M A T E M A T I C K O - F Y Z I K Á L N Í F A K U L T A

P R A H A

PRAŽSKÝ ZÁVISLOSTNÍ KORPUS MLUVENÉ ČEŠTINY

REKONSTRUKCE STANDARDIZOVANÉHO TEXTU Z MLUVENÉ ŘEČI

Jan Hajič, Marie Mikulová, Martina Otradovcová,
Petr Pajas, Petr Podveský, Zdeňka Urešová

úfal/ckl technical report
TR-2006-33

U N I V E R S I T A S C A R O L I N A P R A G E N S I S

Pražský závislostní korpus mluvené češtiny

Rekonstrukce standardizovaného textu z mluvené řeči

Jan Hajič
Marie Mikulová
Martina Otradovcová
Petr Pajas
Petr Podveský
Zdeňka Urešová

ÚFAL/CKL Technická zpráva TR-2006-33
ISSN 1214-5521

Copies of ÚFAL/CKL Technical Reports can be ordered from:

Institute of Formal and Applied Linguistics (ÚFAL MFF UK)
Faculty of Mathematics and Physics, Charles University
Malostranské nám. 25, CZ-11800 Prague 1
Czech Republic

or can be obtained via the Web: http://ufal.mff.cuni.cz

Pražský závislostní korpus mluvené češtiny: Rekonstrukce standardizovaného textu z mluvené řeči

Jan Hajič, Marie Mikulová, Martina Otradovcová, Petr Pajas, Petr Podveský a Zdeňka Urešová

Abstrakt

Technická zpráva shrnuje dosavadní poznatky získané při budování Pražského závislostního korpusu mluvené češtiny (Prague Dependency Treebank of Spoken Czech; PDTSC) v Ústavu formální a aplikované lingvistiky MFF UK Praha. PDTSC bude prvním korpusem mluvené řeči, který bude obsahovat anotace významu promluv. Výzkum ukázal, že před vlastní hloubkovou analýzou mluvené řeči je nezbytné transkribované segmenty mluvené řeči nejprve standardizovaným způsobem převést na gramaticky správné věty, tj. provést tzv. rekonstrukci standardizovaného textu z mluvené řeči.

Obsah

Seznam obrázků

Seznam tabulek

Acknowledgement

Technická zpráva vznikla za finanční podpory projektů: LC 536, GAČR 405/06/0589 a GAUK 375/2005 a GAUK 352/2005.

Úvod

V lingvistické analýze mluvené češtiny postoupila česká lingvistika poměrně daleko. Na řadě bohemistických pracovišť se systematicky shromažďují autentické nahrávky mluvené řeči a vznikají korpusy mluvené češtiny (viz 1.2.1 – „Korpusy mluvené češtiny“). Nahrávky jsou přepisovány podle různých transkripčních pravidel s různou mírou přídavné anotace. Odlišují se od sebe svým rozsahem, hloubkou, zaměřením tematickým, žánrovým, teritoriálním atd. Korpusy se používají pro rozmanitý lingvistický výzkum (například fonetický, morfologický, syntaktický, stylistický, interakční). Vznikla i řada teoretických prací v této oblasti (viz 1.2.2 – „Lingvistické práce o mluvené češtině“).

Na rozdíl od výzkumu psaného textu se však výzkum mluveného jazyka soustřeďuje většinou jen na přepis akustického signálu do textové podoby. Rozsah lingvistické anotace těchto transkripcí je nevelký, anotace se obvykle zastavuje na morfologické rovině či se přidává prozodické značkování; další zpracování směrem k významu se neprovádí. Tomu odpovídá i stav zdrojů, tj. jazykových dat, vhodných pro pravděpodobnostní trénování a strojové učení za účelem plného porozumění mluvené řeči. Dosud neexistuje (ani ve světovém měřítku) systematicky a ve větším rozsahu manuálně anotovaný korpus na úrovni syntaktické, tím méně na úrovni jazykového významu (na úrovni hloubkově syntaktické). Více viz 1.1 – „Výzkum mluvené řeči ve světě“.

V Ústavu formální a aplikované lingvistiky (na MFF UK Praha) byl v roce 2005 zahájen projekt Pražského závislostního korpusu mluvené češtiny (Prague Dependency Treebank of Spoken Czech, dále PDTSC). Cílem projektu je vybudovat komplexně anotovaný korpus mluvené češtiny. Základním východiskem pro hloubkovou analýzu mluvené češtiny je hloubková anotace korpusu psaných textů, která je zpracována v projektu Pražského závislostního korpusu verze 2.0 (viz 3.1 – „Východiska pro syntakticko-sémantickou analýzu mluvené řeči“). Předpokládá se, že zkušenosti získané budováním korpusu mluvené češtiny budou využity i pro budování komplexně anotovaných korpusů dalších jazyků.

Původním záměrem při budování PDTSC bylo syntakticko-sémanticky anotovat mluvenou řeč podle pravidel pro anotaci psaných textů Pražského závislostního korpusu a tato pravidla pouze upravovat a rozšiřovat pro zvláštnosti mluvené řeči. Lingvistická analýza mluvené řeči a první pokusy se syntaktickou anotací však ukázaly, že vzhledem ke specifičnosti mluveného projevu (přeřeknutí, koktání, smích, opakování slov atp.) je tento původní záměr zřejmě neschůdný. Na základě těchto zjištění a v souladu s celosvětově novými směry výzkumu v oblasti zpracování mluvené řeči (viz 2 – „*Standardizace mluvené řeči*“) byl proto stanoven nový postup: před vlastní hloubkovou analýzou se segmenty mluvené řeči nejprve převedou na gramaticky správné věty, tj. provede se tzv. rekonstrukce standardizovaného textu z mluvené řeči. Klíčové přitom je zachování (anotování) vazeb mezi původní transkripcí mluvené řeči a rekonstruovaným standardizovaným textem. Standardizovaný text bude následně anotován podle stávajících pravidel tektogramatické anotace, aniž by se tato pravidla musela nějak výrazně upravovat.

Tento nový postup je navíc podepřen tím, že i pro automatickou závislostní analýzu je nezbytný jazykově korektní vstup, neboť současné metody analýzy vět přirozeného jazyka na gramaticky a lexikálně „nesprávném“ textu nedávají dobré výsledky.

Jako první data pro práci na projektu byla zvolena data z projektu Malach (viz 4.1 – „Data“). Pro manuální rekonstrukci standardizovaného textu z mluvené řeči byla vytvořena první verze speciálního softwarového nástroje MEd (viz 4.2 – „Anotační nástroj MEd“). Zároveň byl vypracována první verze manuálu pro anotátory.

Kapitola 1. Výzkum mluvené řeči

1.1. Výzkum mluvené řeči ve světě

Podporou a propagací výzkumu mluvené řeči se od roku 1988 zabývala organizace ESCA (European Speech Communication Association), která se později (v roce 1999) změnila v mezinárodní organizaci ISCA (International Speech Communication Association; http://www.isca-speech.org/index.html). ISCA organizuje od roku 2000 pravidelné mezinárodní vědecké konference Interspeech (http://www.interspeech2007.org/), které vznikly spojením původních konferencí Eurospeech (European Conference on Speech Communication and Technology) a ICSLP (International Conference on Spoken Language Processing). Konference Interspeech zahrnuje všechny aspekty vědy, výzkumu a technologií v oblasti mluvené řeči.

Důležitou složkou výzkumu mluvené řeči jsou mluvené korpusy. Pro jejich anotování je třeba vytvářet normy a standardizovat anotační uspořádání mluvených korpusů. Těchto úkolů se ujala organizace EAGLES (Expert Advisory Group on Language Engineering Standards; http://www.ilc.cnr.it/EAGLES96/home.html) a US-ISLE Spoken Language Group (International Standards in Language Engineering; http://www.ldc.upenn.edu/sb/isle.html). Zásadní publikací anotačních norem je zejména Handbook of Standards and Resources for Spoken Language Systems (viz Gibbon — Moore — Winski, 1997).

K projektům, které usilovaly o vývoj nástrojů pro anotaci mluvené řeči, patří například projekt MATE (Multilevel Annotation Tools Engineering; http://mate.nis.sdu.dk/) nebo projekt NITE (Natural Interactivity Tools Engineering; http://nite.nis.sdu.dk/). Široce používaným nástrojem je například Praat, který byl vytvořen v Institute of Phonetics na Univerzitě v Amsterodamu (http://www.fon.hum.uva.nl/praat/). V tomto nástroji je možné poslouchat a vizualizovat řečový signál a ve stejné době vytvářet a kontrolovat ortografickou transkripci. Jiným nástrojem je Speech Filing System (SFS; http://www.phon.ucl.ac.uk/resource/sfs/) pocházející z Ústavu fonetiky a lingvistiky na University College London nebo anotační nástroj Transcriber vyvinutý ve výzkumných centrech ve Francii a volně dostupný na http://trans.sourceforge.net.

K nejznámějším světovým korpusům mluveného jazyka patří zejména korpusy, které byly publikovány v Linguistic Data Consortium (http://www.ldc.upenn.edu/Catalog/byType.jsp#speech), ale i jiné. V následujícím seznamu uvádíme přehled shromážděných korpusů mluveného jazyka (seznam zcela jistě není úplný).

Přehled mluvených korpusů

- **Switchboard Corpus**, angličtina (http://www.cis.upenn.edu/~treebank/switchboard-sample.html);
- **CHILDES Database**, angličtina a další jazyky (http://childes.psy.cmu.edu/data/);
- **International Corpus of English** (ICE), angličtina (http://www.ucl.ac.uk/english-usage/ice/index.htm);
- **Corpus Gesproken Nederlands**; holandština (http://lands.let.kun.nl/cgn/ehome.htm);
- **Tübingen Treebank of Spoken German** (TüBa-D/S), němčina (http://www.sfs.uni-tuebingen.de/en_tuebads.shtml);
- **Tübingen Treebank of Spoken English** (TüBa-E/S), angličtina (http://www.sfs.uni-tuebingen.de/en_tuebaes.shtml);
- **Tübingen Treebank of Spoken Japanese** (TüBa-J/S), japonština (http://www.sfs.uni-tuebingen.de/en_tuebajs.shtml);
- **Swedish Treebank**, švédština (http://w3.msi.vxu.se/~nivre/research/st.html);
- **British National Corpus** (BNC), angličtina (http://www.natcorp.ox.ac.uk/);
- **British Academic Spoken English (BASE) corpus**, angličtina (http://www2.warwick.ac.uk/fac/soc/celte/research/base/);

- **Saarbrücken Corpus of Spoken English** (SCoSE), angličtina (http://www.uni-saarland.de/fak4/norrick/scose.htm);
- **London Lund Corpus**, angličtina (http://khnt.hit.uib.no/icame/manuals/LONDLUND/INDEX.HTM);
- **Bergen Corpus of London Teenage Language** (COLT), angličtina (http://www.hf.uib.no/i/Engelsk/COLT/);
- **Limerick corpus of Irish English** (L-CIE), angličtina (http://www.ul.ie/~lcie/);
- **COBUILD — The Bank of English**, angličtina (http://www.titania.bham.ac.uk/docs/svenguide.html);
- **Machine Readable Spoken English Corpus** (MARSEC), angličtina (http://www.rdg.ac.uk/AcaDepts/ll/speechlab/marsec/);
- **Lancaster/IBM Spoken English Corpus** (SEC), angličtina (http://www.comp.leeds.ac.uk/amalgam/tagsets/sec.html);
- **TRAINS Spoken Dialog Corpus**, angličtina (http://www.cs.rochester.edu/research/cisd/resources/trains.html);
- **Santa Barbara Corpus of Spoken American English** (SBCSAE), angličtina (http://projects.ldc.upenn.edu/SBCSAE/);
- **Corpus of Spoken, Professional American-English**, angličtina (http://www.athel.com/cspatg.html);
- **Boston University Radio Speech Corpus**, angličtina (http://ssli.ee.washington.edu/projects/radio.html);
- **Michigan Corpus of Academic Spoken English** (MICASE), angličtina (http://www.lsa.umich.edu/eli/micase/index.htm);
- **Australian National Database of Spoken Language** (ANDOSL), angličtina (http://andosl.anu.edu.au/andosl/);
- **Wellington Corpora of New Zealand English** (WSC), angličtina (http://www.vuw.ac.nz/lals/corpora/index.aspx);
- **Spanish spoken corpus of youth language**, španělština (COLA) (http://www.corpus.bham.ac.uk/PCLC/cl-195-pap-COLA.doc);
- **Hypermedia Corpus of Spoken Japanese**, japonština (http://www.env.kitakyu-u.ac.jp/corpus/docs/index.html);
- **Multilingual Spoken Corpus**, turečtina, francouzština, španělština, japonština (http://www.coelang.tufs.ac.jp/english/language_function.html);
- **Danish Speech Corpus** (BySoc), dánština (http://www.ling.gu.se/projekt/nordtalk/members_resources/BySoc.html);
- **The Danish Phonetically Annotated Spontaneous Speech** (DannPASS), dánština (http://www.cphling.dk/~ng/danpass_webpage/danpass.htm);
- **Corpus of Spoken Norwegian** (NoTa), norština (http://www.tekstlab.uio.no/nota/english/index.html);
- **Big Brother-korpuset**, norština (http://www.tekstlab.uio.no/talespraak/bigbrother/);
- **Göteborg Spoken Language Corpus** (GSLC), švédština (http://www.ling.gu.se/projekt/tal/index.cgi?PAGE=3);
- **Corpus of Spoken Estonian**, estonština (http://www.cl.ut.ee/suuline/Korpus.php);
- **Corpus of Spoken Bulgarian**, bulharština (http://www.hf.uio.no/ilos/studier/studenttjenester/Nettressurser/bulg/mat/Aleksova/);
- **Spoken Corpus of Slovene**, slovinština (http://helmer.hit.uib.no/batmult/Janas_Final_Report.htm).

Základem mluvených korpusů je manuální, nebo automatická fonetická transkripce. Pokud jsou mluvené korpusy obohaceny o další anotaci, je to zejména anotace slovního druhu (part of speech), někdy zahrnující i více či méně

úplnou morfologickou informaci. Gramatické značkování na úrovni vyšší než slovnědruhové lze najít například v korpusu London Lund Corpus, kde byla provedena anotace týkající se promluvových (textových) ukazatelů; jinou anotací je vyznačování anaforických vztahů v textu nebo anotace prozódie.

Jen málo korpusů mluvené řeči je anotováno syntakticky. Syntaktická anotace byla zjištěna u prvních osmi uvedených korpusů (více k syntakticky anotovaným korpusům viz následující sekci 1.1.1 – „Syntaktické korpusy mluvené řeči“).

1.1.1. Syntaktické korpusy mluvené řeči

Syntaktická anotace v mluvených korpusech je většinou prováděna automaticky s využitím nástrojů natrénovaných na korpusech psaných textů nebo na manuálně oanotované části mluveného korpusu (například International Corpus of English, http://www.ucl.ac.uk/english-usage/ice/index.htm; Switchboard, viz 1.1.1.1 – „The Switchboard corpus“; Childes database, viz 1.1.1.4 – „The Childes database“). Manuální anotace byla zjištěna v korpusech projektu Verbmobil (viz 1.1.1.3 – „The Verbmobil treebanks“) a v Korpusu mluvené holandštiny (viz 1.1.1.2 – „The Spoken Dutch Corpus“); u všech těchto korpusů šlo o anotaci poloautomatickou za pomoci anotačního nástroje Annotate, vyvinutého na univerzitě v Saarbrücken.

1.1.1.1. The Switchboard corpus

Známý **Switchboard korpus** (http://www.ldc.upenn.edu/Catalog/docs/switchboard/) vznikl v rámci projektu PennTreebank (http://www.cis.upenn.edu/~treebank/home.html) na University of Pennsylvania.

Switchboard korpus tvoří 2 438 nahrávek telefonních rozhovorů dlouhých od 5 do 10 minut, namluvených 520 dospělými mluvčími (každý mluvčí se objevuje průměrně v devíti nahrávkách). Jednotliví mluvčí mluví nejrůznějšími dialekty americké angličtiny. Celý korpus představuje 240 hodin mluvené řeči — přibližně 3 milióny slov.

Manuálně transkribovaný korpus Switchboard je následně používán k dalším anotacím a výzkumům na různých pracovištích. V rámci projektu PennTreebank byla do korpusu přidána anotace slovního druhu (part-of-speech tagging; viz Santorini, B., 1990) a následně byl korpus automaticky syntakticky oanotován (syntactic parsing). Jako přípravný krok k těmto dvěma anotacím byla nejprve provedena anotace tzv. dysfluencí, tj. v mluveném textu byly označeny jevy specifické pro mluvenou řeč: hezitační zvuky, zakoktání, fragmenty, opakovaní slov aj., a to z toho důvodu, že všechny tyto jevy znamenají výjimku nebo překážku pro syntaktickou analýzu (podrobněji viz 2.2.1 – „M. Meteer et al.: Dysfluency Annotation“).

Syntaktická anotace. Syntaktická anotace korpusu Switchboard byla provedena automaticky, podle stejných pravidel, jaká se uplatňují při anotaci psaných textů v PennTreebanku (viz Bies et al., 1995). Vstupem pro syntaktický parsing byla transkribovaná data s označenými dysfluencemi. Výstupem je složkový strom, tzv. predikátovo-argumentová struktura, která je rozšířením původní „pouze“ frázové struktury (rozlišující NP — jmenná fráze, VP — slovesná fráze apod.). Základní principy anotačního schématu jsou:

- označení frázových složek standardními frázovými atributy (viz tab. 1.1).
- přidání značky identifikující syntakticko-sémantickou roli složek ve struktuře, jako je subjekt, časové určení apod. Od původního záměru rozlišit těmito značkami argumenty slovesa od adjunktů bylo pro nedostatek nosných rozlišujících kritérií upuštěno. Značky (viz tab. 1.2) se připojují pomocí znaku „-“ k značkám frázovým.
- do struktury se přidávají „nulové elementy“, které jsou koindexovány s lexikálně obsazenými složkami, jež nulové elementy zastupují. Tyto elementy se do struktury doplňují při pasivu, v otázkách, při infinitivních konstrukcích. Pomocí nulových elementů jsou zachyceny případy diskontinuitních složek, tj. případy, kdy dvě slova, která k sobě patří, nestojí ve větě vedle sebe, a to jak případy pravidelné — gramatické: v otázkách, v pasivních konstrukcích aj., tak případy nepravidelné (tzv. pseudo-attach).
- speciálním způsobem jsou vyřešeny elipsy v případech syntaktického paralelismu, tj. v případech, kdy v jedné větě je obsažena plná, neelidovaná klauze a zároveň elidovaná paralelní klauze (např.: *Mary likes Bach and Susan Beethoven*.). Argumenty v elidované klauzi jsou pomocí číselných indexů namapovány na odpovídající argumenty v plné klauzi.
- dysfluencím oanotovaným v předchozí anotaci je přiřazována speciální značka `EDITED`.

Tabulka 1.1. Switchboard: Phrasal labels

S	simple declarative clause	INTJ	interjection	PRC	reduced relative clause
SBAR	clause introduce by a subordinating conjunction	LST	list marker	UPC	unlike coordinated phrase
SBARQ	direct question introduced by a wh-word or wh-phrase	NAC	not a constituent	VP	verb phrase
SINV	inverted declarative sentence	NP	noun phrase	WHADJP	wh-adjective phrase
SQ	inverted yes/no question or main clause of a wh-question	NX	used within certain complex noun phrase	WHADVP	wh-adverb phrase
ADJP	adjective phrase	PP	prepositional phrase	WHNP	wh-noun phrase
ADVP	adverb phrase	PRN	parenthetical	WHPP	wh-prepositional phrase
CONJP	conjunction phrase	PRT	particle	X	unknown, uncertain, or unbracketable
FRAG	fragment	QP	quantifier phrase	EDITED	dysfluency

Tabulka 1.2. Switchboard: Function tags

-AVD	adverbial	-TPC	topicalized	-PRP	purpose of reason
-NOM	nominal	-VOC	vocative	-TMP	temporal
-DTV	dative	-BNF	benefactive	-CLR	closely related
-LGS	logical subject	-DIR	direction	-CFL	cleft
-PRD	predicate	-EXT	extent	-HLN	headline
-PUT	locative complement of *put*	-LOC	locative	-TTL	title
-SBJ	surface subject	-MNR	manner		

Tabulka 1.3. Switchboard: Other labels

T; (NP *); 0; *U*; *?*; *NOT*	null elements
EXP; *ICH*; *PPA*; *RNR*	pseudo-attach

Obrázek 1.1. Příklad syntaktické anotace v korpusu Switchboard

```
( (CODE SpeakerA4 .))
( (S (INTJ Well)
     ,
     (EDITED (RM [)
             (NP-SBJ I)
             ,
             (IP +))
     (NP-SBJ I)
     (RS ])
     (VP think
         (SBAR 0
               (S (NP-SBJ it)
               (VP 's (NP-PRD a
                           (ADJP pretty good)
                           idea)))))
     .
     E_S))
( (S (NP-SBJ I)
     (VP think
         (SBAR 0
               (S (NP-SBJ-1 they)
                  (VP should
                      (VP either
                          (VP do
                             (NP that))
                          ,
                          (EDITED (RM [)
                                  or
                                  ,
                                  (IP +))
                          or
                          (RS ])
                          (VP afford
                             (NP some time)
                             (PP-DTV to
                                     (NP (NP the military)
                                         ,
                                         (EDITED (RM [)
                                                 or
                                                 ,
                                                 (IP +))
                                         or
                                         (RS ])
                                         (S-NOM (NP-SBJ *-1)
                                                (VP helping
                                                    (NP elderly people)))))))))))
     .
     E_S))
```

1.1.1.2. The Spoken Dutch Corpus

Korpus mluvené holandštiny (CGN, Corpus Gesproken Nederlands; http://lands.let.kun.nl/cgn/ehome.htm) byl vybudován v letech 1998 až 2004 na univerzitách v Belgii (University of Leuven a University of Ghent) a v Nizozemí (University of Utrecht a University of Nijmegen). Korpus představuje rozsáhlou databázi současné standardní holandštiny, kterou mluví dospělí lidé v Nizozemí a ve Flandrech. Obsahuje přibližně 800 hodin řeči, což je téměř 9 miliónů slov. Data pocházejí z různého prostředí: telefonní dialogy, spontánní konverzace, politické diskuze,

debaty, živé komentáře, hlasité čtení aj. Dvě třetiny materiálu byly nasbírány v Nizozemí, jedna třetina v holandsky mluvící části Belgie — ve Flandrech.

