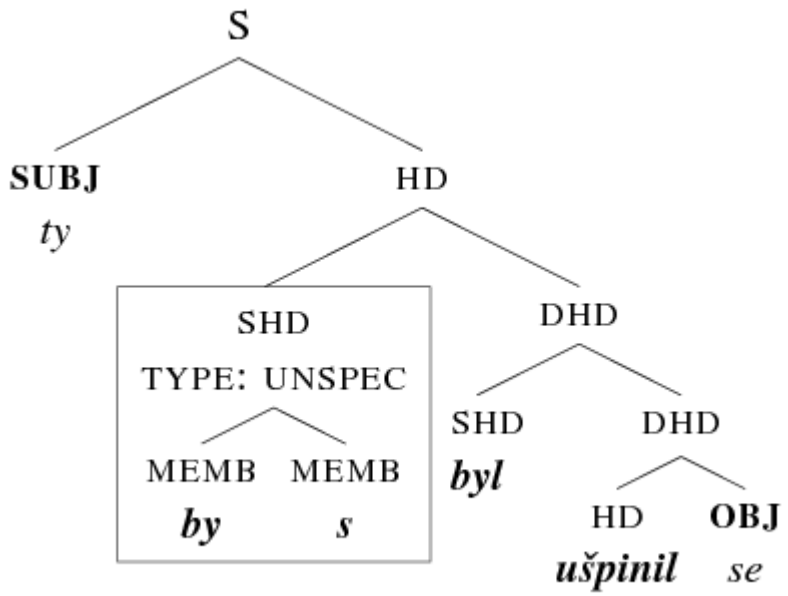
(5) Dvě možné struktury s omezeními na kategorie:
#1 = [7], & X=*nom*, Y=*acc*
#2 = [7], & X=*acc*, Y=*nom*, & [1]→[4], [4]→[1]

2.3. Povrchová a hloubková struktura

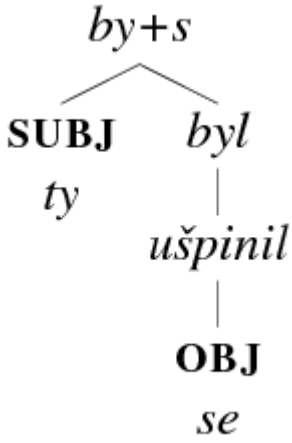
Složky mohou být označeny jako hlavy, nebo nesou označení své syntaktické funkce. Hlavy se mohou dále dělit na povrchové a hloubkové. Funkční slova jako předložky nebo pomocná slovesa jsou označeny jako povrchové hlavy a jejich sestry jsou označeny jako hloubkové hlavy. Toto rozdělení umožňuje odvodit povrchové i hloubkové závislosti z jediné struktury. Koordinace a jiné podobné konstrukce žádnou hlavu nemají. Příklad:

(6) Ty by ses byl ušpinil.

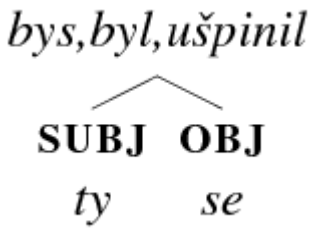
(7) Složková struktura z příkladu 6 s vyznačeným podmětem, předmětem, hlavou, povrchovou hlavou a hloubkovou hlavou:



(8) Povrchová závislostní struktura odvozená z příkladu 7:



(9) Hloubková závislostní struktura odvozená z příkladu 7:



2.4. Oddělení grafémiky, morfologie a syntaxe

Ve vnitřní reprezentaci struktur je každá věta zachycena na třech rovinách, vzájemně provázaných odkazy: rovině grafémické (slova zapsaná podle ortografických pravidel), morfologické (syntaktická slova) a syntaktické (stromy). První, grafémická rovina umožňuje zaznamenat spřežky a jiné čistě ortografické jevy. Na morfologické rovině se spřežky interpretují, tj. dělí na jednotlivé části (např. *ses* na zvratné zájmeno/částici *se* a pomocné sloveso *jsi*), zvratná *se*, u kterých došlo k haplologii, jsou zrekonstruována (viz příklad 10). K dalším rozdílům v počtu elementů dochází na přechodu mezi morfologickou a syntaktickou rovinou, na které je mj. vynechána interpunkce.

