Průvodce ČAK 2.0

Barbora Vidová Hladká
Jan Hajič
Jirka Hana
Jaroslava Hlaváčová
Jiří Mírovský
Jan Raab
Kiril Ribarov

Průvodce ČAK 2.0

Barbora Vidová Hladká, Jan Hajič, Jirka Hana, Jaroslava Hlaváčová, Jiří Mírovský, Jan Raab a Kiril Ribarov

Obsah

Seznam obrázků

Seznam tabulek

Kapitola 1. Předmluva

Rodina pražských anotovaných korpusů se rozrůstá o dalšího člena, a to o Český akademický korpus verze 2.0, morfologicky a syntakticky ručně anotovaný korpus češtiny (ČAK 2.0). Příznačné je označení *staronový* člen, protože druhé verzi předchází verze první "pouze" s morfologickými anotacemi; první verze byla publikována před dvěma lety a dá se proto v jistém slova smyslu chápat jako stará. To nové s sebou přinášejí syntaktické anotace, jimiž je akademický korpus obohacen dalším přívlastkem příznačným pro pražské korpusy, a sice přívlastkem *závislostní*.

Průvodce ČAK 2.0 je, podobně jako v případě ČAK 1.0, průvodce CD-ROMem. Obsah průvodce je koncipován tak, že čtenář nemusí být předem seznámen s průvodcem ČAK 1.0, a přesto se vše potřebné o projektu dozví. Pokud ho budou zajímat podrobnosti historie projektu Českého akademického korpusu a podrobnosti přípravy první verze, může si samozřejmě průvodce ČAK 1.0 otevřít. Čtenář, který je s průvodcem ČAK 1.0 seznámen, se bude v předkládaném průvodci orientovat velmi snadno, protože jsme se v něm přidrželi stejného členění kapitol do třech tématických celků.

První celek, kapitola 2, podává základní charakteristiku Českého akademického korpusu 2.0, popisuje strukturu anotací v něm obsažených a dokumentuje dílčí kroky spojené se syntaktickými anotacemi. Přehledně uvádí porovnání ČAK 2.0 a Pražského závislostního korpusu 2.0, nejstaršího člena rodiny pražských korpusů.

Druhý celek, kapitoly 3 a 6, se věnuje samotnému CD-ROM, tj. jeho datové komponentě, komponentě nástrojů, bonusů a tutoriálů. V oddíle 3.2 je korpus představen jako datový soubor s vnitřní reprezentací. Přiměřená pozornost je věnována nástrojům pro prohlížení korpusu - Bonito (oddíl 3.3.1) a Netgraph (oddíl 3.3.4); pro editaci anotací - LAW (oddíl 3.3.2) a TrEd (oddíl 3.3.3); pro morfologicko-syntaktické zpracování českých textů (oddíl 3.3.5. Kapitola 4 je mašličkou dvou dárků-bonusů, a to elektronické cvičebnice češtiny STYX a modulu pro hlasové ovládání TrEd (oddíl 4). Ke všem předloženým nástrojům s grafickým rozhraním jsou nabídnuty tutoriály ve formě demosnímků - jejich přehled je uveden v oddíle 5. V kapitole 6 jsou vyjmenovány instrukce pro instalaci jednotlivých komponent CD-ROM..

Kapitoly 8 a 9 třetího celku věnují pozornost personálnímu a finančnímu zabezpečení projektu. Doplněno je pět příloh: příloha A předkládá výčet zdrojů, z kterých byly čerpány texty do korpusu; pro pohodlnou orientaci v morfologických anotacích je předložena příloha B s popisem struktury lemmat a příloha C s popisem struktury morfologických značek; příloha D napomáhá k orientaci v syntaktických anotacích; příloha E je přehledem internetových odkazů, které průvodce doplňují.

CD-ROM je vydáváno v závěrečném roce řešení projektu „Data a nástroje pro informační systémy“, id. č. 1ET101120413, financovaného Grantovou agenturou Akademie věd České republiky. Kolektivu projektu se tak podařilo uceleně prezentovat výsledky dosažené během pěti let řešení.

Kapitola 2. Úvod

2.1. Co je Český akademický korpus 2.0

Český akademický korpus verze 2.0 je morfologicky a syntakticky ručně anotovaným korpusem češtiny o objemu 650 tisíc slov.

Český akademický korpus (ČAK) vznikl před více než dvaceti lety v letech 1971-1985 jako podklad pro sestavení frekvenčního slovníku češtiny té doby – původně nesl zcela „věcný“ název Korpus věcného stylu. Nezávisle na korpusu ČAK byla v roce 1996 zahájena anotace Pražského závislostního korpusu (PZK). Při práci na jeho již druhé verzi [12] [1] se objevila myšlenka převést vnitřní formát a anotační schémata korpusu ČAK tak, aby byl zcela kompatibilní s PZK, tedy aby se dal přímo do PZK začlenit. Konverze vnitřního formátu a morfologického anotačního schématu vyústila publikováním první verze ČAK (Vidová Hladká a kol., 2007). Představovaná druhá verze obohacuje ČAK 1.0 v tom smyslu, že obsahuje navíc převedené syntakticko-analytické anotace.

ČAK 2.0 nabízí

- jazykovědcům-teoretikům datový materiál, který reflektuje reálné použití jazyka,
- počítačovým lingvistům nástroje a další data nezanedbatelného objemu, která by měla přispět ke zlepšení výsledků aplikací přirozeného jazyka, jenž se bez morfologického a syntaktického zpracování textů neobejdou,
- uživatelům anotačního nástroje TrEd možnost ovládat tento nástroj hlasem,
- pedagogům, jejich žákům a studentům zajímavou pomůcku do hodin češtiny, při kterých se procvičuje české tvarosloví.

2.2. Zdroje textů

Dokumenty v ČAK jsou nezkrácené články z širokého spektra novin a časopisů a nezkrácené přepisy mluvené řeči z řady rozhlasových a televizních pořadů, a to z oblasti novinářské, vědecké a administrativní. Texty pocházejí ze 70. a 80. let 20. století. Úplný výčet použitých zdrojů je uveden v příloze A.

2.3. Roviny anotace

O anotovaném korpusu se nedá hovořit, aniž by se specifikovalo, čeho se anotace týkají. Jinými slovy, z pohledu jazykovědné teorie, musí se specifikovat tzv. rovina anotace. Anotace ČAK 2.0 pokrývají dvě roviny – morfologickou a syntakticko-analytickou. Abychom byli úplně korektní vzhledem k vnitřnímu formátu ČAK 2.0 (viz kapitola 3, oddíl 3.2.1), musíme doplnit, že operujeme ještě s jednou rovinou, a to s rovinou slovní. Slovní rovina je ve skutečnosti rovinou neanotační (pro pohodlí o ní budeme nadále hovořit jako o rovině anotační), obsahuje pouze původní text rozdělený na slovní jednotky (slova, čísla zapsaná ciframi, interpunkce), popř. dokumenty a odstavce. Slovním jednotkám jsou přiřazeny jednoznačné identifikátory. Věty nebyly zmíněny zcela záměrně, protože slovní rovina neobsahuje segmentaci textu na věty; ta je až na morfologické rovině. Nadále budeme slovní rovinu zkráceně označovat jako w-rovinu (z anglického *word*), morfologickou rovinu jako m-rovinu a syntakticko-analytickou rovinu jako a-rovinu.

Anotace na m-rovině znamená, že slovním jednotkám textu jsou přiřazovány údaje (anotace), které charakterizují jejich morfologické vlastnosti, tedy lemma (základní tvar slova), slovní druh a morfologické kategorie (pád, číslo, čas, osoba, ...). Formálně jsou slovní druhy společně s morfologickými

[1] Vedle bibliografických citací uvádíme v textu i citace internetové – číslo v hranaté závorce, které odkazuje do seznamu internetových adres vyjmenovaných v příloze E.

kategoriemi reprezentovány jako znakové řetězce, tzv. morfologické značky nebo také tagy. V ČAK 2.0 jsou použity značky navržené v PZK jako řetězce pevné délky, a to délky 15 znaků, kde každá pozice jednoznačně odpovídá právě jedné kategorii – hovoříme o tzv. pozičních značkách; jejich popis je k nahlédnutí v příloze C.

Příklad: Slovní forma *Prahu* se analyzuje jako afirmativní (11. pozice) substantivum (1. a 2. pozice) ženského rodu (3. pozice) ve tvaru akuzativu (5. pozice) singuláru (4. pozice). Na všech ostatních pozicích je správně symbol „-“, který signalizuje nerelevantnost příslušné morfologické kategorie danému slovnímu druhu. Například u substantiv se neurčuje osoba (8. pozice).

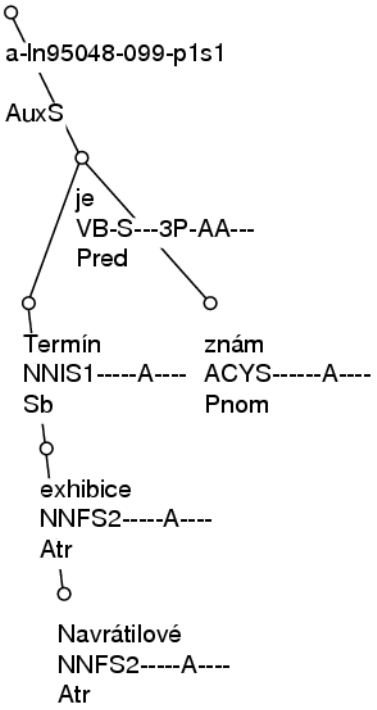
Tabulka 2.1. Příklady lemmat a značek

slovní jednotka	lemma	značka	popis
Prahu	Praha	`NNFS4-----A----`	substantivum, femininum, singulár, akuzativ, afirmativum
123	123	`C=-------------`	číslovka zapsaná číslicemi
)	)	`Z:-------------`	interpunkční znaménko (pravá kulatá závorka)

Anotace na a-rovině znamená, že slovním jednotkám jsou přiřazeny anotace, které charakterizují jejich syntaktické vlastnosti, tedy jejich závislost na ostatních členech věty (tzv. větný rozbor) a jejich funkci ve větě. Formálně jsou závislosti ve větě reprezentovány závislostním stromem. Funkce slovní jednotky ve větě je vyjádřena tzv. analytickou funkcí, jejichž seznam je uveden v příloze D.

Příklad: Obrázek 2.1 zobrazuje syntaktickou anotaci věty *Termín exhibice Navrátilové je znám.* Výsledný strom obsahuje tolik uzlů, kolik je slovních jednotek ve větě, tedy v našem příkladě pět. Abychom byli úplně korektní, tak musíme dodat, že z technických důvodů obsahuje každý strom navíc jeden uzel, který je technickým kořenem stromu. Koncepce anotace vychází z tradice pražské lingvistické školy, která chápe predikát (přísudek; nejčastěji sloveso) jako hlavní člen věty. Proto sloveso *je* je kořenem stromu, který je zavěšen na technickém kořeni. Na kořeni jsou závislé dva větné členy - *termín* a *znám*. Všimněte si, že u každého uzlu stromu na obrázku je zobrazena slovní jednotka, její morfologická značka a její analytická funkce. Zastavíme-li se u uzlu jednotky *termín*, vidíme, že se jedná o podstatné jméno rodu mužského neživotného, v jednotném čísle a v prvním pádě a že tato jednotka je podmětem (Subj) věty.

Obrázek 2.1. Ukázka anotace na a-rovině

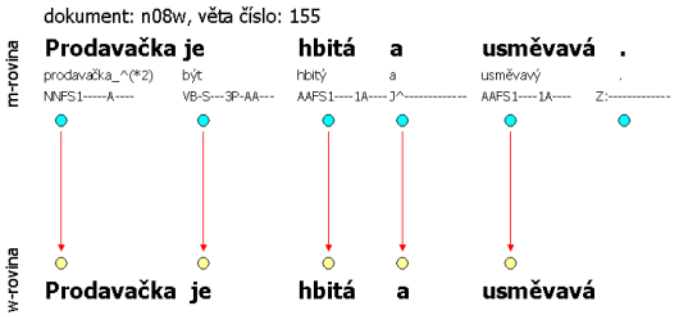


Koncepce vnitřního formátu ČAK 2.0 zachází s anotacemi na rovinách odděleně, tj. každé rovině anotace dokumentu odpovídá jeden soubor. Vztaženo na ČAK 2.0 to znamená, že pro každý dokument existují tři soubory, jeden pro w-rovinu, druhý pro m-rovinu a třetí pro a-rovinu. Nicméně zmíněná oddělenost neznamená, že soubory pro jednotlivé roviny anotace nejsou propojené. Jak vzápětí ukážeme, je tomu právě naopak.

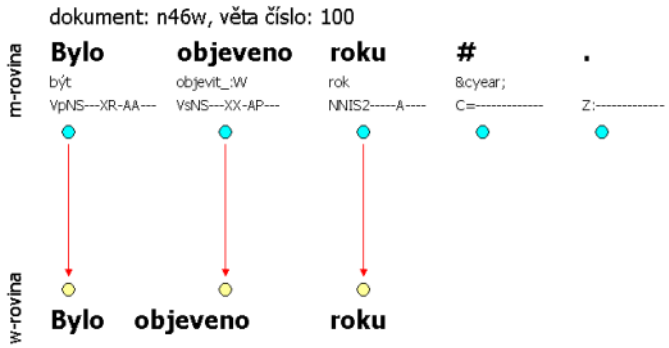
Jak už bylo naznačeno výše, text je segmentován do vět až na m-rovině. To znamená, že m-rovina obsahuje navíc oproti w-rovině koncovou (větnou) interpunkci. Kromě toho se může lišit i počet slovních jednotek na obou rovinách, což může znamenat spojení nesprávně rozdělených slov do jednoho nebo naopak rozdělení omylem spojených slov do více jednotek. Na m-rovině už by měl být správně napsaný text.

Příklad: Následující tři obrázky dokládají propojenost w-roviny a m-roviny, tedy i souborů, ve smyslu počtu slovních jednotek (propojenost naznačují šipky). Všechny tři příklady jsou úmyslně vybrány z ČAK 2.0, aby mohl uživatel přímo nahlédnout do souborů (pro každou větu je uveden název dokumentu a číslo věty). Obrázek 2.2 ilustruje poměr 1:1 – až na koncovou interpunkci se roviny neliší. Obrázek 2.3 ilustruje situaci, kdy byla do textu vložena slovní jednotka – zde evidentně v textu chybělo určení roku. Pro korektora bylo téměř nemožné doplnit konkrétní rok, proto je uveden symbol „#“, který nemá svůj „protipól“ na w-rovině. Naopak obrázek 2.4 ilustruje situaci, kdy více jednotek m-roviny má stejný „protipól“ na w-rovině – slovní jednotka *pedagogicko-psychologické* je rozdělena na tři samostatné jednotky.

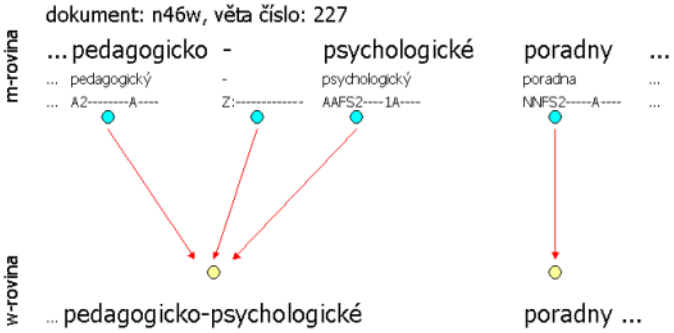
Obrázek 2.2. Technická propojenost w-roviny a m-roviny: žádné změny, až na větnou interpunkci



Obrázek 2.3. Technická propojenost w-roviny a m-roviny: vložení slovní jednotky



Obrázek 2.4. Technická propojenost w-roviny a m-roviny: rozdělení slovní jednotky



Propojenost mezi a-rovinou a m-rovinou znamená, že každé slovní jednotce m-roviny odpovídá právě jeden uzel závislostního stromu a-roviny a naopak, až na jednu výjimku, a tou je technický kořen, který nemá svůj protějšek na m-rovině. Výše uvedený obrázek 2.1 dokládá popsanou propojenost.

2.4. Vývoj projektu

Projekt Českého akademického korpusu prochází nejen staletími, ale i tisíciletími, jak je podrobně popsáno v příspěvku (Hladká, Králík, 2006). Cestě, která vyústila vydáním první verze akademického korpusu, se zde věnovat nebudeme. Je jí věnován Průvodce ČAK 1.0 (Vidová Hladká a kol., 2007). Zde zrekapitulujeme cestu k druhé verzi, a to pro každou anotační rovinu zvlášť.

2.5. Na cestě k ČAK 2.0: m-rovina

Rozsáhlé poloautomatické kontroly morfologické anotace byly navrženy již při přípravě ČAK 1.0. Kontroly byly motivovány obdobnými kontrolami, které probíhaly při tvorbě Pražského závislostního korpusu 2.0. Jejich podrobný popis byl podán v průvodci ČAK 1.0.

Při přípravě dat pro ČAK 2.0 byly provedeny další poloautomatické kontroly morfologické anotace. Automatické skripty procházely data a označovaly podezřelá místa; ta pak ručně prošel, zkontroloval a případně opravil anotátor. Jednalo se především o kontroly shody jednotlivých morfologických kategorií mezi původní morfologickou značkou ČAK a poziční morfologickou značkou ČAK 1.0. Například v pádu u podstatných jmen skripty našly 1258 podezřelých pozic, ze kterých anotátor našel 332 pozice chybné (a opravil). V pádu přídavných jmen našly skripty 177 podezřelých pozic, z nich byla 41 pozice opravena.

2.6. Na cestě k ČAK 2.0: a-rovina

V případě syntaktických anotací jsme stáli před otázkou, jakým způsobem mapovat původní anotace do anotací koncipovaných v projeku Pražského závislostního korpusu. Otázku *Jakým způsobem?* jsme dle zkušeností s morfologickými anotacemi převedli na tři podotázky, a sice *Automaticky? Poloautomaticky? Ručně?* Hledání odpovědí je podrobně popsáno v příspěvku (Ribarov, Bémová, Hladká, 2006). Autoři příspěvku došli k závěru, který možná mnohé čtenáře překvapí: zcela odhlédnout od původních anotací, ČAK 1.0 (tedy ručně morfologicky anotované texty)) zpracovat automatickou procedurou (tzv. parser), která každé větě přiřadí závislostní strom s určenými větnými členy, a následně stromy s větnými členy ručně zkonrolovat (i nadále říkáme anotovat). Použili jsme *maximum spanning tree* parser (MST parser), o kterém informuje podrobněji dále (viz 3.3.5).