Celý korpus obsahuje anotaci slovního druhu (part of speech) a morfologických kategorií (tagování), celý korpus je též lematizován a jsou identifikovány víceslovné jednotky. Pro přibližně 250 000 slov je k dispozici prozodická anotace. Podstatná část korpusu — 1 milión slov — je anotována i syntakticky.

Syntaktická anotace. Syntaktická anotace v Korpusu mluvené holandštiny je anotace povrchové podoby věty. Obsahuje dva typy informace:

- informaci o slovním druhu slovní jednotky (tzv. kategoriální atributy),
- informaci o funkci jednotky v syntaktické struktuře (tzv. závislostní atributy). Syntaktická struktura je reprezentována jako struktura složková.

Formálně je syntaktická struktura v Korpusu mluvené holandštiny označkovaný orientovaný acyklický graf (ne strom): množina uzlů a hran. Uzly nesou kategoriální atributy a hranám náleží atributy závislostní.

V grafu se rozlišují tzv. atomické a složené struktury. Atomické struktury jsou jednoduché uzly — listy grafu, kterým je přiřazena jedna z 50 hodnot nesoucí informaci o slovním druhu (POS-značky; viz Hoekstra et al., 2001); jde o redukované morfologické značky. Základem složené struktury je uzel-rodič nesoucí frázovou značku (viz tab. 1.4). Uzel-potomek má přiřazenou buď opět frázovou značku, nebo pokud jde o list stromu, nese informaci o slovním druhu (POS-značku).

Závislostními atributy (viz tab. 1.5), které náležejí hranám, jsou rozlišeny „heads“, „complements“ a „modifiers“ (tj. přibližně řídící členy, aktanty a volná doplnění).

Pro jevy typické v mluvené řeči byly zavedeny ve všech sadách značek značky speciální.

Pro syntaktickou anotaci byl vytvořen manuál (viz Hoekstra et al., 2003; dostupný pouze v holandštině).

Příklad syntaktické anotace v Korpusu mluvené holandštiny viz obr. 1.2.

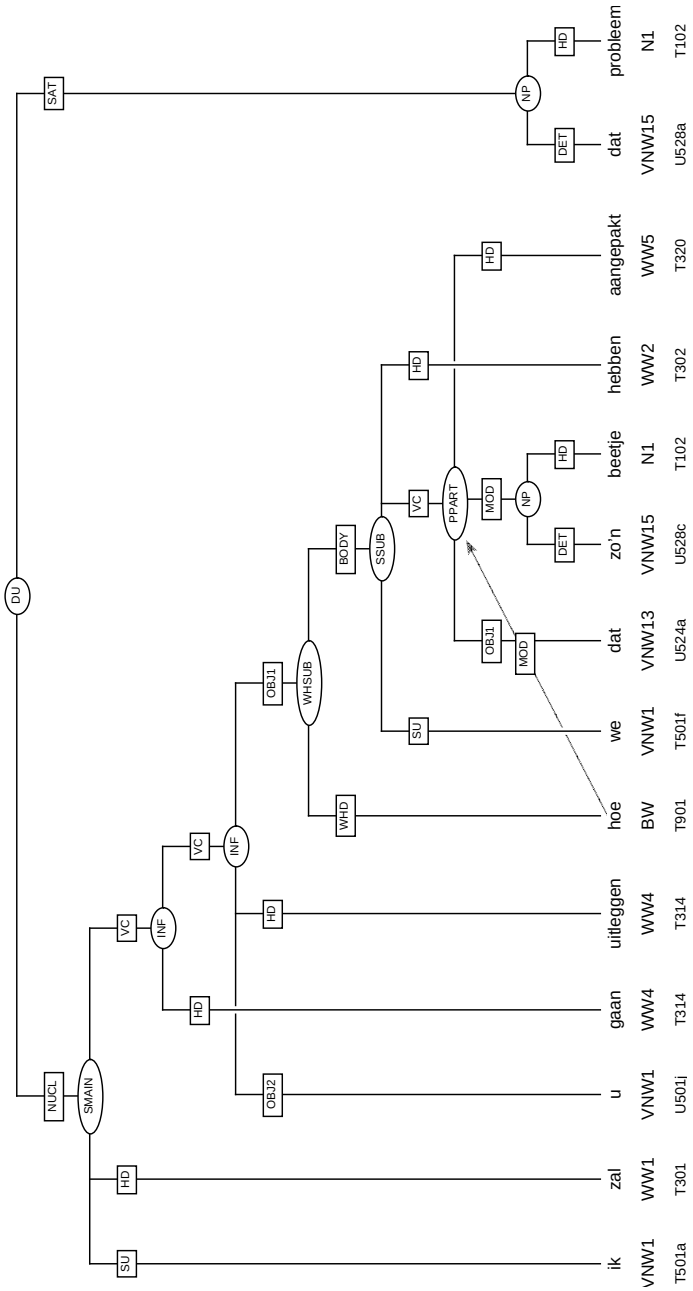
Tabulka 1.4. CGN: Category labels

SMAIN	main clause	OTI	long infinitive group headed by om (_ for to)	WHREL	headless relative
SSUB	subordinate clause	AHI	long infinitive group headed by aan het (a Dutch progressive form)	WHQ	WH-question
SV1	any sentence with a sentence-initial inflected verb	ADVP	adverbial phrases	WHSUB	embedded WH-question
INF	short infinitive group	DETP	determiner group	CONJ	conjunction
PPART	past/passive participle group	AP	adjectival group	DU	discourse-unit
PPRES	present participle group	PP	prepositional group	LIST	asyndetic conjunction
CP	clause headed by any kind of complementizer	NP	nominal group	COMPP	various comparative constructions
MWU	merged-word-unit (used for complex numbers and names)	SVAN	subordinate clause headed by van		
TI	long infinitive group	REL	relative clause		

Tabulka 1.5. CGN: Edge labels

HD	head	PC	prepositional complement	NUCL	nuclear clause
HDF	second part of a circumposition	VC	verbal complement	SAT	satelite
DET	determiner	LD	locational or directional complement	TAG	tag
PART	partitive	ME	measure complement	DP	any part of a DU
SU	subject	CMP	grammatical complementizer	PRT	any part of a particle group
SUP	provisional subject	RHD	complementizer heading (headless) relative	OBCOMP	comparative complement
OBJ1	complement	WHD	complementizer heading WH question	APPOS	apposition
POBJ1	provisional direct or first object	BODY	body of subordinate clause	LP	any part of a LIST
OBJ2	secondary object	PREDM	secondary predicate	DLINK	discourse particles joining discourse fragments
SE	obligatory reflexive object	MOD	modifier	MWP	any part of a MWU
SVP	verbal particle	CRD	coordinator		
PREDC	predicative complement	CNJ	member of conjunction		

Obrázek 1.2. Příklad syntaktické anotace v Korpusu mluvené holandštiny



1.1.1.3. The Verbmobil treebanks

Tři syntakticky anotované korpusy mluvené řeči vznikly v rámci projektu Verbmobil (http://verbmobil.dfki.de/), který probíhal ve dvou fázích v letech 1993 — 2000. Prvořadým cílem projektu byl vývoj nástroje pro automatický překlad spontánní mluvené řeči v němčině, angličtině a japonštině. V současné fázi je vyvinutý systém s to zvládnout takové situace jako telefonní úmluva pracovní schůzky, zjištění časového harmonogramu cesty a rezervování hotelu.

Projekt probíhal na 16 výzkumných pracovištích, univerzitách i u průmyslových partnerů po celém Německu. Korpusy zahrnutých jazyků (mluvené angličtiny, němčiny a japonštiny) byly vybudovány na univerzitě v Tübingen. Jsou to:

The Tübingen Treebank of Spoken German (TüBa-D/S; viz 1.1.1.3.1 – „The Tübingen Treebank of Spoken German“),

The Tübingen Treebank of Spoken English (TüBa-E/S; viz 1.1.1.3.2 – „The Tübingen Treebank of Spoken English“),
The Tübingen Treebank of Spoken Japanese (TüBa-J/S; viz 1.1.1.3.3 – „The Tübingen Treebank of Spoken Japanese“).

Všechny korpusy se blíží počtu 30 000 vět. Všechny tři korpusy byly manuálně transkribovány a jsou kompletně syntakticky anotovány. Syntaktická anotace byla též provedena manuálně (angličtina a němčina za pomoci anotačního nástroje Annotate).

Anotační schéma všech tří korpusů je obdobné. Při navrhování anotačních schémat byla vždy speciálně zvažována otázka větné segmentace a při stanovování sady gramatických a syntaktických značek bylo přihlédnuto k tomu, aby:

- inventář gramatických a syntaktických kategorií nebyl aplikovatelný pouze na data v korpusech projektu Verbmobil, ale na mluvená data obecně.
- inventář gramatických a syntaktických značek sloužil primárně svému budoucímu účelu: natrénování nástrojů pro strojový překlad, čímž je zde myšleno, že je třeba se zaměřit na vlastnosti povrchové struktury věty, zejména na pravidla povrchového slovosledu a na distribuční vlastnosti slov a frází, a naopak užívání prázdných kategorií a zachycování syntaktických struktur neprojektivními závislostmi je třeba se vyhnout.
- inventář gramatických a syntaktických značek byl teoreticky nezávislý a minimální.

Výsledkem anotace je složkový strom, který je formálně definován jako orientovaný acyklický graf, kde:

- jeden uzel je vyznačen jako kořen stromu. Z kořene vede vždy cesta do každého uzlu stromu.
- každý uzel kromě kořene je potomkem pouze jednoho uzlu.
- potomci každého uzlu jsou lineárně uspořádány odleva doprava.

Všechny tři korpusy jsou založeny na následujících obecných anotačních principech:

- **Flat Clustering Principle**. Princip „flat clustering“ říká, že počet úrovní v syntaktické struktuře má být co nejmenší, to znamená, že co nejvíce možných složek je seskupeno pod jednu složku vyšší; jinými slovy je žádoucí co největší stupeň větvení.
- **Longest Match Principle**. Princip „longest match“ řeší otázku hranice věty v mluvené řeči. Vyžaduje, aby stromová struktura (věta) zahrnovala co nejvíce možných složek za předpokladu, že výsledný strom (věta) bude utvořen jak syntakticky, tak sémanticky správně.
- **High Attachment Principle**. Princip „high attachment“ předepisuje, že v případech syntaktické nebo sémantické nejednoznačnosti v připojení nějaké složky jsou tyto složky zachyceny ve stromové struktuře v co nejvyšší možné úrovni.

Specifické jevy mluvené řeči: přeřeknutí, opakování slov, výplňková slova, hezitační zvuky jsou v co největší možné míře strukturovány (většinou až na úroveň frázových kategorií), ale nejsou zpravidla připojovány k okolním složkám, tj. nejsou zanořeny do stromové struktury věty, ve které se vyskytly.

1.1.1.3.1. The Tübingen Treebank of Spoken German

Tübingenský korpus mluvené němčiny (TüBa-D/S; http://www.sfs.uni-tuebingen.de/en_tuebads.shtml) je manuálně syntakticky anotovaný korpus spontánních dialogů, který čítá přibližně 38 000 vět (360 000 slov).

Syntaktická anotace. Anotační schéma pro Tübingenský korpus mluvené němčiny bylo vyvinuto se speciálním zřetelem k částečně volnému slovosledu v německých větách, ve kterých je postavení slovesných doplnění velmi volné, na druhou stranu však lze v němčině vyčlenit tři typy klauzí na základě fixované pozice určitého slovesa: věty se slovesem na první pozici, věty se slovesem na druhé pozici a věty se slovesem na poslední pozici ve větě. Toto nepravidelné slovosledné uspořádání je dobře popsáno v německých syntaktických teoriích — v konceptu

tzv. topologických polí („topological field“), ve kterém jsou rozlišeny například tři následující struktury klauzí (podle T. N. Höhla):

- (KOORD) — (C) — X — VK — Y
- (KOORD) — (KL) — FINIT — X — VK — Y
- (KOORD or PARORD) — (KL) — K — FINIT — X — VK - Y

Vysvětlení zkratek: VK: verb complex; FINIT: element denoting categories of finitness; KOORD: coordinating particles; PARORD: non-coordinating particles; X, Y: sequence of any number of constituents; C: complementizer; K: one constituent; KL: nominativus pendens, resumptive construction.

Koncept topologických polí byl zahrnut do anotačního schématu jako primární princip pro rozdělení věty do složek: na první úrovni pod uzlem pro kořen věty se věta nejprve rozdělí do složek, které odpovídají jednotlivým topologickým polím. Topologická pole charakterizují pořadí slov uvnitř různých typů německých klauzí. Ve výsledném stromě tak sice nevzniknou překřížené hrany, nicméně negativním důsledkem tohoto kroku je, že koncept topologických polí neumožňuje popsat hierarchickou strukturu složek na nejvyšší úrovni. Hierarchická (syntaktická) struktura složek je popisována jen v rámci jednotlivých polí až na nižších úrovních stromu.

Syntaktická struktura v Tübingenském korpusu mluvené němčiny se skládá z následujících typů anotace:

- anotace slovního druhu (part-of-speech tags). Pro značkování byl použit Stuttgart-Tübingen Tagset (STTS; viz Stegmann et al., 2000).
- anotace frázové struktury. Frázové značky (viz tab. 1.6) jsou přiřazovány uzlům.
- anotace gramatických funkcí. Značky gramatických funkcí (viz tab. 1.7) jsou přiřazovány hranám.
- anotace topologických polí. Značky topologických polí viz tab. 1.8 jsou přiřazovány uzlům.
- přiřazení značky kořeni stromu určující typ věty (viz tab. 1.9).

Příklady syntakticky anotovaných vět (obsahujících specifické řečové jevy) v Tübingenském korpusu mluvené němčiny viz obr. 1.3.

Tabulka 1.6. TüBa-D/S: Phrase node labels

NX	noun phrase	VXFIN	finite verb phrase
PX	prepositional phrase	VXINF	infinite verb phrase
AVX	adverbial phrase	DP	determiner phrase
ADJX	adjectival phrase		

Tabulka 1.7. TüBa-D/S: Edge labels

HD	head	OPP	prepositional object	VPT	separable verb prefix
-	non-head	OADVP	adverbial object	APP	apposition
ON	nominative object	OADJP	adjectival object	MOD	ambiguous modifier
OD	dative object	PRED	predicate	ON-MOD; OA-MOD; OD-MOD; MOD-MOD; V-MOD; OPP-MOD; PRED-MOD; FOPP-MOD	modifiers modifying complemnets or modifiers; e.g. V-MOD = modifier of verb
OA	accusative object	OV	verbal object	ONK; ODK; OAK; OPPK; FOPPK; OADJPK; PREDK; MODK; OPP-MODK; PREDMODK; MOD-MODK	second conjunct in split-up coordinations; e.g. ONK = second conjunct of a nominative object
OS	sentential object	FOPP	optional prepositional object	ref1	first verbal object in VC selected by a verbal object

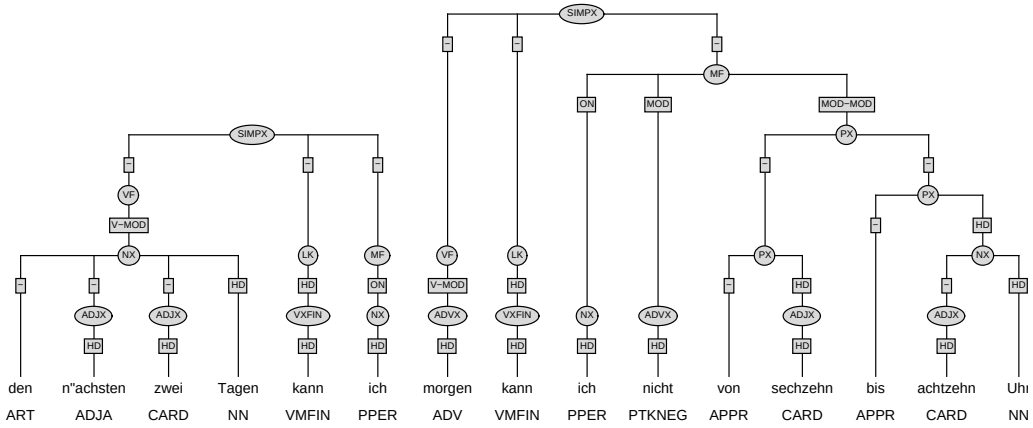
Tabulka 1.8. TüBa-D/S: Topological fields node labels

VF	initial field (Vorfeld)	NF	final field (Nachfeld)	PARORD	field for coordinating particles
LK	left sentence bracket (Linke (Satz-)Klammer)	LV	resumptive constructions (Linksversetzungsfeld)	FKOORD	coordination consisting of conjuncts of fields
MF	middle field (Mittelfield)	C	complementizers field (C-Feld)	LKM; LKMNC; LKMVCN; LKMN; LKVCN; LKN; MVC; MVCN; MN; VCN; CM; CMVC	combinations of fields (V = VF, M = MF, N = NF; e. g. LKM = LK + MF)
VC	verb complex (Verbkomplex)	KOORD	field for coordinating particles		

Tabulka 1.9. TüBa-D/S: Root node labels

SIMPX	simplex clause	P-SIMPX	paratactic construction of simplex clauses
R-SIMPX	relative clause	DM	discourse marker

Obrázek 1.3. Příklad syntaktické anotace v Tübingenském korpusu mluvené němčiny



1.1.1.3.2. The Tübingen Treebank of Spoken English

Tübingenský korpus mluvené angličtiny (TüBa-E/S; http://www.sfs.uni-tuebingen.de/en_tuebaes.shtml) je manuálně syntakticky anotovaný korpus spontánních dialogů, který obsahuje přibližně 30 000 vět (310 000 slov).

Syntaktická anotace. Anotační schéma Tübingenského korpusu mluvené angličtiny je založeno na gramatice Head-Driven Phrase Structure Grammar (http://hpsg.stanford.edu/). Syntaktická struktura se skládá z následujících typů anotace:

- anotace slovního druhu (part-of-speech tags). Sada značek použitá v Tübingenském korpusu mluvené angličtiny je stejná jako sada značek používaná v PennTreebanku (viz Santorini, B., 1990).
- anotace frázové struktury. Frázové značky (viz tab. 1.10) jsou přiřazovány uzlům.
- anotace gramatických funkcí. Značky gramatických funkcí (viz tab. 1.11) jsou přiřazovány hranám.
- přiřazení značky kořeni stromu určující typ věty (viz tab. 1.12).

Příklady syntakticky anotovaných vět (obsahujících specifické řečové jevy) v Tübingenském korpusu mluvené angličtiny viz obr. 1.4.

Tabulka 1.10. TüBa-E/S: Phrasal node labels

AP	ajectival phrase	NP	noun phrase
APS	ajectival phrase heading a small clause	NPS	noun phrase heading a small clause
ADVP	adverbial phrase	PP	prepositional phrase
DGP	degree phrase	PPS	prepositional phrase heading a small clause
DTP	determiner phrase	VP	verb phrase

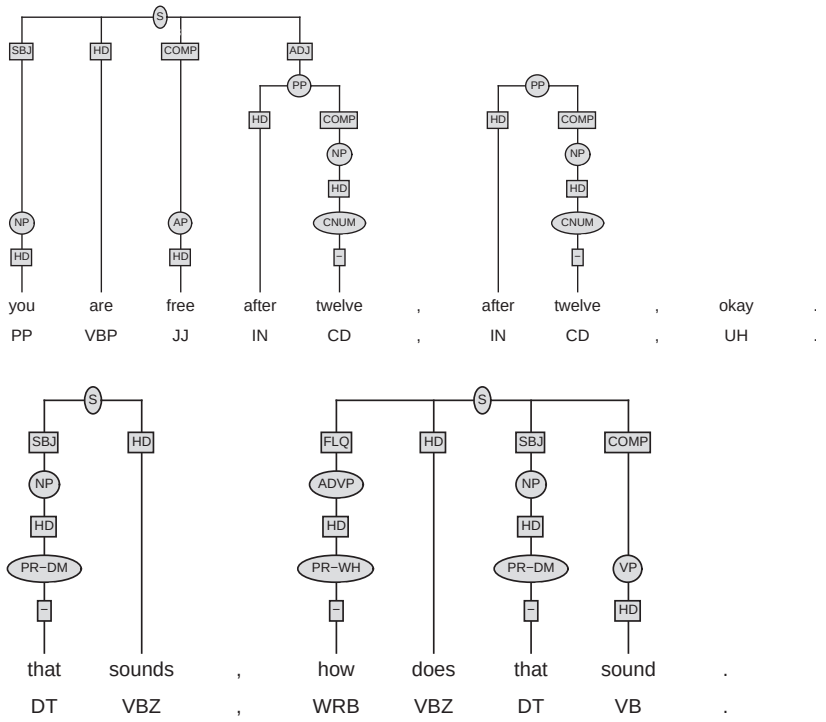
Tabulka 1.11. TüBa-E/S: Edge labels

HD	head	SBR	subject, REL	FLR	filler, REL
COMP	complement	ADJ	adjunct	MRK	marker
SPR	specifier	ADJ?	adjunct? (ambiguous attachment)	-	intentionally empty edge label
SBJ	subject	FLL	filler		
SBQ	subject, WH	FLQ	filler, WH		

Tabulka 1.12. TüBa-E/S: Root node labels

S	sentence
SUGG	suggestion; for utterance started *how about*, *what about*, *what if*
phrasal labels: NP, VP, PP, ADVP	sentence external, isolated phrase occuring without verb

Obrázek 1.4. Příklad syntaktické anotace v Tübingenském korpusu mluvené angličtiny



1.1.1.3.3. The Tübingen Treebank of Spoken Japanese

Tübingenský korpus mluvené japonštiny (TüBa-J/S; http://www.sfs.uni-tuebingen.de/en_tuebajs.shtml) je manuálně syntakticky anotovaný korpus spontánních dialogů, který obsahuje přibližně 18 000 vět (160 000 slov).

