Sborník konference Čeština v mluveném korpusu (ČMK2007), Praha, 2007 (vyjde 2008)

Anotace strukturálních metadat ve spontánní mluvené češtině

Jáchym Kolář

Západočeská univerzita, Fakulta aplikovaných věd, Univerzitní 8, 306 14 Plzeň
jachym@kky.zcu.cz

1. Úvod

Spontánní mluvená řeč, jako protipól řeči čtené nebo předem připravené, je nejpřirozenější formou lidské komunikace. Spontánní produkce řeči ale také představuje složitý mentální proces, při kterém se mluvčí obvykle nejsou schopni vyhnout značnému množství různých selhání, která narušují plynulost a znepřehledňují syntaktickou a prozodickou stavbu jejich promluvy. Proto je automatické zpracování spontánní mluvené řeči, které je v současné době jedním ze zásadních problémů řešených v oblasti automatického zpracování přirozeného jazyka (Natural Language Processing, NLP), velice obtížné.

Pro odhad parametrů automatických systémů je třeba mít k dispozici dostatek dat ve formě anotovaných řečových korpusů. Při vytváření spontánních řečových databází však není možné využít běžné postupy navržené pro řeč čtenou. Tyto standardní anotační techniky nám totiž neumožňují adekvátně zachytit strukturu spontánního mluveného projevu. Její zachycení je však velmi důležité, protože holé řetězce slov nenesou kompletní informaci, kterou se řečník snaží posluchači předat. Strukturální informace za úrovní slov (tzv. strukturální metadata) jsou pro pochopení vyřčeného sdělení neméně důležitá.

V databázích čtené řeči se strukturální informace typicky zachycuje ve formě zjednodušené standardní interpunkce. Ta ale není pro popis struktury spontánní řeči z několika důvodů vhodná. Za prvé, neexistují žádná pravidla, jak používat interpunkci v defektních syntaktických konstrukcích, které jsou velmi časté. Za druhé, standardní interpunkce nemůže zachytit všechny typické jevy obsažené ve spontánní řeči, která často není plynulá a obsahuje mnoho nedokončených výpovědních útvarů. Za třetí, interpunkční znaménka nemají jednoznačný význam – např. čárka může označovat několik různých jevů (hranici klauze, apozici, vsuvku aj.). Za čtvrté, i pro psaný text nejsou pravidla pro použití interpunkčních znamének zcela exaktní; např. psaní čárky je v některých případech nepovinné. Výše uvedené nedostatky představují klíčové problémy pro (počítačové) porozumění mluvené řeči, proto je žádoucí použít pro zachycení její struktury speciální anotační systém.

Z tohoto důvodu definovalo Linguistic Data Consortium[1] (LDC) anotační standard, který se označuje jako MDE[2] (Strassel 2004). Zjednodušeně řečeno, tyto anotace zahrnují identifikaci konkrétních jevů spontánní řeči (oprav a výplní) a vložení syntakticko-sémantických předělů do spojitého proudu řeči. MDE bylo nejprve navrženo pro angličtinu, naší snahou bylo ho převést do formy použitelné pro konverzační češtinu. Tato transformace však není úplně snadná, protože nelze jen přeložit anotační pravidla z angličtiny a přímo je

[1] http://www.ldc.upenn.edu
[2] Zkratka MDE značí MetaData Extraction.

aplikovat na češtinu. Tato pravidla musí být citlivě upravena tak, aby zohledňovala specifické vlastnosti cílového jazyka, zejména jeho syntaxi. Čeština není jediným jazykem, do kterého bylo MDE převedeno; paralelně s vytvářením pravidel pro češtinu probíhala stejná práce i na arabštině a mandarínské čínštině (Strassel 2005).

Převod MDE do češtiny není nikterak snadný, protože klasická česká lingvistika se na popis syntaxe mluveného projevu nikdy příliš nezaměřovala, a tak ani nevznikla ustálená terminologie použitelná pro konverzační jazyk. Ucelenější popis skladby mluveného textu dává až monografie (Müllerová 1994). Přestože tato práce poskytuje velmi cenný popis některých typických jevů přítomných v mluvené češtině, dívá se na problematiku z trochu jiného úhlu a neklade si za cíl vytvořit *jednoznačný* systém popisu struktury spontánní řeči.

Výzkumná skupina při Západočeské univerzitě v Plzni získala své zkušenosti s automatickým zpracováváním spontánní řeči zejména během řešení projektu MALACH (Psutka et al. 2004). Jelikož ale výpovědi svědků holocaustu, které tvoří tento korpus, nemohly být volně distribuovány, bylo rozhodnuto o vytvoření nového českého spontánního korpusu. Tato nová databáze je tvořena nahrávkami rozhlasového diskusního pořadu Radiofórum. V tomto živém pořadu pozvaní hosté spontánně odpovídají na otázky jednoho nebo dvou moderátorů. V současné době je strukturálními metadaty označeno 52 nahrávek, které byly pořízeny od 12.2. do 6.6.2003. To odpovídá 24 hodinám přepsané řeči.

