Rekonstrukce standardizovaného textu z mluvené řeči

Marie Mikulová, Zdeňka Urešová

Příspěvek shrnuje dosavadní poznatky získané při budování Pražského závislostního korpusu mluvené češtiny (Prague Dependency Treebank of Spoken Czech; dále PDTSC) v Ústavu formální a aplikované lingvistiky MFF UK. PDTSC bude prvním korpusem mluvené řeči, který nabídne i syntakticko-sémantickou anotaci promluv. Východiskem projektu PDTSC je syntakticko-sémantická anotace korpusu psaných textů, která již byla zpracována v projektu Pražského závislostního korpusu 2.0 (dále PDT).

Lingvistická analýza mluvené řeči a první pokusy se syntakticko-sémantickou anotací ukázaly, že vzhledem ke specifičnosti mluvených projevů není možné anotovat mluvenou řeč přímo podle pravidel anotace psaných textů, a tato pravidla případně pouze upravovat a rozšiřovat pro zvláštnosti mluvené řeči. Byl proto zvolen zcela nový postup, kdy se před vlastní syntakticko-sémantickou anotací jednotlivé segmenty mluvené řeči nejprve převádí na tzv. standardizovaný text, tj. provádí se tzv. rekonstrukce standardizovaného textu z mluvené řeči.

V tomto článku blíže vysvětlujeme v počítačové lingvistice nový pojem standardizace mluvené řeči. Přinášíme vlastní vymezení standardizovaného textu a popis základních principů a pravidel naším pracovištěm navržené rekonstrukce standardizovaného textu z mluvené řeči. Tyto principy a pravidla stanovují, jakým způsobem se z mluvené řeči vytváří syntakticko-sémanticky anotovatelný standardizovaný text, aniž by se přitom ztratily původní významy vyřčených segmentů a zároveň aby zůstaly zachovány vazby mezi původní transkripcí mluvené řeči a rekonstruovaným standardizovaným textem.

1. Syntakticko-sémanticky anotovaný korpus mluvené řeči

1.1. Existující syntakticky anotované korpusy mluvené řeči

Výzkum mluvené řeči se v oblasti počítačového zpracování jazyka na rozdíl od výzkumu psaného textu soustřeďuje většinou jen na přepis akustického signálu do textové podoby. Rozsah lingvistické anotace těchto transkripcí je nevelký, anotace se obvykle zastavuje na morfologické rovině, popřípadě se přidává prozodické značkování. Další zpracování směrem k významu v korpusech nenajdeme. 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 syntakticko-sémantické).

Syntaktická anotace v několika málo existujících syntakticky anotovaných mluvených korpusech byla většinou provedena automaticky s využitím nástrojů natrénovaných na korpusech psaných textů nebo na manuálně anotované části mluveného korpusu (například International Corpus of English, korpus Switchboard, Childes database). Poloautomaticky je anotován Korpus mluvené holandštiny a korpusy projektu Verbmobil (Tübingenské korpusy mluvené němčiny, angličtiny a japonštiny). Ve všech zmíněných případech se jedná o korpusy, ve kterých je mluvená řeč anotována na úrovni povrchové syntaxe (nikoli na úrovni syntakticko-sémantické).

1.2 Syntakticko-sémantická anotace mluvené řeči

Syntakticko-sémantická anotace je strukturní anotace, která jednoznačně zachycuje jazykový význam napsané/vyřčené věty, popřípadě jeden z jejích významů (je-li věta víceznačná). Syntakticko-sémantický zápis věty obsahuje veškerou informaci o zapsaném/vyřčeném významu, kterou gramatická stavba věty a její lexikální obsazení dává a která je nutná pro převod tohoto zápisu na zápis nižších rovin (až do původní napsané/vyřčené podoby).