Syntaktická anotace. Anotační schéma Tübingenského korpusu mluvené japonštiny je založeno na gramatice Head-Driven Phrase Structure Grammar (http://hpsg.stanford.edu/). Skládá se z následujících typů anotace:

- anotace slovního druhu (part-of-speech tags; viz Kawata — Barteles, 2000).
- anotace frázové struktury. Frázové značky (viz tab. 1.13) jsou přiřazovány uzlům.
- anotace gramatických funkcí. Značky gramatických funkcí (viz tab. 1.14) jsou přiřazovány hranám.
- přiřazení značky kořeni stromu určující typ věty. Značky přiřazované kořeni stromu nejsou speciálně definovány, neboť segmentem mluvené řeči může být fragment. Kořen stromu proto vždy nese některou z frázových značek.

Specifické jevy mluvené řeči jsou označeny speciální značkou `err` pro uzel a značkou `-` náležející hraně.

Tabulka 1.13. TüBa-J/S: Node labels

NP	noun phrase	VP.foc	verb phrase (focus)	ADVP	adverb phrase
NPloc	noun phrase (location)	PPacc	postpositional phrase (accusative)	ADVP.foc	adverb phrase (focus)
NPper	noun phrase (person)	PPgen	postpositional phrase (genitive)	S	sentence
NPtmp	noun phrase (temporal)	PPnom	postpositional phrase (nominative)	SS	subordinated sentence
NP.foc	noun phrase (focus)	PP	postpositional phrase	GR	greeting
VPN	verbal noun phrase	PP.foc	postpositional phrase (focus)	ITJ	interjective expression
VP	verb phrase	AP	adjective phrase	err	false start, speech error
VPcnd	verb phrase (conditional)	APcnd	adjective phrase (conditional)		
VPfin	verb phrase (finite)	AP.foc	adjective phrase (focus)		

Tabulka 1.14. TüBa-J/S: Edge labels

HD	head	ADJ	adjunct	MRK	marker
COMP	complement	SBJ	subject	-	unspecified

1.1.1.4. The Childes database

Cílem projektu Childes (Child Language Data Exchange System; http://childes.psy.cmu.edu/), který se realizuje na Carnegie Mellon University v Pittsburghu, je zachytit jazyk těch, kdo se jej teprve učí, a poskytnout bohatý materiál těm, kdo jazyk dětí studují (tj. lingvistům, psychologům, pedagogům aj.).

Korpus projektu Childes obsahuje transkribované spontánní dialogy dětí s rodiči nebo se sourozenci. Část korpusu představují přepisy dialogů bilingválních dětí, dětí s různými jazykovými dysfunkcemi a afáziemi, jsou zde i nahrávky starších dětí a dospělých, kteří se učí druhému jazyku. Korpus tvoří přepisy nahrávek ve 26 jazycích, přičemž podstatná část materiálu je v angličtině (přes 100 megabytů). Celková velikost korpusu je v současné době 300 miliónu znaků (300 megabytů). Všechny transkribované dialogy jsou opatřeny anotací slovního druhu (part-of-speech).

Syntaktická anotace. Syntaktická anotace je v korpusu Childes realizována pouze v jeho anglické části. 10 000 slov bylo oanotováno manuálně. Manuální syntaktická anotace byla východiskem pro vytvoření nástroje pro automatickou syntaktickou anotaci celého korpusu (realizovanou v roce 2005).

Syntaktická anotace v korpusu Childes představuje závislostní (nikoli složkovou) analýzu povrchové struktury věty. Vztah mezi řídící a závislou jednotkou je vyjádřen označením závislostních vztahů, které korespondují s gramatickými vztahy, jakými je například subjekt, objekt nebo adjunkt. Platí, že každé slovo ve větě musí být závislé na jednom jiném řídícím slově (ale řídící slovo může mít více slov závislých). Jedinou výjimkou z tohoto pravidla je, že každá věta má jeden kořen, který není závislý na žádném slově ve větě. Kvůli zachování konzistence je tento kořen zachycen jako závislý na speciálním prázdném slově připojeném na začátek každé věty (toto prázdné slovo je označováno jako LeftWall).

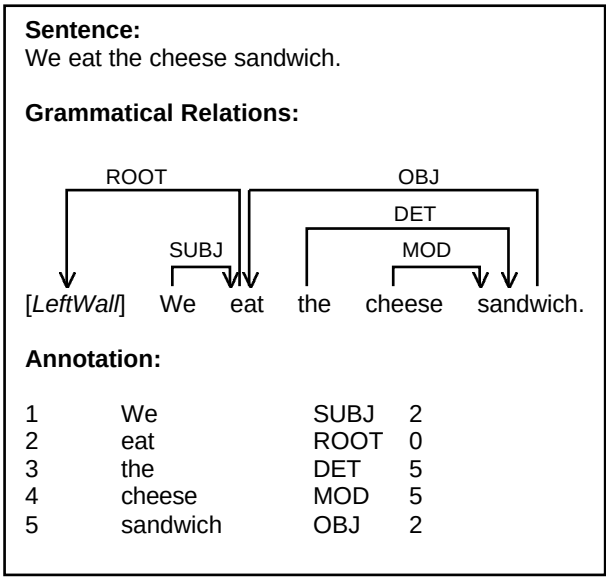
Značka gramatického vztahu (vyjadřující vztah slova k slovu nadřízenému) je přiřazena každému slovu ve větě. Rozlišuje se třicet různých značek pro gramatické vztahy. Sada značek (viz tab. 1.15) byla stanovena se speciálním zřetelem k následnému výzkumu dětské řeči.

Příklad syntaktické anotace v korpusu Childes viz obr. 1.5.

Tabulka 1.15. Childes: Labels of grammatical relations

SUBJ	subject	CPRED	clausal predicative finite	DET	determiner
ESUBJ	expletive subject	XPRED	clausal predicative non-finite	QUANT	quantifier
CSUBJ	clausal subject finite	JCT	adjunct	POBJ	prepositional object
XSUBJ	clausal subject non-finite	CJCT	clausal adjunct finite	PTL	verbal particle (of phrasal verb)
OBJ1	object	XJCT	clausal adjunct non-finite	CPZR	complementizer
OBJ2	second object	MOD	nominal modifier	COM	communicator
IOBJ	indirect object	CMOD	clausal nominal modifier finite	INF	infinitival particle
COMP	clausal complement finite	XMOD	clausal nominal modifier non-finite	VOC	vocative
XCOMP	clausal complement non-finite	AUX	auxiliary	COORD	coordination
PRED	predicative	NEG	negation	ROOT	special relation for the top node

Obrázek 1.5. Příklad syntaktické anotace v korpusu Childes



1.2. Výzkum mluvené řeči u nás

1.2.1. Korpusy mluvené češtiny

V českém prostředí existuje hned několik korpusů mluvené češtiny:

Pražský mluvený korpus,
Brněnský mluvený korpus,
Český mluvený korpus ORAL2006,
korpus Dialog – mluvená čeština v televizních debatách,
Korpus věcného stylu
a korpusy používané pro automatické rozpoznávání mluvené řeči.

1.2.1.1. Pražský mluvený korpus

Prvním uceleným korpusem mluvené češtiny je **Pražský mluvený korpus** (http://ucnk.ff.cuni.cz/pmk_bonito.html). Zachycuje autentickou mluvenou češtinu, hlavně obecnou a tématicky nespecializovanou, z oblasti Prahy a jejího okolí. Korpus tvoří magnetofonové nahrávky (v počtu 304), které pocházejí z let 1988 - 1996. Obsahuje přibližně 675 000 slovních jednotek (bez interpunkce).

Nahrávky Pražského mluveného korpusu byly pořizovány tak, aby ve vyvážených proporcích obsáhly čtyři sociolingvistické proměnné, všechny pro jednoduchost dělené pouze na dvě hodnoty: pohlaví mluvčího (muž — žena), věk mluvčího (mladší, tj. 20 — 35 let, - starší), vzdělání (nižší, zahrnující jak základní školu, tak vzdělání maturitní — vyšší, vztahující se ke vzdělání vysokoškolskému) a typ promluvy (formální promluva, tj. monolog vytvářený sledem odpovědí na otázky kladené nahrávajícím - neformální promluva, tj. dialogický soubor promluv dvou mluvčích, kteří se znají, téma jejich rozhovoru nebylo nijak předem určováno). Každá nahrávka je dále doplněna zpřesňujícími informacemi o mluvčích, o roku svého vzniku, případně relevantními údaji o situaci promluvy.

Pro manuální přepis nahrávek nebyl použit přepis dialektologický (fonetický), nýbrž účelová kombinace fonetického zápisu a standardních pravopisných norem, a to tak, aby přepisy zachycovaly mluvenou řeč co nejvěrněji a nejsrozumitelněji. V jednotlivých přepisech se proto pochopitelně objevuje kolísání, odrážející mj. individuální přístupy přepisovatelů (zčásti studentů). Korpus obsahuje kromě přepisu i velmi detailní anotaci na rovině morfologické a lexikální, nikoli však syntaktické.

V červenci 2001 byl Pražský mluvený korpus zpřístupněn běžným uživatelům. V současné době se dokončuje rozsáhlé manuální kódování a tagování korpusu, a lze proto zatím pracovat jen s jeho texty čistými.

Pražský mluvený korpus je součástí Českého národního korpusu, spravovaného na Filozofické fakultě Univerzity Karlovy. Autory Pražského mluveného korpusu jsou v různých proporcích především: Anna Adamovičová, František Čermák, Jiří Pešička, Josef Šimandl, Jitka Šonková, Petr Savický a Zdena Smetanová; s nahrávkami pomáhala řada studentů.

1.2.1.2. Brněnský mluvený korpus

Brněnský mluvený korpus (http://ucnk.ff.cuni.cz/bmk.html) je v rámci Českého národního korpusu prvním korpusem mluvené češtiny z oblasti Moravy. Zaznamenává autentickou tematicky nespecializovanou mluvenou řeč města Brna. Brněnský mluvený korpus je přepisem 250 magnetofonových nahrávek z let 1994-1999 zachycujících 294 mluvčích. Obsahuje přibližně 490 000 slovních jednotek (bez interpunkce).

Brněnský mluvený korpus byl pořizován v souladu se zásadami Pražského mluveného korpusu (viz 1.2.1.1 – „Pražský mluvený korpus“). Pravidla přepisu v základních rysech odpovídají pravidlům užívaným v Pražském mluveném korpusu, diference spočívají především v pokusu o nahrazení tradiční interpunkce interpunkcí „pauzovou“ (pauza je podle délky trvání označena pomocí jedné, dvou nebo tří pomlček) a v zachycení simultánnosti dialogických promluv. Pravidla přepisu jsou k dispozici na http://ucnk.ff.cuni.cz/popis_bmk.html.

Gramatické značkování Brněnského mluveného korpusu probíhá na Fakultě informatiky a na Filozofické fakultě Masarykovy Univerzity v Brně. Využívána je metoda poloautomatické analýzy, práci však komplikuje velká hlásková, tvarová i lexikální variabilita brněnské mluvy (prolínání dialektických, interdialektických, obecně českých i spisovných podob). V Českém národním korpusu lze proto s Brněnským mluveným korpusem pracovat zatím pouze jako s čistými texty.

Autory Brněnského mluveného korpusu jsou v různých proporcích především: Zdeňka Hladká, Dana Hlaváčková, Daniel Jedlička a Táňa Vykypělová z Filozofické fakulty Masarykovy univerzity v Brně; s pořizováním nahrávek pomáhala řada studentů.

1.2.1.3. Český mluvený korpus ORAL2006

Mluvený korpus ORAL2006 (http://ucnk.ff.cuni.cz/ORAL2006.html) je v pořadí třetím mluveným korpusem, který je dostupný v rámci projektu Český národní korpus. Zachycuje mluvenou češtinu z celé oblasti českých nářečí v užším slova smyslu. Jedná se o přepis 221 nahrávek z let 2002-2006. Všechny nahrávky vznikaly

v neformálních situacích, to znamená, že mluvčí se vzájemně znali a měli k sobě přátelský vztah. Celkem bylo nahráno 6 693 minut, tj. asi 111 a půl hodiny, a v jejich rámci byl zaznamenán přibližně 1 milión slov (bez interpunkce) od 754 mluvčích.

Způsob pořizování nahrávek, jejich přepis a označování probíhal v souladu se zásadami předchozích mluvených korpusů (Pražského a Brněnského mluveného korpusu). Z tohoto důvodu bylo zachováno i označování sociolingvistických kategorií mluvčích: pohlaví: muž (M) a žena (Z), věk do 35 let (I) a věk 35 let a více (V), vzdělání základní a středoškolské (B) a vysokoškolské, a to i započaté (A). Navíc lze o mluvčích v tomto korpusu zjistit jejich přesný věk, vzdělání (ZŠ, SŠ, VŠ) a nářeční oblast jejich pobytu v dětství — tedy v době, kdy se formoval základ jejich individuálního jazykového úzu.

Způsob přepisu vychází z pravidel přepisu pro Pražský mluvený korpus. Pravidla přepisu jsou k dispozici na http://ucnk.ff.cuni.cz/ORAL2006pravidla.html.

Hlavními koordinátory korpusu ORAL2006 jsou Marie Kopřivová a Martina Waclawičová. Na pořizování a přepisu nahrávek se podíleli především studenti pražských vysokých škol a další spolupracovníci Ústavu Českého národního korpusu.

1.2.1.4. Korpus Dialog – mluvená čeština v televizních debatách

Korpus Dialog – mluvená čeština v televizních debatách (viz Čmejrková — Jílková — Kaderka, 2004) je specializovaný korpus mluvené češtiny Ústavu pro jazyk český Akademie věd České republiky. Tento korpus soustřeďuje nahrávky řeči mediální, tj. především nahrávky dialogických mluvených projevů (interview, diskuse, debata, polemika, talk-show aj.) z televizních pořadů Sedmička, Nedělní partie, Na plovárně s..., Krásný ztráty a další.

Archiv korpusu Dialog tvoří asi 360 videokazet většinou o délce 240 minut; což představuje přibližně 80 000 minut dialogické řeči. Část nahrávek je přepsána: existuje asi 350 přepisů představujících textový korpus čítající podle odhadu asi jeden a půl až dva milióny slov.

Nahrávky byly ručně přepisovány podle jednotných transkripčních zásad, jejichž specifickým rysem je soubor značek pro zachycení prozodických rysů mluveného projevu. Transkripční systém umožňuje zachytit jednak sekvenční průběh rozhovoru (tj. střídání účastníků rozhovoru v řeči včetně simultánního mluvení více mluvčích, okamžitého, bezpauzového navázání na předcházející repliku apod.), jednak segmentální složku řečového projevu (včetně odchylek od noremního vyjadřování) a vybrané jevy suprasegmentální složky (především kadence, pauzy a zdůraznění slova či slabiky). Své místo má v transkripčním systému i záznam parajazykových jevů (hezitačních a responzních zvuků, smíchu, hlasitých nádechů/výdechů apod.), zaznamenat lze i mimojazykové jevy (např. střih). Transkripční zásady užité v korpusu Dialog se tak odlišují od zásad užitých ve výše popsaných mluvených korpusech Českého národního korpusu.

Za účelem dalšího počítačového zpracování byly původní přepisy nahrávek ve formátu DOC převedeny do standardizovaného formátu XML. Konvertovaná data byla posléze automaticky lematizována (každému slovnímu tvaru bylo přiřazeno lema, tj. základní tvar slova) a obohacena o morfologické značky (podle systému J. Hajiče, viz Hajič, 2004).

Pro odbornou veřejnost je na internetu zpřístupněna část korpusu pod názvem DIALOG 0.1 (byl vytvořen ve spolupráci s Ústavem formální a aplikované lingvistiky MFF UK Praha). Obsahuje revidované přepisy politické talk-show Sedmička (starší název pořadu je 7 čili Sedm dní) soukromé televize Nova z let 1999–2005. Korpus existuje ve dvou verzích: (a) morfologicky neanotovaná verze obsahuje 10 přepisů o celkové velikosti 92 000 slov, (b) morfologicky anotovaná verze obsahuje 5 přepisů o velikosti 45 000 slov. Tato část korpusu byla anotována ručně.

Pro další využívání korpusu Dialog je podstatné, že původní nahrávky na videokazetách VHS se postupně digitalizují. V budoucnu tak bude možné pohodlně pracovat nejen s transkripcí, ale i s transkripcemi synchronně propojenými s audiovizuálním záznamem.

Autory korpusu Dialog jsou v různých proporcích především: Světla Čmejrková, Lucie Jílková, Petr Kaderka, Jana Klímová, Kamila Mrázková, Zdeňka Svobodová. Autorem technického řešení projektu je Nino Peterek (z Ústavu formální a aplikované lingvistiky MFF UK Praha).

1.2.1.5. Korpus věcného stylu

Rovněž v Ústavu pro jazyk český Akademie věd České republiky byl již v 70. letech (pod vedením dr. Marie Těšitelové) vyvinut tzv. **Korpus věcného stylu**, který jako jediný obsahuje ve své malé části přepis mluvené (byť ne spontánní) řeči se základní (částečnou) syntaktickou anotací.

Korpus věcného stylu je korpusem textů o celkovém objemu 550 000 slov anotovaných na morfologické a na povrchově syntaktické rovině. V době vzniku korpusu existovaly ještě další dva anotované korpusy — Brown Corpus of Standard American English a Lancaster-Oslo/Bergen Corpus of British English. Oba dva korpusy měly v té době dvojnásobný objem oproti Korpusu věcného stylu, ale jejich anotace zachycovaly pouze morfologické informace, nikoli informace o větné stavbě. V tomto kontextu je na místě zdůraznit ojedinělost českého korpusu. Bohužel politická situace tehdejšího Československa nedovolila, aby se Korpus věcného stylu dostal do podvědomí světové korpusové lingvistiky.

Korpus věcného stylu obsahuje písemné a mluvené texty (celkem 180 textů) stylu publicistického (33 % textů), administrativního (11 % textů) a vědeckého (56 % textů). Mluvené texty (psaná podoba rozhlasových reportáží a rozhovorů, televizních komentářů a zpráv, proslovených přednášek) představují třetinu celkového objemu (asi 120 tisíc slov) a pro češtinu reprezentují vůbec první morfologicky i syntakticky anotovaný mluvený datový materiál.

Tento korpus se v současné době v Ústavu formální a aplikované lingvistiky (MFF UK Praha) ve spolupráci s Ústavem pro jazyk český Akademie věd ČR převádí do moderní podoby, do formátu PML, ve kterém je Pražský závislostní korpus 2.0. Vydán bude pod názvem **Český akademický korpus**.

1.2.1.6. Korpusy používané pro automatické rozpoznávání mluvené řeči

Existuje i řada korpusů používaných pro automatické rozpoznávání mluvené řeči, které však zpravidla obsahují pouze textovou transkripci. Vzhledem k jejich velikosti jsou však dalším cenným materiálem. Takové korpusy vznikají zejména v Oddělení umělé inteligence na Západočeské univerzitě v Plzni. Na tomto pracovišti byly vytvořeny například tyto korpusy:

- **Czech Broadcast News Speech and Transcripts** (viz Radová et al., 2004) - korpus obsahující zhruba 25 hodin řečových nahrávek vysílání českých rozhlasových a televizních stanic. Získané soubory řečových nahrávek se přepisují podle určitých pravidel tak, aby výsledný přepis byl co možná nejpřesnějším obrazem vyslovené promluvy (včetně přeřeknutí, nádechů a dalších šumů a neřečových událostí, jako např. zakašlání, smích a podobně).

  Korpus Czech Broadcast News Speech and Transcripts je společný projekt s Johns Hopkins University v Baltimore. Hlavními autory korpusu jsou: Byrne W., Radová V., Psutka J., Müller L., Ircing P., Matoušek J. Korpus je distribuován prostřednictvím Linguistic Data Consortium.

- **Korpus projektu Malach** (viz Psutka et al., 2002; Psutka et al., 2003) — korpus spontánních a emocionálních svědectví z doby holocaustu. Celý archiv videonahrávek shromážděný nadací Shoah Visual History Foundation čítá více než 52 tisíc výpovědí ve 32 jazycích. Jedním z nich je i čeština. Archiv nahrávek je zpracováván v rámci mezinárodního projektu Malach (http://malach.umiacs.umd.edu/). Čeština, ruština, slovenština, polština a další jazyky jsou zpracovávány v Oddělení umělé inteligence na Katedře kybernetiky ZČU v Plzni ve spolupráci s Ústavem formální a aplikované lingvistiky (MFF UK Praha). Více viz 4.1.1 – „Korpus projektu Malach“.

1.2.2. Lingvistické práce o mluvené češtině

V této sekci uvádíme z různých zdrojů získaný (zcela jistě ne úplný) soupis lingvistické literatury o mluvené češtině.

Bachmanová, J.: **Zvukový archiv nářečních textů**. Naše řeč, 82, 1999, s. 47-48.

Čmejrková, S.: **Slovo psané a mluvené**. Slovo a slovesnost, 54, 1993, s. 51-58.

Čmejrková, S.: **Televizní interview a jiné duely: mediální dialog jako žánr veřejného projevu**. Slovo a slovesnost, 60, 1999, s. 247-268.

Čmejrková, S.: **Emotions in language and communication**. In E. Weigand (ed.): Emotion in Dialogic Interaction: Advances in the Complex. Amsterdam – Philadelphia, John Benjamins, 2004, s. 37-57.

Čmejrková, S.; Daneš, F.; Havlová, E. (eds.): **Writing vs. Speaking: Language, Text, Discourse, Communication**. Tuebingen, Gunter Narr Verlag, 1994.

Čmejrková, S.; Jílková, L.; Kaderka, P.: **Mluvená čeština v televizních debatách: korpus DIALOG**, Slovo a slovesnost, 2004, 65, s. 243-269.

Daneš, F.: **Intonace a věta ve spisovné češtině**. Praha, 1975.

Hála, B.; Sovák, M.: **Hlas, řeč, sluch**. Praha, 1955.

Hausenblas, K.: **Výstavba jazykových projevů a styl**. Praha, 1972.

Hoffmannová, J.: **Intertextové, metatextové, terminologické problémy české analýzy diskurzu**. In P. Žigo (ed.): Philologica 56: Zborník Filozofickej fakulty Univerzita Komenského. Bratislava, Univerzita Komenského, 2003, s. 67-72.