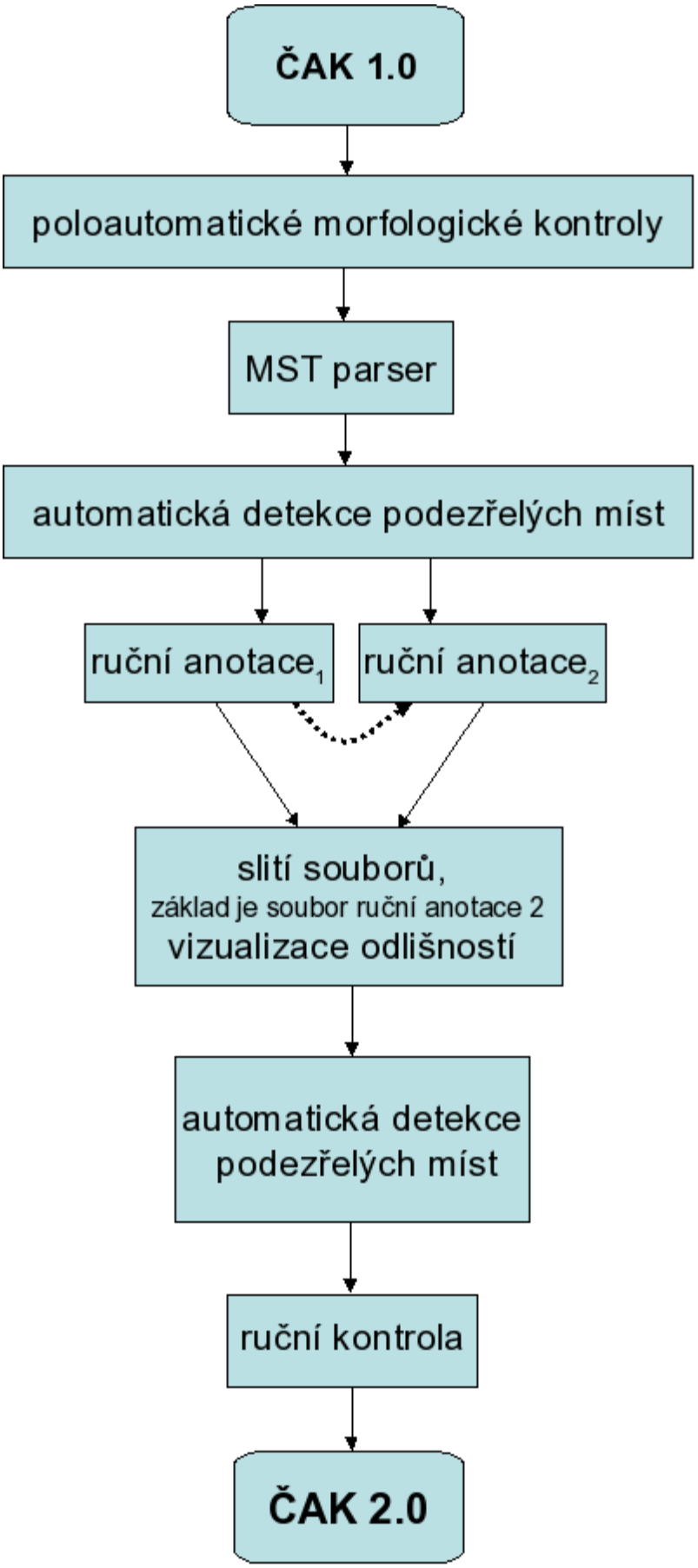
Tímto závěrem vyvstala další otázka ohledně stupně kontroly automaticky vygenerovaných stromů a klasifikovaných větných členů. Na syntakticko-analytických anotacích Pražského závislostního korpusu pracovali přímo jazykovědci. Z této skupiny byla pro náš projekt k dispozici jedna anotátorka, která se stala hlavním arbitrem. Dále byli k dispozici studenti filologických oborů - dvě české anotátorky a tři slovenští anotátoři, kteří měli za sebou zkušenost z anotování Slovenského národního korpusu pod vedením pražských jazykovědců ´natrénovaných´na PZK. Výsledně byla anotace ČAK dvoustupňová: anotátor, arbitr. Z počátku pracovali anotátoři paralelně, tj. jeden dokument byl anotován dvěma anotátory. Jejich anotace byly automaticky porovnány a podstoupeny arbitrovi. Jakmile se anotátoři uspokojivě (dle arbitra) zacvičili, anotovali každý dokument právě jednou. Při druhém stupni kontroly, arbitr procházel celý dokument větu po větě, tj. v případě paralelních anotací se nesoustředil pouze na odlišnosti. Mezi jednotlivými stupni anotací byly dokumenty zpracovány automatickými kontrolními skripty.

Stejně jako u morfologických anotací byly automatické skripty inspirovány obdobnými kontrolami prováděnými při přípravě PZK 2.0. Skripty procházely data a označovaly podezřelá místa. Kontrolovala se jednak přípustnost některých vztahů několika uzlů na analytické rovině, jednak přípustnost kombinace morfologické značky a analytické značky jednotlivých uzlů. Podezřelá místa byla označena a anotátoři je při práci se stromy viděli zvýrazněna, spolu se stručným popisem nesrovnalosti. Chyba pak mohla být jak na analytické rovině, tak někdy i na rovině morfologické.

Příkladem analyticko-morfologické kontroly může být jeden ze skriptů, který ověřoval anotaci slovní formy `se`. Skript ověřil u každého takového uzlu podmínku: Každý uzel se slovní formou `se` je buď zvratné zájmeno s analytickou funkcí `AuxT` nebo `AuxR`, nebo je to vokalizovaná předložka s analytickou funkcí `AuxP`. Další skripty ověřovaly shodu morfologických kategorií nebo přípustnou kombinaci analytických funkcí dvou uzlů, řídícího a závislého (jako např. vazbu předložky a pád podstatného jména, které na ní visí, nebo možné umístění uzlu označeného jako Subjekt apod.). Uzly označené jako `AuxP` nebo `AuxC` by neměly být bezdětné (až na speciální případy) atd.

Obrázek 2.5 souhrnně ukazuje, jaké operace probíhaly na datech od vydání ČAK 1.0 až do vydání ČAK 2.0.

Obrázek 2.5. Práce s daty při přípravě ČAK 2.0


ČAK 1.0
poloautomatické morfologické kontroly
MST parser
automatická detekce podezřelých míst
ruční anotace1
ruční anotace2
slití souborů,
základ je soubor ruční anotace 2
vizualizace odlišností
automatická detekce
podezřelých míst
ruční kontrola
ČAK 2.0

2.7. Kvantitativní údaje

V tabulce 2.2 jsou souhrnně uvedeny kvantitativní charakteristiky korpusu ČAK 2.0. Ještě podrobnější údaje jsou uvedeny dále v tabulce 3.8.

Tabulka 2.2. Kvantitativní charakteristiky ČAK 2.0

styl	forma zdroje	počet souborů	počet vět	počet slovních jednotek
publicistický	psaná	52	10 234	189 435
publicistický	mluvená	8	1433	28 737
odborný	psaná	68	11 113	245 175
odborný	mluvená	32	4576	115 853
administrativní	psaná	16	3362	58 697
administrativní	mluvená	4	989	14 235
celkem	**psaná**	**136**	**24 709**	**493 307**
celkem	**mluvená**	**44**	**6998**	**158 825**
celkem	**psaná a mluvená**	**180**	**31 707**	**652 132**

Tabulka 2.3. Kvantitativní charakteristiky ČAK 2.0 – vložené symboly

styl	forma	počet výskytů '#' (v počtu vět)	počet výskytů '?' (v počtu vět)	počet výskytů '#' nebo '?' (v počtu vět)	počet vět bez '#' nebo '?'
publicistický	psaná	1769 (1187)	925 (680)	2694 (1563)	8671
publicistický	mluvená	5 (5)	25 (25)	30 (30)	1403
odborný	psaná	2149 (1222)	2230 (1418)	4379 (2030)	9083
odborný	mluvená	9 (9)	131 (108)	140 (113)	4463
administrativní	psaná	901 (611)	635 (476)	1536 (915)	2447
administrativní	mluvená	0 (0)	16 (15)	16 (15)	974

Vzhledem k motivaci přičlenit ČAK k PZK předkládáme tabulku 2.4 základního srovnání obou korpusů. K přímému začlenění ČAK do PZK dojde někdy v budoucnu při vydání další verze PZK.

Tabulka 2.4. Srovnání ČAK 2.0 a PZK 2.0

charakteristika	**PZK 2.0**		**ČAK 2.0**	
	počet slov (mil.)	**počet vět (tis.)**	**počet slov (tis.)**	**počet vět (tis.)**
anotace morfologická	2	116	116	31
anotace syntaktická	1,5	88	650	31
anotace větná	1,5	88	650	31
anotace tektogramatická	0,8	49	--	--
	dokumenty		dokumenty	
texty písemné	100%		75%	
texty mluvené	--		25%	
	dokumenty		dokumenty	
texty novinové	81%		33%	
texty administrativní	--		11%	
texty odborné	9%		56%	

Kapitola 3. CD ROM Český akademický korpus 2.0

3.1. Adresářová struktura

V této části nabízíme popis adresářové struktury CD, a to až do druhé, příp. třetí úrovně zanoření, viz tabulka 3.1. Pokud v textu odkazujeme na obsah CD, který je zanořen hlouběji, explicitně na to upozorňujeme uvedením úplné cesty.

Tabulka 3.1. Adresářová struktura CD-ROM ČAK 2.0

`index.html`	# průvodce ČAK 2.0 česky (html formát)
`index-en.html`	# průvodce ČAK 2.0 anglicky (html formát)
`Install-on-Linux.pl`	# instalační skript pro Linux (anglicky)
`Install-on-Windows.exe`	# instalační program pro MS Windows (anglicky)
`Instaluj-na-Linuxu.pl`	# instalační skript pro Linux (česky)
`Instaluj-na-Windows.exe`	# instalační program pro MS Windows (česky)
`bonus-tracks/`	# bonusový materiál
`  STYX/`	# elektronická cvičebnice češtiny
`data/`	
`  csts/`	# ČAK 2.0 ve formátu CSTS (soubory `[ans][0-9][0-9][sw].csts`)
`  pml/`	# ČAK 2.0 ve formátu PML (soubory `[ans][0-9][0-9][sw].[amw]`)
`  schemas/`	# PML schémata
`doc`	
`  cac-guide/`	# průvodce ČAK 2.0 česky a anglicky (pdf formát)
`tools/`	# nástroje
`  Bonito/`	# korpusový manažer
`  Java/`	# Java Runtime Environment 6 Update 3 pro Linux a MS Windows
`  LAW/`	# anotační morfologický editor
`  TrEd/`	# anotační syntaktický editor včetně modulu TrEdVoice pro hlasové ovládání
`  Netgraph/`	# korpusový prohlížeč
`  tool_chain/`	# automatické zpracování českých textů
`    tool_chain`	# skript pro spouštění tokenizace a/nebo morf. analýzy a/nebo tagování a/nebo parsingu
`    ...`	
`tutorials/`	# tutoriály k nástrojům

3.2. Data

Organizace ČAK 2.0 do souborů, jejichž jména podléhají jisté konvenci, a samotná vnitřní reprezentace souborů jsou předmětem následující sekce.

3.2.1. Formát dat

Hlavním datovým formátem ČAK 2.0 je formát nazvaný *Prague Markup Language* (PML), založený na XML a navržený pro bohatou reprezentaci lingvistické anotace textů. Každé zvolené rovině anotace odpovídá jeden samostatný soubor. Návrh PML probíhal souběžně s tektogramatickou anotací PZK 2.0.

Vedlejším datovým formátem ČAK 2.0 je formát *CSTS*. Jde o formát SGML používaný v PZK 1.0 a rovněž v Českém národním korpusu. Důvody jeho použití v ČAK 2.0 jsou jeho snadná čitelnost člověkem, jeho snadné zpracování jednoduchými nástroji a rovněž to, že některé z nástrojů ČAK 2.0 pracují výhradně s CSTS. K dispozici je též nástroj pro převod mezi těmito dvěma formáty.

Následující oddíl obsahuje stručný přehled hlavních vlastností formátu PML; podrobné informace jsou publikovány v technické zprávě (Pajas, Štěpánek, 2005). V dalším oddíle uvádíme stručný přehled hlavních vlastností formátu CSTS. Podrobnější informace je možno nalézt v dokumentaci PZK 2.0 [10].

3.2.1.1. Formát PML

V PML se mohou jednotlivé oddělené roviny anotace překrývat a mohou být konzistentně propojeny jak mezi sebou, tak i s dalšími zdroji dat. Každá rovina anotace je popsána v souboru *PML schéma*, který je jakousi formalizací abstraktního anotačního schématu pro tu kterou rovinu anotace. PML schéma popisuje, které elementy se na dané rovině vyskytují, jak jsou spojovány, vnořovány a strukturovány, hodnoty jakého typu se v nich mohou vyskytovat a jakou roli hrají v anotačním schématu (tato informace o tzv. *PML-roli* může být využívána i aplikacemi ke správnému určení způsobu zobrazení PML dat uživateli). Z PML schématu mohou být automaticky generována další schémata, jako je Relax NG, díky čemuž může být konzistence dat ověřena pomocí běžných nástrojů pro XML. Obě verze schémat jsou k dispozici v adresáři `data/schemas/`. Pro ilustraci uvádíme v tabulce 3.2 část PML schématu w-roviny dat ČAK (`data/schemas/wdata_schema.xml`), která specifikuje, že odstavec (typ `para`, v případě ČAK 2.0 vždy celý dokument) sestává z posloupnosti elementů typu `w-node.type`; tento typ je níže definován jako struktura obsahující mimo jiné dva povinné elementy: `id` (jednoznačný identifikátor s rolí `#ID`) a `token` (slovní jednotku).

Tabulka 3.2. PML schéma w-roviny ČAK 2.0

```
<type name="w-para.type">
    <sequence>
        <...
        <element name="w" type="w-node.type"/>
    </sequence>
</type>
<type name="w-node.type">
    <structure name="w-node">
        <member as_attribute="1" name="id" role="#ID" required="1">cdata format
        <member name="token" required="1"><cdata format="any"/></member>
        <member name="no_space_after" type="bool.type"/>
    </structure>
</type>
...
```

Každý PML soubor začíná hlavičkou odkazující na PML schéma souboru. V hlavičce jsou uvedeny všechny externí zdroje, na které je z tohoto souboru odkazováno, spolu s několika dalšími informacemi, potřebnými pro správné vyhodnocení odkazů. Zbytek souboru obsahuje vlastní anotaci. Část hlavičky souboru m-roviny (`n01w.m`), kde se odkazuje na PML-schéma tohoto souboru (`mdata_schema.xml`) a na příslušný soubor w-roviny (`n01w.w`), uvádíme jako příklad v tabulce 3.3.

Tabulka 3.3. Část hlavičky souboru `n01w.m`

```
<head>
    <schema href="mdata_schema.xml" />
    <references>
        <reffile id="w" href="n01w.w" name="wdata" />
    </references>
</head>
...
```

Obdobně tabulka 3.4 ukazuje referenční část hlavičky souboru a-roviny (`n01w.a`), kde se odkazuje na PML-schéma tohoto souboru (`adata_schema.xml`) a na příslušný soubor m-roviny (`n01w.m`) a w-roviny (`n01w.w`):

Tabulka 3.4. Část hlavičky souboru `n01w.a`

```
<head>
    <schema href="adata_schema.xml" />
    <references>
        <reffile id="m" href="n01w.m" name="mdata" />
        <reffile id="w" href="n01w.w" name="wdata" />
    </references>
</head>
...
```

Anotace je vyjádřena pomocí XML elementů a atributů, pojmenovaných a použitých v souladu s příslušným PML schématem. Pro ilustraci uvádíme v tabulce 3.5 příklad anotace části věty `Váš boj je i naším bojem.` na m-rovině. Otvírací značka elementu `s` obsahuje identifikátor celé věty, stejně tak otvírací značky elementu `m` obsahují identifikátory dané anotace odpovídajících slovních jednotek w-roviny, na které se odkazuje z elementu `w.rf`. Další elementy obsahují formu (`form`), morfologickou značku (`tag`) a lemma (`lemma`) a element `src.rf` udává zdroj anotace, v tomto případě ruční.

Tabulka 3.5. Ukázka anotace věty na m-rovině ve formátu PML

```
<s id="m-n01w-s14">
    <m id="m-n01w-s14W1">
        <src.rf>manual</src.rf>
        <w.rf>w#w-n01w-s14W1</w.rf>
        <form>Váš</form>
        <lemma>tvůj_^(přivlast.)</lemma>
        <tag>PSYS1-P2-------</tag>
    </m>
    <m id="m-n01w-s14W2">
        <src.rf>manual</src.rf>
        <w.rf>w#w-n01w-s14W2</w.rf>
        <form>boj</form>
        <lemma>boj</lemma>
        <tag>NNIS1-----A----</tag>
    </m>
    <m id="m-n01w-s14W3">
        <src.rf>manual</src.rf>
        <w.rf>w#w-n01w-s14W3</w.rf>
        <form>je</form>
        <lemma>být</lemma>
        <tag>VB-S---3P-AA---</tag>
    </m>

    ...

    <m id="m-n01w-s14W7">
        <src.rf>manual</src.rf>
        <form_change>insert</form_change>
        <form>.</form>
        <lemma>.</lemma>
        <tag>Z:-------------</tag>
    </m>
</s>
```

Tabulka 3.6 ukazuje příklad anotace věty `Váš boj je i naším bojem.` na a-rovině. Pro přehlednost jsou vynechány méně důležité elementy. Závislostní struktura věty je zachycena ve struktuře vnořovaných elementů. Synovské uzly jsou obaleny elementem `children`. Každý uzel je dále obalen elementem `LM`, jehož atributem je identifikátor tohoto uzlu; výjimkou jsou jednoprvkové seznamy uzlů, kde tento element může být vynechán, identifikátor uzlu je pak atributem elementu `children`. Element `m.rf` odkazuje na příslušný prvek nižší roviny, element `afun` obsahuje analytickou funkci uzlu. Element `ord` obsahuje pořadí uzlu ve stromu zleva doprava.

Tabulka 3.6. Ukázka anotace věty na a-rovině ve formátu PML

```
<LM id="a-n01w-s14">
    <s.rf>m#m-n01w-s14</s.rf>
    <afun>AuxS</afun>
    <ord>0</ord>
    <children>
        <LM id="a-n01w-s14W3">
            <afun>Pred</afun>
            <m.rf>m#m-n01w-s14W3</m.rf>
            <ord>3</ord>
            <children>
                <LM id="a-n01w-s14W2">
                    <afun>Sb</afun>
                    <m.rf>m#m-n01w-s14W2</m.rf>
                    <ord>2</ord>
                    <children id="a-n01w-s14W1">
                        <afun>Atr</afun>
                        <m.rf>m#m-n01w-s14W1</m.rf>
                        <ord>1</ord>
                    </children>
                </LM>
                <LM id="a-n01w-s14W6">
                    <afun>Pnom</afun>
                    <m.rf>m#m-n01w-s14W6</m.rf>
                    <ord>6</ord>
                    <children id="a-n01w-s14W5">
                        <afun>Atr</afun>
                        <m.rf>m#m-n01w-s14W5</m.rf>
                        <ord>5</ord>
                        <children id="a-n01w-s14W4">
                            <afun>AuxZ</afun>
                            <m.rf>m#m-n01w-s14W4</m.rf>
                            <ord>4</ord>
                        </children>
                    </children>
                </LM>
            </children>
        </LM>
        <LM id="a-n01w-s14W7">
            <afun>AuxK</afun>
            <m.rf>m#m-n01w-s14W7</m.rf>
            <ord>7</ord>
        </LM>
    </children>
</LM>
```

XML elementy všech souborů patří do vyhrazeného jmenného prostoru `http://ufal.mff.cuni.cz/pdt/pml/` (pouze název jmenného prostoru, nejedná se o smysluplný odkaz). Formát PML poskytuje jednotnou reprezentaci většiny běžných anotačních konstrukcí, jako jsou struktury atribut-hodnota, seznam alternativních hodnot určitého typu (atomického nebo dále strukturovaného), odkazy v rámci PML souboru, odkazy mezi různými PML soubory (v ČAK 2.0 použité k odkazům mezi rovinami) nebo do dalších externích zdrojů typu XML.