Budování syntakticko-sémanticky anotovaného korpusu mluvené řeči s sebou však přináší řadu otázek, které se při budování korpusu psaných textů neobjeví (nebo jejich řešení je nasnadě). Jde především o problém segmentace mluvené řeči do vět a způsob zachycení specifických jevů mluvené řeči (otázka zachycení významu segmentů mluvené řeči vůbec). Řešení je třeba vždy hledat se zřetelem k účelu, ke kterému je korpus určen. Jinak se při syntaktické anotaci přistoupí k jevům mluvené řeči, pokud bude cílem korpusu zachytit specifickou strukturu mluvené řeči, například pro lingvistický či psychologický výzkum (jako v projektu Childes). Budujeme-li korpus vhodný pro pravděpodobnostní trénování a strojové učení za účelem plného porozumění mluvené řeči (to je účel PDTSC i ostatních syntaktických korpusů zmíněných v 1.1), je otázka, do jaké míry je nutné při syntakticko-sémantické anotaci specifické jevy mluvené řeči zohledňovat.

Podle J. B. Johannessenové a F. Jørgensena (2005) jsou v zásadě tři možnosti, jak naložit se specifickými jevy mluvené řeči při syntaktické anotaci:

A) zohlednit všechny jevy mluvené řeči,
B) zohlednit jen některé vybrané jevy mluvené řeči a ostatní ignorovat,
C) ignorovat všechny specifické 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 při budování PDTSC bylo původním záměrem syntakticko-sémanticky anotovat mluvenou řeč přímo, podle pravidel pro anotaci psaných textů Pražského závislostního korpusu 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. Uskutečnily se též první pokusy se syntaktickou anotací mluvené řeči. Ukázalo se, že právě vzhledem ke specifičnosti mluvených projevů je tento původní záměr zřejmě neschůdný.

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; příklady lze nalézt v části Českého akademického korpusu) přes odchylky viditelné a místy značné (korpus Dialog, televizní politické debaty, korpus svědectví o holocaustu Malach) až po odchylky zcela zásadní (části Pražského mluveného korpusu, volně dostupné zdroje). 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. Musela by tudíž být doslova vymýšlena „gramatická“ (spíše však „negramatická“) pravidla syntaktické anotace na všechno, co lze při věrném přepisu toho, co lidé říkají, očekávat – včetně výplňkových slov, opakování, zakoktání, nových začátků, nedořečených klauzí, anakolutů i tzv. neřečových událostí, jako je zakašlání, smích apod.

Původní myšlenka úpravy specifikace syntakticko-sémantické reprezentace pro mluvenou řeč při zachování stylu a všech zásad anotace byla proto opuštěna jako nereálná a byl stanoven nový postup: před vlastní syntakticko-sémantickou anotací se segmenty mluvené řeči nejprve převedou na gramaticky správné věty, tj. provede se tzv. rekonstrukce standardizovaného textu z mluvené řeči. Převod mluvené řeči na standardizovaný text umožní, že při syntakticko-sémantické anotaci budeme moci pracovat s obdobnými pravidly a značkami jako při anotaci psaného textu.

Klíčové při tomto postupu je ovšem zachování (anotování) vazeb mezi vstupní transkripcí mluvené řeči a rekonstruovaným standardizovaným textem; jinými slovy: při syntakticko-sémantické anotaci sice pracujeme s textem, který je již zbavený specifických jevů mluvené řeči, ale pamatujeme si, jak jsme tento text z původní transkripce mluvené řeči získali, a jsme schopni jej na základě anotovaných vazeb zpětně zkonstruovat. Základní princip rekonstrukce, princip zachování významu, pak zaručuje, že z původního významu vyřčeného segmentu nebylo rekonstrukcí nic ztraceno. Princip zachování významu znamená, že nově rekonstruovaná standardizovaná podoba věty a původní vyřčený segment mají stejný význam. Je založen na myšlence, že tentýž význam (zapsatelný tímtéž syntakticko-sémantickým zápisem) lze prezentovat za použití pravidel pro psaný text v psané podobě, nebo za použití jiných prostředků v podobě mluvené. Specifické jevy mluvené řeči (přeřeknutí, zakoktání, nadměrné množství elips atp.) na straně jedné a specifické jevy psaného textu (členění do odstavců, velká a malá písmena atp.) jsou jen důsledkem zvolené prezentace významu.