Hoffmannová, J.: **K univerzálnosti a specifičnosti pojmu dialog**. In Z. Hladká; P. Karlík (eds.): Čeština – univerzália a specifika 2: Sborník z konference. Brno, Masarykova univerzita, 2000, s. 53-61.

Hoffmannová, J.: **Metodologie „konverzační analýzy" a transkripční symboly**. In J. Stachová (ed.): Symbol v lidském vnímání, myšlení a vyjadřování: Sborník příspěvků. Praha, Filozofický ústav ČSAV, 1992, s. 234-241.

Hoffmannová, J.; Müllerová, O.: **Dialog v češtině**. München, Verlag Otto Sagner, 1999.

Hoffmannová, J.; Müllerová, O.: **Jak vedeme dialog s institucemi**. Praha, Academia, 2000.

Hoffmannová, J.; Müllerová, O.; Zeman, J.: **Konverzace v češtině při rodinných a přátelských návštěvách**. Praha, Trizonia, 1999.

Karhanová, K.: **Rhetorical question in polemical media dialogue (based on material drawn from Czech TV political debates)**. In A. Betten; M. Dannerer (eds.): Dialoganalyse IX: Dialogue in Literature and the Media. Tübingen, Max Niemeyer Verlag, 2005.

Kořenský, J. a kol.: **Komplexní analýza komunikačního procesu a textu**. České Budějovice, 1987.

Koťátko, P.: **Význam a komunikace**. Praha, 1998.

Kraus, J.: **Argumentation in Czech political debates**. In M. Bondi; S. Stati (eds.): Dialogue Analysis 2000. Tübingen, Max Niemeyer Verlag, 2003, s. 277-282.

Müllerová, O.: **Mluvený text a jeho syntaktická výstavba**. Praha, Academia, 1994.

Müllerová, O.: **Mluvená čeština v autentických textech**. Jinočany, 1995.

Müllerová, O.; Hoffmannová, J.: **Kapitoly o dialogu**. Praha, Pansofia, 1994.

Müllerová, O.; Nekvapil, J.: **Pauzy v mluveném textu**. Slovo a slovesnost, 47, 1986, s. 105-113.

Nekvapil, J.; Müllerová, O.: **K pauzám v komunikačním procesu**. Slovo a slovesnost, 49, 1998, s. 202-208.

Sgall, P.: **Jan Firbas, Functional sentence perspective in written and spoken communication**. Review of Jan Firbas, Functional sentence perspective in written and spoken communication. Journal of Pragmatics, 2000, s. 639-644.

Sgall, P.: **Problémy mluvené češtiny v Praze**. In Ondrejovič, S. (ed.): Město a jeho jazyk. Bratislava, Veda, vydavateľstvo Slovenskej akadémi vied, v edici Sociolinguistica Slovaca, 2000, s. 75-83.

Sgall, P.: **Spoken Czech revisited**. In Where One's Tongue Rules Well. A Festschrift for Charles E. Townsend, Slavica Publishers, 2002, s. 299-309.

Těšitelová, M. (ed.): **Psaná a mluvená odborná čeština z kvantitativního hlediska**. Linguistica IV., Praha, 1983.

Vaňková, I.: **Mlčení a řeč v komunikaci, jazyce a kultuře**. Praha, 1996.

Kapitola 2. Standardizace mluvené řeči

Korpusy mluvené řeči představují rozsáhlý materiál poskytující věrný obraz současného mluveného jazyka. Odlišují se od sebe svým rozsahem, hloubkou, zaměřením tematickým, žánrovým, teritoriálním atd. a používají se pro rozmanitý lingvistický výzkum (například fonetický, morfologický, syntaktický, stylistický, interakční). Na rozdíl od výzkumu psaného textu se však výzkum mluveného jazyka soustřeďuje většinou jen na přepis akustického signálu do textové podoby. Rozsah lingvistické anotace těchto transkripcí je nevelký, anotace se obvykle zastavuje na morfologické rovině, či se přidává prozodické značkování; další zpracování směrem k významu se obvykle neprovádí. Dosud neexistuje (ani ve světovém měřítku) systematicky a ve větším rozsahu manuálně anotovaný korpus na úrovni syntaktické, tím méně na úrovni jazykového významu (na úrovni hloubkově syntaktické). Tomu odpovídá i stav zdrojů, tj. jazykových dat, vhodných pro pravděpodobnostní trénování a strojové učení za účelem plného porozumění mluvené řeči. Více viz 1.1 – „Výzkum mluvené řeči ve světě“.

Na základě morfologicky a prozodicky anotovaných dat byly již vytvořeny aplikace pro automatické rozpoznávání slovních jednotek v mluvené řeči a jejich přepis do psané podoby. I přes to, že tyto programy zvládnou s poměrně velkou přesností určit, o jaké slovo se jedná, nedokáží pracovat s jeho významem. Proto je dnešním úkolem vytvořit takový systém analýzy mluvené řeči, který zvládne řeč identifikovat až na úroveň významu. Potom bude možné provádět automatický přepis mluvené řeči do podoby řeči psané, vyhledávat informace v rozsáhlých audio nahrávkách, či dokonce uskutečnit automatický překlad z mluvené řeči jednoho jazyka do mluvené řeči jiného jazyka.

2.1. Potřeba standardizace mluvené řeči

Budování **syntakticko-sémanticky anotovaného korpusu mluvené řeči** však s sebou přináší řadu problémů, které se při budování korpusu psaného jazyka neobjeví (nebo jejich řešení je nasnadě). Jsou dány značnou rozdílností mluvené řeči od řeči psané.

Mluvená řeč, zejména ve své spontánní podobě, nedodržuje často ani elementární gramatická pravidla a zvyklosti. Tyto odchylky se pohybují od vcelku nepatrných (čtený nebo pečlivě připravený mluvený projev) přes viditelné a místy značné (korpus Dialog, televizní politické debaty, korpus projektu Malach) až po zcela zásadní (části Pražského mluveného korpusu, volně dostupné zdroje). Na následující ukázce jsou zřetelně vidět některé specifické problémy mluvené řeči:

ale kdyby náhodou tam byl nějakej ten ale mají tam zachariáš s tím radkem bejblem vole mají tam žlutý karty … aspoň desetník na kartu

výplňková slova: *vole*,
nové užití standardních slov: *náhodou*,
nespisovná (obecná) slova: *nějakej*,
opakování částí věty: *mají tam*,
elipsa (vynechaná slova, včetně slov pro význam klíčových: *nějakej ten [faul (?)]*, *aspoň [ti dám] desetník na kartu*.

Hlavní otázkou při budování syntakticko-sémantického korpusu je, jak tyto nejrůznější specifické jevy mluvené řeči zachytit, tj. jakým způsobem budou při syntakticko-sémantické anotaci zohledněna přeřeknutí, opakování slov, falešné začátky, nedokončené věty, koktání, neřečové události jako smích nebo kašlání atd. Odpověď na tuto otázku je třeba hledat v souladu s účelem, ke kterému je korpus určen. Jinak se při syntaktické anotaci přistoupí ke specifickým jevům mluvené řeči, pokud bude cílem korpusu zachytit specifickou strukturu mluvené řeči, například pro lingvistický či psychologický výzkum (viz projekt Childes — 1.1.1.4 – „The Childes database“). Budujeme-li korpus vhodný pro pravděpodobnostní trénování a strojové učení za účelem plného porozumění mluvené řeči (to je účel PDTSC i ostatních korpusů popsaných v 1.1.1 – „Syntaktické korpusy mluvené řeči“), je otázka, do jaké míry je nutné při syntakticko-sémantické anotaci specifické jevy mluvené řeči zohledňovat. Podle J. B. Johannessen a F. Jørgensen (viz Johannessen — Jørgensen, 2005) jsou v zásadě tři možnosti, jak naložit se specifickými jevy mluvené řeči při syntaktické anotaci:

A. vzít všechny jevy mluvené řeči vážně.

B. vzít vážně jen některé vybrané jevy mluvené řeči a ostatní ignorovat.

C. ignorovat všechny jevy mluvené řeči.

Rozhodnutí pro jednu z možností A, B, C s sebou nese různé důsledky pro podobu anotačního schématu na syntaktické rovině. Možnost A a B znamená především rozšíření inventáře anotačních značek i celého souboru anotačních pravidel tak, aby byly aplikovatelné i na specifickou strukturu mluvené řeči. V současné době však neexistuje manuálně syntakticko-sémanticky anotovaný korpus mluvené řeči, který by potvrzoval, že je tento způsob možný.

I původním záměrem při budování PDTSC bylo syntakticko-sémanticky anotovat mluvenou řeč přímo, podle pravidel pro (tektogramatickou) anotaci psaných textů Pražského závislostního korpusu (viz 3.1.2 – „Pražský závislostní korpus 2.0“) a tato pravidla pouze upravovat a rozšiřovat pro zvláštnosti mluvené řeči. V prvním roce projektu probíhala lingvistická analýza dat z existujících, elektronicky dostupných mluvených korpusů (Pražský mluvený korpus, korpus Dialog, korpus projektu Malach) a dalších vzorků získaných z volně dostupných textů na internetu. Uskutečnily se též první pokusy se syntaktickou anotací mluvené řeči. A ukázalo se, že právě vzhledem ke specifičnosti mluvených projevů je původní záměr „přímé“ syntakticko-sémantické anotace neschůdný. To, co lidé spontánně říkají, má do gramatičnosti, jak je obvykle při formálním popisu jazyka chápána, velmi daleko a odchylky se mohou vyskytnout v mluvené promluvě prakticky kdykoliv a kdekoliv. Nejde je systematicky popsat a říci, co do mluvené řeči ještě patří a co již nikoli. Pro „přímou“ syntakticko-sémantickou anotaci by tudíž musela být doslova vymýšlena „gramatická“ (spíše však „negramatická“) pravidla anotace na všechno, co lze při věrném přepisu toho, co lidé říkají, očekávat.

Navíc pro automatickou závislostní analýzu a pro navazující analýzu významu se počítá s víceméně jazykově korektním (spisovným) vstupem; současné metody analýzy vět přirozeného jazyka, přestože používají statistické (a tedy inherentně velmi robustní) metody, na neúplném, nevhodném, nebo gramaticky a lexikálně velmi „nesprávném“ textu nedávají dobré výsledky. Současné běžné metody rozpoznávání mluvené řeči (automatic speech recognition, ASR) jsou založené na doslovném přepisu toho, co mluvčí řekl (včetně všech přeřeknutí, zakašlání apod.). Ovšem, čím lepší jsou v takovém systému výsledky (měřené shodou s originální promluvou), tím paradoxně přidělávají práci v případě, že danou promluvu chceme dále zpracovávat metodami analýzy textu. Je totiž lhostejné, že systém rozpoznávání řeči precizně určí všechna přeřeknutí, výplňková slova, slova ve špatném tvaru a podobně, pokud udělá chybu ve slově klíčovém pro pochopení významu věty. I současné metody hodnocení pomocí tzv. poměrné slovní chybovosti (Word Error Rate, WER), kde se za slovo považuje skutečně každé vyřčené slovo nebo jeho část (a to včetně tzv. neřečových událostí), nerozlišují dobře mezi systémy z výše uvedeného hlediska kvalitními a systémy ostatními (méně kvalitními). Stále více se tudíž ukazuje, že současné metody rozpoznávání mluvené řeči se staly „obětí vlastního úspěchu“ a že jejich výstupy nejsou pro následnou závislostní analýzu vhodné. Jiné možnosti však prakticky neexistují, a tedy nejsou k dispozici ani z literatury (a to ani z nejnovějších sborníků konferencí).

Je zřejmé, že situace dozrála pro změnu „paradigmatu“ výzkumu na pomezí mluveného a psaného jazyka. Z celosvětového hlediska je novým směrem výzkumu v této oblasti tzv. standardizace mluvené řeči. **Standardizace mluvené řeči** představuje nový způsob definice rozhraní mezi doslovnou transkripcí mluvené řeči a přepisem mluvené řeči určeným pro další následnou syntaktickou analýzu. Vychází se z přesvědčení, že pro syntaktickou analýzu není nutné zohledňovat specifické jevy mluvené řeči (výše uvedená možnost C). Standardizace mluvené řeči znamená, že před vlastní syntaktickou analýzou/anotací se segmenty mluvené řeči nejprve převedou na tzv. „standardizovaný text“, tj. na text více či méně blízký textu psanému. Takový převod pak umožní při syntaktické analýze mluvené řeči pracovat se stejnými pravidly a nástroji jako při analýze textu psaného.

2.2. První pokusy se standardizací mluvené řeči

První pokusy se standardizací byly provedeny v Computer and Information Science Department na University of Pennsylvania na datech korpusu Switchboard (viz 2.2.1 – „M. Meteer et al.: Dysfluency Annotation“). Tzv. „dysfluency annotation“ spočívá však pouze v označení specifických jevů mluvené řeči: v transkribované mluvené řeči se pomocí speciálních značek označí všechna místa, kde došlo k přeřeknutí, zakoktání, opakovaní, falešnému začátku, nedořečené výpovědi, výplňkovému slovu, neřečové události apod. Segmenty mluvené řeči se však dále

neupravují a v drtivé většině případů tak zůstávají pouze fragmenty s neúplnou syntaktickou strukturou, které není možné plně syntakticko-sémanticky analyzovat, přestože význam segmentu je z toho, co bylo vyřčeno, patrný.

Rovněž na datech korpusu Switchboard provádějí v Center for Speech and Langauge Processing na Johns Hopkins University v Baltimore tzv. „rekonstrukci“ (speech reconstruction). Jde zřejmě o první projekt svého druhu (viz 2.2.2 – „Erin Fitzgerald: Speech reconstruction“). Námi navrhovaná pravidla rekonstrukce i způsob zachycení ze systému pravidel navrženého na Johns Hopkins University vycházejí a v řadě bodů jej přejímají.

2.2.1. M. Meteer et al.: Dysfluency Annotation

Na datech korpusu Switchboard (v rámci projektu PennTreebank) se před vlastní syntaktickou anotací provedla tzv. anotace dysfluencí. Tato anotace spočívá v označení míst, která nějakým způsobem narušují plynulost mluvené řeči. Po provedení anotace je možné tato označená místa z původního záznamu mluvené řeči vypustit („cleaning“), a získat tak mluvenou řeč zbavenou všech specifických jevů. Dysfluence jsou při anotaci typovány, tj. určuje se i typ konkrétního jevu narušujícího plynulost mluvené řeči.

V manuálu anotace dysfluencí pro korpus Switchboard (viz M. Meteer et al., 1995) se rozlišují následující **typy dysfluencí**:

a. **filled pause**: slovo, které má neomezenou distribuci a žádný lexikální význam (výplňkové slovo); například: *uh*, *um*, *huh*.

b. **explicit editing term**: slovo nebo fráze, která sice má nějaký význam, ale objevuje se obvykle mezi restartem a následující opravou; například: *I mean*; *sorry*, *excuse me*.

c. **discourse marker**: slovo nebo fráze, která má širokou distribuci a na rozdíl od výplňkových slov má i nějaký význam; například: *well*; *you know*; *like*; *so*; *actually*; *now*; *see*.

d. **coordinating conjunction**: koordinační spojky (například: *and*; *and then*; *and so*; *but*; *but anyway*; *either*; *or*; *so*) užité na začátku větných celků.

e. **aside**: fráze, kterou je přerušena plynulá posloupnost věty a věta později opět navazuje tam, kde byla přerušena.

f. **restarts**: falešné začátky, které mají následující strukturu: opakované slovo (**reparandum**) — bod přerušení (**interruption point**) - opravující výraz **(interregnum**) — oprava (**repair**). Příklady:

   repetition; například: *I think I think two or three minutes*;
   substitution; například: *I put it in on the table*;
   deletion; například: *I used to work and when I had two children*;
   chaining: *I really I I like it*;
   nesting: *I liked, uh, I, I liked it a lot*.

Součástí anotace je také určování hranic segmentů, tzv. **slash-units** (název podle lomítka, kterým se segmenty od sebe oddělují). Segmentem může být maximálně věta, ale jako „slash-units“ mohou být být označeny i úseky menší než věta (nevětné výpovědi), jestliže je anotátor interpretuje jako kompletní jednotky. Jako „slash-units“ jsou označovány:

continuers (například: *Right.*; *Oh.*),
phrases (například: *A cradle for it.*; *Especially with the kids.*),
simple sentences,
subordinate clauses standing alone,
complex sentences.

Jako **incomplete slash-units** jsou označeny všechny segmenty, které nepředstavují kompletní jednotku.

Přehled značek pro anotaci dysfluencí je v tab. 2.1; příklad anotace je v tab. 2.2.

Tabulka 2.1. Switchboard: Labels for dysfluency annotation

/	slash-units (boundary marker)	+	interruption point in restart	{D ...}	discourse marker
-/	incomplete slash-unit (boundary marker)	{A ...}	aside	{E ...}	explicit editing term
[...]	restart	{F ...}	filler	{C ...}	coordinating conjunction

Tabulka 2.2. Switchboard: Example of dysfluency annotation

```
B.1:  Okay. /

A.2:  Okay. /

B.3: {D Well } what do you think about the idea of, {F uh, } kids having to do
public service work for a year? / Do you think it's a , -/

@A.4: {D Well, } [ I, + I ] think it's a pretty good idea. / I think they
should either do that, [ or, + or ] afford some time to the military, [ or, +
or ] helping elderly people. /

B.5:  Yes, / yes, / def-, -/

A.6: [ [ I, + I, ] + {D you know, } I ] think that we have a bunch of elderly
folks in the country that could use some help / {C and } I think that before we
expend all our young talent overseas [ and, + and ] helping other countries we
ought to perhaps give a little bit of our help to our own folks at home / {C
and } --
```

2.2.2. Erin Fitzgerald: Speech reconstruction

Při „speech reconstruction“, kterou navrhuje E. Fitzgerald (viz Fitzgerald, 2006), je mluvená řeč každého mluvčího uložena ve dvou souborech — v jednom je zaznamenán originální transkribovaný text (w-rovina), ve druhém je pak uložen text rekonstruovaný (m-rovina). Oba soubory jsou mezi sebou propojeny různě pojmenovanými typy odkazů (vedoucích z jednotek m-roviny na jednotky w-roviny). Odkazy popisují, jakými akcemi se z původního textu získal text rekonstruovaný. Rozlišují se následující typy akcí (typy odkazů mezi m-rovinou a w-rovinou):

A. **sentence boundary actions**: akce, kterými je anotována správná segmentace vět. Umožňují spojit dva neúplné segmenty v jeden úplný a naopak (v případě, že jsou spojeny dvě myšlenky spolu nesouvisející) rozdělit jeden segment na dva.

B. **deletion actions**: jednotky v originálním transkribovaném textu se ruší bez náhrady:

   delete co-reference: smazání nadbytečných deiktických slov.
   delete filler/interregnum/discourse marker: smazání výplňkových slov a frází, slov nebo frází, která se objevují mezi restartem a následující opravou (interregnum); například: *you know*; *uh*, *see*.
   delete leading coordinations: smazání spojek užitých na začátku větných celků; zejména: *and*, *but*; *because*; *or*.
   delete unnecessary function words: smazání nadbytečných funkčních slov.
   delete reparandum: smazání slov, která se opakují; například: *I really really hope so.*
   delete fragment: smazání fragmentů — myšlenek, které nebyly dokončeny.

C. **insertion actions**: do rekonstruovaného textu jsou přidávány nové slovní jednotky (které v originálním textu nejsou):

   insert function word: vložení chybějící předložky, spojky, členu, relativního zájmena.

insert neutral noun: vložení neutrálního jména *_NOUN_* na pozici chybějícího argumentu; například: *_NOUN_ still wants to party*
insert neutral verb: vložení neutrálního slovesa *to be* nebo *to have*; například: *I am actually working in Jersey.*

D. **substitution actions**: změna tvaru slova:

substitute tense/number change: změna tvaru slova z důvodu nesprávně vyjádřené gramatické kategorie času nebo čísla.
substitute: transcript error: změna tvaru slova kvůli chybě v transkripci.

E. **phrase movement actions**: změny ve slovosledu:

phrase movement: adjunct: změna slovosledného postavení adjunktu vůči slovesu.
phrase movement: argument: změna slovosledného postavení argumentu vůči slovesu.
phrase movement: grammar: změna slovosledného postavení funkčního slova vůči slovesu.

Každé rekonstruované větné jednotce je na konci anotačního procesu přiřazen její typ, který určuje ne/kompletnost rekonstrukce (tj. zda se větu ne/podařilo rekonstruovat) a příspěvek větné jednotky k obsahu celého textu (tj. zda věta jako celek ne/nese nějaký obsah). Rozlišují se následující typy:

backchannel: kladné přitakání mluvčího, bez obsahu; například: *Yes.*; *I know.*
good sentence, no repairs: plynulá, kompletní, gramatická věta, kterou nebylo třeba upravovat.
good sentence, post-repairs: ve větě byly provedeny úpravy; výsledná věta je plynulá, kompletní a gramatická.
clean fragment with content: segment obsahuje posloupnost plnovýznamových slov, která může být částí gramatické věty, nicméně některé elementy ve větě chybí (například sloveso) a úplná rekonstrukce není možná; například: *I remember*
fragment, no content: posloupnost neplnovýznamových slov, nenese žádný význam; například: *I could; Which was that was that no*
cannot fix sentence: větu nelze rekonstruovat, protože její význam (to, co chtěl mluvčí říci) je nejasný.

Erin Fitzgerald, která v USA na tomto projektu pracuje společně s profesorem F. Jelínkem, pobývala v roce 2006 v Praze a k anotaci používá software vyvinutý na Ústavu formální a aplikované lingvistiky MFF UK Praha.

Kapitola 3. Pražský závislostní korpus mluvené češtiny (PDTSC)

Pražský závislostní korpus mluvené češtiny (PDTSC) bude syntakticko-sémanticky anotovaný korpus mluvené češtiny pro účely pravděpodobnostního trénování a strojového učení s cílem plného porozumění mluvené řeči. Při syntakticko-sémantické anotaci mluvené řeči budeme vycházet z hloubkové (tektogramatické) anotace psaných textů, která je zpracovávána v projektu Pražského závislostního korpusu verze 2.0 (více viz 3.1 – „Východiska pro syntakticko-sémantickou analýzu mluvené řeči“).