3.2.1.2. Formát CSTS

V CSTS formátu jsou všechny roviny anotace uchovány v jednom souboru.

CSTS soubor začíná (nepovinnou) hlavičkou (element `h`) a dále obsahuje alespoň jeden element `doc`. Element `doc` sestává z hlavičky (element `a`) a obsahu (element `c`). Element `c`) pak sestává z posloupnosti odstavců (element `p`) a vět v těchto odstavcích (element `s`).

Každá slovní jednotka věty je na samostatném řádku souboru (element `f`, resp. `d` pro interpunkci), dále na tomto řádku následuje anotace této slovní jednotky na všech rovinách. Element `l` obsahuje lemma slovní jednotky, element `t` obsahuje její morfologickou značku. Element `A` obsahuje analytickou značku slovní jednotky. Jednoznačný identifikátor slovní jednotky v rámci věty je uložen v elementu `r`, na nějž odkazují hodnoty elementu `g`, které obsahují identifikátor řídícího uzlu slovní jednotky.

Tabulka 3.7 ukazuje, jak ve formátu CSTS vypadá kompletní anotace věty `Váš boj je i naším bojem.`

Tabulka 3.7. Ukázka anotace věty ve formátu CSTS.

```
<s id=n01w-s14>
<f id=n01w-s14W1>Váš<l>tvůj_^(přivlast.)<t>PSYS1-P2-------<r>1<g>2<A>Atr
<f id=n01w-s14W2>boj<l>boj<t>NNIS1-----A----<r>2<g>3<A>Sb
<f id=n01w-s14W3>je<l>být<t>VB-S---3P-AA---<r>3<g>0<A>Pred
<f id=n01w-s14W4>i<l>i<t>J^-------------<r>4<g>5<A>AuxZ
<f id=n01w-s14W5>naším<l>můj_^(přivlast.)<t>PSZS7-P1-------<r>5<g>6<A>Atr
<f id=n01w-s14W6>bojem<l>boj<t>NNIS7-----A----<r>6<g>3<A>Pnom
<D>
<d id=n01w-s14W7>.<l>.<t>Z:-------------<r>7<g>0<A>AuxK
```

DTD soubor pro formát CSTS se nachází v adresáři `data/schemas/`. Podrobný popis tohoto formátu je možno nalézt v dokumentaci PZK 2.0 [10].

V adresářích `tools/tool_chain/csts2pml/` a `tools/tool_chain/pml2csts/` jsou k dispozici převodní skripty mezi oběma fomáty.

3.2.2. Konvence pojmenování souborů

Každý datový soubor ČAK 2.0 odpovídá jednomu dokumentu anotovanému na jedné rovině anotace. První znak jména souboru indikuje styl textu: `n` označuje novinové články (publicistiku), `s` označuje vědecké texty (odborný styl), `a` označuje texty administrativní. Následuje dvoumístné pořadové číslo dokumentu v rámci skupiny dokumentů jednoho stylu; písmeno za číslem udává, zda jde o původně psaný text (písmeno `w`) nebo o přepis mluvené řeči, tzv. mluvený text (písmeno `s`). Jména souborů jsou také součástí identifikátorů vět a prvků vět, obsažených v těchto souborech, např. `<m id="m-n01w-s1W1">` v tabulce 3.5. V příloze A jsou uvedena jména souborů pro jednotlivé dokumenty.

Příklad: Soubory se jménem podle šablony `a[0-9][0-9]s*` obsahují přepisy mluvené řeči s administrativním obsahem.

Přípona souboru vyjadřuje rovinu anotace dokumentu. Přípona `.w` označuje w-rovinu, `.m` označuje m-rovinu a `.a` označuje a-rovinu. Mluvíme potom o w-souborech, m-souborech a a-souborech. Ke každému a-souboru existuje právě jeden m-soubor a právě jeden w-soubor. Z každého a-souboru vedou odkazy do příslušného m-souboru a w-souboru a z každého m-souboru vedou odkazy do příslušného w-souboru (viz výše). Z tohoto důvodu by soubory neměly být přejmenovány. Z w-souboru do m-souboru (ani do a-souboru) odkazy nevedou, stejně tak nevedou odkazy z m-souboru do a-souboru.

Příklad: `s17w.a` označuje soubor obsahující anotace na a-rovině odborného psaného dokumentu. Ze souboru vedou odkazy do souborů `s17w.m` a `s17w.w`, ze souboru `s17w.m` vede odkaz do souboru `s17w.w`.

3.2.3. Velikost dat

ČAK 2.0 sestává ze 180 ručně anotovaných textových dokumentů, obsahujících celkem 31 707 vět s 652 132 slovními jednotkami (tyto údaje jako i všechny ostatní jsou počítány z m-souborů). Slovních jednotek bez interpunkce je 570 761. Slovních jednotek bez interpunkce a bez čísel zapsaných číslicemi je 565 928. V tabulce 3.8 jsou uvedeny velikosti jednotlivých částí dat rozdělených podle stylu a podle formy.

Tabulka 3.8. Velikost jednotlivých částí ČAK 2.0 podle stylu a formy

styl	forma	počet souborů	počet vět	počet slovních jednotek	počet slovních jednotek bez interpunkce	počet slovních jednotek bez interpunkce a bez čísel zapsaných číslicemi
publicistický	psaná	52	10 234	189 435	165 469	163 700
publicistický	mluvená	8	1433	28 737	24 864	24 859
odborný	psaná	68	11 113	245 175	216 281	214 132
odborný	mluvená	32	4576	115 853	100 281	100 272
administrativní	psaná	16	3362	58 697	51 431	50 530
administrativní	mluvená	4	989	14 235	12 435	12 435
celkem	**psaná**	**136**	**24 709**	**493 307**	**433 181**	**428 362**
celkem	**mluvená**	**44**	**6998**	**158 825**	**137 580**	**137 566**
celkem	**psaná a mluvená**	**180**	**31 707**	**652 132**	**570 761**	**565 928**

Pro úplnost dodáváme, že každý zveřejněný experiment provedený na datech ČAK 2.0 by měl obsahovat informaci o tom, jaká část dat byla pro jaký účel v experimentu použita.

Souhrnně konstatujeme, že anotace ČAK 2.0 je rozdělena do tří rovin, a to do roviny slovní (w-rovina), morfologické (m-rovina) a analytické (a-rovina). Každá z těchto rovin má vlastní PML schéma (v adresáři `data/schemas/` soubory `wdata_schema.xml`, `mdata_schema.xml`, `adata_schema.xml`). Adresář `data/pml/` obsahuje celkem 540 souborů, a to 180 w-souborů, 180 m-souborů a 180 a-souborů.
Data jsou k dispozici též ve formátu CSTS v adresáři `data/csts/`, který obsahuje celkem 180 souborů.

3.3. Nástroje

Anotování dat, opravy anotací, vyhledávání v anotacích a zpracování dat automatickými procedurami zprostředkovává celá řada nástrojů. Vzhledem k tomu, že ČAK 2.0 je anotovaným korpusem na m-rovině a a-rovině, nabízíme nástroje, které umožňují práci s daty, tedy i s ČAK, právě na těchto dvou rovinách. Tabulka 3.9 je úvodní navigací v nástrojích, které jsou součástí CD-ROM. Pro každý nástroj je uvedena jeho základní charakteristika spolu s tipy, na jaký druh práce je nástroj vhodný. V dalších oddílech pak následují podrobnější popisy nástrojů.

Tabulka 3.9. Přehled nástrojů

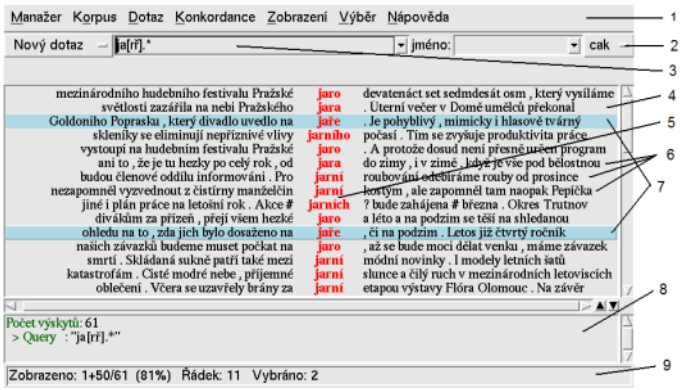
nástroj	popis	tipy
Bonito	korpusový manažer	• vyhledávání v textech ČAK 2.0 • vyhledávání v morfologických anotacích ČAK 2.0 • vyhledávání v anotacích analytických funkcí ČAK 2.0 • statistické výpočty nad ČAK 2.0
LAW	anotační editor	• morfologické anotování (ruční zjednoznačnění morfologické analýzy)
TrEd	anotační editor	• syntaktické anotování (určování větných členů a syntaktických závislostí mezi nimi)
Netgraph	browser	• vyhledávání v syntaktických strukturách ČAK 2.0
tool_chain	automatická procedura pro zpracování českých textů	• tokenizace • morfologická analýza • tagování (automatické zjednoznačnění morf. analýzy) • parsování (automatická syntaktická analýza s určováním větných členů)

3.3.1. Grafický nástroj Bonito

Grafický nástroj Bonito ulehčuje uživatelům práci s jazykovými korpusy, zejména při vyhledávání a při základních statistických výpočtech nad vyhledanými daty. Bonito je nadstavbou korpusového manažeru Manatee, který provádí nejrůznější operace nad korpusovými daty. Podrobná dokumentace k nástroji Bonito je součástí nástroje samotného a vyvolá se z hlavního menu `Nápověda`.

Hlavní obrazovku Bonito ilustruje obrázek 3.1. Základní ovládání nástroje ukážeme na konkrétních příkladech.

Obrázek 3.1. Bonito: hlavní obrazovka



Vysvětlivky k obrázku 3.1

- 1 Hlavní menu

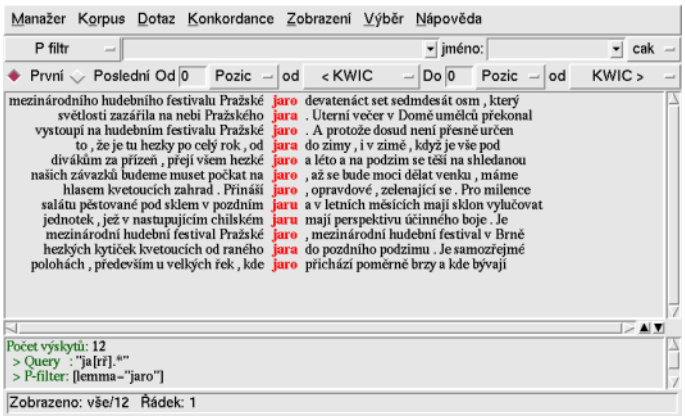
- 2 Tlačítko pro výběr korpusu
- 3 Dotazovací řádek
- 4 Hlavní okno pro zobrazení výsledků dotazu
- 5 Sloupec s výskyty odpovídajícími dotazu
- 6 Konkordanční řádky
- 7 Vybrané konkordanční řádky
- 8 Vedlejší okno pro zobrazení historie dotazu a širšího kontextu
- 9 Stavový řádek

Uživatele často zajímá, v jakých kontextech se slova v korpusu vyskytují. Například ho může zajímat, v jakých kontextech se vyskytuje slovo *jaro*. Zadáním tohoto slova do okénka "3" (dotazovacího řádku) a stiskem klávesy `Enter` se odešle dotaz a vzápětí se odpověď zobrazí v hlavním okně "4" ve formě tzv. konkordancí, tedy výskytů zadaného slova v kontextech tak, jak se nachází v celém korpusu. Zobrazeným řádkům říkáme konkordance nebo také konkordanční řádky ("6").

Jako dotaz může sloužit i jednoduchý regulární výraz. Například všechny tvary slova *jaro* lze získat zadáním regulárního výrazu *ja[rř].** (viz obrázek 3.1). Uvedený dotaz sice vyhledá všechny tvary, které požadujeme, ale může se stát a v tomto případě se stane, že vyhledá i něco nežádoucího. Třeba tvary přídavného jména *jarní* nebo i podstatného jména *jarmark*. Není-li výsledek příliš rozsáhlý a je-li nežádoucích konkordancí málo, můžeme je vybrat kliknutím levého tlačítka myši a poté vymazat pomocí příkazu `Smazání vybraných` z menu `Konkordance` (`Konkordance | Smazání vybraných`). Také lze napřed výběr invertovat, tj. zaměnit vybrané řádky za nevybrané a naopak (`Výběr | Inverze`), a teprve potom invertovaný výběr smazat.

Lepší je ale buď upravit regulární výraz tak, aby výsledkem byly jenom tvary, které uživatel chce (např. *ja(ro|ra|ře|ru|rem|r|rech|ry)*), což může být zbytečně složité, nebo výsledek zúžit pomocí volby `P filtr` (pozitivní filtr) či `N filtr` (negativní filtr). Filtry zvolíme kliknutím na tlačítko `Nový dotaz` a výběrem příslušného filtru. Jestliže se např. do negativního filtru zapíše dotaz *jarn.+*, odstraní se z vyhledaných konkordančních řádků všechny výskyty odpovídající uvedenému regulárnímu výrazu, tedy v našem případě tvary slova *jarní*. Lepším řešením je `P filtr`, do kterého zapíšeme dotaz `[lemma="jaro"]`. Tímto způsobem omezíme vyhledaný výsledek jen na ty výskyty, u nichž se atribut lemma rovná řetězci "jaro" - viz obrázek 3.2. Na stejném obrázku si všimněte historie dotazu, která se zapisuje do spodního okna "8". Samozřejmě jsme mohli dotaz `[lemma="jaro"]` zadat hned na začátku. Výsledek by se ale od právě popsaného postupu lišil. Dosud jsme totiž vyhledávali jen tvary lemmatu *jaro* začínající malým písmenem. Dotaz na lemma vyhledá všechny možné zápisy tvarů lemmatu, včetně těch, které obsahují velké písmeno, např. na začátku věty. Při vyhledávání lemmat je potřeba si uvědomit, že některá lemmata mají k sobě přiřetězeny některé další informace, především rozlišení významů u homonymních lemmat (např. lemma *stát-1_^(státní_útvar)* a *stát-2_^(něco_se_přihodilo)*. Při zadání dotazu `[lemma="stát"]` se nenalezne nic, je třeba dotaz formulovat s pomocí regulárního výrazu `[lemma="stát.*"]`, zvláště když si nejsme jisti, jak přesně lemma vypadá. Vyhledají se nám všechna lemmata, která začínají řetězcem stát, tedy obě výše uvedená, ale i např. lemma *státní*. Požadované konkordance potom získáme upřesněním dotazu pomocí P nebo N filtru (viz výše).

Obrázek 3.2. Bonito: použití P filtru



Jak je zřejmé z dosud uvedených příkladů, je možné formulovat i složité dotazy a kombinovat hodnoty všech atributů, které jsou v korpusu definovány. O které atributy se jedná, zjistíme výběrem položky `Souhrnné informace` v menu `Korpus`. Kromě operátoru = pro rovnost je možno použít i operátoru `!=` pro nerovnost. Popis jednotlivých atributů následuje v tabulce 3.10.

Tabulka 3.10. Bonito: popis atributů ČAK 2.0

jméno atributu	popis
word	slovní forma
lemma	základní slovní forma, lemma
tag	morfologická značka
num	pořadí slovní formy ve větě
afun	analytická funkce
tparentform	přímý rodič
tparentnum	pořadí přímého rodiče ve větě
eparentform	efektivní (lingvistický) rodič
eparentnum	pořadí efektivního (lingv.) rodiče ve větě

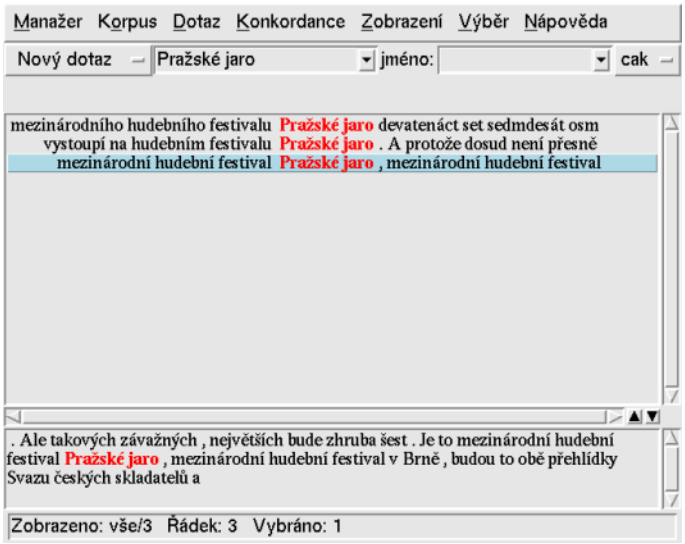
Implicitním atributem, který si však může uživatel sám změnit, je atribut slovo. Proto stačilo v případě dosavadních příkladů zapsat do dotazovacího řádku pouze slovo *jaro*, aniž by se specifikovalo, o který atribut se jedná. K vyhledání je možné použít i kombinaci atributů. Následující dotaz `[lemma="jaro" & tag="NNN.6.+" & word="j.+"]` najde všechny výskyty lemmatu *jaro*, které se vyskytují v 6. pádě (jednotného nebo i množného čísla, protože ve značce je na pozici čísla tečka) a začínají malým písmenem - dotaz na lemma totiž vyhledá i výskyty daného lemmatu na začátku věty, což se nám původním dotazem, kdy jsme hledali slovo, ne lemma, nepodařilo.

Dotazy je třeba vytvářet velmi pečlivě. Vynechání hranatých závorek, uvozovek, přidání mezer, to vše může způsobit, že nenajdete to, co hledáte. Pro názornější a bezchybné vytváření dotazů slouží grafický editor dotazů, který vyvoláme pomocí `Dotaz | Grafické vytváření`. Rychlejší je ale zadávání dotazu přímo do dotazovacího řádku.