Pro syntakticko-sémantickou anotaci mluvené řeči tudíž preferujeme výše uvedenou možnost C, kdy se při syntaktické anotaci již nepracuje se specifickými jevy mluvené řeči.

1.3 Standardizace mluvené řeči

Tzv. standardizace mluvené řeči je i z celosvětového hlediska nový směr výzkumu. Ten vychází z názoru, že současné běžné metody automatického rozpoznávání řeči se staly „obětí vlastního úspěchu“. Tyto metody 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 nám paradoxně přidělávají práci v případě, že danou promluvu chceme dále zpracovávat metodami automatické analýzy textu. Je totiž lhostejné, že systém rozpoznávání řeči precizně určí všechna přeřeknutí, výplňková slova, slova ve špatném tvaru apod., pokud udělá chybu ve slově klíčovém pro pochopení významu věty.

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 jsou používány statistické (a tedy velmi robustní) metody, na neúplném, nevhodném, nebo gramaticky a lexikálně velmi „nesprávném“ textu nedávají dobré výsledky. Stále více se tudíž ukazuje, že dosavadní způsoby doslovné transkripce pomocí systému automatického rozpoznávání nejsou pro další zpracování textu vhodné. Jiné možnosti ovšem prakticky neexistují, a tedy nejsou k dispozici ani z literatury (a to ani z nejnovějších sborníků konferencí). Přitom je zřejmé, že situace dozrála pro změnu „paradigmatu“ výzkumu na pomezí mluveného a psaného jazyka.

Standardizace mluvené řeči znamená, že nad doslovně přepsaným záznamem mluvené řeči se před další (syntaktickou) anotací provede „přípravná“ anotace, která se nějakým způsobem vypořádá se specifickými jevy mluvené řeči.

První pokusy se standardizací byly provedeny v Computer and Information Science Department na University of Pennsylvania na datech korpusu Switchboard. Tato tzv. „dysfluency annotation“ (M. Meteer et al., 1995) 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 atd. Segmenty mluvené řeči se však dále neupravují, a v drtivé většině případů tak zůstávají pouze fragmenty s neúplnou syntaktickou strukturou, kterou není možné plně syntakticko-sémanticky analyzovat, přestože význam segmentu je z toho, co bylo vyřčeno, patrný.

Rekonstrukce ve vlastním slova smyslu rovněž na datech korpusu Switchboard se provádí v Center for Speech and Langauge Processing na Johns Hopkins University v Baltimore (E. Fitzgerald, 2006). Jde zřejmě o první projekt svého druhu. Námi navrhovaná pravidla rekonstrukce i způsob jejího zachycení ze systému pravidel navrženého na Johns Hopkins University vycházejí a v řadě bodů jej přejímají. 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 v našem Ústavu formální a aplikované lingvistiky MFF UK Praha.

2. Principy 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 pouze zachovat 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 2.1), věty různě upravuje, některá slova maže, jiná přidává, mění jejich pořadí (viz 2.2). Rozhovor těmito modifikacemi dostává psanou podobu (tj. podobu, která 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,
- všechny 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 následující m-rovině. Provedené modifikace budou zachyceny systémem pojmenovaných odkazů mezi m-uzly a w-uzly (více k systému rovin v PDTSC viz 3).

Příklad vstupního segmentu mluvené řeči a jeho standardizované podoby (na prvním řádku je uveden původní transkribovaný text, za šipkou na druhém řádku je rekonstruovaná standardizovaná podoba):
no tak já s- chtěl sem jenom říct jak- jaký byli mezi náma hrdinové že
→ Chtěl jsem jenom říci, jací byli mezi námi hrdinové.