Před vlastní syntakticko-sémantickou analýzou se však segmenty mluvené řeči nejprve převedou na gramaticky správné věty, tj. provede se tzv. rekonstrukce standardizovaného textu z mluvené řeči (více viz 3.2 – „Rekonstrukce standardizovaného textu z mluvené řeči“). Převod mluvené řeči na standardizovaný text umožní, že při syntaktické anotaci mluvené řeči budeme moci pracovat s obdobnými pravidly a značkami jako při anotaci psaného textu v PDT 2.0. Klíčové při tomto postupu bude zachování (anotování) vazeb mezi vstupní transkripcí mluvené řeči a rekonstruovaným standardizovaným textem; jinými slovy: při syntakticko-sémantické analýze sice pracujeme s textem, který je zbavený specifických jevů mluvené řeči, ale pamatujeme si, jak jsme tento text z původní transkripce mluvené řeči získali, a jsme schopni jej zpětně zkonstruovat.

PDTSC bude mít strukturu analogickou struktuře korpusu PDT 2.0 (viz 3.3 – „Systém rovin“): hierarchický systém vzájemně propojených rovin anotace. Na rozdíl od systému rovin v PDT 2.0 bude v systému rovin v PDTSC zavedena jedna rovina navíc. Nová rovina bude sloužit pro uchování automaticky rozpoznané mluvené řeči a bude nejnižší rovinou systému.

3.1. Východiska pro syntakticko-sémantickou analýzu mluvené řeči

Syntakticko-sémantická analýza mluvené řeči v PDTSC bude vycházet z hloubkové (tektogramatické) anotace psaných textů (viz 3.1.3 – „Tektogramatická reprezentace“), která je zpracovávána v projektu Pražského závislostního korpusu verze 2.0 (viz 3.1.2 – „Pražský závislostní korpus 2.0“). Projekt Pražského závislostního korpusu je založen na teoretickém pojmovém rámci funkčního generativního popisu (viz 3.1.1 – „Funkční generativní popis“).

3.1.1. Funkční generativní popis

Funkční generativní popis (viz zejména Sgall, 1967; Sgall et al., 1986; Sgall — Hajičová — Panevová, 1986; Panevová, 1980) navazuje na tradici Pražské školy a kombinuje ji s aktuálními poznatky světové komputační lingvistiky; rozvíjí se od 60. let 20. století. Jde o závislostní typ formalismu, který byl navržen pro účely teoretického popisu struktury českých vět pomocí generativní procedury.

Základní charakteristikou funkčního generativního popisu je stratifikační přístup k popisu jazyka: popis jazyka (vztah mezi jazykovým výrazem a významem) je rozčleněn na několik kroků podle rovin jazykového systému. Každá z rovin je množinou zápisů vět. Každá rovina má svou vlastní syntax (různé elementární jednotky se skládají v jednotky komplexní). Nejvyšší rovina odpovídá významovému plánu jazyka, nejnižší plánu výrazovému. Jednotky sousedních rovin jsou ve vzájemném vztahu formy a funkce — ve vztahu reprezentace (tj. jednotka vyšší roviny je funkcí jednotky nižší roviny, jednotka nižší roviny je její formou).

3.1.2. Pražský závislostní korpus 2.0

Pražský závislostní korpus 2.0 (Prague Dependency Treebank 2.0, dále PDT 2.0; viz http://ufal.mff.cuni.cz/pdt2.0) je komplexně anotovaný korpus českých textů. Data používaná pro anotaci byla zvolena náhodně (v blocích) z textů Českého národního korpusu. Jde o psané texty; 60 % korpusu tvoří publicistické články (politika, sport, kultura atd.), 20 % populárně vědecké texty a 20 % ekonomické zprávy a analýzy. Korpus tedy neobsahuje, až na výjimky typu citací, přímé řeči v textu apod., přepis spontánního mluveného jazyka.

Data jsou v PDT 2.0 anotována na třech rovinách:

- **na morfologické rovině** (2 milióny slovních jednotek). Na morfologické rovině je každé slovní jednotce (tj. každému slovnímu tvaru, číslu i interpunkčnímu znaménku) přiřazeno lema (základní slovní forma) a tag (morfologické kategorie).

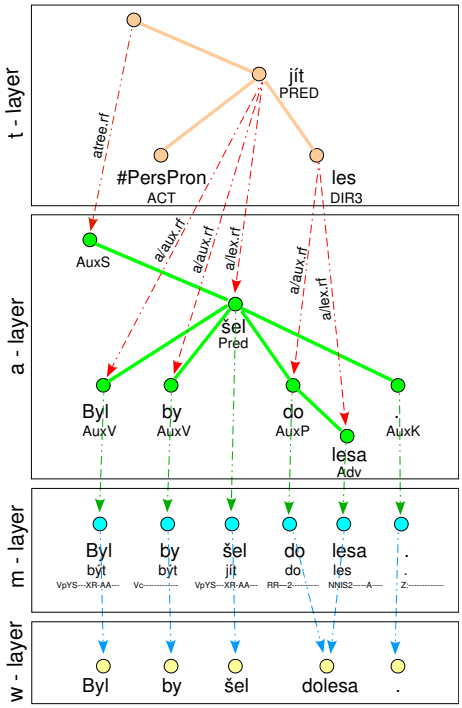
- **na analytické rovině** (tj. na rovině povrchové syntaxe; 1,5 miliónu slovních jednotek). Analytická rovina zhruba odpovídá větnému rozboru z hlediska závislostní syntaxe. Věta je reprezentována kořenovým stromem s ohodnocenými hranami a uzly. Každá slovní jednotka odpovídá právě jednomu uzlu. Hranou stromu je naznačen závislostní vztah mezi dvěma uzly a zároveň je uveden syntaktický typ tohoto vztahu (tzv. analytická funkce: podmět — Sb, příslovečné určení – Adv, předložka — AuxP apod.).

- **na tektogramatické rovině** (tj. na rovině hloubkové syntaxe; 0,8 miliónu slovních jednotek). Dosud nejvyšší, tektogramatická rovina tvoří jakýsi přechod mezi systémem jazyka a sémantikou, popisuje hloubkové syntaktické vztahy. Více viz 3.1.3 – „Tektogramatická reprezentace“.

Ve skutečnosti existuje ještě jedna, neanotační rovina reprezentující „surový text“. Na této rovině, zvané **slovní rovina**, je text rozdělen do dokumentů a odstavců. Jsou tu rozlišeny slovní jednotky (slova, čísla, interpunkce) a jsou opatřeny jednoznačnými identifikátory. Slovní rovina je také nazývána **w-rovina**, morfologická: **m-rovina**, analytická: **a-rovina** a tektogramatická: **t-rovina**.

Na obr. 3.1 jsou znázorněny vztahy mezi sousedními rovinami, jak jsou anotovány a reprezentovány v datech PDT 2.0. Zobrazená česká věta *Byl by šel dolesa.* obsahuje minulý čas podmiňovacího způsobu slovesa *jít* a tiskovou chybu.

Obrázek 3.1. Propojení rovin anotace v Pražském závislostním korpusu 2.0



3.1.3. Tektogramatická reprezentace

Tektogramatická anotace je založena na teoretickém pojmovém rámci funkčního generativního popisu (viz 3.1.1 – „Funkční generativní popis“).

Tektogramatická anotace je anotací strukturní, závislostní, zachycuje tzv. hloubkovou, významovou strukturu věty. Na tektogramatické rovině má každá (správně tvořená) věta alespoň jeden zápis, který jednoznačně charakterizuje význam této věty, popřípadě jeden z jejích významů (je-li věta i z hlediska svého širšího kontextu výjimečně

víceznačná). Zápis na tektogramatické rovině obsahuje veškerou informaci, kterou stavba věty a její lexikální obsazení dává a která je nutná pro převod tektogramatického zápisu na zápis nižších rovin i pro její interpretaci ve smyslu intenzionální sémantiky.

Tektogramatický zápis věty tedy obsahuje kromě zachycení vlastní hloubkové struktury věty a funkcí jednotlivých členů této struktury i řadu dalších údajů, jako jsou různé druhy tzv. gramatémů, informace o gramatické a textové koreferenci a o aktuálním členění (včetně hloubkového slovosledu, tj. stupně výpovědní dynamičnosti).

Tektogramatický strom (jakožto datová struktura) má tyto základní vlastnosti:

a. **Tektogramatický strom** je datová struktura, jejímž základem je kořenový strom (ve smyslu definice teorie grafů): skládá se z množiny uzlů a z množiny hran a jeden uzel je vyznačen jako kořen stromu.

b. **Uzel** tektogramatického stromu reprezentuje významovou jednotku věty, tj. autosémantické slovo s jeho slovy pomocnými (s předložkou, podřadicí spojkou, pomocnými slovesy). Uzlem jsou reprezentovány i významové jednotky nepřítomné v povrchové podobě věty (například strom obsahuje uzly pro elidovaná obligatorní valenční doplnění vyplývající z valence řídícího slova). Tento aspekt bude mít silný vliv na řešení projektu PDTSC vzhledem k tomu, že elips je v mluvené řeči podstatně více než v jazyce psaném.

c. **Atributy uzlu**. Každý uzel je sám o sobě komplexní, uvnitř strukturovaná jednotka. Lze ji chápat jako množinu atributů, přesněji řečeno jako množinu uspořádaných dvojic jméno atributu — hodnota atributu. Přítomnost nebo nepřítomnost jednotlivých atributů v daném uzlu vyplývá z typu uzlu.

   Atributy uzlu můžeme třídit do několika skupin. Základními atributy uzlu tektogramatického stromu jsou tektogramatické lema, gramatémy a funktor. Tektogramatické lema zachycuje lexikální význam uzlu. Gramatémy odpovídají především významům morfologických kategorií. Funktory odpovídají druhům syntaktické závislosti mezi autosémantickými výrazy, syntaktickým funkcím. U uzlů jsou dále uvedeny také hodnoty atributů, které podávají informaci o koreferenci, aktuálním členění a hloubkovém slovosledu.

d. **Hrany** tektogramatického stromu slouží primárně k zachycení závislostních vztahů mezi uzly (respektive mezi významovými jednotkami).

e. Uzly tektogramatického stromu jsou **lineárně uspořádané**, lineární uspořádání uzlů slouží k reprezentaci hloubkového slovosledu věty.

Definice a konvence anotace na tektogramatické rovině jsou podrobně popsány v příručce pro anotátory (viz Mikulová et al., 2005; též http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/html/index.html).

3.2. Rekonstrukce standardizovaného textu z mluvené řeči

Rekonstrukce standardizovaného textu z mluvené řeči představuje nový způsob definice rozhraní mezi systémy automatického rozpoznávání řeči a systémy hloubkové (významové) analýzy (psaného) textu. Vychází z přesvědčení, že při syntakticko-sémantické analýze, tj. při zachycování významu promluv, není nutné zohledňovat specifické jevy mluvené řeči, ale nezbytně nutné je zachovat pouze významy původních vyřčených segmentů a tyto významy zachytit v anotaci.

Práci anotátora při rekonstrukci lze přirovnat k redaktorovi, který zpracovává nahraný rozhovor k otištění v časopise: vstupní přepis mluvené řeči anotátor rozčleňuje do vět (více viz 3.2.1 – „Segmentace mluvené řeči do vět“), věty různě upravuje, některá slova maže, jiná přidává, mění jejich pořadí (viz 3.2.2 – „Úpravy segmentů mluvené řeči“). Rozhovor těmito modifikacemi dostává psanou podobu (tj. dodržuje pravidla psané řeči), ta by měla být potenciálnímu čtenáři nejen srozumitelná, ale měla by se tomuto čtenáři též dobře číst.

Výstupem anotace je tzv. **standardizovaný text**, který vymezujeme na základě následujících podmínek:

- text neobsahuje neřečové události,
- specifické jevy mluvené řeči jsou z textu odstraněny,

- proud mluvené řeči je rozčleněn do vět,
- text je celkově srozumitelný a dobře se čte,
- věty mají gramatický slovosled a běžnou českou syntax,
- použity jsou jen spisovné tvary slov,
- text je napsán v souladu s pravidly českého pravopisu.

Pro rekonstrukci standardizovaného textu z původních segmentů mluvené řeči platí dva základní principy:

A. **Princip zachování významu**: provedené modifikace původních segmentů mluvené řeči nesmějí zasahovat do významu (obsahu); jinými slovy: platí, že významy (obsahy) sdělované původní mluvenou řečí a významy (obsahy) obsažené ve standardizovaném textu jsou tytéž.

B. **Princip minimálního počtu úprav**: provádí se jen tolik modifikací, kolik jich původní segmenty mluvené řeči nutně vyžadují, aby bylo dosaženo standardizovaného textu.

Rozdíly, kterými se vstupní segmenty mluvené řeči liší od svých standardizovaných verzí, tj. provedené modifikace, jsou zachyceny **ve vztazích mezi jednotkami obou textů**. V PDTSC budou vstupní segmenty mluvené řeči reprezentovány na w-rovině korpusu, standardizovaný text bude reprezentován na m-rovině. Provedené modifikace budou zachyceny systémem pojmenovaných odkazů mezi m-uzly a w-uzly (více viz 3.3.4.2 – „Vztahy mezi m-rovinou a w-rovinou“).

Tabulka 3.1. Příklad vstupního segmentu mluvené řeči a jeho standardizované verze

no tak já s- chtěl sem jenom říct jak- jaký byli mezi náma hrdinové že
➜ *Chtěl jsem jenom říct, jací byli mezi námi hrdinové.*

3.2.1. Segmentace mluvené řeči do vět

Jedním z hlavních problémů při budování syntaktického korpusu mluvené řeči je problém segmentace mluvené řeči, totiž stanovení kritérií, podle nichž bude různě přerušovaný (krátkými a dlouhými pauzami, kašláním, nádechy, smíchem), nebo naopak nepřerušovaný (rychlá překotná mluva) proud mluvené řeči rozčleněn na syntaktické jednotky odpovídající v psaném textu větám.

V automaticky transkribovaném textu je segmentace mluvené řeči (která je vždy výsledkem nějaké automatické procedury v rámci použitého rozpoznávače řeči) zpravidla provedena podle výskytu neřečových událostí v proudu mluvené řeči (tj. například podle delších úseků ticha, ale třeba i v místě zakašlání nebo smíchu). Výsledné segmenty zhruba odpovídají větám, ne však nutně. Pro následnou syntaktickou anotaci ovšem může tato segmentace být použita jen s velkými obtížemi.

Při manuální transkripci mluvené řeči se otázka segmentace většinou řeší nějakým jednoduchým pravidlem (srov. například Radová, V., 2002) — proud mluvené řeči je členěn na základě pauz tak, aby výsledné úseky „zhruba odpovídaly jedné větě“. Taková pravidla však nejsou příliš přesná a vedou k nekonzistentním anotacím, nehledě na to, že specifické jevy mluvené řeči (zejména opakování celých úseků textu, falešné začátky, nedokončené věty) takovou segmentaci často úplně znemožňují. Problém segmentace mluvené řeči tak mimo jiné ukazuje na potřebu standardizace, neboť segmentaci do „opravdových“ vět je možné v úplnosti provést vždy jen s dalšími úpravami mluvených segmentů.

V PDTSC se při určování větné hranice řídíme:

- **principem nejdelší možné klauze**: klauze zahrnuje co nejvíce potenciálních větných členů za podmínky, že výsledná věta je ještě utvořena jak syntakticky, tak sémanticky správně.

někteří lidé mně utkvěli velmi v paměti z toho koncentračního tábora

➜ *Někteří lidé z koncentračního tábora mně velmi utkvěli v paměti.*

Tento princip „longest match“ je převzat z pravidel syntaktické anotace mluvené řeči pro korpusy projektu Verbmobil (viz 1.1.1.3 – „The Verbmobil treebanks“). V korpusech projektu Verbmobil však nebyla před syntaktickou anotací provedena žádná standardizace a je zřejmé, že uplatnění tohoto principu na přerývanou, syntakticky často porušenou strukturu mluvené řeči bez možnosti jakékoli úpravy původního přepisu mluvené řeči nemůže vést k úplnému rozčlenění proudu mluvené řeči na celky, které budou odpovídat větám.

Při rekonstrukci standardizovaného textu jsou vytvářeny větné celky, které odpovídají obvyklým pravidlům pro psaný text. Výsledná (rekonstruovaná) věta může být i neúplná (jde-li například o nedokončenou myšlenku), musí však odpovídat jednomu ze čtyř typů klauzí popsaných v tektogramatickém manuálu (Mikulová et. al., 2005), tj. musí jít o:

- slovesnou klauzi (i elidovanou): *(Kdy přijdeš?) V pátek odpoledne.*
- nominativní klauzi: *Nesmysly.*
- vokativní klauzi: *Pane Barňák!*
- citoslovečnou klauzi: *Pozor!*

nebo o spojení jedné nebo více těchto klauzí.

Rekonstruovaná věta je obsahové sdělení (tj. má nějaký obsah), bezobsažné úseky textu (obsažené v proudu mluvené řeči) nemají ve standardizovaném textu svůj protějšek.

3.2.1.1. Typování segmentů

Každá věta je ohodnocena z hlediska obsahové důležitosti v kontextu celého textu. Určuje se, jakého druhu je obsah dané věty, tj. zda daná věta přináší novou informaci, nebo je otázkou po takové informaci, příkazem, přitakáním mluvčího apod.

Z tohoto hlediska rozlišujeme následujících osm typů vět:

`information` Hodnota náleží větám, které do výsledného rekonstruovaného textu přináší podstatné nové informace. Věty s hodnotou `information` za žádných okolností nelze ze standardizovaného textu vypustit. Z formálního hlediska jde primárně o věty oznamovací (případně věty přací, zvolací, řečnické otázky); např.: *Je mi osmdesát let.*

`question` Hodnota náleží primárně zjišťovacím a doplňovacím otázkám, tedy otázkám po informaci, nikoli otázkám zvolacím, řečnickým a otázkám, které jsou ve skutečnosti žádostí. Z formálního hlediska jde o věty tázací; např.: *Kolik je vám let?*

`instruction` Hodnota náleží větám, které vyjadřují příkaz, žádost, přání jednoho mluvčího, aby druhý mluvčí něco vykonal, řídil se jeho pokyny. Z formálního hlediska jde primárně o věty rozkazovací a některé otázky vyjadřující žádost; např.: *Řekněte, jak jste strávil dětství.*

`confirmation` Hodnota náleží větám, které vyjadřují kladné přitakání druhého mluvčího (posluchače) k obsahu projevu prvního mluvčího, aniž by tento projev prvního mluvčího byl danou větou nějak přerušován; první mluvčí na základě přitakání nemění směr hovoru. V konverzaci jsou tyto věty naprosto běžné, nenesou žádnou informaci, nepřispívají k obsahu konverzace, mohou být z textu i vypuštěny a jeho informační hodnota se tím neztratí; např.: *Aha.*

`surprise` Hodnota náleží větám, které v širokém smyslu vyjadřují překvapení posluchače nad novou informací, kterou mu mluvčí sděluje; např.: *Vážně?*

`disbelief` Hodnota náleží větám, které signalizují, že posluchač nevěděl o tom, co mluvčí říká, nebo si myslel opak a nové informaci příliš nevěří, není o ní ještě přesvědčen; např.: *To není možný.*

`repetition` Hodnota náleží větám, které znova opakují celou předcházející myšlenku, například proto, aby byl zvýrazněn její význam nebo aby byla potvrzena její platnost. (Nemusí jít o doslovné opakování; např.: *(Pět korun jsme dostali.) Dostali jsme pět korun.*)

`other` Hodnota slouží pro označení ostatních nedefinovaných případů.

3.2.2. Úpravy segmentů mluvené řeči

Nejdůležitější částí anotace jsou různé typy modifikací vstupní transkripce za účelem vytvoření standardizovaného textu. Rozlišujeme následující typy modifikací:

A. **Ortografické modifikace**:

a. pravopisné úpravy textu (vložení interpunkce, velká písmena),

b. přepis slov pomocí nealfabetických znaků (zejména číslic).

B. **Vlastní modifikace**:

a. vymazání slovní jednotky,

b. vložení slovní jednotky,

c. substituce slovní jednotky,

d. změna ve slovosledu.

C. **Interpretace neřečových událostí**:

a. odstranění obsahově nerelevantních neřečových událostí,

b. zachycení obsahově relevantních neřečových událostí.

Všechny modifikace jsou vždy prováděny se zřetelem ke kontextu celého textu a za přísného dodržování obou principů rekonstrukce (uvedených v úvodu kapitoly 3.2 – „Rekonstrukce standardizovaného textu z mluvené řeči“).

Jednotlivé typy úprav podrobněji popisujeme v následujících sekcích.

3.2.2.1. Ortografické modifikace

Ortografické modifikace představují pravidelné úpravy vstupního textu, vyplývající ze základního požadavku na standardizovaný text, totiž že standardizovaný text splňuje obecné charakteristiky psaného textu a jsou v něm dodržena pravidla českého pravopisu. K ortografickým modifikacím patří:

pravopisné úpravy,
přepis slov pomocí nealfabetických znaků.

Ve standardizovaném textu jsou dodržována všechna pravopisná pravidla pro psaný text (přijaté transkripční zásady pro zápis segmentů mluvené řeči přitom tato pravidla dodržovat nemusí). K pravopisným úpravám patří zejména vložení interpunkčních znamének a náhrada malých písmen za velká (v případech, kdy jsou v transkripci používána jen malá písmena).

on řekl byl sem tam ale nikdo mu nevěřil

➜ *On řekl: „Byl jsem tam,“ ale nikdo mu nevěřil.*

Při transkripci mluvené řeči je zpravidla vše, co bylo řečeno, zaznamenáváno slovně (pomocí písmen). V psaném textu však často s výhodou užíváme k zápisu některých slov nealfabetických znaků (číslic a jiných symbolů). Ve standardizovaném textu jsou proto například letopočty, data, časové údaje zapisovány pomocí číslic.

bylo to v roce čtyřicet pět

➜ *Bylo to v roce 1945.*

3.2.2.2. Vlastní modifikace

Nejdůležitější částí anotace jsou tzv. vlastní modifikace vstupního transkribovaného textu, představují na rozdíl od ortografických modifikací podstatný zásah do podoby vstupního textu. K dispozici jsou čtyři typy vlastních modifikací:

vymazání slovní jednotky,
vložení nové slovní jednotky,
substituce slovní jednotky,
změny ve slovosledu.