Vyhledávat lze i více slov najednou. Stačí je zapsat jako nový dotaz. Jednotlivá slova je třeba v dotazu oddělit mezerami (viz obrázek 3.3). Pozor při vyhledávání nealfanumerických znaků, které mají svůj význam pro regulární výrazy. Jedná se především o otazník a tečku. Chcete-li nalézt skutečně znak ?, který v korpusu ČAK 2.0 zastupuje vynechané slovo, je třeba zapsat do dotazovacího řádku `\?`. Podobně pro tečku `\.`. Zadáním samotné tečky by se vyhledaly všechny pozice korpusu, na nichž je jediný znak (tečky mezi nimi budou také).

Dvojitým kliknutím na konkordanční řádek se ve spodním okně zobrazí širší kontext (viz obrázek 3.3). Zde je možné kontext rozšiřovat směrem vpřed či vzad pomocí šipky nahoru nebo dolů (napřed je třeba do spodního okna kliknout levým tlačítkem myši). Výběr atributů se provede pomocí `Zobrazení | Atributy`. Navíc lze u každé řádky zobrazit název zdrojového textu, ze kterého konkordance pochází (`Zobrazení | Reference`).

Obrázek 3.3. Bonito: zobrazení širšího kontextu



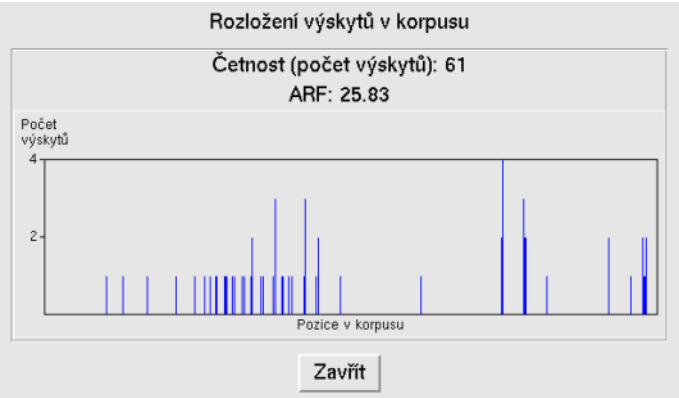
Pomocí `Zobrazení | Kontext` lze nastavit velikost kontextu – tzn. kolik slov, znaků či vět má být zobrazeno okolo každého nalezeného výskytu. Položka `Zobrazení | Rozsah` zase umožňuje automatický výběr jen určitého počtu řádků. To je užitečné zejména pro výsledky obsahující mnoho řádků, které je náročné projít ručně. Nejčastěji se asi použije zvolení náhodného vzorku dat.

Konkordance je možné také třídit, a to jednak podle samotného nalezeného slovního tvaru, jednak podle slov vyskytujících se v levém či pravém kontextu (`Konkordance | Jednoduché třídění`). Třídění může být i poměrně složité podle více kritérií (`Konkordance | Obecné třídění`). Po zvolení uvedených funkcí se objeví okna, do nichž je třeba vyplnit příslušné parametry pro třídění.

Výsledek jakkoli upraveného, promazaného a setříděného dotazu v hlavním okně je možné uložit do souboru pro pozdější použití (`Konkordance | Uložení`).

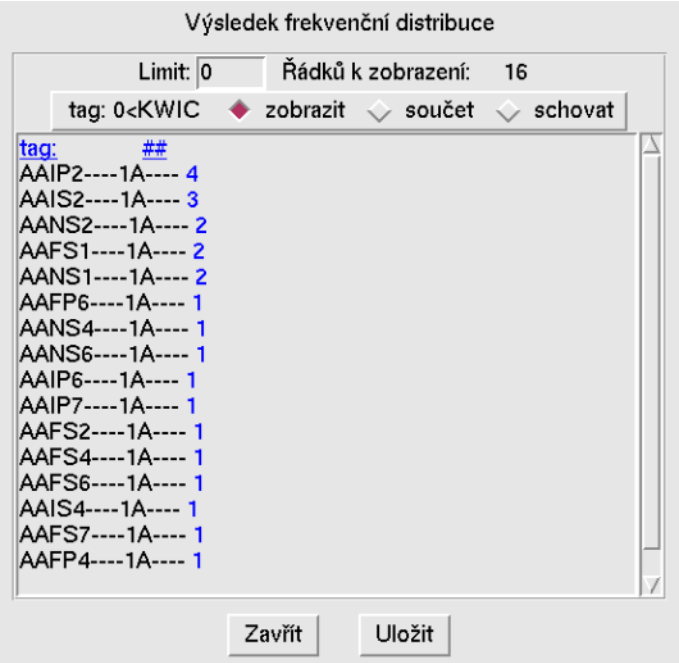
Nakonec se zmíníme o užitečných statistických funkcích, které jsou přístupné přes menu `Konkordance | Statistiky`. Položka `Rozložení` zobrazí nové okno s obrázkem znázorňujícím rozložení vyhledaných konkordancí v rámci celého korpusu (viz obrázek 3.4). Z obrázku je na první pohled vidět, zda jsou výskyty rozloženy rovnoměrně či nikoliv. V okně se zobrazí i číslo vyjadřující průměrnou redukovanou četnost (Savický, Hlaváčová, 2002), což je objektivnější vyjádření (ne)rovnoměrnosti rozložení výskytů.

Obrázek 3.4. Bonito: rozložení



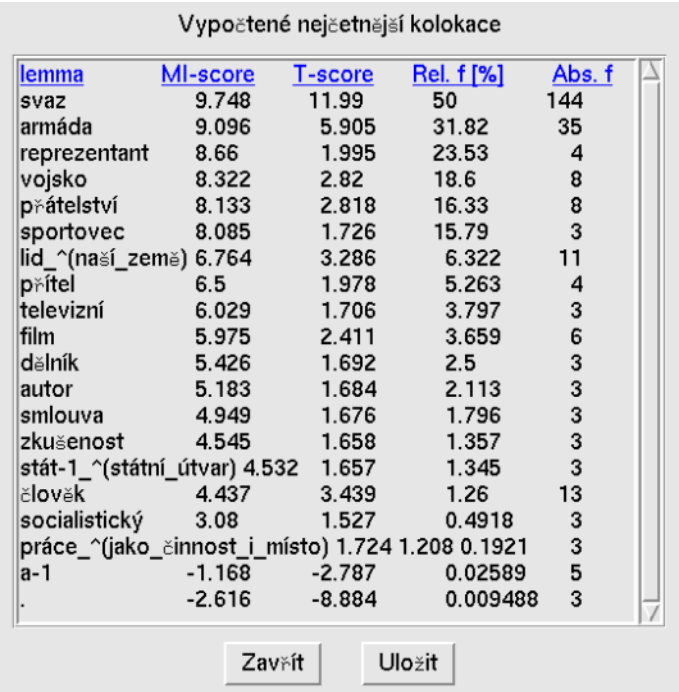
`Frekvenční distribuce` zobrazí vybrané atributy nalezených hodnot spolu s četnostmi. Na obrázku 3.5 je vidět frekvenční rozložení morfologických značek pro lemma *jarní*.

Obrázek 3.5. Bonito: frekvenční distribuce

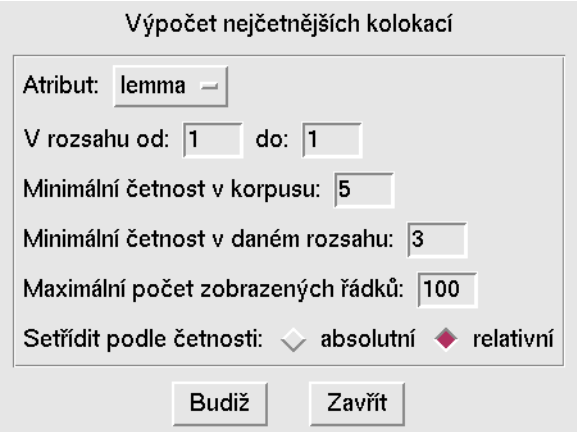


Poslední významnou statistickou funkcí jsou `Kolokace`. Pomocí této funkce lze zobrazit slova (nebo lemmata nebo značky), která se vyskytují v zadaném okolí nalezených výskytů (viz obrázek 3.7). Výsledkem je tabulka udávající pro každé slovo ze zadaného okolí jeho frekvenci v rámci nalezených konkordancí, relativní frekvenci, MI-score a T-score. Kliknutím na kategorii v záhlaví tabulky se změní řazení řádků podle vybrané kategorie, přičemž nejvýznamnější kolokace jsou vždy nahoře. Obrázek 3.6 obsahuje kolokace k lemmatu *sovětský*.

Obrázek 3.6. Bonito: kolokace

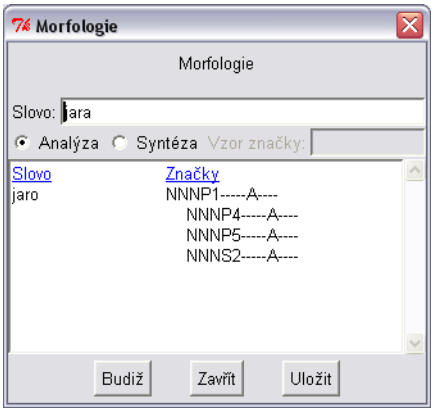


Obrázek 3.7. Bonito: zobrazení nejčetnejších kolokací



Z korpusového nástroje Bonito je možné vyvolat morfologickou analýzu, a to z menu `Manažer | Morfologie`. Nové okno, které tak otevřete, si můžete nechat připravené po celou dobu práce s korpusovým nástrojem. Můžete z něj spouštět morfologickou analýzu, nebo syntézu (generování). Morfologická analýza zadaného slova vypíše všechna možná lemmata a jim příslušné značky. Při zaškrtnutí syntézy zase dostanete všechny možné slovní tvary, které lze ze zadaného lemmatu vytvořit, spolu s jejich morfologickými značkami. Viz obr. 3.8.

Obrázek 3.8. Bonito: volání morfologické analýzy



Přepínání mezi anglickým a českým pracovním prostředím je možné z menu volbou `Manažer | Změna jazyka` (`Manager | Change language`).

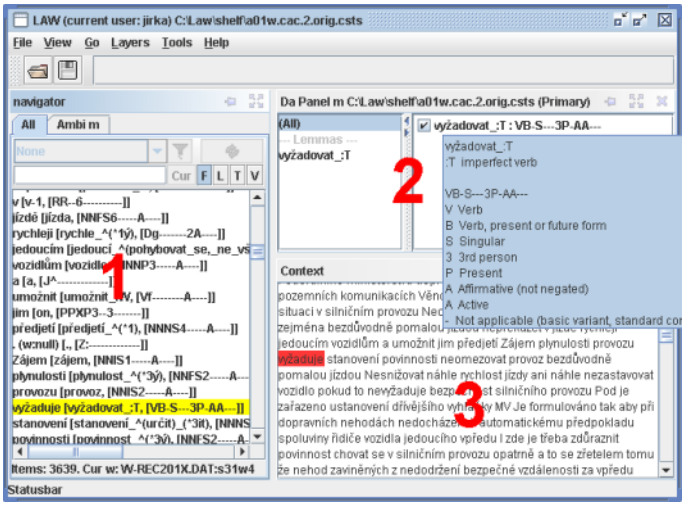
3.3.2. Morfologický anotační editor LAW

LAW (Lexical Annotation Workbench) je integrované prostředí pro morfologické anotování. Podporuje přímou morfologickou anotaci (tj. přiřazování lemmatu a značky danému slovu), porovnání anotací jednoho textu (kupříkladu více anotátory), vyhledávání slov, značek atd. Editor pracuje na všech operačních systémech, které podporují Javu, včetně systémů Windows a Linux. LAW je otevřeným systémem rozšiřitelným prostřednictvím externích modulů – např. pro různá zobrazení dat, import/export souborů a nápovědy. LAW podporuje formáty PML [11], csts [10] a TNT [24].

3.3.2.1. Hlavní části programu

Nástroj se skládá ze tří hlavních částí, jak je vidět na obrázku 3.9.

Obrázek 3.9. LAW: hlavní obrazovka



1. *Navigátor* – zobrazuje seznamy slov dokumentu filtrované podle různých kritérií a umožňuje výběr určitého slova pro disambiguaci.

2. *Da Panely* – zobrazují morfologickou informaci o slovu a umožňují její disambiguaci, tzn. výběr správného lemmatu a značky. Panel se skládá ze dvou oken – seznamu skupin a seznamu položek. Seznam položek zobrazuje všechna lemmata a značky přiřazené danému slovu (na dané m-rovině). Pomocí seznamu skupin lze položky omezit jen na určité lemma, slovní druh, podrobný slovní

druh nebo rod. Jeden Da Panel je vždy *hlavní* (*primary*), určité akce se pak týkají jen tohoto panelu (např. **Ctrl-T** aktivuje seznam lemmat a značek v hlavním panelu).

3. *Kontextová okna* – kontextové informace, např. text dokumentu, syntaktické struktury atd.

3.3.2.2. Obvyklý způsob práce

Anotační proces probíhá následovně:

1. Otevřete m-soubor, který chcete anotovat: `File | Open` (**Ctrl-O**). Odpovídající w-soubor se otevře automaticky.

2. Přepněte se v Navigátoru do seznamu nejednoznačných slov (Ambi + jméno daného m-souboru), ve kterém se zobrazí nejednoznačná slova, tj. slova, pro která morfologická analýza nabízí více možností, a vyberte ze seznamu první slovo.

3. Zmáčkněte `Enter`. Kurzor se přesune do hlavního Da Panelu. Vyberte správné lemma a značku a opět zmáčkněte `Enter`. Kurzor se přesune na další nejednoznačné slovo.

   Pokud uděláte chybu, přepněte se v Navigátoru do seznamu všech položek (All), nalezněte chybně anotované slovo a vyberte jej. Příslušná anotace se zobrazí v Da Panelu. Vyberte správné lemma a značku a přepněte se zpátky do seznamu nejednoznačných slov (Ambi X).

4. Uložte výsledek anotování: `File | Save` (**Ctrl-S**).

3.3.3. TrEd

TrEd (Tree Editor) je integrovaným prostředím primárně navrženým pro syntaktické anotování vět, při kterém je větě přiřazena stromová struktura. Zároveň může být použit i k prohlížení dat a obsahuje také několik druhů vyhledávacích funkcí.

TrEd podporuje celou řadu vstupních a výstupních formátů - určitě PML a CSTS, které jsme čtenářům již představili v části 3.2.1. TrEd je zároveň modulárním systémem, proto je snadné doplnit podporu pro další formáty.

TrEd nabízí různé možnosti nastavení v závislosti na požadavcích uživatele. Jeho funkčnost může být dále rozšířena pomocí uživatelsky definovaných maker v jazyce Perl, které je možné vyvolat buď přímo klávesovou zkratkou nebou přes položku v menu.

Programátorsky orientované uživatele bude jistě zajímat varianta nástroje TrEd bez grafického rozhraní, a to nástroj btred pro dávkové zpracování dat (Batch-mode Tree Editor). Dalšín doplňkem je nástroj NTrEd, který umožňuje paralelizovat procesy spuštěné nástrojem btred jejich distribucí na více výpočetních strojů.

Pro otevření souborů v TrEdu zvolte menu `File | Open`. Vyberte jakýkoliv soubor *.a (tj. soubor se syntaktickou anotací nějakého dokumentu), TrEd jej otevře a ihned zobrazí strom pro první větu daného souboru.

Typický vzhled TrEdu je na na obrázku 3.10. Jde o větu *Problémy motivace jsou tak staré jako lidstvo.* - vysvětlivky následují.

- 1 Okno se stromem, který reprezentuje syntaktickou anotaci věty.

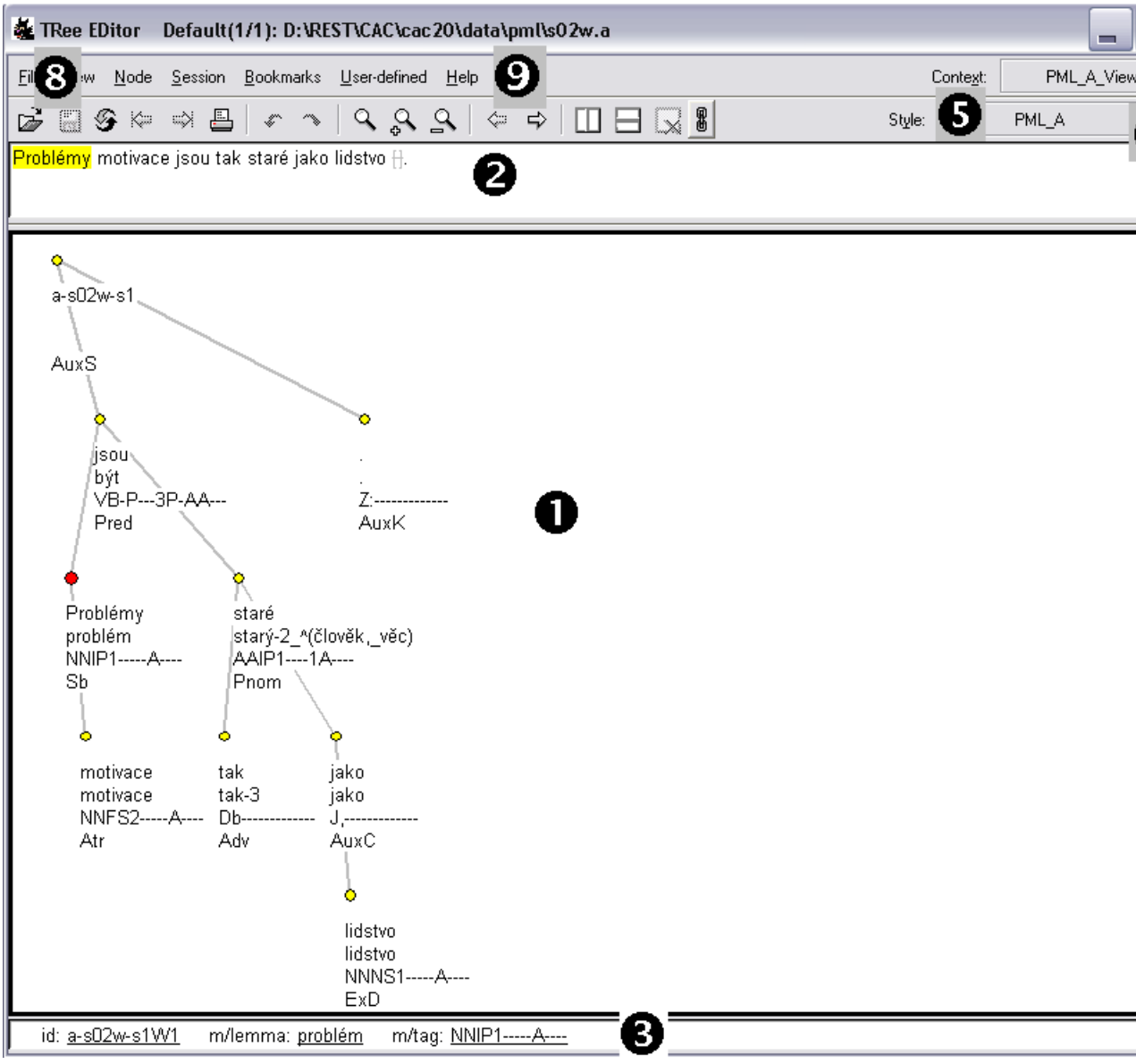
- 2 Textová forma věty.