(10) Rozhodl se umýt.

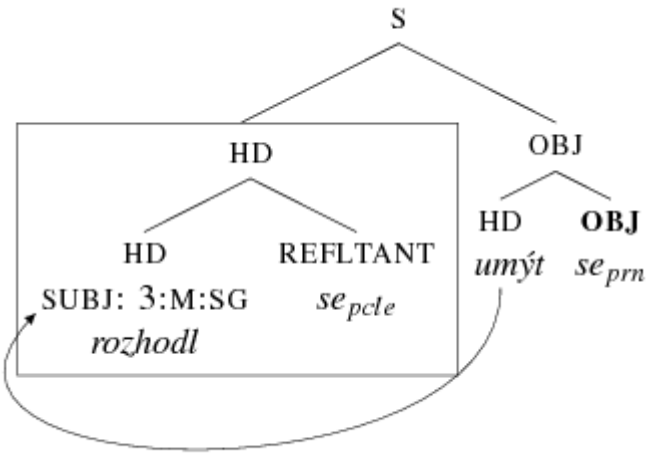
Haplologizovaná forma *se* je zvratná částice náležející ke zvratnému slovesu *rozhodnout se* a zároveň zvratné zájmeno náležející jako předmět ke slovesu *umýt*. Na morfologické rovině tak má dvě interpretace.

(11)

GRAFÉMIKA	rozhodl	se		umýt
MORFOLOGIE	*minulé par.t, mask, sg*	*částice*	*zájm. refl., akuzativ*	*infinitiv*

Dvě interpretace zvratného *se* se projeví jako dva uzly v syntaktické struktuře (příklad 12). Složka v rámečku vyznačuje reflexivní sloveso *rozhodnout se*.

(12)



Snahou je minimalizovat rozdíly v počtu uzlů mezi jednotlivými rovinami, uzly se přidávají jen v nezbytných případech. S výjimkou haplologizovaných tvarů nejsou na syntaktické rovině žádné uzly, které by neměly reprezentaci na rovině grafémické nebo morfologické (např.

elidované větné členy). V původním návrhu syntaktické roviny se počítalo s uzly pro vypuštěné podměty, které měly sloužit jako cíl pro odkazy na podměty doplňků a závislých infinitivů, ale později byla tato strategie přehodnocena a odkazování se řeší jiným způsobem (viz níže).

Kromě odkazů na podmět jsou ve struktuře navrženy i další druhy odkazů, jejichž cílem je zaručit, aby kategorie podléhající shodě sdílely stejné hodnoty a shodující se slovní formy byly ve struktuře dohledatelné. V lineárním zobrazení pak mohou být kongruentní tvary vyznačeny (např. podtržením jako v příkladu 1).

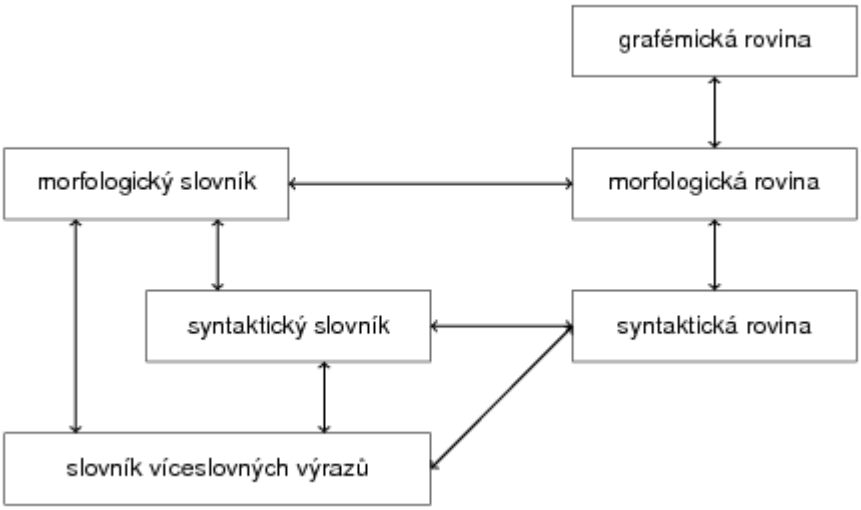
Syntaktická struktura sama o sobě abstrahuje od slovosledu, při specifikaci zobrazení syntaktické struktury je však možné určit, jakým způsobem se mají informace o slovosledu z nižších rovin projevit. Je tedy možné např. přísně zachovávat povrchový slovosled i za cenu nespojitých/neprojektivních struktur, řídit se povrchovým slovosledem až na určené případy, nebo stanovit úplně jiná kritéria pro horizontální uspořádání uzlů v syntaktickém stromě.