Modifikace mazání. Ve standardizovaném textu jsou obsaženy jen takové slovní jednotky, které mají význam, tj. přispívají k vyjádření obsahu sdělení. Slovní jednotky i celé úseky textu, které nenesou žádný význam a nepřispívají k obsahu věty, nebo jinak porušují plynulost textu jsou při rekonstrukci standardizovaného textu ze vstupní transkripce odstraňovány. Jde o slovní jednotky obsahově nerelevantní.

K obsahově nerelevantním slovním jednotkám řadíme zejména:

- výplňková slova (*hledali nějakýho ubožáčka že jo*),
- výplňkové fráze (*to bylo v Praze na já myslím na Vánoce*),
- nadbytečná deiktická slova (*jel sem do té Prahy*),
- nadbytečné konektory (*a tam to trvalo dva roky*),
- nadbytečná a nesprávně užitá gramatická slova (*on je sedumadvacátý únor*),
- restarty (*a to byli většinou to byl většinou ten personál*),
- opakující se úseky textu (*to bylo poslední poslední jídlo*),
- fragmenty (*v pátek sem <cough> Barňák pak odešel*).

my sme tam dostávali v Bratislavě podporu že asi deset korun denně sme dostávali že

➜ *V Bratislavě jsme dostávali podporu asi deset korun denně.*

Modifikace vkládání. Standardizovaný text může obsahovat i slovní jednotky, které nebyly vyřčeny, ale které jsou nezbytné pro vytvoření gramaticky i lexikálně správné věty (standardizovaného textu).

Takovými při rekonstrukci vkládanými jednotkami jsou zejména:

- chybějící gramatická slova,
- nevyjádřená plnovýznamová slova.

revolverem mu takle začali před nos jo

➜ *Revolverem mu takhle začali dělat před nosem.*

Modifikace substituce. Ve standardizovaném textu jsou užívána jen slova spisovná a též jen správně utvořené tvary slov. Při rekonstrukci jsou proto měněny vstupní nespisovné a nesprávně utvořené formy slov a v případě slov užitých nesprávně z hlediska vyjadřovaného významu též i celá slova.

Forma slovních jednotek se mění z následujících důvodů:

- forma slova je nespisovná (*takže vy ste se znali*),
- forma slova je nesprávně utvořená; vyjadřuje nesprávně hodnotu nějaké gramatické kategorie (*tyto auta se vracely prázdné*).

Lema slovních jednotek se mění z následujících důvodů:

- lema slova je zvoleno nesprávně z hlediska vyjadřovaného významu (*tak jsem začal mluvit jaký má krásný obrazy*),
- lema slova je zvoleno nesprávně z hlediska vyjadřované vazby (*architekt zelenka má velikou zásluhu o tuto činnost*),
- lema slova je zvoleno nesprávně z hlediska zachování koherence textu (*nalil mi kávu do hrnku pak si nabral omáčku a podal mi ji*)

architekt Zelenka má velikou zálohu o tuto činnost

➜ *Architekt Zelenka má velikou zásluhu na této činnosti.*

Změny ve slovosledu. Rekonstruované věty mají gramatický slovosled, který nenarušuje plynulost textu, dochází tedy i ke změnám ve slovosledném uspořádání.

po pěti sme leželi

➜ *Leželi jsme po pěti.*

3.2.2.3. Interpretace neřečových událostí

Standardizovaný text, ve kterém se řídíme pravidly psaného textu, primárně neobsahuje značky pro neřečové události (jejich přehled viz tab. 3.2). Obsahově nerelevantní neřečové události (většina) se při rekonstrukci bez náhrady odstraňují. Obsahově relevantní neřečové události, tj. takové, které nesou nějaký význam, kterým přispívají k obsahu sdělení (souhlas, protest), zachycujeme ve standardizovaném textu primárně prostředky textu psaného, tj. zejména pomocí interpunkčních znamének, slovosledu.

Význam pro obsah sdělení může mít ale celá řada neřečových událostí, které jen pomocí běžných prostředků psaného textu nezachytíme (ironický smích, šeptání, náhlé zvýšení hlasu aj.). Takové neřečové události zaznamenáváme ve standardizovaném textu speciální značkou („textem v závorkách“), kterou tyto obsahově relevantní neřečové události interpretujeme.

<mouth> <inhale> tak možná že bych ještě něco řek <breath> <uh> <silence>

➜ *Tak možná, že bych ještě něco řekl.*

[spk1] *no a je to tak* [spk2] *<silence>*

➜ [spk1] *Je to tak?* [spk2] *<nejspíš kývnul>*

Tabulka 3.2. Přehled značek pro neřečové události

breath	zvuk dechu
click	mlaskání jazykem
cough	kašlání
laugh	smích
mouth	mlaskání rty
noise	konec hluku v pozadí
inhale	nádech
silence	ticho, pauza
uh	hezitační zvuk
unintelligible	nesrozumitelný úsek

3.3. Systém rovin

PDTSC bude mít strukturu analogickou korpusu PDT 2.0: **hierarchický systém vzájemně propojených rovin anotace**. V systému rovin PDTSC však bude jedna rovina navíc. Tato rovina, označovaná jako z-rovina (z angl. „zero“), bude nejnižší rovinou systému, teprve nad ní bude postavena w-rovina, m-rovina, a-rovina a t-rovina, tj. roviny zavedené již v PDT 2.0. Nově definovaná (oproti PDT 2.0) bude též w-rovina a m-rovina.

Věrný přepis mluvené řeči (který mj. sleduje proud řeči v čase a lineárně odpovídá vstupnímu akustickému signálu) bude v PDTSC zachován (pro účely trénování systémů automatického rozpoznávání řeči) na dvou nejnižších rovinách: na z-rovině (výstup z automatického rozpoznávače) a na w-rovině (manuální transkripce). Segmenty mluvené řeči budou však již na úrovni morfologické, na m-rovině, „standardizovány“ ve smyslu použití standardních spisovných slovních tvarů a slov vůbec, gramatického slovosledu a běžné české syntaxe, tj. mezi w-rovinou a m-rovinou bude provedena tzv. rekonstrukce standardizovaného textu z mluvené řeči (viz 3.2 – „Rekonstrukce standardizovaného textu z mluvené řeči“).

Standardizované věty mluvené řeči budou na syntaktických rovinách anotovány podle obdobných pravidel jako psané texty v korpusu PDT 2.0. Předpokládá se, že stávající pravidla anotace budou měněna minimálně.

Základním datovým formátem pro PDTSC je stejně jako pro PDT 2.0 formát označovaný jako PML (Prague Markup Language), který je založený na XML. Více k tomuto formátu viz http://ufal.mff.cuni.cz/jazz/PML. K PML schématům a aktuálním verzím PML-schémat viz http://ufal.mff.cuni.cz/pdt/pml/schema/.

3.3.1. Z-rovina

Nově zavedenou **z-rovinu** definujeme jako neanotační rovinu určenou pro reprezentaci tokenů (bez dalších atributů) získaných automatickým procesem z jiného zdroje. Do tohoto externího zdroje mohou vést z této roviny odkazy.

Na z-rovině v PDTSC bude zachován přepis mluvené řeči v původním znění, které je výstupem automatického rozpoznávače. Ze z-roviny povedou odkazy do externích souborů obsahujících digitalizované audio nahrávky.

Základní jednotky z-roviny (**z-uzly**) jsou dvojího typu: token a tzv. proluka (`gap`). **Tokenem** (z-tokenem) rozumíme každý rozpoznaný prvek transkribovaného textu, každou rozpoznanou slovní jednotku. **Proluka** (**gap**) vyplňuje časový úsek mezi tokeny, může se jednat o pouhé ticho, ale i o nejrůznější neřečové události a též i o nerozpoznané úseky mluveného textu. Sekvence tokenů a proluk na z-rovině je časově spojitá.

Tokeny a proluky jsou reprezentovány jako komplexní jednotky, které obsahují strukturu atributů popsaných v tab. 3.3 a v tab. 3.4.

Tabulka 3.3. Atributy z-uzlu `token`

`id`	unikátní identifikátor z-uzlu v rámci celého korpusu
`start_time`	časový údaj označující začátek trvání tokenu na časové ose audio nahrávky (odkaz do audia)
`end_time`	časový údaj označující konec trvání tokenu na časové ose audio nahrávky (odkaz do audia)
`token`	rozpoznaný token (slovní jednotka)

Tabulka 3.4. Atributy z-uzlu `gap`

`id`	unikátní identifikátor z-uzlu v rámci celého korpusu
`start_time`	časový údaj označující začátek trvání proluky na časové ose audio nahrávky (odkaz do audia)
`end_time`	časový údaj označující konec trvání proluky na časové ose audio nahrávky (odkaz do audia)

Lineární, časově spojitá posloupnost z-uzlů (tokenů a proluk) je na z-rovině rozčleněna do **segmentů** (z-segmentů). Hranice segmentů jsou určovány automatickým rozpoznávačem na základě delších úseků ticha, ale i na základě jiných neřečových událostí. Každý z-segment má v rámci PDTSC jednoznačný identifikátor.

Z-souboru jako celku náleží též atribut popisující zdroj, ze kterého byla automaticky rozpoznaná data získána (`original_format`), a dále atribut identifikující jazyk rozpoznaného textu (`lang`).

3.3.2. W-rovina

W-rovina bude v PDTSC též reprezentovat věrný přepis mluvené řeči (se všemi jejími zvláštnostmi — přeřeknutí, opakování slov, neřečové události aj.), ale půjde již o přepis manuálně upravený anotátorem podle přesně nadefinovaných pravidel pro manuální transkripci řeči (viz 4.1.1.1 – „Pravidla pro manuální transkripci zvukového signálu“). W-rovina bude představovat transkripci mluvené řeči zbavenou automatických transkripčních omylů.

Struktura w-roviny (PML-schéma) zohledňuje všechny informace, které při manuální transkripci umožňuje anotační nástroj Transcriber (http://trans.sourceforge.net). PML-schéma w-roviny je navrženo tak, aby bylo převoditelné na XML-zápis manuální transkripce v nástroji Transcriber. Přitom může být některá informace vynechána (například odkazy do z-roviny). A naopak: XML-zápis manuální transkripce v nástroji Transcriber je převoditelný do PML. Chybějící (i povinná) informace pro PML je při tomto převodu dodávána.

Manuální transkripce mluvené řeči je na w-rovině reprezentována jako sekvence různých jednotek, tzv. **událostí** (**`events`**), též **w-uzlů**. Jsou to:

`w` (manuálně) rozpoznaná slovní jednotka.

`nonspeech` (manuálně) rozpoznaná neřečová událost, která nepřekrývá mluvenou řeč (tj. která není hlukem na pozadí).

`sync` časový údaj, který je odkazem do audio nahrávky. Je povinný na začátku každé repliky. Uvnitř repliky je pak umísťován přibližně po každých deseti rozpoznaných slovních jednotkách (událostí typu `w`).

`background_begin` začátek trvání rozpoznaného hluku na pozadí mluvené řeči. Hluk probíhá na pozadí všech rozpoznaných událostí následujících v sekvenci událostí bezprostředně za událostí `background_begin` do té události `background_end`, která na danou událost `background_begin` odkazuje.

`background_end` konec trvání rozpoznaného hluku na pozadí mluvené řeči.

`speaker` identifikace mluvčího v rámci jedné repliky. Do promluvy mluvčího identifikovaného událostí `speaker` patří všechny rozpoznané události následující v sekvenci událostí bezprostředně za událostí `speaker` do nejbližší další události `speaker`.

`comment`	komentáře a anotátorské poznámky k transkribovanému textu, které se z výsledné prezentace anotace odstraní.

Události (w-uzly) dělíme na:

- **obsahové události**: `w`, `nonspeech`, `background_begin` a `background_end`;
- **referenční události**: `sync`, `speaker`,
- **pomocné události**: `comment`.

Sekvence tokenů a neřečových událostí není na w-rovině nutně časově spojitá, některé kratší úseky ticha, drobné zvuky nemusí být na w-rovině reprezentovány w-uzlem (událostí).

Všechny události jsou reprezentovány jako komplexní jednotky, které obsahují strukturu atributů popsaných v následujících tabulkách.

Tabulka 3.5. Atributy události `w`

`id`	Unikátní identifikátor události (w-uzlu) v rámci celého korpusu.
`token`	Rozpoznaná slovní jednotka (forma).
`pronounced_as`	V atributu je zapsána neobvyklá výslovnost rozpoznané slovní jednotky.
`is_slip`	Hodnoty: `0`, `1`. Atribut označuje, zda rozpoznaná slovní jednotka je/není přeřeknutí.
`z.rf`	Hodnotou je seznam, jehož každý prvek je PML odkaz. Atribut obsahuje identifikátory všech z-uzlů, které se na časové ose plně nebo i jen částečně překrývají s rozpoznaným tokenem.

Tabulka 3.6. Atributy události `nonspeech`

`id`	Unikátní identifikátor události (w-uzlu) v rámci celého korpusu.
`desc`	Hodnoty: `click`, `mouth`, `cough`, `laugh`, `breath`, `inhale`, `uh`, `unintelligible`, `noise`, `silence`. Hodnoty označují různé typy rozpoznaných neřečových událostí.
`z.rf`	Hodnotou je seznam, jehož každý prvek je PML odkaz. Atribut obsahuje identifikátory všech z-uzlů, které se na časové ose plně nebo i jen částečně překrývají s rozpoznanou neřečovou událostí.

Tabulka 3.7. Atributy události `sync`

`id`	Unikátní identifikátor události (w-uzlu) v rámci celého korpusu.
`audio.rf`	Odkaz do audia. Obsahuje strukturu atributů `reffile.rf` a `time`.
`audio.rf/`**`reffile.rf`**	Název audio souboru.
`audio.rf/`**`time`**	Časový údaj (odkaz do audia).

Tabulka 3.8. Atributy události `background_begin`

`id`	Unikátní identifikátor události (w-uzlu) v rámci celého korpusu.
`desc`	Hodnoty: `click`, `mouth`, `cough`, `laugh`, `breath`, `inhale`, `uh`, `unintelligible`, `noise`, `silence`. Hodnoty označují různé typy rozpoznaných hluků v pozadí.

Tabulka 3.9. Atributy události `background_end`

`background_begin.rf`	PML odkaz. Obsahuje identifikátor té události `background_begin`, která označuje začátek hluku na pozadí, jehož trvání daná událost `background_end` ukončuje.

Tabulka 3.10. Atributy události `speaker`

`id`	Unikátní identifikátor události (w-uzlu) v rámci celého korpusu.
`speaker.rf`	PML odkaz. Obsahuje identifikátor mluvčího (který je popsán v hlavičce souboru).

Tabulka 3.11. Atributy události `comment`

`type`	Hodnoty: `punctuation`, `compatibility`, `other`. Hodnoty určují typ komentáře, anotátorské poznámky. Hodnota `punctuation` slouží pro zachování případné interpunkce, která může být v některých manuálních transkripcích získaných z jiných zdrojů. Hodnota `compatibility` slouží pro zachování hodnot atributů, jejichž význam neznáme, ale které mohou být v některých manuálních transkripcích získaných z jiných zdrojů, všechny ostatní typy komentářů a poznámek mají hodnotu `other`.
`text`	Text komentáře.

Posloupnost událostí je na w-rovině rozdělena do **replik** (**`turn`**). Každá replika obsahuje sekvenci různých typů událostí, přičemž první událostí každé takové sekvence je vždy událost `speaker`, za kterou následuje událost `sync`. Replika je vymezena mluvčím. Každá replika má primárně jednoho mluvčího, může jich však být i více v případě, že promluva hlavního mluvčího je překrývána promluvami ostatních mluvčích, kteří se účastní dialogu (více mluvčí mluví současně, skáčou si do řeči).

Jednotlivé repliky jsou sdruženy do **dokumentů** (**`doc`**). Pro případnou obsahovou anotaci, tj. anotaci témat, o kterých se v nahrávaných rozhovorech mluví, jsou v hlavičce dokumentu připraveny příslušné atributy.

Replice jako celku náleží atributy popsané v tab. 3.12. Dokumentu jako celku náleží atributy popsané v tab. 3.13. Souboru náleží atributy popsané v tab. 3.14.

Tabulka 3.12. Atributy náležející replice (`turn`)

`id`	Unikátní identifikátor repliky v rámci celého korpusu.
`audio.rf`	Odkaz do audia. Obsahuje strukturu atributů: `reffile.rf`, `start_time`, `end_time`, které jsou popsány dále.
`audio.rf`/**`reffile.rf`**	Název audio souboru.
`audio.rf`/**`start_time`**	Časový údaj udávající začátek trvání repliky.
`audio.rf`/**`end_time`**	Časový údaj udávající konec trvání repliky.
`mode`	Hodnoty: `spontaneous`, `planned`. Atribut určuje, zda replika byla pronesena spontánně (hodnota `spontaneous`), nebo zda byla dopředu připravená (`planned`).
`speakers.rf`	Seznam, jehož každý prvek je PML odkaz. Identifikátory mluvčích, kteří mluví v rámci repliky. Primárně má replika jednoho mluvčího, při souběžné mluvě více mluvčích však může mít mluvčích více.
`fidelity`	Hodnoty: `high`, `medium`, `low`. Atribut určuje kvalitu nahrávky dané repliky, jak dobře je nahrávce rozumět (velmi vysoká kvalita nahrávky — hodnota `high`, průměrná kvalita nahrávky — `medium`, nízká kvalita nahrávky — `low`).
`channel`	Hodnoty: `telephone`, `studio`. Atribut určuje, v jakém prostředí byla replika pronesena, zda ve studiu (hodnota `studio`), nebo zda jde o záznam z telefonu (`telephone`).

Tabulka 3.13. Atributy náležející dokumentu (`doc`)

`topics`	V atributu je uložen seznam témat, o kterých byly rozhovory v souboru vedeny. Obsahuje seznam struktur atributů: `id`, `desc`, které jsou popsány dále.
`topics/`**`id`**	Unikátní identifikátor tématu v rámci celého korpusu.
`topics/`**`desc`**	Vymezení, popis tématu.
`topic.rf`	PML odkaz. Obsahuje identifikátor tématu, o kterém se v dokumentu mluví (témata jsou popsána v hlavičce souboru).
`speakers`	V atributu je uložen seznam mluvčích, kteří v souboru vystupují. Obsahuje seznam struktur atributů: `id`, `check`, `dialect`, `name`, `accent`, `type`, `scope`, které jsou popsány dále.
`speakers/`**`id`**	Unikátní identifikátor mluvčího v rámci celého korpusu.
`speakers/`**`check`**	Hodnoty: `yes`, `no`. Atribut je převzat ze struktury, kterou poskytuje program Transcriber, a jeho význam se nepodařilo zjistit.
`speakers/`**`dialect`**	Hodnoty: `native`, `nonnative`. Atribut udává, zda jazyk, kterým mluvčí mluví, je pro mluvčího jazykem rodným (`native`), či nikoli (`nonnative`).
`speakers/`**`name`**	Jméno mluvčího.
`speakers/`**`accent`**	Poznámka o přízvuku.
`speakers/`**`type`**	Hodnoty: `male`, `female`, `child`, `unknown`. Atribut udává, zda mluvčím je muž (`male`), žena (`female`), dítě (`child`), případně, že o pohlaví či věku mluvčího není nic známo (`unknown`).
`speakers/`**`scope`**	Hodnoty: `local`, `global`. Atribut je převzat ze struktury, kterou poskytuje program Transcriber, a jeho význam se nepodařilo zjistit.
`audio.rf`	Odkaz do audia. Obsahuje strukturu atributů: `reffile.rf`, `start_time`, `end_time`, které jsou popsány dále.
`audio.rf/`**`reffile.rf`**	Název audio souboru.
`audio.rf/`**`start_time`**	Časový údaj udávající začátek trvání dokumentu.
`audio.rf/`**`end_time`**	Časový údaj udávající konec trvání dokumentu.
`type`	Hodnoty: `report`, `nontrans`, `filler`. Atribut je převzat ze struktury, kterou poskytuje program Transcriber, a jeho význam se nepodařilo zjistit.

Tabulka 3.14. Atributy náležející w-souboru jako celku

original_format	Informace o zdroji, ze kterého byl získán manuálně rozpoznaný text. Obsahuje strukturu atributů: desc, version_date, version, které jsou popsány dále.
original_format/**desc**	Popis (název) formátu, zdroje.
original_format/**version**	Číslo verze.
original_format/**version_date**	Datum získané verze.
annotator	Informace o anotátorovi. Obsahuje atribut name, který je popsaný dále.
annotator/**name**	Jméno anotátora.
lang	Jazyk manuálně rozpoznaného textu.
air_date	Datum pořízení nahrávky.
program	Nahrávaná událost (název nahrávaného programu, vymezení události).
elapsed_time	Čas, který z nahrávané události uplynul před započetím nahrávání.

3.3.3. M-rovina

Transkribovaný proud mluvené řeči je na m-rovině nahrazen standardizovaným textem.

Automatická segmentace mluvené řeči je na m-rovině nahrazena manuální segmentací do vět. Věta je na m-rovině v PDTSC identifikována pomocí **s-elementu**. S-element obsahuje atributy popsané v tab. 3.15 a sekvenci **m-uzlů**.

Základními jednotkami m-roviny jsou slovní jednotky, kterými rozumíme každý prvek rekonstruované standardizované věty: slovo, číslo, interpunkční znaménko. Slovní jednotky jsou reprezentované **m-uzly typu m**. Speciálními **m-uzly typu nontext** jsou pak zachyceny obsahově relevantní neřečové události. Každému m-uzlu typu m náleží atributy popsané v tab. 3.16 a každému m-uzlu typu nontext náleží atributy popsané v tab. 3.17.

M-uzlům se při morfologické anotaci přiřazuje lemma (základní slovní forma) a tag (morfologické kategorie). Pravidla této anotace nejsou předmětem této technické zprávy; jsou k dispozici na http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/m-layer/html/index.html.