- 3 Stavový řádek. Zobrazuje pro vybrané slovo (červený uzel, zde *Problémy*) různé informace dle kontextu (zde id uzlu, lemma a morfologická značka)

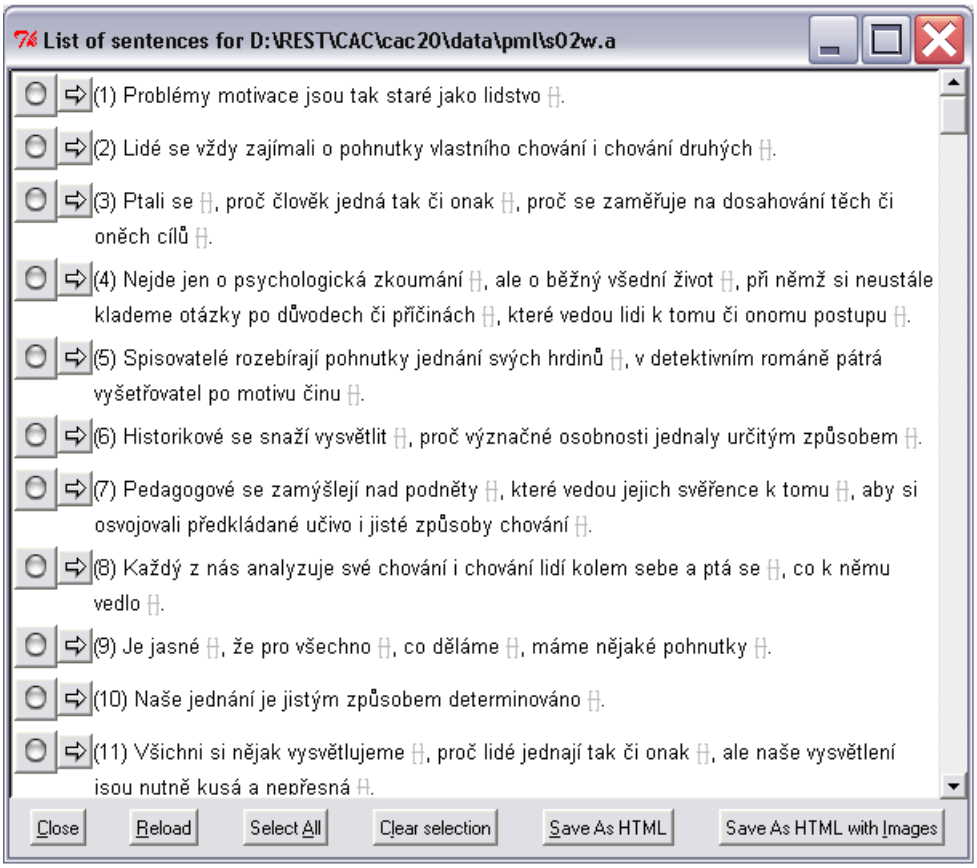
- 4 Aktuální kontext. Kontext je možné změnit kliknutím na jméno aktuálního kontextu a následným výběrem nového kontextu ze zobrazeného seznamu (např. PML_A_Edit).

- 5 Aktuální zobrazovací styl. Může být změněn podobným způsobem jako kontext.
- 6 Editace zobrazovacího stylu.
- 7 Zobrazit seznam všech vět aktuálního souboru - viz obrázek 3.11. Nad tlačítkem je zobrazeno pořadí aktuálního stromu (tj. aktuální věty) v aktuálním souboru.
- 8 Tlačítka pro otevření, uložení a opětovné otevření souboru.
- 9 Tlačítka pro přesunutí na předchozí/následující strom v aktuálním souboru a pro správu oken.

Obrázek 3.10. TrEd: hlavní obrazovka



Obrázek 3.11. TrEd: obrazovka s větami souboru



Implicitně jsou soubory ČAK 2.0 otevřeny v kontextu PML_A_View, který neumožňuje jejich editaci, pouze prohlížení. Pokud si přejete soubory měnit, přepněte se do kontextu PML_A_Edit. V obou kontextech je k dispozici jeden zobrazovací styl - PML_A. V libovolném kontextu můžete zobrazit seznam všech maker definovaných v daném kontextu a jejich klávesové zkratky, a to vybráním menu `View | List of Named Macros`.

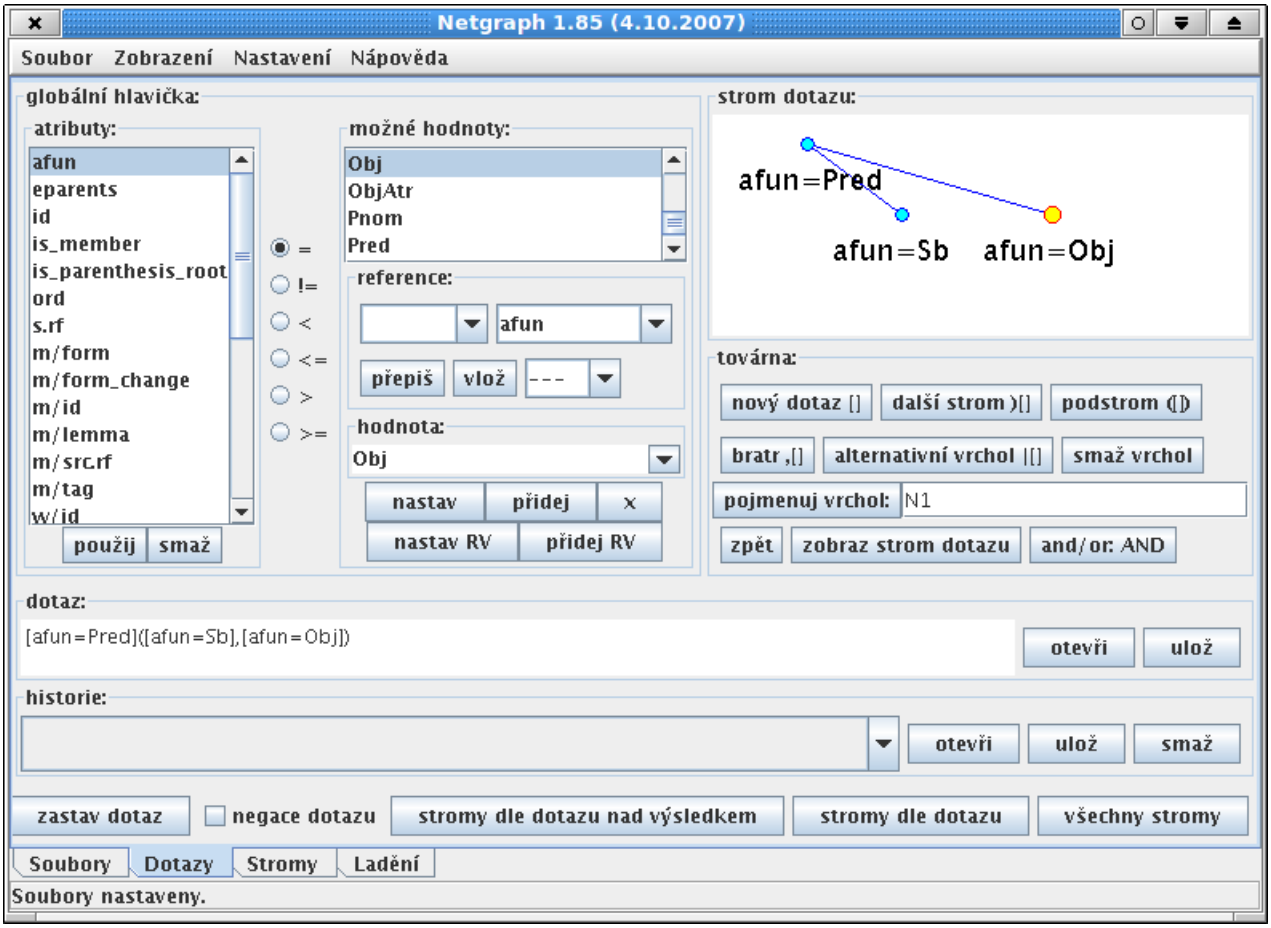
3.3.4. Netgraph

`Netgraph` je aplikace typu klient-server, která umožňuje prohledávat ČAK 2.0 současně několika uživateli, připojenými přes internet. `Netgraph` je navržený tak, aby prohledávání bylo co nejjednodušší a intuitivní, při zachování vysoké síly dotazovacího jazyka.

Dotaz v `Netgraphu` je jeden uzel nebo strom s uživatelem definovanými vlastnostmi, který má být vyhledán v korpusu. Prohledání korpusu pak znamená hledat věty (samozřejmě ve formě anotovaných stromů), které obsahují dotaz jako svůj podstrom. Uživatel má možnost zadat dotazy nejrůznější složitosti, od těch nejjednodušších (jako je hledání všech stromů korpusu, které obsahují dané slovo), po pokročilejší (jako např. hledání všech vět, obsahujících sloveso rozvinuté adresátem, který není ve třetím pádě, *a* nejméně jedním příslovcem udávajícím směr, atd.). Dotazy mohou být dále rozšířeny tzv. *meta atributy*, které umožňují vyhledávat ještě složitější konstrukce.

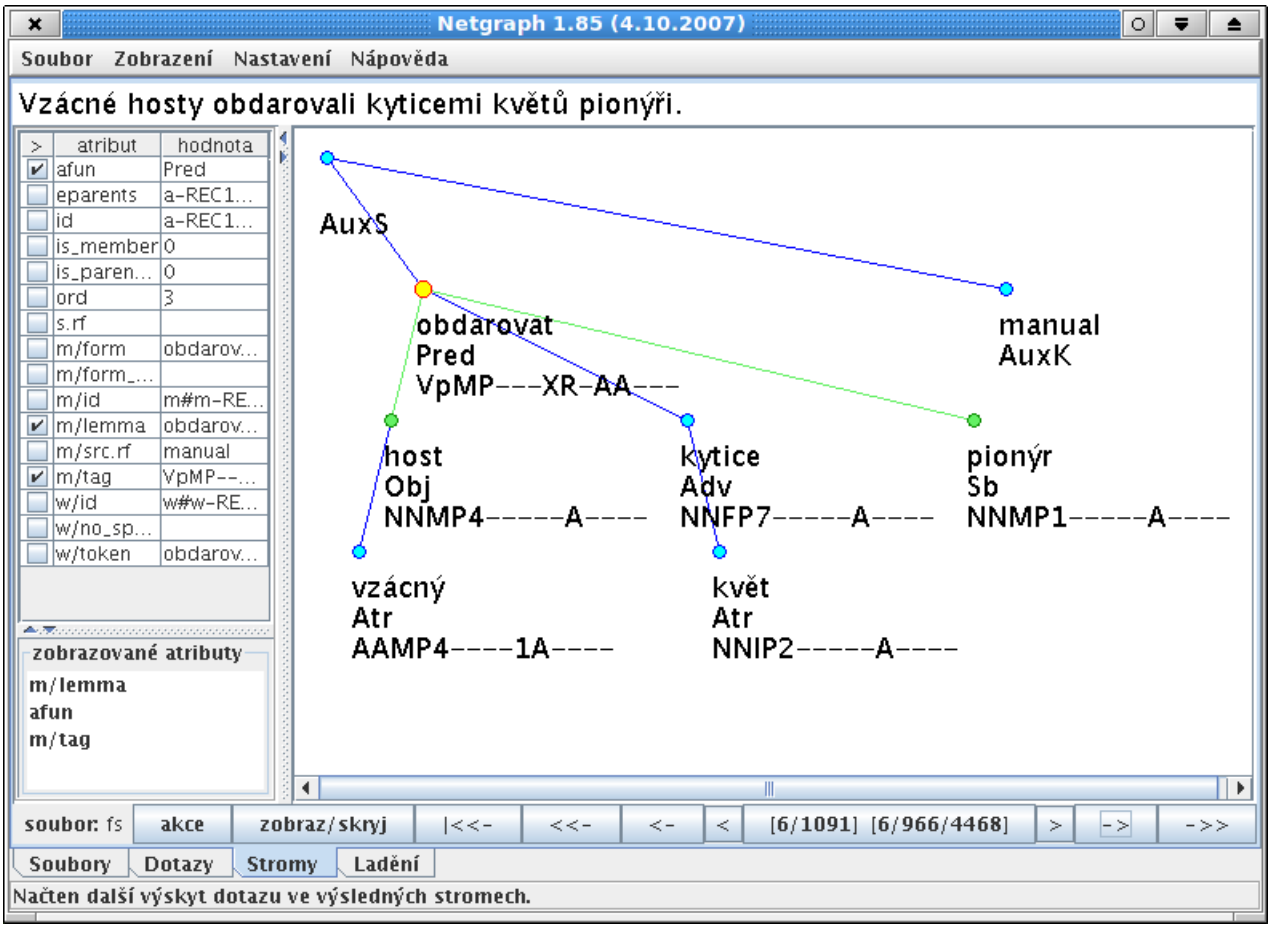
Dotazy se v `Netgraphu` vytvářejí v uživatelsky přívětivém grafickém prostředí. Příkladem je dotaz na obrázku 3.12. V tomto jednoduchém dotazu hledáme všechny stromy, které obsahují uzel označený jako predikát, rozvitý nejméně dvěma uzly, označenými jako subjekt a objekt. Pořadí těchto uzlů v nalezených stromech není v dotazu nijak omezeno.

Obrázek 3.12. Vytváření dotazu v `Netgraphu`



Jedním z výsledků, zaslaných zpět serverem, může být strom z obrázku 3.13.

Obrázek 3.13. Nalezený strom v `Netgraphu`



Uživatel vždy používá klientskou část programu Netgraph. Tímto klientem se může připojit k veřejnému serveru `quest.ms.mff.cuni.cz` na portu `2001`. Pokud si uživatel nainstaluje i serverovou část Netgraphu, může se připojit lokálně k tomuto vlastnímu serveru a prohledávat korpus bez přístupu k internetu.

3.3.5. Automatické zpracování textů

Paralelně s prací nad daty jsou vyvíjeny aplikace pro morfologické a syntaktické zpracování českých textů. Součástí CD jsou dvě základní morfologické aplikace – morfologická analýza a tagování – a jedná syntaktická aplikace – parsování. Tyto tři aplikace doplňuje tokenizace.

Tokenizace je proces, který rozkouskuje daný text na jednotlivá slova. Výsledkem je tzv. vertikála, tedy soubor, který obsahuje každé slovo nebo interpunkční znaménko na zvláštním řádku. Pod pojmem tokenizace se často rozumí i tzv. segmentace, což je označení začátků odstavců a vět. I naše tokenizace současně text segmentuje.

V našem pojetí je však tokenizace ještě něco navíc -- převádí vertikálu do formátu CSTS (viz část 3.2.1). Převod do tohoto formátů spočívá hlavně v přidání hlavičky před vertikálu a označení jednotlivých slov jednoduchými značkami, které rozlišují vlastnosti slov viditelné přímo z jejich ortografického zápisu. Jde zejména o odlišení interpunkčních znamének, slov skládajících se z číslic, nebo číslice obsahující. Dále se speciální značkou označí slova začínající velkým písmenem a slova celá složená z velkých písmen. Výsledek tokenizace, tedy vertikála ve formátu CSTS, je potom přímo vstupním formátem pro další zpracování.

Morfologická analýza zpracovává jednotlivé slovní formy a určuje pro ně lemmata (základní tvary, např. pro podstatná jména první pád jednotného čísla, pro slovesa infinitiv) a možné morfologické interpretace.

Základem morfologické analýzy je morfologický slovník, který obsahuje tvaroslovné informace o českých slovech a jejich odvozeninách. Každý slovní tvar má přiřazeno lemma a morfologickou značku (morfologický tag), která danému tvaru přísluší. V použitém morfologickém slovníku mají mnohá lemmata doplňující informace o stylu, sémantice nebo o způsobu odvození. V případě zkratek bývají lemmata opatřena komentářem s vysvětlujícím textem (viz příloha B).

Vzhledem k vysoké homonymii češtiny náleží většině slovních tvarů více morfologických značek, občas i více lemmat. Např. slovní tvar *pekla* má dvě různá lemmata – podstatné jméno *peklo* a sloveso *péci*. Obě lemmata dále mohou pro uvedený slovní tvar mít několik různých morfologických značek. Při morfologické analýze se probírají jednotlivé slovní formy z celého korpusu a porovnávají se se slovními formami obsaženými v morfologickém slovníku. V případě shody se danému slovnímu tvaru přiřadí příslušná lemmata a morfologické značky. Výsledkem morfologické analýzy pro konkrétní slovo je tedy množina dvojic lemma – morfologická značka.

Na morfologickou analýzu navazuje tagování (někdy také desambiguace nebo disambiguace). Během této fáze se vybere ze všech možných lemmat a morfologických značek, přiřazených v předchozí fázi, jediná dvojice, která by měla být v konkrétním kontextu správná. Vzhledem k obtížnosti úlohy není možné navrhnout takovou metodu tagování, která by pracovala se stoprocentní úspěšností. Program, který tagování provádí, je označován jako *tagger*.

Použitý tagger je založen na *skrytých markovovských modelech* s využitím *průměrovaného perceptronu* (Collins, 2002). Jedná se o metodu statistickou. Vstupem taggeru je text, který pro každé slovo obsahuje množinu všech možných morfologických značek a lemmat (výstup z morfologické analýzy). Na výstupu pak k těmto datům přidává jednoznačně vybranou značku a jí odpovídající lemma. Tagger byl natrénován na datech z PZK 2.0 a jeho úspěšnost (procento správně určených morfologických značek) na ČAK 1.0 je 91,8 %. Část chyb je ovšem způsobena rozdíly mezi PZK a ČAK, což v důsledku vede k tomu, že morfologická analýza pro některá slova nenabízí správné značky. Systematicky se to stává pro číslovky zapsané ciframi (v ČAK reprezentovány jako #) a neznámá slova (reprezentována znakem `?`). Pokud tyto systematické rozdíly nebereme v potaz, je výsledná úspěšnost značkování 93,1%.

Parsování představuje další úroveň zpracování textů, která navazuje na tagování. Při parsování se pro každé slovo věty určí jeho syntaktická závislost na jiném slově věty a přiřadí se mu analytická funkce. Program, který parsování provádí, je označován jako *parser*.

Použitý parser je založen na stejné metodologii jako použitý tagger. Vstupem parseru je text, který pro každé slovo textu obsahuje dvojici (lemma, morfologická značka). Na výstupu je stromová struktura s analytickými funkcemi. Parser byl natrénován na trénovacích datech PZK 2.0 a jeho úspěšnost na ČAK 2.0 je ...%.