2.5. Slovník a gramatika

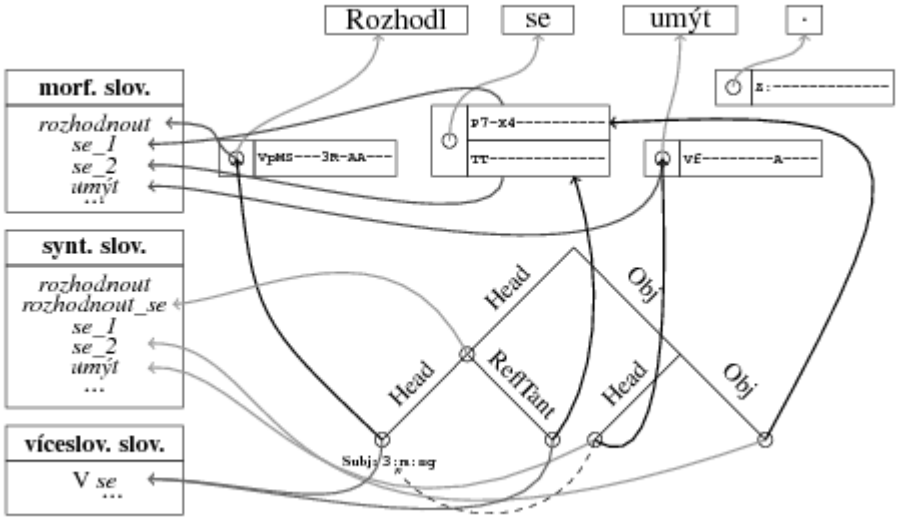
V zájmu konzistentní anotace musí všechny syntaktické struktury v korpusu vyhovovat specifikaci přípustných konstrukcí, definovaných pomocí formální gramatiky. Tento požadavek v sobě zahrnuje i podmínku, že slova a složky mají všechny náležité vlastnosti (potenciálně podspecifikované). Všechna slova z korpusu a také všechny typy složených a víceslovných výrazů jsou obsaženy ve slovníku. Slovník spolu s gramatikou představují abstraktní popis korpusových dat.

2.6. Celkové schéma anotace

Na obrázku 1 je znázorněna celková architektura korpusu a na obrázku 2 je na větě z příkladu 10 předvedeno, jak jsou jednotlivé informace uloženy.



Obrázek 1: Architektura systému



Obrázek 2: Grafické znázornění uložení věty z příkladu 10

3. Kódování anotace

Pro kódování treebanku je na výběr několik možností, mezi kterými není snadné si vybrat. Pro přehled existujících standardů i s jejich hodnocením viz (Przepiórkowski a Bański 2009) a (Bański a Przepiórkowski 2010).

Jako nejsnadnější řešení se jevil formát výstupu stochastického parseru, používaný v systému TectoMT (Popel a Žabokrtský 2010)[1]. Formát vychází z PML (Prague Markup Language), což je anotační jazyk založený na XML, používaný pro anotaci PDT (Prague Dependency Treebank)[2]. Tento jazyk je však schopen zachytit i složkové struktury a dá se upravit i pro jiné účely. Na druhou stranu tento jazyk neumožňuje zachycení některých vlastností námi navržené struktury jako např. distinkci mezi povrchovou a hloubkovou hlavou, podspecifikovanou nebo víceznačnou strukturu, složku bez struktury aj.

Dalším standardem, který přicházel v úvahu, jsou specifikace TEI (Text Encoding Initiative)[3], ale ani ty nenabízejí řešení pro všechny naše požadavky, a standard by bylo nutné rozšířit.