Tabulka 3.15. Atributy s-elementu

id	Unikátní identifikátor s-elementu v rámci celého korpusu.
w-begin.rf	PML odkaz. Obsahuje identifikátor první (obsahové) události na w-rovině (prvního w-uzlu), která byla použita pro rekonstruovanou větu.
w-end.rf	PML odkaz. Obsahuje identifikátor poslední (obsahové) události na w-rovině (posledního w-uzlu), která byla použita pro rekonstruovanou větu.
w-speaker.rf	PML odkaz. Obsahuje identifikátor mluvčího, který je původcem rekonstruované věty (odkaz do seznamu mluvčích na w-rovině).
stype	Hodnoty: `information`, `question`, `instruction`, `confirmation`, `surprise`, `disbelief`, `repetition`, `other`. Atributem určuje, jakého druhu je obsah dané věty, tj. zda daná věta přináší novou informaci, nebo je otázkou po takové informaci, příkazem, přitakáním mluvčího apod.
is_modified	Hodnoty: `0`, `1`. Atribut určuje, zda výsledná rekonstruovaná věta je vůči původnímu textu na w-rovině modifikovaná (hodnota `1`), či nikoliv (hodnota `0`).

Tabulka 3.16. Atributy m-uzlu typu `m`

id	Unikátní identifikátor m-uzlu (slovní jednotky) v rámci celého korpusu.
scr.rf	PML odkaz. Atribut ukazuje na zdroj morfologické anotace.
wrefs	Odkazy do w-roviny. Atribut obsahuje seznam struktur atributů `w.rf` a `type`, které jsou popsány dále.
`wrefs/`**w.rf**	PML odkaz. Obsahuje identifikátor w-uzlu, kterému m-uzel odpovídá.
`wrefs/`**type**	Hodnoty: `basic`, `substitution`, `num`, `nonspeech`. Atribut zachycuje typ odkazu.
form	Atribut obsahuje formu reprezentované slovní jednotky (tvar slova, číslo, interpunkci).
lemma	Atribut obsahuje lema, tj. základní slovní formu slova přiřazenou při morfologické anotaci. Homonymní lemata s různým významem jsou rozlišena číselnými sufixy. Součástí některých lemmat je i informace o jejich významu a stylovému příznaku.
tag	Atribut obsahuje morfologický tag, který kóduje morfologické kategorie relevantní pro reprezentovanou slovní jednotku. Dvojice lema a tag jednoznačně určuje slovní formu.

Tabulka 3.17. Atributy m-uzlu typu `nontext`

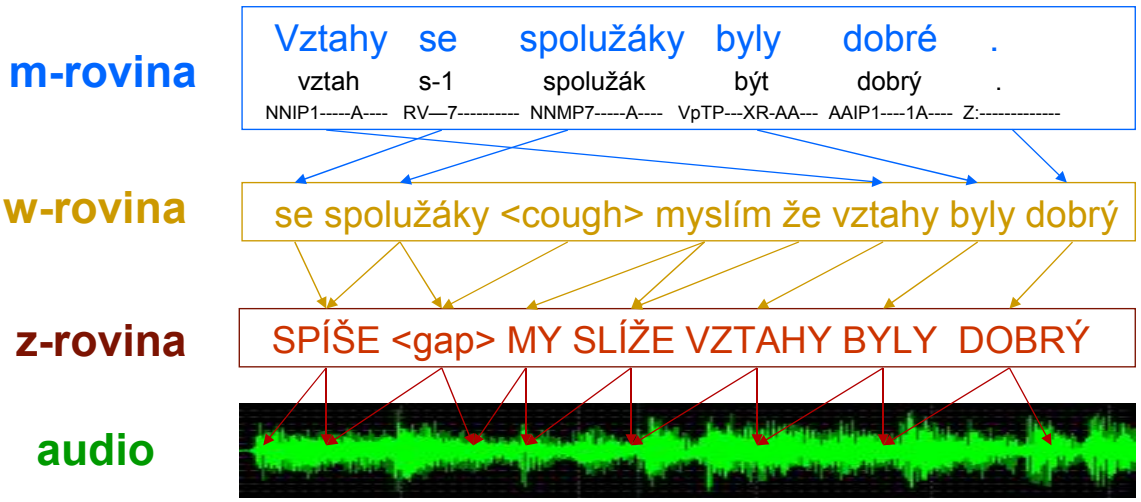
`id`	Unikátní identifikátor m-uzlu v rámci celého korpusu.
`type`	Atribut obsahuje popis obsahově relevantní neřečové události.
`wrefs`	Odkazy do w-roviny. Atribut obsahuje seznam struktur atributů `w.rf` a `type`, které jsou popsány dále.
`wrefs/`**`w.rf`**	PML odkaz. Obsahuje identifikátor w-uzlu, kterému m-uzel odpovídá.
`wrefs/`**`type`**	Hodnoty: `basic`, `substitution`, `num`, `nonspeech`. Atribut zachycuje typ odkazu.

3.3.4. Vztahy mezi rovinami

Podobně jako v korpusu PDT 2.0 budou i v PDTSC jednotlivé roviny mezi sebou propojeny systémem odkazů vedoucích vždy z uzlů roviny vyšší na uzly následující roviny nižší. Způsob realizace odkazů z nejvyšší t-roviny do a-roviny a z a-roviny do m-roviny bude stejný jako v PDT 2.0. V této sekci proto popisujeme pouze reprezentaci vztahů mezi nejnižšími rovinami.

Propojení nejnižších rovin v PDTSC je schematicky znázorněno na obr. 3.2.

Obrázek 3.2. Propojení nejnižších rovin v PDTSC



3.3.4.1. Vztahy mezi w-rovinou a z-rovinou

Ve vztazích mezi w-uzly a z-uzly jsou zachyceny opravy automatické transkripce.

Platí, že **z každého w-uzlu vede alespoň jeden odkaz do z-roviny**. Z w-uzlu vedou odkazy na všechny z-uzly, se kterými se na časové ose (plně nebo jen částečně) překrývá.

Pořadí w-uzlů na w-rovině a pořadí odpovídajících z-uzlů na z-rovině je shodné, tj. odkazy do z-roviny se nekříží. Výjimkou jsou pouze případy, kdy více mluvčích hovoří najednou, pak se odkazy z w-uzlů na z-uzly mohou křížit.

Na jeden z-uzel může vést více odkazů. **Neplatí však, že na každý z-uzel vede nějaký odkaz z w-roviny**. Zejména některým z-uzlům typu `gap`, které vyplňují jen krátká ticha mezi slovy nebo větami, nemusí na w-rovině odpovídat žádný w-uzel.

Mezi w-uzly a z-uzly **mohou nastat** následující typy vztahů:

a. **w-uzel → z-uzel**

W-uzel odkazuje na jeden z-uzel, pokud hranice manuálně (anotátorem) rozpoznaného w-uzlu odpovídají hranicím nějakého automaticky rozpoznaného z-uzlu. (Manuálně rozpoznaný w-token může být s automaticky rozpoznaným z-tokenem lexikálně shodný, nebo upravený.)

b. **w-uzel → z-uzly**

W-uzel odkazuje na více z-uzlů, pokud hranice manuálně rozpoznaného w-uzlu odpovídají (na časové ose) více automaticky rozpoznaným z-uzlům. (V tomto případě se, pokud jde o tokeny, lexikální obsazení tokenů zcela jistě neshoduje.)

Z pohledu z-roviny:

a. **z-uzel ← w-uzel**

Na z-uzel vede jeden odkaz z w-roviny, pokud hranice automaticky rozpoznaného z-tokenu odpovídají hranicím nějakého manuálně rozpoznaného w-uzlu. (Automaticky rozpoznaný z-token může být s manuálně rozpoznaným w-tokenem lexikálně shodný, nebo upravený.)

b. **z-uzel ← w-uzly**

Na z-uzel vede více odkazů z w-roviny (z více w-uzlů), pokud automaticky rozpoznaný z-uzel odpovídá na w-rovině více manuálně rozpoznaným w-uzlům. Případ, kdy na jeden z-uzel vede více odkazů z w-roviny, je signálem nesprávného automatického rozpoznání (slovních) jednotek v mluvené řeči.

c. **z-uzel ← ∅**

Na z-uzel nevede žádný odkaz z w-roviny. W-rovina (na rozdíl od z-roviny) není časově spojitá, některá krátká ticha, případně drobné nepatrné zvuky (reprezentované z-uzly) nemusí být reprezentovány w-uzlem.

Mezi w-uzly a z-uzly **nemůže nastat** vztah:

a. w-uzel → ∅ (w-uzel neodkazuje na žádný z-uzel).

Odkazy z w-uzlů na z-uzly jsou realizovány pomocí atributu **`z.rf`**, který náleží každému w-uzlu. Hodnotou atributu `z.rf` je seznam identifikátorů všech odpovídajících z-uzlů.

3.3.4.2. Vztahy mezi m-rovinou a w-rovinou

Rozdíly mezi rekonstruovaným standardizovaným textem a vstupní transkripcí mluvené řeči jsou zachyceny ve vztazích mezi m-rovinou a w-rovinou pomocí propojení segmentů na m-rovině (s-elementů) s odpovídajícími úseky w-uzlů na w-rovině (viz 3.3.4.2.1 – „Vztahy mezi segmenty“) a též pomocí propojení m-uzlů a w-uzlů (viz 3.3.4.2.2 – „Vztahy mezi uzly“).

3.3.4.2.1. Vztahy mezi segmenty

Ve vztazích mezi segmenty je zachycena provedená segmentace proudu mluvené řeči do vět (zachycených na m-rovině s-elementem).

Platí, že **z každého s-elementu vedou dva odkazy do w-roviny**, a sice:

- odkaz na první obsahovou událost (w-uzel), která byla použita jako vstup pro rekonstruovanou větu.
- odkaz na poslední obsahovou událost (w-uzel), která byla použita jako vstup pro rekonstruovanou větu.

... <uh> <cough> <inhale> tak já teda začnu jo <inhale> to se stalo ...

➜ *Tak já začnu.*

Jako vstup rekonstruované věty byl použit celý úsek *<uh> <cough> <inhale> tak já teda začnu jo*, odkazy z s-elementu povedou na w-uzly reprezentující obsahové události *<uh>* a *jo*.

Odkazy ze dvou různých s-elementů se mohou křížit (a to v případě překrývání mluvčích). Na w-rovině mohou být obsahové události, které nebyly použity jako vstup pro žádnou rekonstruovanou větu (a to v případě, že nějaký úsek textu je interpretován jako bezobsažný.

Odkazy z s-elementu jsou realizovány pomocí atributů `w-begin.rf` a `w-end.rf`, jejichž hodnotou je identifikátor příslušného w-uzlu.

Každému s-elementu náleží též atribut `w-speaker.rf`, ve kterém je uložen identifikátor mluvčího, který danou větu pronesl. Odkaz vede do hlavičky dokumentu na w-rovině, do atributu `speakers`, ve kterém jsou identifikováni jednotliví mluvčí (viz tab. 3.13).

3.3.4.2.2. Vztahy mezi uzly

Ve vztazích mezi m-uzly a w-uzly jsou zachyceny provedené modifikace vstupní transkripce za účelem vytvoření standardizovaného textu (mazání, vkládání, substituce, změny ve slovosledu, zachycení obsahově relevantních neřečových událostí).

Platí, že **z m-uzlu, kterému odpovídá nějaký w-uzel na w-rovině, vede na tento w-uzel odkaz**.

Jádro odkazů mezi m-rovinou a w-rovinou tvoří odkazy mezi m-uzly typu `m` (reprezentujícími tokeny na m-rovině) a w-uzly typu `w` (reprezentujícími tokeny na w-rovině).

Mezi m-uzly typu `m` a w-uzly typu `w` mohou nastat následující typy vztahů:

a. **m-uzel typu `m` → w-uzel typu `w`**

   Z m-uzlu typu `m` vede odkaz na jeden w-uzel typu `w`, pokud tento w-uzel danému m-uzlu lexikálně odpovídá (úplně nebo částečně).

b. **m-uzel typu `m` → w-uzly typu `w`**

   Z m-uzlu typu `m` může vést více odkazů (na více než jeden w-uzel typu `w`). Takový případ nastává tehdy, když jeden m-uzel typu `m` reprezentující například zkratku nebo číslo (například: *1945*, *tzv.*) ve skutečnosti představuje více slov reprezentovaných na w-rovině více uzly (tedy: *devatenácet čtyřicet pět*; *tak zvaný*).

c. **m-uzel typu `m` → ∅**

   Z m-uzlu typu `m` nemusí vést žádný odkaz do w-roviny, pokud reprezentuje vložené gramaticky a obsahově nezbytné slovní jednotky, kterým neodpovídá žádný w-uzel na w-rovině. M-uzel typu `m`, ze kterého nevede žádný odkaz do w-roviny, nazýváme **vložený m-uzel**.

Z pohledu w-roviny:

a. **w-uzel typu `w` ← m-uzel typu `m`**

   Na w-uzel typu `w` vede odkaz z m-uzlu typu `m`, pokud w-uzlu na m-rovině nějaký m-uzel lexikálně (úplně nebo částečně) odpovídá.

b. **w-uzel typu `w` ← m-uzly typu `m`**

   Na w-uzel typu `w` vede více odkazů z jednoho m-uzlu typu `m`, pokud w-uzlu na m-rovině lexikálně (úplně nebo částečně) odpovídá více m-uzlů (například řadová číslovka *dvacátý* je na m-rovině nahrazena číslem a tečkou).

c. **w-uzel** ← ∅

Pokud w-uzlu typu `w` na m-rovině žádný m-uzel lexikálně neodpovídá, nevede na něj žádný odkaz. W-uzel, na který nevede žádný odkaz z m-roviny, představuje odstraněné neřečové události a vymazané obsahově nerelevantní slovní jednotky. Hovoříme o **smazaném w-uzlu**.

Specifickým typem odkazů jsou odkazy z m-uzlů typu `nontext` (reprezentujících obsahově relevantní neřečové události) na w-uzly typu `nonspeech` (reprezentující neřečové události), případně na w-uzly typu `background_begin` a `background_end` (reprezentující hluky na pozadí).

Mezi m-uzly typu `nontext` a w-uzly typu `nonspeech` (případně `background`) mohou nastat následující typy vztahů:

a. **m-uzel typu `nontext` → w-uzel/w-uzly typu `nonspeech`**

Z m-uzlu typu `nontext` vede odkaz na jeden nebo více w-uzlů typu `nonspeech`, pokud m-uzel typu `nontext` zachycuje obsahově relevantní neřečovou událost reprezentovanou na w-rovině jedním nebo více w-uzly.

b. **m-uzel typu `nontext`** → ∅

Z m-uzlu typu `nontext` nevede žádný odkaz do w-roviny, pokud m-uzel typu `nontext` zachycuje obsahově relevantní neřečovou událost, která není reprezentována na w-rovině.

Odkazy mezi m-uzly a w-uzly jsou typovány. Typ odkazu signalizuje provedenou modifikaci. Rozlišujeme následující typy odkazů:

A. **Typy odkazů z m-uzlů typu `m` na w-uzly typu `w`**:

- **`basic`**: forma m-uzlu se rovná tokenu w-uzlu, nebo byly provedeny pouze tzv. ortografické modifikace.
- **`num`**: mezi w-uzlem a m-uzlem byla provedena ortografická modifikace čísel (číslo zapsané slovy bylo nahrazeno číslem zapsaným pomocí číslic).
- **`substitution`**: forma nebo lema m-uzlu bylo vůči odpovídajícímu w-uzlu upraveno (byla provedena modifikace substituce).

B. **Typy odkazů z m-uzlů typu `nontext` na w-uzly typu `nonspeech`** (případně `background`):

- **`nonspeech`**.

Přehled odkazů mezi m-uzly a w-uzly je uveden v tabulce tab. 3.18.

Tabulka 3.18. Přehled odkazů z m-uzlů na w-uzly

Typ m-uzlu	Odkazované typy w-uzlů	Typ hrany
m-uzel typu `m`	w-uzel typu `w`	`basic`
		`num`
		`substitution`
	∅	-
m-uzel typu `nontext`	w-uzel typu `nonspeech`	`nonspeech`
	w-uzel typu `background`	`nonspeech`
	∅	-

Odkazy z m-uzlů do w-uzlů jsou realizovány pomocí atributu `wrefs`, který náleží každému m-uzlu. Atribut `wrefs` obsahuje seznam struktur atributů `w.rf` a `type`. Atribut `wrefs/w.rf` obsahuje identifikátor odpovídajícího w-uzlu. Atribut `wrefs/typo` pojmenovává typ odkazu.

Pořadí m-uzlů na m-rovině nemusí odpovídat pořadí w-uzlů typu na w-rovině. Odkazy z m-roviny do w-roviny se mohou různě křížit. Rozdílným uspořádáním uzlů na obou rovinách jsou zachyceny změny ve slovosledu.

Kapitola 4. Data a nástroje

4.1. Data

Předpokládáme, že data korpusu PDTSC budou data z existujících mluvených korpusů (viz 1.2.1 – „Korpusy mluvené češtiny“), ke kterým se podaří získat přístup. Vybírána budou tak, aby pokrývala široké spektrum různých typů mluvených projevů: spontánní neformální mluvené projevy, diskuze, přednášky, interview aj.

Jako první data pro anotaci byla zvolena data z projektu Malach (viz 4.1.1 – „Korpus projektu Malach“), která leží zhruba v polovině škály mezi zcela spontánní komunikací ve skupině (kde jsou i technické problémy dosud plně nedořešené i v oblasti zpracování akustiky, což není součástí projektu) a řečí čtenou, která by nepřinesla k poznání mluvené řeči prakticky nic nového.

Předzpracování dat bylo velmi náročné vzhledem k nutnosti získat nahrávky z korpusu projektu Malach a spárovat je s jejich transkripcí a částečnou standardizací z hlediska některých tvaroslovných koncovek, která byla provedena na Západočeské univerzitě v Plzni. Data byla vyčištěna od neúplných párů a poškozených záznamů a převedena do formátu PML.

4.1.1. Korpus projektu Malach

Po natočení filmu Schindlerův seznam založil režisér Steven Spielberg nadaci Shoah Visual History Foundation (VHF, www.vhf.org) s cílem shromáždit archiv videonahrávek svědectví lidí, kteří přežili holocaust. Nesmírné utrpení obětí holocaustu prezentované v těchto výpovědích se má stát studijním materiálem pro současnou i nastupující generaci a má být základem výchovy k toleranci lidí na celém světě. Do současné doby shromáždila VHF více než 52 tisíc výpovědí v 32 jazycích (přes 100 000 hodin videozáznamu). K vyhledávání důležitých informací ve svědeckých výpovědích budou využity nejnovější technologie z oblasti automatického rozpoznávání mluvené řeči. Tyto problémy řeší od roku 2001 prestižní projekt MALACH (Multilingual Access to Large Spoken Archives, viz http://malach.umiacs.umd.edu/) financovaný americkou grantovou agenturou NSF (National Science Foundation). K řešení projektu byla přizvána i Západočeská univerzita v Plzni, která se má společně s MFF UK v Praze podílet na zpracování jazyků střední a východní Evropy. Dalšími řešiteli projektu MALACH jsou IBM, Johns Hopkins University v Baltimore, University of Maryland a Shoah Visual History Foundation.

Česká část korpusu Malach, kterou mají na Západočeské univerzitě v Plzni k dispozici, představuje 561 výpovědí. Průměrná délka jedné výpovědi je asi 2 hodiny 15 minut. Výpovědi jsou zaznamenány na videonahrávkách, přičemž jedna videonáhrávka obsahuje 30 minut výpovědi (celá výpověď je tak zaznamenána na více videonahrávkách). Práce, které probíhají na Západočeské univerzitě v Plzni, jsou zaměřeny zejména na zpracování akustické části videonahrávek, na přípravu systému rozpoznávání řeči a provedení základních testů.

Zvukové nahrávky byly transkribovány za použití speciálního anotačního nástroje Transcriber 1.4.1, který je volně k dispozici na http://trans.sourceforge.net/en/presentation.php. Pro natrénování akustických modelů bylo ručně anotováno (transkribováno) vždy prvních patnáct minut z druhé půlhodinové nahrávky jedné výpovědi. Takto oanotovaných výpovědí je celkem 336 (tj. celkem 84 hodin). Jako testovací data bylo deset výpovědí manuálně transkribováno kompletně (výpovědi mají různou délku - dohromady představují přibližně 23 hodin). Tyto vybrané úseky výpovědí jsou anotovány podle pravidel popsaných v 4.1.1.1 – „Pravidla pro manuální transkripci zvukového signálu“. Automatickými nástroji bylo rozpoznáno a transkribováno zatím 335 výpovědí, ale plánuje se zpracovat všech 561 výpovědí.

Řeč na nahrávkách je spontánní a je velmi ovlivněna emocionálním stavem řečníka i jeho stářím (průměrný věk řečníků je cca 75 let), obsahuje často šepot a řadu neřečových událostí jako zakašlání, smích, pláč. Ve spontánní řeči českých výpovědí je také velké množství hovorových slov a gramaticky nesprávných vazeb, což velmi znesnadňuje práci spojenou s přípravou jazykového modelu (textový materiál vhodný pro trénování jazykových modelů spontánní české řeči v podstatě neexistuje). I pro relativně objemný slovník cca 50 000 slov stále velké množství neznámých slov velmi znesnadňuje funkci rozpoznávače. Při zpracování českých výpovědí byly proto navrženy speciální způsoby konstrukce slovníků i postupy při zpracování jazykových modelů. Současné výsledky rozpoznávání českých výpovědí jsou plně srovnatelné s výsledky, kterých dosahuje IBM při zpracování anglických výpovědí.