Aby uživatel studovat, jak nástroje instalovat a spustit, vytvořili jsme skript `tool_chain`, který prostřednictvím základních přepínačů dokumentovaných v tabulce 3.11 spustí požadovaný nástroj. Řetězením přepínačů je možné spustit více nástrojů za sebou.

Příklad: Chceme-li surový text zpracovat morfologickou analýzou, spustíme `tool_chain -t -A`

Přípomenutí: Při práci s formátem PML musí být se souborem, který vstupuje do skriptu `tool_chain`, v adresáři i soubory, na které se z něho odkazuje. Pokud je vstupním souborem m-soubor, musí ho "doprovázet" i w-soubor.

Tabulka 3.11. Skript `tool_chain`

parametr	formát vstupního souboru	formát výstupního souboru	popis
`-t`	surový text	CSTS	tokenizace
`-A`	CSTS	PML m-soubor, CSTS	morfologická analýza
`-T`	PML m-soubor, CSTS (výstup morfologické analýzy)	PML m-soubor, CSTS	tagování
`-P`	PML m-soubor, CSTS	PML a-soubor, CSTS	parsování

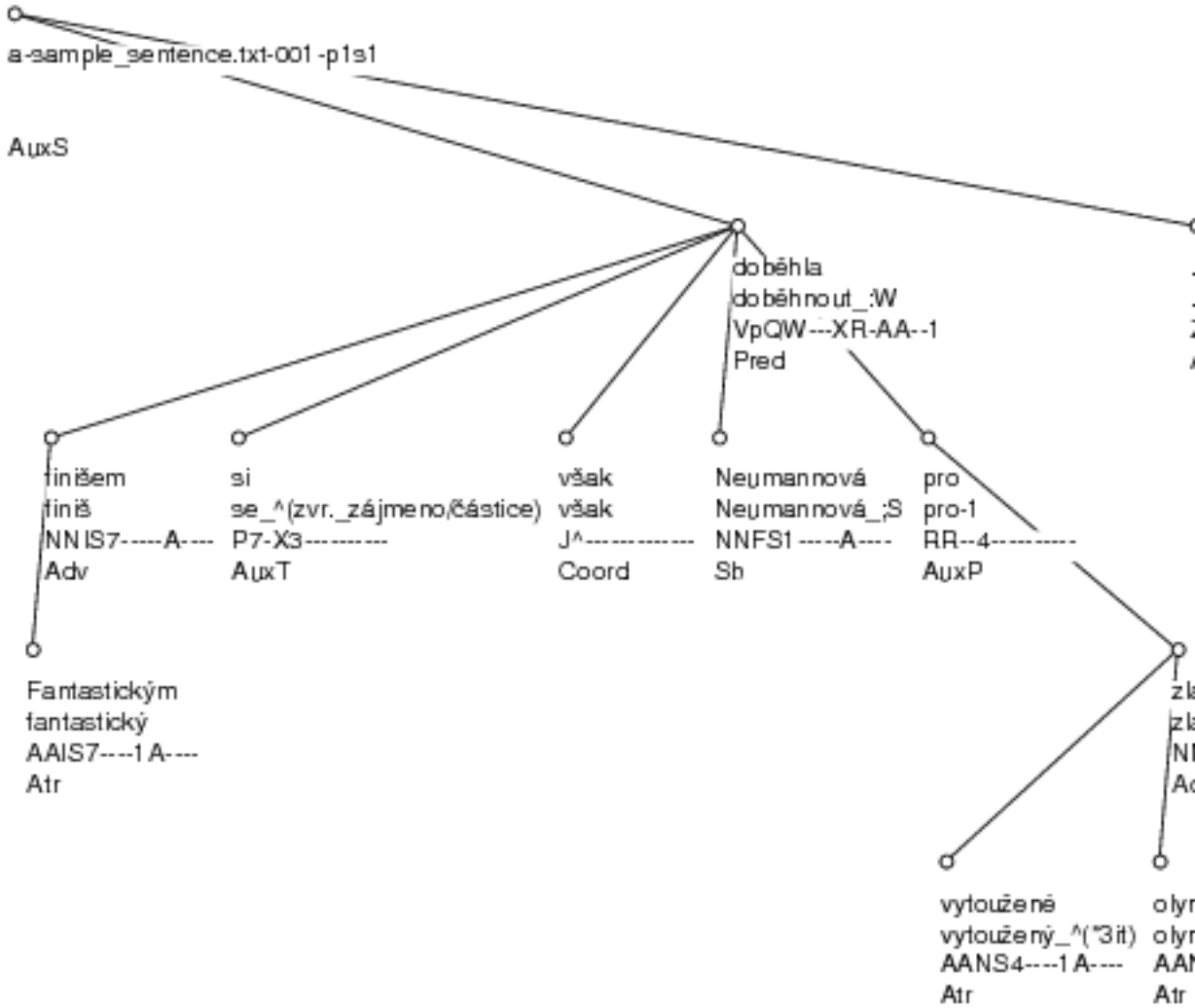
Nástroje jsou implementovány v programovacích jazycích C/C++ a Perl a hlavní skript `tool_chain` v jazyku bash. Vzhledem k autorským právům neposkytujeme zdrojové C/C++ kódy. Spustitelné programy jsou kompilovány pro operační systém Linux běžící na architektuře i386.

Příklad: Ukážeme zpracování textu *Fantastickým finišem si však Neumannová doběhla pro vytoužené olympijské zlato.* Výsledky morfologické analýzy (spuštěním `tool_chain -t -A`)a tagování (spuštěním `tool_chain -T`) jsou souhrnně uvedeny v tabulce 3.12. V případě více možných základních tvarů slovní formy (např. slovní forma *si* je analyzována buď jako sloveso *být*, nebo jako zvratná částice *se*) jsou tyto základní tvary odděleny symbolem svislého lomítka „|“. Abychom usnadnili čtenáři pátrání po chybách, kterých se tagger dopustil, potvrzujeme, že se tagger žádných chyb nedopustil. k danému kontextu. Výsledek parsování (spuštěním `tool_chain -P`) zobrazuje obrázek 3.14. U každého uzlu stromu je zobrazena slovní forma, zjednoznačněné lemma, zjednoznačněná morfologická značka a analytická funkce. Abychom i v tomto případě usnadnili čtenáři pátrání po chybách, potvrzujeme, že se parser nedopustil žádné chyby.

Tabulka 3.12. Ukázka textu zpracovaného morfologickou analýzou a tagováním

text	morfologická analýza	tagování
Fantastickým	fantastický AAFP3----1A---- AAIP3----1A---- AAIS6----1A---7 AAIS7----1A---- AAMP3----1A---- AAMS6----1A---7 AAMS7----1A---- AANP3----1A---- AANS6----1A---7 AANS7----1A----	fantastický AAIS7----1A----
finišem	finiš NNIS7-----A----	finiš NNIS7-----A----
si	být VB-S---2P-AA--7|se_ ^(zvr._zájmeno/částice) P7-X3----------	se_^(zvr._zájmeno/částice) P7-X3----------
však	však J^------------	však J^------------
Neumannová	Neumannová_;S NNFS1-----A---- NNFS5----A----	Neumannová_;S NNFS1----A----
doběhla	doběhnout_:W VpQW---XR-AA--1	doběhnout_:W VpQW---XR-AA--1
pro	pro-1 RR--4----------	pro-1 RR--4----------
vytoužené	vytoužený_^(*3it) AAFP1----1A---- AAFP4----1A---- AAFP5----1A---- AAFS2----1A---- AAFS3----1A---- AAFS6----1A---- AAIP1----1A---- AAIP4----1A---- AAIP5----1A---- AAMP4----1A---- AANS1----1A---- AANS4----1A---- AANS5----1A----	vytoužený_^(*3it) AANS4----1A----
olympijské	olympijský AAFP1----1A---- AAFP4----1A---- AAFP5----1A---- AAFS2----1A---- AAFS3----1A---- AAFS6----1A---- AAIP1----1A---- AAIP4----1A---- AAIP5----1A---- AAMP4----1A---- AANS1----1A---- AANS4----1A---- AANS5----1A----	olympijský AANS4----1A----
zlato	zlato NNNS1-----A---- NNNS4-----A---- NNNS5-----A----	zlato NNNS4-----A----
.	. Z:----------	. Z:----------

Obrázek 3.14. Ukázka zpracování věty parsováním



Fantastickým finišem si však Neumannová doběhla pro vytoužené olympijské zlato.

Doporučujeme uživatelům, aby si vybrali libovolný český text a zpracovali jej skriptem `tool_chain -t -A` . Následně výstup skriptu otevřte v nástroji LAW a začněte zjednoznačňovat značky slov.

Soubor s ručně zjednoznačněnými značkami zpracujte skriptem `tool_chain -P`. Následně výstup skriptu otevřte v nástroji TrEd a můžete začít opravovat závislosti a analytické funkce.

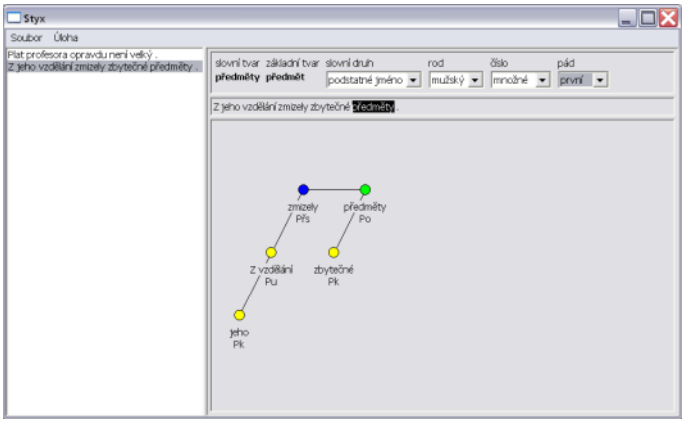
Kapitola 4. Bonusový materiál

4.1. Elektronická cvičebnice STYX

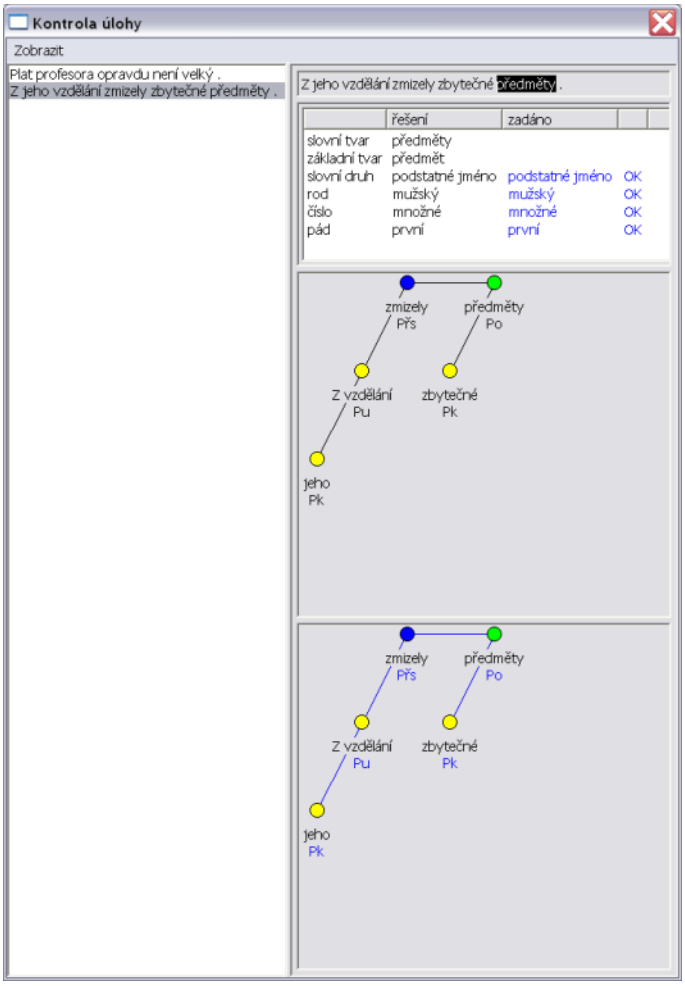
Bonusový materiál je určen žákům devátých tříd, středoškolákům a jejich pedagogům. Jedná se o elektronickou cvičebnici českého tvarosloví a české syntaxe pojmenovanou STYX [22]. Nejpozoruhodnější vlastností cvičebnice STYX je počet vět, který je nabídnut (vice než jedenáct tisíc), a možnost okamžité kontroly rozborů. Cvičebnice byla sestavena z vět a jejich anotací, které jsou součástí PZK. Je třeba zdůraznit, že akademické pojetí české syntaxe (jak je prezentováno v PZK 2.0) se v některých jevech liší od pojetí vyučovaného ve školách. Podrobně jsou rozdíly dokumentovány v (Kučera, 2006). V rámci jednoho cvičení se komplexně, dle školního pojetí české syntaxe, zpracuje libovolný počet vět – pro každé slovo věty se provede tvaroslovný rozbor a pro celou větu větný rozbor spolu s určováním větných členů. Abychom uživatele nezahltili, nepředkládáme mu všech 11 tisíc vět, ale jen nepatrný zlomek – 50 vět (viz `bonus-tracks/STYX/sample.styx`).

Procvičování v nástroji STYX ilustruje obrázek 4.1. Při tvaroslovném rozboru se pro každé slovo určí slovní druh a odpovídající morfologické kategorie (horní část pravého okna). Na začátku větného rozboru jsou slova věty seřazena vedle sebe a postupným přenášením uzlů se provádí větný rozbor. Následně se určí větné členy včetně základní skladební dvojice. Obrázek 4.2 demonstruje vyhodnocení rozborů. V našem případě uživatel zaslouží pochvalu, protože vše analyzoval bezchybně.

Obrázek 4.1. Procvičování v nástroji Styx



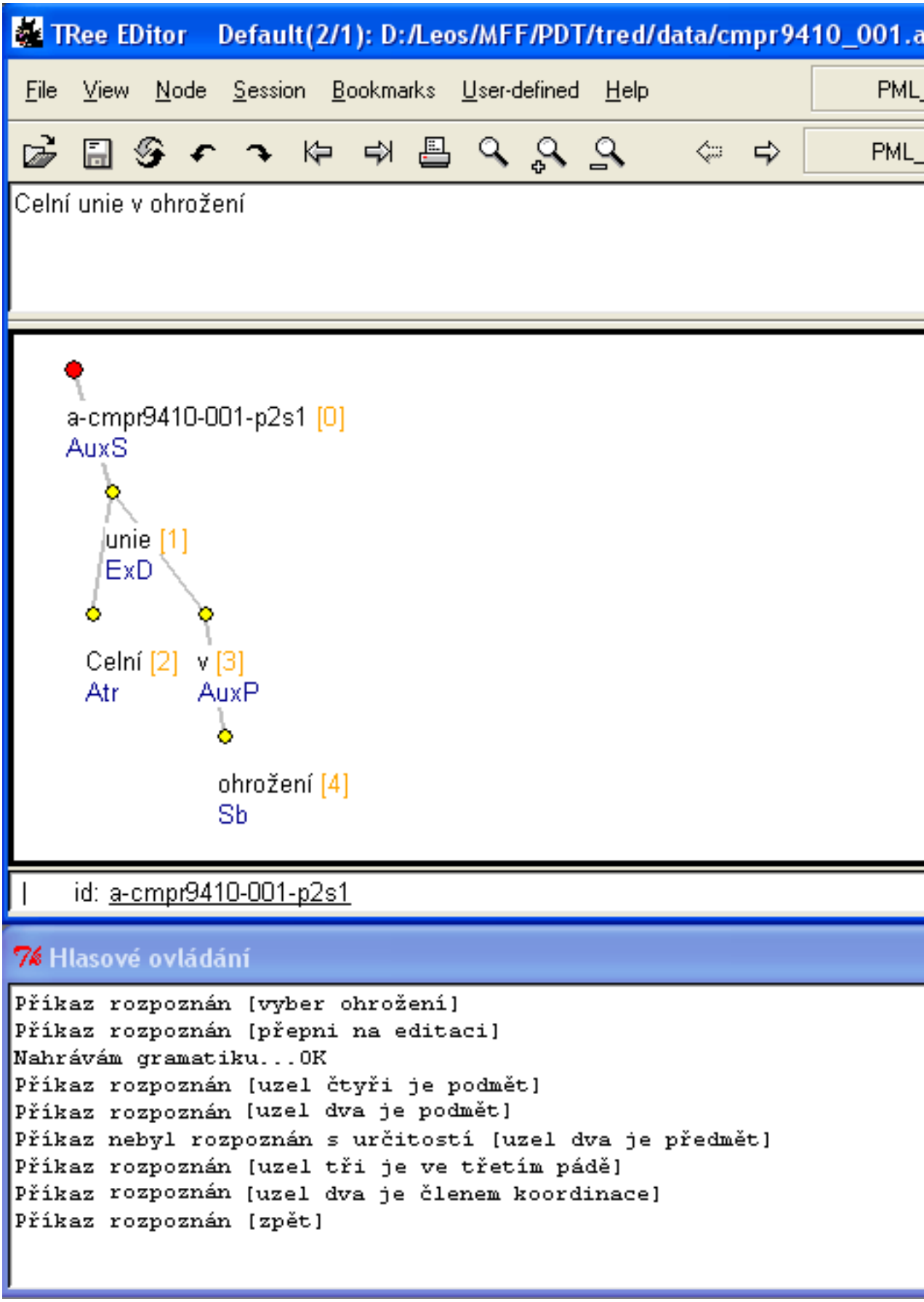
Obrázek 4.2. Vyhodnocení cvičení v nástroji Styx



4.2. TrEdVoice: hlasové ovládání anotačního editoru TrEd

Základním anotačním nástrojem používaným k syntakticko-analytické anotaci ČAK je anotační editor TrEd (doplnit odkaz do textu). Hned po svém vzniku obsahoval poměrně hodně funkcí a maker a postupem času se jejich množství dále zvyšovalo. Vzhledem k tomu, že by pro uživatele bylo časově neúnosné neustále hledat všechny funkce v menu, jsou většině funkcí přiřazeny klávesové zkratky. To ovšem zase klade poměrně velké nároky na uživatele, který si musí toto velké množství zkratek pamatovat. Jeden ze způsobů, jak uživateli usnadnit práci, je přidat další způsob ovládání, a to ovládání hlasem, které není u počítačových programů příliš rozšířené. Proto přicházíme s modulem TrEdVoice . Tento modul si neklade za cíl vytvořit kompletní hlasové ovládání všech funkcí TrEd tak, aby bylo možné program ovládat úplně bez použití klávesnice a myši. Jedná se spíše o vhodný doplněk ke stávajícím způsobům ovládání (menu, klávesové zkratky a myš). Obrázek 4.3 zachycuje hlavní obrazovku editoru TrEd při zapnutém hlasovém ovládání. Pro samotné rozpoznávání povelů je použit modul rozpoznávání mluvené řeči (tzv. ASR modul) vytvořený na Katedře kybernetiky Západočeské univerzity v Plzni [5] (Müller, Psutka, Šmídl, 2000), který není přímo začleněn do TrEdVoice, ale je spouštěn samostatně jako ASR server a komunikace s ním probíhá pomocí síťového protokolu TCP/IP. Modul ASR pracuje na statistickém principu a je nezávislý na řečníkovi, tj. dokáže rozpoznávat řeč "libovolného" člověka. Více o rozpoznávání mluvené řeči viz (Psutka, Müller, Matoušek, Radová, 2006).