Nakonec jsme přistoupili k vytvoření vlastního anotačního formátu. Stejně jako výše uvedené možnosti je založen na jazyku XML. Reflektuje všechny požadavky plynoucí z navrženého anotačního schématu, včetně výše zmíněných tří rovin (grafémické, morfologické a syntaktické). Na grafémické rovině jsou zachycena slova anotovaná na dalších úrovních. Morfologická rovina obsahuje morfologicky anotovaná slova. Slovo se může skládat z několika textových řetězců (v případě ustrnulých frází bez struktury), nebo naopak může být jedno textové slovo rozděleno na několik slov morfologických. Jeden řetězec výrazů z grafémické roviny může být interpretován více způsoby, přičemž je zajištěno, že slova obsažená v jedné analýze se nepřekrývají s jinou analýzou.

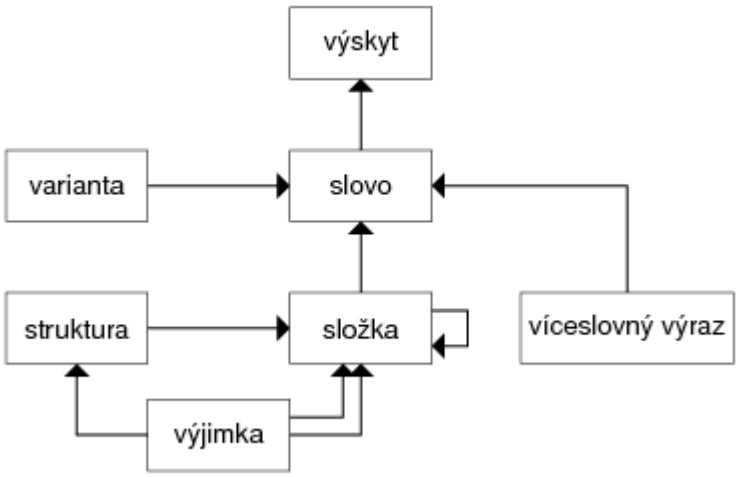
Na obrázku 3 je schématicky znázorněno, jak se zachází s variantami a na obrázku 4 je úryvek kódu XML, který zachycuje větu z příkladu 13, víceznačnou na morfologické i syntaktické rovině.

[1] viz http://ufal.mff.cuni.cz/tectomt

[2] viz http://ufal.mff.cuni.cz/jazz/PML

[3] viz http://www.tei-c.org

(13) Ohlas to.
#1 Měla bys to ohlásit.
#2 Ty jsi to ohla.



Obrázek 3: Zachycení variant

```
<?xml version="1.0" encoding="utf-8"?>
<sentence>
<head />
<graphemics>
<token order="1000" value="ohlas" type="word"/>
<token order="2000" value="to" type="word"/>
</graphemics>
<morfology>
<word id="1" order="1000" wcl="verb.imper" pers="2" num="sg" lemma="ohlásit" />
<word id="2" order="1000" wcl="verb.lpple" gend="fem" num="sg" lemma="ohnout" />
<word id="3" order="1010" wcl="verb.fin" pers="2" num="sg" lemma="být">
<token_ref order="1000"/>
</word>
<word id="4" order="2000" wcl="ppron" case="acc" num="sg" gend="neut" lemma="to" />
<variant id="1">
<word_ref id="1" />
<word_ref id="4" />
</variant>
<variant id="2">
<word_ref id="2" />
<word_ref id="3" />
<word_ref id="4" />
</variant>
</morfology>
<syntax>
<constituent id="1">
<word_ref id="1" function="head" />
<word_ref id="4" function="obj" />
</constituent>
<constituent id="2">
<word_ref id="5" function="sb" />
<constituent_ref id="3" function="head" />
</constituent>
<constituent id="3">
<word_ref id="3" function="shead" />
<constituent_ref id="4" function="dhead" />
</constituent>
<constituent id="4">
<word_ref id="2" function="head" />
<word_ref id="4" function="obj" />
</constituent>
<structure id="1" constituent_ref="1" rating="1000" />
<structure id="2" constituent_ref="2" rating="1000" />
```

```
</syntax>
</sentence>
```
Obrázek 4: XML kód zachycující příklad 13

Syntaktická rovina se skládá ze složek odkazujících ke slovům a dalším složkám v nich obsaženým. Každá složka je označena syntaktickou funkcí, mezi které zahrnujeme i hloubkovou a povrchovou hlavu. Tak jako na morfologické rovině, i zde může být jedna posloupnost slov interpretována několika různými způsoby. Abychom zabránili množení variantních syntaktických struktur v případech, kdy je ve větě několik lokálních víceznačností, může mít vnořená struktura několik variant a zbytek struktury je touto víceznačností nedotčen.