4.1.1.1. Pravidla pro manuální transkripci zvukového signálu

Při transkripci zvukových nahrávek provedené na Západočeské univerzitě v Plzni za použití anotačního nástroje Transcriber platila následující anotační pravidla:

- Zvukové nahrávky jsou rozděleny do segmentů tak, že každý segment odpovídá přibližně jedné větě.
- Začátek segmentu je označen značkou `<b ti>`, kde `ti` udává čas (v sekundách), ve kterém segment začíná.
- Okamžik, ve kterém se mění mluvčí, je označen `<t ti> <<spk#, n, g>>`, kde `ti` je opět čas v sekundách, `spk#` je identifikátor mluvčího (`spk1` pro tazatele, `spk2` pro dotazovaného, `spk3` pro jiného mluvčího), `n` je jméno a příjmení mluvčího (je-li známo), a na pozici `g` je buď písmeno `m` značící mluvčího-muže, nebo písmeno `f` značící, že mluvčím je žena.
- Situace, kdy mluvčí mluví jeden přes druhého, je označena:

  `<t ti><<spk_1, n_1, g_1 + spk_2, n_2, g_2>>`
  `SPEAKER1`: transkripce toho, co řekl mluvčí `spk_1`
  `SPEAKER2`: transkripce toho, co řekl mluvčí `spk_2`
- Všechno, co bylo řečeno, se transkribuje slovy, nejsou používána žádná čísla.
- Věta začíná malým písmenem. Pouze vlastní jména a akronyma jako *IBM, NATO* jsou zapsána velkými písmeny. Je-li nějaké slovo ve výpovědi spelováno, jsou jednotlivá spelovaná písmena zapsána velkými písmeny a oddělena mezerou.
- V transkripci se neužívají žádná interpunkční znaménka.
- Jestliže se mluvčí zakoktá a řekne například *tř třicet*, je takové zakoktání transkribováno jako *tř- třicet*. Znak „-“ se též užívá pro případy, kdy nějaké pronesené slovo je neúplné z jiných důvodů. V takových případech se znak „-“ umísťuje na začátek, nebo na konec proneseného výrazu v závislosti na tom, která část slova byla pronesena, zda začátek, nebo jeho konec. Jestliže za znakem „-“, ani před znakem „-“ není v transkripci mezera, je znak „-“ chápán jako součást zapsaného slova.
- Části výpovědí pronesené jiným jazykem než českým jsou uzavřeny v hranatých závorkách [ ].
- Ty části výpovědí (segmentů), u kterých si anotátor není jistý, zda jim dobře porozuměl, jsou uzávorkovány v kulatých závorkách. Například, jestliže si anotátor myslel, že mluvčí řekl *vypadá jako toto*, ale nebyl si jist, transkriboval daný úsek následovně: *(vypadá jako toto)*. Je-li nějaký úsek kompletně nesrozumitelný, tj. nelze-li rozpoznat jednotlivá slova, transkripce je: `<unintelligible>`.
- Neřečové události, jako jsou mlaskání jazykem nebo rty, kašlání, smích, dýchání, hluboké nádechy, se zachycují značkami: `<breath>`, `<click>`, `<cough>`, `<laugh>`, `<inhale>`, `<mouth>` aj.
- Hluk v pozadí výpovědi (projíždějící auto, štěkot psa) je zachycen následovně: jestliže žádné slovo není tímto hlukem překryto, je takový hluk označen jako `<noise>`. Pokud však některá slova pronáší mluvčí za hluku odehrávajícího se v pozadí, je značka `<noise_begin>` užita před prvním slovem, které je hlukem zasaženo, a značka `<noise_end>` je umístěna za poslední slovo překryté hlukem.
- Jiná narušení plynulosti projevu jsou označována `<UH>`, `<UM>`, `<UH-HUH>`, `<UH-HUM>`.
- Pauzy a delší přerušení jsou označeny jako `<silence>`.

V tab. 4.1 je ukázka transkribovaného textu.

Tabulka 4.1. Ukázka transkripce

<t 26.800> <<spk2, f>>
<mouth><inhale> to vám neřeknu data já si absolutně nepamatuju
<t 31.747> <<spk1, f + spk2, f>>
SPEAKER1: aspoň roční období
SPEAKER2: <mouth><inhale>
<t 33.372> <<spk2, f>>
roční tož to mohlo být v třiaštyrc- dvaaštyrycet už třiaštyrycátém roce
<b 40.838>
<noise begin> protože to byl čas vždycky ten odstup <inhale><noise end>
<b 45.525>
<inhale> jak ty chlapy odvedly tak sme zůstali jenom s maminkama
<b 53.172>
<inhale> v ty [Modělevi] já sem <inhale> utíkala z teho <noise> lágru

4.2. Anotační nástroj MEd

Pro anotaci rekonstrukce standardizovaného textu (pro převod vstupní transkripce zachycené na w-rovině na standardizovaný text na m-rovině) byla vytvořena první pracovní verze softwarového nástroje nazvaného **MEd**. Tento nástroj umožňuje:

- přesouvat libovolně slovní jednotky na m-rovině z hlediska jejich pořadí ve větě;
- rozdělit nevhodně segmentovaný proud řeči (a původní transkripci) na segmenty vhodné pro strukturální syntaktickou anotaci;
- slovní jednotky vymazat, vložit, spojit, jinak modifikovat, a to včetně jejich atributů (forma, lema, morfologická značka);
- propojit slovní jednotky na m-rovině a w-rovině tak, aby bylo zřejmé, se kterými slovními jednotkami původní transkripce (tj. slovními jednotkami w-roviny) konkrétní slovní jednotka m-roviny souvisí (ze kterých „vznikla“), a určit typ propojení;
- použít připravená (předzpracovaná) data tak, aby anotace byla i časově efektivní (tj. zamezit rutinním úkonům, které lze udělat automaticky před anotací a nebo v jejím průběhu);
- poslech původní audionahrávky, který je často nutný v případech, kdy ani původní transkripce (například vzhledem k absenci prozodické informace, informace o délce pauz a vzhledem k další „ztrátě informace“), ani její kontext neumožňuje anotátorovi rozhodnout o vhodné modifikaci.

Literatura

Bies, A.; Feguson, M.; Katz, K.; MacIntyre, R. (1995): Bracketing Guidelines for Treebank II Style. Penn Treebank Project.

Čmejrková, S.; Jílková, L.; Kaderka, P. (2004): Mluvená čeština v televizních debatách: korpus Dialog. Slovo a slovesnost, 65.

Fitzgerald, E. (2006): Speech Reconstruction Annotation Guide for Conversational Telephone Speech Conversations. Draft.

Gibbon, D.; Moore, R.; Winski, R. (eds.) (1997): Handbook of Standards and Resources for Spoken Language Systems. Mouton de Gruyter; Bk&CD-Rom edition, Berlin. ISBN: 3110153661.

Godfrey, J., Holliman, E. (1997): Switchboard-1 Release 2. Sada CDROM. LDC Catalog No. LDC97S62. Linguistic Data Consortium, Philadelphia, PA, USA.

Godfrey, J.; Holliman, E.; McDaniel, J. (1992): SWITCHBOARD: Telephone speech corpus for research and development. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Graff, D.; Bird, S. (2000): Many uses, many annotations for large speech corpora: Switchboard and TDT as case studies. Proceedings of the Second International Conference on Language Resources and Evaluation, pp 427-433.

Hajič, J. (2004): Disambiguation of Rich Inflection. Karolinum, Charles University Press, Prague.

Hajič, J. et al. (2006): Prague Dependency Treebank 2.0. CD ROM. CAT: LDC2006T01, 1-58563-370-4. Linguistic Data Consortium, Univ. of Pennsylvania, Philadelphia, USA.

Hoekstra, H.; Moortgat, M.; Renmans, B.; Schouppe, M.; Schuurman, I.; van der Wouden, T. (2003): CGN Syntactische annotatie.

Hoekstra, H.; Moortgat, M.; Schuurman, I.; van der Wouden, T. (2001): Syntactic Annotation for the Spoken Dutch Corpus Project. In Daelemans, W.; Sima'an, K.; Veenstra. J.; Zavrel, J. (eds.): Computational Linguistics in the Netherlands 2000. Amsterdam/New York, Rodopi, pp. 73-87.

Johannessen, J. B.; Jørgensen, F. (2005): Annotation of spoken language data. Paper read at NODALIDA 2005, Joensuu. URL: http://www.hf.uio.no/tekstlab/treebank_workshop/program.htm

Kawata, Y.; Barteles, J. (2000): Stylebook for Japanese Treebank in Verbmobil. Technical Report 240, Verbmobil, Eberhard-Karls-Universität, Tübingen.

Kordoni, V. (2000): Stylebook for the English Treebank in Verbmobil. Technical Report 241, Verbmobil. Eberhard-Karls-Universität, Tübingen.

Marcus, M.P.; Santorini, B.; Marcinkiewicz, M. A. (1993): Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics, vol. 19, pp. 313-330.

Marcus, M. P.; Kim, G.; Marcinkiewicz, M. A.; MacIntyre, R.; Bies, A.; Ferguson, M.; Katz, K.;Schasberger, B. (1994): The Penn Treebank: Annotating Predicate Argument Structure. In Proceedings of the ARPA Human Language Technology Workshop. Princeton.

Meteer et al. (1995): Dysfluency Annotation Stylebook for the Switchboard Corpus. University of Pennsylvania, Department of Computer and Information Science. [ftp.cis.upenn.edu/pub/treebank/swbd/doc/DFL-book.ps].

Mikulová et al. (2005): Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka. Technická zpráva ÚFAL TR-2005-28. MFF UK, Praha.

Nebeská, I. (1983): Kvantitativní charakteristiky souvětí v psaných a mluvených odborných projevech. In Těšitelová, M. (ed.). Psaná a mluvená odborná čeština z kvantitativního hlediska. Linguistica IV. Praha, s. 99-120.

Panevová, J. (1980): Formy a funkce ve stavbě české věty. Academia, Praha.

Psutka, J.; Ircing, P.; Psutka, J. V.; Radová, V.; Byrne, W.; Hajič, J.; Mírovský, J.; Gustman, S. (2003): Large Vocabulary ASR for Spontaneous Czech in the MALACH Project. In EUROSPEECH 2003 Proceedings (8th European Conference on Speech Communication and Technology), pp. 1821-1824. ISCA.

Psutka, J.; Ircing, P.; Psutka, J. V.; Radová, V.; Byrne, W.; Hajič, J.; Gustman, S; Ramabhadran, B.; (2002): Automatic Transcription of Czech Language Oral History in the MALACH Project: Resources and Initial Experiments. In Text, Speech and Dialogue. 5th International Conference, TSD 2002, pp. 253-260. Springer.

Radová, V. (2002): Pokyny pro zpracování nahrávek Holocaustu pomocí programu Transcriber. Interní dokument. KKY ZČU, Plzeň.

Radová, V. et al. (2004): Czech Broadcast News Speech and Transcripts. Sada CDROM. LDC Catalog No. LDC2004S01 a LDC2004T01. Linguistic Data Consortium, Philadelphia, PA, USA.

Sagae, K.; MacWhinney, B.; Lavie, A. (2004): Adding syntactic annotations to transcripts of parentchild dialogs. In Proceedings of the 4th Conference on Language Resources and Evaluation, Lisbon, Portugal, pp. 1815-1818.

Santorini, B. (1990): Part-of-speech Tagging Guidelines for Penn Treebank Poject.

Sgall, P. (1967): Generativní popis jazyka a česká deklinace. Academia, Praha.

Sgall, P. a kol. (1986): Úvod do syntaxe a sémantiky. Praha.

Sgall, P.; Hajičová, E.; Buráňová, E. (1980): Aktuální členění věty v češtině. Academia, Praha.

Sgall, P.; Hajičová, E.; Panevová, J. (1986): The Meaning of the Sentence in Its Semantic and Pragmatic Aspects. Prague, Academia a Dordrecht, Reidel.

Schuurman, I.; Goedertier, W.; Hoekstra, H.; Oostdijk, N.; Piepenbrock, R.; Schouppe, M. (2004): Linguistic annotation of the Spoken Dutch Corpus: If we had to do it all over again... In Proceedings of the 4th Conference on Language Resources and Evaluation. Lisbon, Portugal, pp 57-60.

Schuurman, I.; Schouppe, M.; van der Wouden, T.; Hoekstra, H. (2003): CGN, an annotated corpus of Spoken Dutch. In Abeillé, A.; Hansen-Schirra, S.; Uszkoreit, H.: Proceedings of 4th International Workshop on Linguistically Interpreted Corpora (LINC-03). Budapest 2003, pp. 101-108.

Stegmann, R.; Telljohann, H.; Hinrichs, E. W. (2000): Stylebook for German Treebank in Verbmobil. Technical Report 239, Verbmobil, Eberhard-Karls-Universität, Tübingen.

Použité odkazy:

http://www.isca-speech.org/index.html
http://www.interspeech2007.org/
http://www.ilc.cnr.it/EAGLES96/home.html
http://www.phon.ucl.ac.uk/resource/sfs/
http://www.ldc.upenn.edu/sb/isle.html
http://mate.nis.sdu.dk/
http://www.fon.hum.uva.nl/praat/
http://nite.nis.sdu.dk/
http://www.ldc.upenn.edu/Catalog/byType.jsp#speech
http://www.cis.upenn.edu/~treebank/switchboard-sample.html
http://www.ucl.ac.uk/english-usage/ice/index.htm
http://childes.psy.cmu.edu/data/
http://lands.let.kun.nl/cgn/ehome.htm
http://www.sfs.uni-tuebingen.de [http://www.sfs.uni-tuebingen.de/en_tuebads.shtml]

http://w3.msi.vxu.se/~nivre/research/st.html
http://www.natcorp.ox.ac.uk/
http://www2.warwick.ac.uk/fac/soc/celte/research/base/
http://www.uni-saarland.de/fak4/norrick/scose.htm
http://khnt.hit.uib.no/icame/manuals/LONDLUND/INDEX.HTM
http://www.hf.uib.no/i/Engelsk/COLT/
http://www.ul.ie/~lcie/
http://www.titania.bham.ac.uk/docs/svenguide.html
http://www.rdg.ac.uk/AcaDepts/ll/speechlab/marsec/
http://www.comp.leeds.ac.uk/amalgam/tagsets/sec.html
http://www.cs.rochester.edu/research/cisd/resources/trains.html
http://projects.ldc.upenn.edu/SBCSAE/
http://www.athel.com/cspatg.html
http://ssli.ee.washington.edu/projects/radio.html
http://www.lsa.umich.edu/eli/micase/index.htm
http://andosl.anu.edu.au/andosl
http://www.vuw.ac.nz/lals/corpora/index.aspx
http://www.corpus.bham.ac.uk/PCLC/cl-195-pap-COLA.doc
http://www.env.kitakyu-u.ac.jp/corpus/docs/index.html
http://www.coelang.tufs.ac.jp/english/language_function.html
http://www.ling.gu.se/projekt/nordtalk/members_resources/BySoc.html
http://www.cphling.dk/~ng/danpass_webpage/danpass.htm
http://www.tekstlab.uio.no/nota/english/index.html
http://www.tekstlab.uio.no/talespraak/bigbrother/
http://www.ling.gu.se/projekt/tal/index.cgi?PAGE=3
http://www.cl.ut.ee/suuline/Korpus.php
http://www.hf.uio.no/ilos/studier/studenttjenester/Nettressurser/bulg/mat/Aleksova/
http://helmer.hit.uib.no/batmult/Janas_Final_Report.htm
http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/annotate.html
http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/
http://hpsg.stanford.edu/
http://nextens.uvt.nl/~conll/
http://ucnk.ff.cuni.cz/pmk_bonito.html
http://ucnk.ff.cuni.cz/bmk.html
http://malach.umiacs.umd.edu/
http://ufal.mff.cuni.cz/pdt2.0
http://ufal.mff.cuni.cz/pdt/pml/schema/
http://www.vhf.org
http://trans.sourceforge.net [http://trans.sourceforge.net/en/presentation.php]

The ÚFAL/CKL Technical Report Series

ÚFAL

ÚFAL (Ústav formální a aplikované lingvistiky; http://ufal.mff.cuni.cz) is the Institute of Formal and Applied linguistics, at the Faculty of Mathematics and Physics of Charles University, Prague, Czech Republic. The Institute was established in 1990 after the political changes as a continuation of the research work and teaching carried out by the former Laboratory of Algebraic Linguistics since the early 60s at the Faculty of Philosophy and later the Faculty of Mathematics and Physics. Together with the "sister" Institute of Theoretical and Computational Linguistics (Faculty of Arts) we aim at the development of teaching programs and research in the domain of theoretical and computational linguistics at the respective Faculties, collaborating closely with other departments such as the Institute of the Czech National Corpus at the Faculty of Philosophy and the Department of Computer Science at the Faculty of Mathematics and Physics.

CKL

As of 1 June 2000 the Center for Computational Linguistics (Centrum komputační lingvistiky; http://ckl.mff.cuni.cz) was established as one of the centers of excellence within the governmental program for support of research in the Czech Republic. The center is attached to the Faculty of Mathematics and Physics of Charles University in Prague.

Technical Reports

The ÚFAL/CKL technical report series has been established with the aim of disseminate topical results of research currently pursued by members, cooperators, or visitors of the Institute. The technical reports published in this Series are results of the research carried out in the research projects supported by the Grant Agency of the Czech Republic, GAČR 405/96/K214 ("Komplexní program"), GAČR 405/96/0198 (Treebank project), grant of the Ministry of Education of the Czech Republic VS 96151, and project of the Ministry of Education of the Czech Republic LN00A063 (Center for Computational Linguistics). Since November 1996, the following reports have been published.

ÚFAL TR-1996-01 Eva Hajičová, *The Past and Present of Computational Linguistics at Charles University*
Jan Hajič and Barbora Hladká, *Probabilistic and Rule-Based Tagging of an Inflective Language – A Comparison*

ÚFAL TR-1997-02 Vladislav Kuboň, Tomáš Holan and Martin Plátek, *A Grammar-Checker for Czech*

ÚFAL TR-1997-03 Alla Bémová at al., Anotace na analytické rovině, *Návod pro anotátory (in Czech)*

ÚFAL TR-1997-04 Jan Hajič and Barbora Hladká, *Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structural Tagset*

ÚFAL TR-1998-05 Geert-Jan M. Kruijff, *Basic Dependency-Based Logical Grammar*

ÚFAL TR-1999-06 Vladislav Kuboň, *A Robust Parser for Czech*

ÚFAL TR-1999-07 Eva Hajičová, Jarmila Panevová and Petr Sgall, *Manuál pro tektogramatické značkování (in Czech)*

ÚFAL TR-2000-08 Tomáš Holan, Vladislav Kuboň, Karel Oliva, Martin Plátek, *On Complexity of Word Order*

ÚFAL/CKL TR-2000-09 Eva Hajičová, Jarmila Panevová and Petr Sgall, *A Manual for Tectogrammatical Tagging of the Prague Dependency Treebank*

ÚFAL/CKL TR-2001-10 Zdeněk Žabokrtský, *Automatic Functor Assignment in the Prague Dependency Treebank*

ÚFAL/CKL TR-2001-11 Markéta Straňáková, *Homonymie předložkových skupin v češtině a možnost jejich automatického zpracování*

ÚFAL/CKL TR-2001-12 Eva Hajičová, Jarmila Panevová and Petr Sgall, *Manuál pro tektogramatické značkování (III. verze)*

ÚFAL/CKL TR-2002-13 Pavel Pecina and Martin Holub, *Sémanticky signifikantní kolokace*

ÚFAL/CKL TR-2002-14 Jiří Hana, Hana Hanová, *Manual for Morphological Annotation*

ÚFAL/CKL TR-2002-15 Markéta Lopatková, Zdeněk Žabokrtský, Karolína Skwarská and Vendula Benešová, *Tektogramaticky anotovaný valenční slovník českých sloves*

ÚFAL/CKL TR-2002-16 Radu Gramatovici and Martin Plátek, *D-trivial Dependency Grammars with Global Word-Order Restrictions*

ÚFAL/CKL TR-2003-17 Pavel Květoň, *Language for Grammatical Rules*

ÚFAL/CKL TR-2003-18 Markéta Lopatková, Zdeněk Žabokrtský, Karolina Skwarska, Václava Benešová, *Valency Lexicon of Czech Verbs VALLEX 1.0*

ÚFAL/CKL TR-2003-19 Lucie Kučová, Veronika Kolářová, Zdeněk Žabokrtský, Petr Pajas, Oliver Čulo, *Anotování koreference v Pražském závislostním korpusu*

ÚFAL/CKL TR-2003-20 Kateřina Veselá, Jiří Havelka, *Anotování aktuálního členění věty v Pražském závislostním korpusu*

ÚFAL/CKL TR-2004-21 Silvie Cinková, *Manuál pro tektogramatickou anotaci angličtiny*

ÚFAL/CKL TR-2004-22 Daniel Zeman, *Neprojektivity v Pražském závislostním korpusu (PDT)*

ÚFAL/CKL TR-2004-23 Jan Hajič a kol., *Anotace na analytické rovině, návod pro anotátory*

ÚFAL/CKL TR-2004-24 Jan Hajič, Zdeňka Urešová, Alevtina Bémová, Marie Kaplanová, *Anotace na tektogramatické rovině (úroveň 3)*

ÚFAL/CKL TR-2004-25 Jan Hajič, Zdeňka Urešová, Alevtina Bémová, Marie Kaplanová, *The Prague Dependency Treebank, Annotation on tectogrammatical level*

ÚFAL/CKL TR-2004-26 Martin Holub, Jiří Diviš, Jan Pávek, Pavel Pecina, Jiří Semecký, *Topics of Texts. Annotation, Automatic Searching and Indexing*

ÚFAL/CKL TR-2005-27 Jiří Hana, Daniel Zeman, *Manual for Morphological Annotation (Revision for PDT 2.0)*

ÚFAL/CKL TR-2005-28 Marie Mikulová a kol., *Pražský závislostní korpus (The Prague Dependency Treebank) Anotace na tektogramatické rovině (úroveň 3)*

ÚFAL/CKL TR-2005-29 Petr Pajas, Jan Štěpánek, *A Generic XML-Based Format for Structured Linguistic Annotation and Its application to the Prague Dependency Treebank 2.0*

ÚFAL/CKL TR-2006-30 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Annotation on the tectogrammatical level in the Prague Dependency Treebank (Annotation manual)*

ÚFAL/CKL TR-2006-31 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Petr Sgall, Magda Ševčíková, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Anotace na tektogramatické rovině Pražského závislostního korpusu (Referenční příručka)*

ÚFAL/CKL TR-2006-32 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Petr Sgall,Magda Ševčíková, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Annotation on the tectogrammatical level in the Prague Dependency Treebank (Reference book)*

ÚFAL/CKL TR-2006-33 Jan Hajič, Marie Mikulová, Martina Otradovcová, Petr Pajas, Petr Podveský, Zdeňka Urešová, *Pražský závislostní korpus mluvené češtiny. Rekonstrukce standardizovaného textu z mluvené řeči*