Obrázek 4.3. Obrazovka editoru TrEd se zapnutým modulem TrEdVoice



Kapitola 5. Tutoriály

Pro snadnější seznámení se s daty a nástroji jsou k dispozici tutoriály dvou forem. První forma tutoriálů jsou videonahrávky spolu s textovými podklady přednášek, které zazněly během tutoriálu o PZK (Prague Treebanking for Everyone: A two-day tutorial [http://ufal.mff.cuni.cz/vmc/?a=ls21_tutorial]) pořádaného na podzim roku 2006. Druhá forma tutoriálů jsou demosnímky, které provádějí uživatele ovládáním nástrojů s grafickým rozhraním. Demosnímky jsou přímou součástí CD-ROM, zatímco videonahrávky jsou uvedeny jako odkazy na externí zdroj. Tabulka 5.1 uvádí přehled tutoriálů vztahujících se k datům, a to konkrétně tutoriály o anotačních rovinách (m-rovina, a-rovina) a tutoriál o vnitřní reprezentaci dat (formát PML). Tutoriály jsou ve formě videonahrávek. Tabulka 5.2 uvádí přehled tutoriálů vztahujících se k nástrojům. Tutoriály jsou ve formě jak videonahrávek, tak demosnímků.

Tabulka 5.1. Tutoriály k datům

videonahrávka
m-rovina [http://ufallab.ms.mff.cuni.cz/video/recordshow/index/17/28]
a-rovina [http://ufallab.ms.mff.cuni.cz/video/recordshow/index/17/29]
PML [http://ufallab.ms.mff.cuni.cz/video/recordshow/index/17/34]

Tabulka 5.2. Tutoriály k nástrojům

videonahrávka	**demosnímek**
Bonito [http://ufallab.ms.mff.cuni.cz/video/recordshow/index/2/24]	Bonito [../../../../tutorials/bonito.htm]
LAW [http://ufallab.ms.mff.cuni.cz/video/recordshow/index/2/22]	LAW [../../../../tutorials/law.htm]
TrEd [http://ufallab.ms.mff.cuni.cz/video/recordshow/index/2/23]	TrEd [../../../../tutorials/tred.htm]
Netgraph [http://ufallab.ms.mff.cuni.cz/video/recordshow/index/2/25]	Netgraph [../../../../tutorials/netgraph.htm]
STYX [http://ufallab.ms.mff.cuni.cz/video/recordshow/index/2/27]	STYX [../../../../tutorials/styx.htm]
---	TrEdVoice [../../../../tutorials/TrEdVoice.htm]

Kapitola 6. Instalace

Pro usnadnění práce uživatelů s ČAK jsou k dispozici "instalační" programy pro operační systémy Linux a MS Windows. Poznamenejme však, že **komponenty CD-ROM budou pouze zkopírovány, ne nainstalovány**! Instalaci nástrojů budou uživatelé muset provést samostatně - v adresáři každého nástroje je k dispozici soubor `README_CZ.txt`, ve kterém jsou uvedeny systémové požadavky nástroje a odkazy na uživatelskou dokumentaci včetně instalačních pokynů. Zároveň je možné většinu částí ČAK 2.0 používat přímo z distribučního CD-ROM nebo z jeho kopie. V tabulce 6.1 je uveden přehled všech nástrojů dostupných na CD-ROM a jejich (ne)spustitelnost pod operačními systémy Linux a MS Windows.

Tabulka 6.1. Spustitelnost nástrojů pod operačními systémy Linux a MS Windows

nástroj	Linux	MS Windows
Bonito	ano	ano
LAW	ano	ano
STYX	ano	ano
TrEd	ano	ano
TrEdVoice	ne	ano
Netgraph	ano	ano
`tool_chain`	ano	ne

"Instalace" se spustí následujícími příkazy:

- **Instalace na Linuxu.** V kořenovém adresáři CD-ROM spusťte program `./Instaluj-na-Linuxu.pl`.
- **Instalace na MS Windows.** "Instalační" program spusťte poklepem na ikonu `Instaluj-na-Windows.exe` v kořenovém adresáři CD.

Na začátku "instalace" si bude uživatel moci vybrat mezi dvěma typy "instalace"; následně bude vyzván k určení cílového adresáře (struktura cílového adresáře bude dodržovat adresářovou strukturu distribučního CD-ROM):

- **základní**, která pokrývá zkopírování kompletní dokumentace, tutoriálů a instalačních balíčků nástrojů Bonito, TrEd včetně modulu TrEdVoice pro hlasové ovládání a STYX.
- **uživatelská**, která pokrývá zkopírování libovolné komponenty CD-ROM dle volby uživatele.

Upozornění pro uživatele CD-ROM ČAK 1.0: "Instalační" programy z CD-ROM ČAK 2.0 pracují nezávisle na na "instalaci" ČAK 1.0. Pokud chcete používat nástroj z distribuce ČAK 2.0, který byl součástí distribuce ČAK 1.0, doporučujeme jeho instalaci provést znova. Distribuce ČAK 2.0 totiž obsahuje aktuálnější verze nástrojů.

Upozornění pro uživatele Bonito: Pro vyhledávání v ČAK 2.0 pomocí nástroje Bonito **není nutné** kopírovat ČAK 2.0 ve formátu `XML` z adresáře `data/pml`.

Upozornění pro uživatele TrEd a TrEdVoice: Modul TrEdVoice pro hlasové ovládání TrEd pracuje pouze pod systémem MS Windows. Pokud si nainstalujete TrEd pod MS Windows instalačním programem, který je součástí distribuce ČAK 2.0 (`tools/TrEd/tred_wininst_en.zip`), zároveň tím nainstalujete TrEdVoice. Pro Linux toto neplatí. Jakkoli nabízíme TrEdVoice jako bonusový materiál, vzhledem k jeho těsné provázanosti s editorem TrEd je jeho uživatelská dokumentace v adresáři `tools/TrEd/docs/`, a ne v adrsari `bonus-tracks/`. **Upozornění pro uživatele Bonito:** Pro vyhledávání v ČAK 2.0 pomocí nástroje Bonito **není nutné** kopírovat ČAK 2.0 ve formátu `XML` z adresáře `data/pml`.

Kapitola 7. Distribuce a licence

Plnou distribuci CD-ROM ČAK 2.0 je možno objednat u vydavatelství Linguistic Data Consortium [http://www.ldc.upenn.edu]; během objednávání budete přesměrováni na formulářovou licenční stránku (text licence je k nahlédnutí na adrese http://ufal.mff.cuni.cz/corp-lic/cac20-reg-cs.html.), kterou musíte vyplnit, aby mohla být objednávka dokončena.

Některé nástroje distribuce mohou být kryty licencí GPL (GNU Public License). U těchto nástrojů je to vždy explicitně uvedeno v souboru `README_CZ.txt`, který se nachází v domovském adresáři nástroje na CD-ROM ČAK 2.0. V těchto případech má GPL přednost před licencí pro užití ČAK 2.0.

Kapitola 8. Osobnosti v projektu

Všichni ti, kteří se o ČAK zasloužili, jsou představeni svým jménem.

- **Český akademický korpus verze 2.0**
  - **Kontrola morfologických anotací:** Jiří Mírovský
  - **Syntakticko-analytické anotace**: Alla Bémova, Katarina Gajdošová, Katarína Kandračová, Ivana Klímová, KK
  - **TBA:**
- **Nástroje**
  - **Bonito:** Pavel Rychlý
  - **LAW:** Jirka Hana
  - **Segmentace a tokenizace českých textů:** TBA
  - **Morfologický analyzátor češtiny:** Jan Hajič, Jaroslava Hlaváčová, David Kolovratník, Pavel Květoň
  - **Tagger:** Jan Raab
  - **Parser:** Ryan McDonald, Václav Novák, Kiril Ribarov
  - **Konverze csts2PML:** Petr Pajas
  - **Skript pro morfologické a syntaktické zpracování českých textů:** Michal Kebrt
  - **TrEd:** Petr Pajas
  - **Netgraph:** Jiří Mírovský
- **Bonusový materiál**
  - **STYX:** Ondřej Kučera
  - **TrEdVoice:** Leoš Přikryl
- **CD-ROM, webová stránka**
  - **Instalační skript:** Ondřej Bojar
  - **Domovská stránka, CD obal:** Michal Šotkovský
- **Průvodce ČAK**
  - **Technický editor:** Jan Raab
  - **Jazyková korektura:** TBA
  - **Překlad do angličtiny:** Alena Chrastová

Kapitola 9. Poděkování

Práce na Českém akademickém korpusu verze 2.0 byla podporována těmito organizacemi:

- *Grantová agentura Akademie věd České republiky*, granty č. 1ET101120413, 1ET101120503,
- *Grantová agentura Univerzity Karlovy*, grant č. 207-10/257559,
- *Ministerstvo školství, mládeže a tělovýchovy České republiky*, granty č. MSM0021620838,
- *Matematicko-fyzikální fakulty Univerzity Karlovy v Praze*,
- *Univerzita Karlova v Praze*.

Kapitola 10. Literatura

[Collins, 2002] Michael Collins: *Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms*. Proceedings of EMNLP´2002, University of Pennsylvania, Philadelphia, USA, 2002.

[Čermák, Blatná, 2005] František Čermák, Renata Blatná: *Jak využívat Český národní korpus*. Nakladatelství Lidové noviny, Praha, 2005.

[Hajič, 2004] Jan Hajič: *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Karolinum, Praha, 2004

[Hajič a kol., 2004] Jan Hajič, Jarmila Panevová, Eva Buráňová, Alevtina Bémová, Jan Štěpánek, Petr Pajas, Jiří Kárník: *Anotace na analytické rovině. Návod pro anotátory*. TR-2004-23, Ústav formální a aplikované lingvistiky, MFF UK, Praha, 2004.

[Hladká, 1994] Barbora Hladká: *Programové nástroje pro anotování velkých textových korpusů*. Diplomová práce, MFF UK, Praha, 1994

[Hladká, Králík, 2006] Barbora Hladká, Jan Králík: *Proměny Českého akademického korpusu*. Slovo a slovesnost, 67:179-194, 2006.

[Jelínek, Bečka, Těšitelová, 1961] Jaroslav Jelínek, Josef Václav Bečka, Marie Těšitelová: *Frekvence slov, slovních druhů a tvarů v českém jazyce (FSSDTČJ)*. SPN, Praha, 1961.

[Kopřivová, Kocek, 2000] Marie Kopřivová, Jan Kocek: *Český národní korpus, úvod a příručka uživatele*. Filozofická fakulta UK, Praha, 2000.

[Kučera, 2006] Ondřej Kučera: *Pražský závislostní korpus jako cvičebnice jazyka českého*. Diplomová práce, MFF UK, Praha, 2006.

[McDonald, Pereira, Ribarov, Hajič, 2005] Ryan McDonald, Fernando Pereira, Kiril Ribarov, Jan Hajič,: *Non-projective Dependency Parsing using Spanning Tree Algorithms*. Proceedings of HLT/EMNLP´2005, Vancouver, Canada, 2005.

[Mikulová a kol., 2006] Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský: *Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka*. TR-2005-28, Ústav formální a aplikované lingvistiky, MFF UK, Praha, 2005.

[Müller, Psutka, Šmídl, 2000] Luděk Müller, Josef Psutka, Luboš Šmídl: *Design of Speech Recognition Engine*. TSD 2000, Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin, Heidelberg, 2000.

[Pajas, Štěpánek, 2005] Petr Pajas, Jan Štěpánek: *A Generic XML-based Format for Structured Linguistic Annotation and its Application to the Prague Dependency Treebank 2.0*. TR-2005-29, Ústav formální a aplikované lingvistiky, MFF UK, Praha, 2005.

[Přikryl, 2007] Leoš Přikryl: *Rozhraní v mluveném jazyce pro korpusové anotační nástroje*. Diplomová práce, MFF UK, Praha, 2007.

[Psutka, Müller, Matoušek, Radová, 2006] Josef Psutka, Luděk Müller, Jindřich Matoušek, Vlasta Radová: *Mluvíme s počítačem česky*. Praha: Academia, 2006.

[Ribarov, 2004] Kiril Ribarov: *Automatic Building of a Dependency Tree - The Rule-Based Approach and Beyond*. Doktorská práce, MFF UK, Praha, 2004

[Ribarov, Bémová, Hladká, 2006] Kiril Ribarov, Alla Bémová, Barbora Hladká: *When a statistically oriented parser was more efficient than a linguist: A case of treebank conversion*. Prague Bulletin of Mathematical Linguistics 86, str. 21-38, 2006.

[Savický, Hlaváčová, 2002] Petr Savický, Jaroslava Hlaváčová: *Measures of Word Commonness*. Journal of Quantitative Linguistics. Swets & Zeitlinger, Vol. 9, No. 3, s. 215-231. 2002.

[Šmilauer, 1972] Vladimír Šmilauer: *Nauka o českém jazyku*. Praha, 1972

[Vidová Hladká a kol., 2007] Barbora Vidová Hladká, Jan Hajič, Jiří Hana, Jaroslava Hlaváčová, Jiří Mírovský, Jan Votrubec: *Průvodce Českým akademickým korpusem 1.0*. Nakladatelství Karolinum, Praha. 2007.

[Votrubec, 2005] Jan Votrubec: *Volba vhodné sady rysů pro morfologické značkování češtiny*. Diplomová práce, MFF UK, Praha, 2005

[Hana, Zeman, 2005] Jiří Hana, Daniel Zeman, Jan Hajič, Hana Hanová, Barbora Hladká, Emil Jeřábek: *Manual for Morphological Annotation*. TR-2005-27, Ústav formální a aplikované lingvistiky, MFF UK, Praha, 2005.

Příloha A. Zdroje textů

Tabulka A.1. Administrativní styl

soubor	psaná forma		soubor	mluvená forma
a01w	Vyhláška č. 100		a16s	Zelená vlna
a02w	Hospodaření s domovním bytovým majetkem		a17s	Zprávy o počasí
a03w	Pracovní řád		a18s	Přehled rozhlasových pořadů
a04w	Národní pojištění 12/1977		a19s	Hlášení v metru
a05w	Kolektivní smlouvy – TIBA			
a06w	Materiál – TIBA			
a07w	Zpráva o činnosti Ústavu pro jazyk český			
a08w	Metodické pokyny			
a09w	Zápisy z porad			
a10w	Závazky			
a11w	Zápisy ze schůzí			
a12w	Pokyny SÚRPMO			
a13w	Pracovní návody, pokyny			
a14w	Oběžníky Ústavu pro jazyk český			
a15w	Zpráva o činnosti oddělení matematické lingvistiky			
a20w	Hlášení v obchodním domě			

Tabulka A.2. Publicistický styl

soubor	psaná forma		soubor	mluvená forma
n01w	Rudé právo		n53s	Rozhlasové reportáže a rozhovory
n02w	Svět práce		n54s	Televizní komentáře
n03w	Práce		n55s	Zprávy čs. rozhlasu
n04w	Československý rozhlas I.		n56s	Televizní diskuse
n05w	Mladá fronta		n57s	Televizní zprávy a reportáže
n06w	Československý rozhlas II.		n58s	Rozhlasová diskuse
n07w	Večerní Praha		n59s	Televizní zprávy a lekce
n08w	Československý sport		n60s	Televizní diskuse a komentáře
n09w	Svobodné slovo			
n10w	Lidová demokracie			
n11w	Obrana lidu			
n12w	Týdeník aktualit			
n13w	Zemědělské noviny			
n14w	Gramorevue G 73			
n15w	Tribuna			
n16w	Záběr			
n17w	Úder			
n18w	Svoboda			
n19w	Služba lidu			
n20w	Zpravodaj TIBY			
n21w	Nové Hradecko			
n22w	Pochodeň			
n23w	Technický týdeník			
n24w	Horník a energetik			
n25w	Sázavan			
n26w	Čelákovický zpravodaj			
n27w	Nové Klatovsko			
n28w	Pravda			
n29w	Průboj			
n30w	Zpravodaj TIBY			
n31w	Krkonošská Pravda			
n32w	Školství a věda			
n33w	Stráž lidu			
n34w	Zbrojovák			
n35w	Nová svoboda			
n36w	Vlasta			
n37w	Mladý svět			
n38w	Naše rodina			
n39w	Ahoj na sobotu			
n40w	Květy			

soubor	psaná forma		soubor	mluvená forma
n41w	Signál			
n42w	Zahrádkář			
n43w	Film a doba			
n44w	Melodie			
n45w	Stadion			
n46w	Věda a technika mládeži			
n47w	Haló sobota			
n48w	Svět socialismu			
n49w	Zahradnické listy			
n50w	Kino			
n51w	Chovatel			
n52w	Zápisník Z'73			

Tabulka A.3. Odborný styl

soubor	psaná forma		soubor	mluvená forma
s01w	Dějiny české hudební kultury		s69s	Divadelní přehlídka
s02w	Motivace lidského chování		s70s	Výklad Zákoníku práce
s03w	Škola – opora socialismu		s71s	Opera o Bratrech Karamazových (prof. dr.Václav Holzknecht)
s04w	Jak rozumíme chemickým vzorcům a rovnicím		s72s	Zpráva o cestě do Belgie (PhDr. Marie Těšitelová, DrSc.)
s05w	Konflikty mezi lidmi		s73s	Obecné otázky jazykové kultury
s06w	Škoda 1000		s74s	Provozní kontrola potrubí
s07w	Pražský vodovod		s75s	Modelování diod
s08w	Nauka o materiálu		s76s	Přenosové parametry
s09w	Tranzistory řízené elektrickým polem		s77s	O počtu koster jednoho grafu
s10w	Pro půvab a eleganci		s78s	Streptokoky
s11w	Tisíciletý vývoj architektury		s79s	Statické zajištění domu U Rytířů
s12w	Polovodičová technika		s80s	Problémy aerodynamiky závodních vozů
s13w	Plazma, čtvrté skupenství hmoty		s81s	Schůze vědecké rady ČSTV
s14w	Nadhodnota a její formy		s82s	Plenární schůze ROH / Pauzy váhání
s15w	Určování efektivnosti za socialismu		s83s	Seminář o houbách
s16w	Stažlivost myokardu		s84s	Česká filharmonie hraje a hovoří (Václav Neumann)
s17w	K biologickým a psychologickým zřetelům výchovy		s85s	Seminář o fotografii
s18w	Poetika		s86s	Působení hromadných sdělovacích prostředků
s19w	Slovo a slovesnost 4/1973		s87s	Ochrany v průmyslových závodech
s20w	Sociologický časopis 3/1973		s88s	Práce se čtenářem
s21w	Teorie a empirie		s89s	Dlouhodobé skladování masa
s22w	Česká literatura		s90s	Personalistika
s23w	Československá informatika		s91s	Archeologické nálezy v Toušeni (Jaroslav Špaček)
s24w	Národopisné aktuality		s92s	Přednáška o geografii
s25w	Vlastivědný sborník moravský		s93s	Úvod do dějin feudalismu
s26w	Český lid		s94s	Filosofie fyziky (RNDr. Jiří Mrázek, CSc.)
s27w	Otázky lexikální statistiky		s95s	O vývoji knihovnictví
s28w	Památková péče 4/1974		s96s	Základní podmínky pro pěstování zeleniny
s29w	Základní a rekreační tělesná výchova 10/1974		s97s	O výchově socialistické inteligence
s30w	Společenské vědy ve škole 2/1974		s98s	Petrologie sedimentů a reziduálních hornin
s31w	Hospodářské právo		s99s	Organizace a řízení vnitřního obchodu

soubor	psaná forma		soubor	mluvená forma
s32w	Sociální jistoty včera a dnes		s00s	Rozbor situace v JZD
s33w	Arbitrážní praxe			
s34w	Filosofický časopis 5/1974			
s35w	Československá psychologie			
s36w	Společenská struktura a revoluce			
s37w	Humanismus v naší filosofické tradici			
s38w	Společnost – vzdělání – jedinec			
s39w	Rozvoj osobnosti a slovesné umění			
s40w	Ke kritice buržoasních teorií společnosti			
s41w	Spisovný jazyk v současné komunikaci			
s42w	Přirozený jazyk v informačních systémech			
s43w	Česká literatura			
s44w	NA			
s45w	Vědeckotechnická revoluce a socialismus			
s46w	Zesilovače se zpětnou vazbou			
s47w	Teorie a počítače v geofyzice			
s48w	Výzkum hlubinné geologické stavby Československa			
s49w	Podstata hypnózy a spánek			
s50w	Nukleární medicína			
s51w	Hutnictví a strojírenství			
s52w	Záruční lhůty potravinářských výrobků			
s53w	Mineralogie			
s54w	Ptáci			
s55w	Elektronický obzor 6/1974			
s56w	Teplárenství			
s57w	Vědecko-technický rozvoj za socialismu			
s58w	Jak na práce se stavebninami			
s59w	NA			
s60w	Obkládáme interiéry a fasády			
s61w	Alpinkářův svět			
s62w	Opravujeme a modernizujeme rodinný domek			
s63w	Jak na práce s kovem			
s64w	Astronomie			
s65w	Pokroky matematiky, fyziky a astronomie			
s66w	Elektrotechnický obzor			
s67w	Hvězdářská ročenka			
s68w	Lékařská fyzika			

Příloha B. Popis lemmat

Obecný tvar lemmatu je *lemma_:P1_;P2_,P3_^(K)*, kde *lemma* je vlastní lemma a *P1*, *P2*, *P3*, *K* jsou nepovinné doplňkové informace.