Některé jevy popisované naším schématem, jako např. zájmenné odkazování, odkazování k podmětu infinitivu nebo doplňku nebo shoda, narušují základní složkovou strukturu. Pro jejich zachycení používáme odkazy (z prostorových důvodů nejsou zachyceny na obrázku 2). Důležitou součástí formátu je slovník víceslovných výrazů, ve kterém budou zachyceny analytické slovesné tvary, nespojité víceslovné výrazy a frazémy.

4. Zlepšování výstupu ze stochastického parseru

V současné době jsou výsledky stochastických parserů lepší, než jakých dosahují syntaktické analyzátory založené na pravidlech. Přesto je i chybovost stochastických parserů ve srovnání s taggery stále poměrně vysoká. Proto pracujeme na spojení lingvistických a stochastických metod. Východiskem je lingvistická analýza nejčastějších chyb stochastického parseru jako podklad pro automatickou proceduru, která je opravuje.

Mezi nejčastější chyby, kterých se stochastický parser dopouští a které lze automaticky identifikovat, patří chybná identifikace podmětu ve větě, chybná klasifikace reflexivního zájmene/částice, vyznačení více subjektů v jedné větné klauzi, analýza přívlastku jako doplnění slovesa, záměna předmětu a příslovečného určení, koordinace a vůbec správná identifikace řídících uzlů apod.

Podrobný popis analýzy a opravné procedury je v kapitole [Tomášova kapitola] tohoto svazku.

5. Závěry

Ukázali jsme, jak navrhnout syntakticky anotovaný korpus, který by nebyl svázaný s jednou syntaktickou teorií a umožňoval prezentaci dat v různých formách. K vybudování velkého korpus s přijatelnou mírou chybovosti vytváříme proceduru na automatickou opravu chyb stochastického parseru. Protože výstupem parseru je vždy jen jedna struktura, dalším úkolem je vytvořit nástroj, který by automaticky vyhledával systémové víceznačnosti. Těžištěm další práce však bude další vývoj nástrojů pro využívání syntakticky anotovaného korpusu podle představ uživatele.

Literatura

Bański P., A. Przepiórkowski, 2010, The TEI and the NCP: the model and its application. In *LREC2010 Workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management*, ELRA, Valletta, 34–38.

Hajič J., 2006, Complex Corpus Annotation: The Prague Dependency Treebank, In *Insight into the Slovak and Czech Corpus Linguistics*, ed M. Šimková, Veda, Bratislava, 54–73.

Hajič J., J. Panevová, E. Buráňová, Z. Urešová, A. Bémová, J. Štěpánek, P. Pajas & J. Kárník, 1999, *Anotace na analytické rovině, Návod pro anotátory*. Ústav formální a aplikované lingvistiky MFF UK, Praha. http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/cz/a-

layer/pdf/a-man-cz.pdf.

Holan T., Z. Žabokrtský, 2006, Combining Czech Dependency Parsers. In *Proceedings of the 9th International Conference on Text, Speech and Dialogue*, Springer, Berlin / Heidelberg, 95–102.

Oliva, K. (2001). On retaining ambiguity in disambiguated corpora. *TAL (Traitement Automatique des Langues)*, 42(2).

Popel M., Z. Žabokrtský, 2010, TectoMT: Modular NLP Framework, In *Proceedings of the 7th International Conference on Advances in Natural Language Processing (IceTAL 2010)*, eds H. Loftsson, E. Rögnvaldsson, S. Helgadottir, Springer, Berlin / Heidelberg, 293–304.

Przepiórkowski A., P. Bański, 2009, Which XML standards for multilevel corpus annotation? In *Human Language Technology. Challenges for Computer Science and Linguistics*, ed Z. Vetulani, Springer, Berlin / Heidelberg, 400–411.

Skut W., B. Krenn, T. Brants, H. Uszkoreit, 1997, An annotation scheme for free word order languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing ANLP-97*, Washington.