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 (např. Radová V., 2002) - proud mluvené řeči je členěn na základě pauz tak, aby výsledné úseky opět „zhruba odpovídaly větám“. 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.

Tento princip („longest match“) jsme převzali z pravidel syntaktické anotace mluvené řeči pro korpusy projektu Verbmobil (např. Kordoni V., 2000). 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.

Příklad použití principu nejdelší možné klauze:
<silence><inhale> někteří lidé mě <noise> utkvěli <inhale> velmi v paměti <silence> z toho koncentračního tábora <silence>
→ Někteří lidé z koncentračního tábora mně velmi utkvěli v paměti.

2.2 Úpravy segmentů mluvené řeči

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

A) ortografické modifikace,
B) vlastní modifikace.

Ortografické modifikace představují pravidelné úpravy vstupního textu vyplývající ze základních podmínek na standardizovaný text, totiž že standardizovaný text splňuje obecné charakteristiky psaného textu a jsou v něm dodržena pravidla českého pravopisu. K ortografickým modifikacím patří jednak odstranění neřečových událostí a jednak vlastní pravopisné úpravy.

Výsledný standardizovaný text neobsahuje žádné značky pro neřečové události, jako je smích, zakašlání, nádechy, pauzy. Pokud je neřečovou událostí vyjadřován nějaký význam (souhlas, protest), je zaznamenán prostředky psaného textu (slovem, grafickými symboly - například vykřičníkem).

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

→ Tak možná, že bych ještě něco řekl.

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.

Nejdůležitější částí anotace jsou tzv. **vlastní modifikace** vstupního transkribovaného textu. Ty 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.

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 (tedy výplňková slova a i celé fráze, nadbytečná deiktická slova, nadbytečné konektory, nadbytečná a nesprávně užitá gramatická slova, restarty, opakující se úseky textu, fragmenty) jsou při rekonstrukci standardizovaného textu ze vstupní transkripce odstraňovány.

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

→ V Bratislavě jsme dostávali podporu asi deset korun denně.

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, ale i nevyjádřená plnovýznamová slova.

<silence> <inhale> revolverem mu takle začali před nos <inhale> jo <silence>

→ Revolverem mu takhle začali dělat před nosem.

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 jsou měněna i celá slova.

architekt zelenka má velikou zálohu o tuto činnost

→ Architekt Zelenka má velikou zásluhu na této činnosti.

Rekonstruované věty mají gramatický slovosled, který nenarušuje plynulost textu (včetně aktuálního členění textu), dochází tedy i ke změnám ve slovosledném uspořádání.

po pěti sme leželi

→ Leželi jsme po pěti.

Všechny modifikace jsou vždy prováděny se zřetelem ke kontextu celého textu a za přísného dodržování obou výše uvedených principů rekonstrukce.

3. Pražský závislostní korpus mluvené češtiny (PDTSC)

PDTSC bude mít strukturu analogickou korpusu PDT 2.0: 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, označovaná jako z-rovina, 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 definované (oproti PDT 2.0 rozšířené) budou 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 a na w-rovině.

Na **z-rovině** bude v PDTSC zachován přepis mluvené řeči v původním znění, které je výstupem automatického rozpoznávače. Ze z-roviny povedou odkazy do externích souborů obsahujících digitalizované audionahrávky. **W-rovina** bude v PDTSC též reprezentovat věrný přepis mluvené řeči (se všemi jejími zvláštnostmi - přeřeky, opakování slov, neřečové události aj.), ale půjde již o přepis manuálně upravený anotátorem podle přesně nadefinovaných pravidel pro manuální transkripci řeči. W-rovina bude představovat transkripci mluvené řeči zbavenou automatických transkripčních omylů. Může tedy sloužit i jako trénovací data pro současné technologie používané v systémech rozpoznávání řeči.