Tabulka B.1. Struktura doplňkových informací lemmat

označení	oddělovač	popis	vysvětlivky
P1	:	morfosyntaktický příznak	slovní druh nebo jeho bližší určení
P2	;	sémantický příznak	obecná sémantická kategorie
P3	,	stylový příznak	stylové zařazení lemmatu
K	^	komentář	vysvětlivka, způsob odvození, jiný komentář

Tabulka B.2. Morfosyntaktické příznaky lemmat

Hodnota	Popis
B	zkratka
T	verbum imperfectum (nedokonavé sloveso)
W	verbum perfectum (dokonavé sloveso)

Tabulka B.3. Sémantické příznaky lemmat

Hodnota	Popis
E	příslušník národa, obyvatel území
G	zeměpisný název
H	chemie
K	společnost, organizace, instituce
L	přírodní vědy
R	výrobek
S	příjmení
U	lékařství
Y	křestní jméno
b	ekonomie, finance
c	výpočetní technika a elektronika
g	technologie
j	právo
m	ostatní vlastní jména
o	specifikace barev
p	politika, vláda, armáda
u	kultura, vzdělávání, umění, další vědy
w	sport
y	koníčky, volný čas, cestování
z	ekologie, životní prostředí

Tabulka B.4. Stylové příznaky lemmat

Hodnota	Popis
a	zastaralé
e	expresivní
h	hovorové
l	slangové, hantýrka
n	nářečí
s	knižní
t	cizí slovo
v	vulgární
x	zastaralý pravopis, pravopisná chyba

Tabulka B.5. Příklady lemmat

lemma	doplňkové informace	vysvětlivky
Abchaz	_;E	příslušník národa
Agned	_;Y_,t	křestní jméno|cizí slovo
dobromysl	_;L	přírodní vědy
dementi	_,t	cizí slovo
FFUK	_:B_;K_;u-^(Filozof._fakulta_Univerzity_Karlovy)	zkratka|instituce|kultura, vzdělávání|vysvětlení zkratky
líně	_^(*1ý)	odvození: odtrhni od konce 1 znak, nahraď ho znakem „ý“ a vznikne lemma, ze kterého je „líně“ odvozeno, tedy „líný“

Příloha C. Popis morfologických značek

Tabulka C.1. Slovní druh

Hodnota	Popis
A	adjektivum (přídavné jméno)
C	číslovka nebo číselný výraz s číslicemi
D	adverbium (příslovce)
I	interjekce (citoslovce)
J	konjunkce (spojka)
N	substantivum (podstatné jméno)
P	pronomen (zájmeno)
V	verbum (sloveso)
R	prepozice (předložka)
T	partikule (částice)
X	neznámý, neurčený, neurčitelný slovní druh
Z	interpunkce, hranice věty

Tabulka C.2. Slovní poddruh

Hodnota	Popis
#	hranice věty
%	autorova signatura (např. „haš-99_:B_;S“)
*	slovo „krát“
,	spojka podřadicí (včetně „aby“, „kdyby“ ve všech tvarech)
}	číslovka psaná římskými číslicemi (např. XIV)
:	interpunkce všeobecně
=	číslo zapsané číslicemi
?	číslovka „kolik“
@	slovní tvar, který není morfologickou analýzou rozpoznán
^	spojka souřadicí
4	zájmeno vztažné/tázací s adjektivním skloňováním (např. „jaký“, „který“, „čí“)
5	zájmeno „on“ ve tvarech po předložce (např. „něj“, „něho“)
6	zájmeno reflexivní „se“ v dlouhých tvarech (např. „sebe“)
7	zájmeno reflexivní „se“, „si“, „ses“, „sis“
8	zájmeno přivlastňovací „svůj“
9	zájmeno vztažné „jenž“, „již“, ... po předložce (n-: „něhož“, „niž“, ...)
A	adjektivum
B	sloveso, tvar přítomného nebo budoucího času
C	adjektivum, jmenný tvar (např. „rád“, „schopen“)
D	zájmeno ukazovací (např. „ten“, „onen“)
E	zájmeno vztažné „což“
F	součást předložky, která nikdy nestojí samostatně (např. „nehledě (na)“, „vzhledem (k)“)
G	adjektivum odvozené od přítomného přechodníku
H	krátké tvary osobních zájmen (např. „mě“, „mi“, „ti“, „mu“)
I	citoslovce
J	zájmeno vztažné „jenž“, „již“, ... bez předložky
K	zájmeno tázací/vztažné „kdo“ vč. tvarů s „-ž“ a „-s“
L	zájmeno neurčité (např. „všechen“, „sám“)
M	adjektivum odvozené od minulého přechodníku
N	substantivum
O	samostatně stojící zájmena (např. „svůj“, „nesvůj“, „tentam“)
P	zájmena osobní (např. „já“, „ty“, „on“) vč. tvarů s „-s“ (např. „tys“)
Q	zájmeno tázací/vztažné „co“, „copak“, „cožpak“
R	předložka (obecná, bez vokalizace)
S	zájmeno přivlastňovací „můj“, „tvůj“, „jeho“, vč. plurálu
T	částice
U	adjektivum přivlastňovací (např. „-ův“, „-in“)
V	předložka vokalizovaná (např. „ve“, „ku“)
W	zájmena záporná (např. „nic“, „nikdo“, „nijaký“, „žádný“)
X	slovní tvar, který byl rozpoznán morfologickou analýzou, ale značka chybí

Hodnota	Popis
Y	zájmeno tázací/vztažné „co“ spojené s předložkou („oč“, „nač“, „zač“)
Z	zájmeno neurčité (např. „nějaký“, „některý“, „číkoli“, „cosi“)
a	číslovka neurčitá (např. „mnoho“, „málo“, „tolik“, „několik“, „kdovíkolik“)
b	příslovce, které se neskloňuje ani ho nelze negovat (např. „pozadu“, „naplocho“)
c	kondicionál slovesa „být“ („by“, „bych“, „bys“, „bychom“, „byste“)
d	číslovka druhová s adjektivním skloňováním (např. „dvojí“, „desatery“)
e	přechodník přítomný
f	infinitiv
g	příslovce, které se skloňuje a časuje
h	číslovka druhová (např. „jedny“, „nejedny“)
i	sloveso ve tvaru rozkazovacího způsobu
j	číslovka druhová větší nebo rovna 4 v substantivním postavení (např. „čtvero“, „desatero“)
k	číslovka druhová větší nebo rovna 4 v adjektivním postavení, krátký tvar (např. „čtvery“)
l	číslovka základní s nesubstantivním skloňováním (např. „jeden“, „dva“, „tři“, „čtyři“, „půl“, „sto“, „tisíc“)
m	přechodník minulý (např. „udělav“)
n	číslovka základní větší nebo rovna 5
o	číslovky násobná neurčitá (např. „mnohokrát“, „tolikrát“)
p	příčestí činné (vč. přidaného „-s“)
q	archaické příčestí činné (zakončení „-ť“)
r	číslovka řadová
s	příčestí trpné (vč. přidaného „-s“)
t	archaický slovesný tvary přítomného a budoucího času (zakončení „-ť“)
u	číslovka tázací násobná „kolikrát“
v	číslovka násobná (např. „pětkrát“)
w	číslovka neurčitá s adjektivním skloňováním (např. „nejeden“, „tolikátý“)
y	zlomek zakončený na „-ina“ (např. „pětina“)
z	číslovka tázací řadová „kolikátý“

Tabulka C.3. Rod

Hodnota	Popis
-	neurčuje se
F	femininum (ženský rod)
N	neutrum (střední rod)
H	femininum nebo neutrum
I	maskulinum inanimatum (rod mužský neživotný)
M	maskulinum animatum (rod mužský životný)
Q	femininum singuláru nebo neutrum plurálu (pouze u příčestí a jmenných adjektiv)
T	maskulinum inanimatum nebo femininum plurálu (pouze u příčestí a jmenných adjektiv)
X	libovolný rod (F/M/I/N)
Y	maskulinum (animatum nebo inanimatum)
Z	nikoli femininum, tj. M/I/N

Tabulka C.4. Číslo

Hodnota	Popis
-	neurčuje se
D	duál (pouze 7. pád feminin)
P	plurál (množné číslo)
S	singulár (jednotné číslo)
W	pouze v kombinaci se jmenným rodem Q (singulár pro feminina, plurál pro neutra)
X	libovolné číslo (D/S/P)

Tabulka C.5. Pád

Hodnota	Popis
-	neurčuje se
1	nominativ (první pád)
2	genitiv (druhý pád)
3	dativ (třetí pád)
4	akuzativ (čtvrtý pád)
5	vokativ (pátý pád)
6	lokál (šestý pád)
7	instrumentál (sedmý pád)
X	libovolný pád (1/2/3/4/5/6/7)

Tabulka C.6. Přivlastňovací rod

Hodnota	Popis
-	neurčuje se
F	femininum
M	maskulinum animatum (pouze u adjektiv)
X	libovolný rod (M/I/F/N)
Z	nikoli femininum, tj. M/I/N

Tabulka C.7. Přivlastňovací číslo

Hodnota	Popis
-	neurčuje se
P	plurál
S	singulár
X	libovolné číslo (S/P)

Tabulka C.8. Osoba

Hodnota	Popis
-	neurčuje se
1	1. osoba
2	2. osoba
3	3. osoba
X	libovolná osoba (1/2/3)

Tabulka C.9. Čas

Hodnota	Popis
-	neurčuje se
F	futurum (budoucí čas)
P	prézens (přítomný čas)
R	préteritum (minulý čas)
H	prézens nebo préteritum (P/R)
X	libovolný čas (F/R/P)

Tabulka C.10. Stupeň

Hodnota	Popis
-	neurčuje se
1	1. stupeň
2	2. stupeň
3	3. stupeň

Tabulka C.11. Negace

Hodnota	Popis
-	neurčuje se
A	afirmativ (bez negativní předpony „ne-“)
N	negace (tvar s negativní předponou „ne-“)

Tabulka C.12. Aktivum/pasivum

Hodnota	Popis
-	neurčuje se
A	aktivum
P	pasivum

Tabulka C.13. Nepoužito

Hodnota	Popis
-	neurčuje se

Tabulka C.14. Nepoužito

Hodnota	Popis
-	neurčuje se

Tabulka C.15. Varianta, stylový příznak apod.

Hodnota	Popis
-	základní tvar
1	varianta k základnímu tvaru
2	řídká, archaická nebo knižní varianta k základnímu tvaru
3	velmi archaický tvar, též hovorový
4	velmi archaický nebo knižní tvar, pouze spisovný
5	hovorový tvar (ve veřejných projevech)
6	hovorový tvar (koncovka obecné češtiny)
7	hovorový tvar, varianta k 6
8	zkratky
9	speciální použití (např. tvary zájmen po předložkách)

Příloha D. Popis analytických funkcí

Tabulka D.1. Analytické funkce v ČAK 2.0

Analytická funkce	Popis	Analytická funkce	Popis	Analytická funkce	Popis	Analytická funkce	Popis
Pred	Predikát, resp. uzel, který nezávisí na jiném uzlu; věší se na #.	Pnom	Predikát nominální, resp. jmenná část přísudku se sponou být.	AuxC	Spojka (podřadicí)	AuxK	
Sb	Subjekt (podmět)	AuxV	Pomocné sloveso být (Auxiliary Verb).	AuxO	Nadbytečný (odkazovací, emotivní) element.	ExD	Náhradní funkce pro technické hrany vedoucí místo od elidovaného členu k "pseudořídícímu" slovu nebo pro hlavní člen věty bez predikátu (Ex-Dependent).
Obj	Objekt (předmět).	Coord	Koordinační uzel (souřadné spojení).	AuxZ	Zdůrazňovací slovo.	AtrAtr	Řídícím slovem atributu může být díky strukturní víceznačnosti kterékoli z bezprostředně předcházejících (syntaktických) substantiv.
Adv	Adverbiale (příslovečné určení, bez dalšího rozlišení).	Apos	Aposice (hlavní uzel).	AuxX	Čárka (ne však nositel koordinace).	AtrAdv	Strukturní víceznačnost mezi závislostí adverbální (příslovečnou) a adnominální (zavěšení na jméno) bez sémantických důsledků.
Atv	Doplněk (jen tzv. určující), technicky zavěšen na neslovesném členu.	AuxT	Zvratné se, neoddělitelné se - reflexivní tantum.	AuxG	Jiné grafické symboly, které neukončují větu.	AdvAtr	Dtto, s opačnou preferencí.
AtvV	Doplněk (jen tzv. určující), visící na slovese (chybí druhý řídící člen).	AuxR	Zvratné se, které není Obj ani AuxT (tvoří pasivum reflexivní).	AuxY	Příslovce a částice, které nelze zařadit jinam.	AtrObj	Strukturní víceznačnost mezi závislostí objektovou a adnominální (zavěšení na substantivum) bez sémantických důsledků.
Atr	Atribut (přívlastek).	AuxP	Předložka primární, části předložky sekundární.	AuxS	Kořen stromu (#).	ObjAtr	Dtto, s opačnou preferencí.

Příloha E. Pavučina

Tabulka E.1. Internetové odkazy

	Název odkazu (stručný popis)	**Adresa**
	PROJEKTY	
1.	Data a nástroje pro informační systémy (projekt zastřešující ČAK 2.0)	http://ufal.mff.cuni.cz/rest
2.	Morfologické značkování češtiny (souhrnný přehled)	http://ufal.mff.cuni.cz/czech-tagging
	INSTITUCE	
3.	Akademie věd České republiky	http://www.av.cz
4.	Grantová agentura Akademie věd České republiky	http://www.gaav.cz
5.	Katedra kybernetiky Západočeské univerzity v Plzni	http://http://www.kky.zcu.cz [http://www.kky.zcu.cz]
6.	Ministerstvo školství, mládeže a tělovýchovy České republiky	http://www.msmt.cz
7.	Univerzita Karlova v Praze	http://www.cuni.cz
8.	Ústav formální a aplikované lingvistiky, Matematicko-fyzikální fakulty Univerzity Karlovy v Praze	http://ufal.mff.cuni.cz
9.	Ústav pro jazyk český Akademie Věd české republiky	http://mam.ujc.cas.cz/ujc
	DATOVÉ ZDROJE, MANUÁLY	
10.	csts DTD (vnitřní formát anotovaných dat založený na SGML)	http://ufal.mff.cuni.cz/pdt2.0/doc/pdt-guide/cz/html/ch03.html#a-data-formats-csts
11.	Prague Markup Language (vnitřní formát anotovaných dat založený na XML)	http://ufal.mff.cuni.cz/jazz/PML [http://ufal.mff.cuni.cz/pml]
12.	Pražský závislostní korpus	http://ufal.mff.cuni.cz/pdt
13.	Manuál morfologického značkování PZK	http://ufal.mff.cuni.cz/pdt2.0/ doc/manuals/en/m-layer/html/index.html [http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/m-layer/html/index.html]
14.	Manuál syntaktickoanalytického značkování PZK	http://ufal.mff.cuni.cz/pdt2.0/ doc/manuals/cz/a-layer/html/index.html [http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/cz/a-layer/html/index.html]

	Název odkazu (stručný popis)	**Adresa**
15.	Manuál tektogramatického značkování PZK	http://ufal.mff.cuni.cz/pdt2.0/ doc/manuals/cz/t-layer/html/index.html [http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/cz/t-layer/html/index.html]
16.	Relax NG (XML schéma)	http://www.relaxng.org
17.	SGML	http://www.w3.org/MarkUp/SGML/
18.	XML	http://www.w3.org/XML
	NÁSTROJE	
19.	Bonito (grafická nadstavba korpusového manažeru)	http://nlp.fi.muni.cz/projekty/bonito/
20.	LAW (anotační morfologický editor)	http://www.ling.ohio-state.edu/ ~hana/law.html [http://www.ling.ohio-state.edu/~hana/law.html]
21.	Netgraph (nástroj pro vyhledávání v závislostních korpusech)	http://quest.ms.mff.cuni.cz/netgraph
22.	STYX (cvičebnice češtiny založená na PZK)	http://ufal.mff.cuni.cz/styx
23.	TrEd (anotační syntaktický editor)	http://ufal.mff.cuni.cz/~pajas/tred
24.	TNT (Trigrams'n'Tags tagger)	http://www.coli.uni-saarland.de/ ~thorsten/tnt/ [http://www.coli.uni-saarland.de/~thorsten/tnt/]