Transkribovaný proud mluvené řeči bude na **m-rovině** (morfologické rovině) nahrazen standardizovaným textem; tj. mezi w-rovinou a m-rovinou bude provedena výše popsaná rekonstrukce standardizovaného textu z mluvené řeči.

Standardizované věty mluvené řeči budou na syntaktických rovinách (na rovině povrchové syntaxe (**a-rovině**) a na rovině syntakticko-sémantické (**t-rovině**)) anotovány podle obdobných pravidel jako psané texty v korpusu PDT 2.0. Předpokládá se, že stávající pravidla anotace budou měněna minimálně.

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. Ve vztazích mezi m-uzly a w-uzly (v systému odkazů mezi jednotlivými jednotkami obou rovin) budou zachyceny mnohačetné rozdíly mezi rekonstruovaným standardizovaným textem a vstupní transkripcí mluvené řeči.

Acknowledgement
Příspěvek vznikl za finanční podpory projektů: LC 536, ME 838 a GA 405/06/0589.
This work was funded in part by the Companions project (www.companions-project.org) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-034434.

LITERATURA

Childes database. Carnegie Mellon University, Pittsburgh.
World Wide Web: http://childes.psy.cmu.edu/

Český akademický korpus 1.0. ÚFAL MFF UK Praha.
World Wide Web: http://ufal.mff.cuni.cz/rest/CAC/doc/cac-guide/cz/html/index.html

Fitzgerald E., 2006, *Speech Reconstruction Annotation Guide for Conversational Telephone Speech Conversations*. Center for Speech and Langauge Processing, Johns Hopkins University, Baltimore. Interní dokument.

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

Hajič J., Mikulová M., Otradovcová M., Pajas P., Podveský P., Urešová Z., 2006, *Pražský závislostní korpus mluvené češtiny. Rekonstrukce standardizovaného textu z mluvené řeči*. Technická zpráva ÚFAL TR-2006-33, MFF UK, Praha.

International Corpus of English. Department of English Language & Literature, University College London.
World Wide Web: http://www.ucl.ac.uk/english-usage/ice/index.htm

Johannessen J. B., Jørgensen F., 2005, *Annotation of spoken language data*. Paper read at NODALIDA, Joensuu.

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.

Korpus Dialog. Ústav pro jazyk český AV ČR.
World Wide Web: http://www.ujc.cas.cz/oddeleni/index.php?page=DIALOG

Korpus mluvené holandštiny. University of Leuven, University of Ghent, University of Utrecht, University of Nijmegen.
World Wide Web: http://lands.let.kun.nl/cgn/ehome.htm

Meteer M. et al., 1995, *Dysfluency Annotation Stylebook for the Switchboard Corpus*. Department of Computer and Information Science, University of Pennsylvania, Philadelphia.

Mikulová M. et al., 2005, *Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka*. Technická zpráva ÚFAL TR-2005-28, MFF UK, Praha.
World Wide Web: http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/

Pražský mluvený korpus. Ústav Českého národního korpusu FF UK Praha.
World Wide Web: http://ucnk.ff.cuni.cz/pmk_bonito.html

Pražský závislostní korpus 2.0. ÚFAL MFF UK Praha.
World Wide Web: http://ufal.mff.cuni.cz/pdt2.0

Psutka J. et al., 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, Springer, pp. 253-260.

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

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

Switchboard. University of Pennsylvania, Philadelphia.
World Wide Web: http://www.ldc.upenn.edu/Catalog/docs/switchboard/

Verbmobil. Eberhard-Karls-Universität, Tübingen.
World Wide Web: <http://www.sfs.uni-tuebingen.de/en_tuebads.shtml>;
<http://www.sfs.uni-tuebingen.de/en_tuebaes.shtml>;
<http://www.sfs.uni-tuebingen.de/en_tuebajs.shtml>.