2. Cíle anotace

Cílem MDE je převést přepis spontánní konverzace do takové formy, která je co nejlépe čitelná pro člověka a vhodná i pro následné zpracování v automatických NLP aplikacích. Takovouto čitelnější formu si například lze představit tak, že z přepisu jsou vymazány výplně a neplynulosti a každá výpovědní jednotka je vytištěna na samostatném řádku.

Anotační pravidla byla vytvořena tak, aby podporovala co největší konzistenci ve smyslu maximální shody mezi různými anotátory. Toho lze dosáhnout vytvořením co nejjednoznačnějších anotačních pravidel. Jelikož zkušenosti s MDE pro jiné jazyky ukázaly, že pravidla založená na sémantických příznacích jsou ve smyslu konzistence nespolehlivá, jsou naše pravidla založena téměř výhradně na povrchových příznacích – syntaxi a prozódii. Vzhledem k omezenému rozsahu článku jsou prezentovaná anotační pravidla spíše jen nastíněna, více informací je možné najít v (Kolář et al. 2005a, Kolář 2005b).

Při samotném značkování mají anotátoři k dispozici jak doslovný přepis konkrétní nahrávky, který byl vytvořen při prvotním zpracování zvukových dat, tak odpovídající audio. Pro zjednodušení anotačního procesu byl vytvořen speciální softwarový nástroj QAn (Quick Annotator), který umožňuje zvýraznění relevantních úseků textu, přehrání odpovídajících částí zvukových nahrávek a zaznamenání umístění MDE symbolů. Ukázka programu je k dispozici na http://www.mde.zcu.cz.

3. Identifikace výplní

Výplň se v MDE systému dá charakterizovat jako úsek řeči, který nenese „žádnou užitečnou“ informaci a je možné ho z přepisu odstranit, aniž by to negativně ovlivnilo jeho pochopení.[3] V tomto projektu se anotují čtyři druhy výplní: vyplněné pauzy (VP), lexikální členící signály (discourse markers, DM), explicitní editační výrazy (EEV) a vsuvky/odbočky (V/O).

[3] U vsuvek je samozřejmě jejich odstranění z přepisu sporné, záleží na konkrétní aplikaci.

3.1. Vyplněné pauzy

VP jsou neartikulované zvuky vydávané mluvčím při váhání nebo při snaze udržet si v konverzaci slovo v době, kdy přemýšlí, co říci dále. V češtině to nejčastěji bývají zvuky podobné protaženému „é“, každý řečník má ale svůj charakteristický styl. Za VP nepovažujeme zvuky, které mají v rozhovoru komunikační funkci (např. odpovědi na otázky se zavřenými ústy – souhlasné *m-hm* či nesouhlasné *m-mm*).

Zatímco zejména pro angličtinu bylo používání VP poměrně široce studováno (např. Clark a Fox Tree 2002, O'Connel a Kowal 2004), pro češtinu jejich podrobná lingvistická studie chybí. Problémem je i skutečnost, že v češtině není ustálený způsob jejich zápisu. V angličtině se zápis VP v textu standardizoval jako *um* (delší výplň s nasální složkou) a *uh* (kratší bez této složky), méně často se používá zápisů *er* a *ah.* Tento zápis ale není možné převzít, protože v každém jazyce zní VP jinak.

Náš způsob notace VP se nakonec ustálil na použití dvou symbolů: *EE* a *MM.* Rozlišování těchto dvou skupin VP je určitým kompromisem mezi přesností a konzistencí zápisu. Použití více kategorií vedlo k malé shodě mezi anotátory, naopak použití jen jedné kategorie sdružovalo příliš odlišné zvuky. Definování dvou kategorií bylo také motivováno tím, že i pro angličtinu se nejčastěji používají právě dvě. Kategorie *EE* odpovídá v našem systému VP, které jsou foneticky bližší samohláskám. Nejčastějším příkladem této skupiny je právě ono známé protažené „*é*“, méně často se vyskytuje protažené „*á*“ či jiné zvuky podobné samohláskám. Kategorie *MM* naproti tomu zahrnuje zvuky, které mají blíže k souhláskám. Nejtypičtějším zástupcem je mumlání trochu podobné opakované hlásce *m* („*mmm*“), často se také vyskytují zvuky podobné „*vvv*“. *MM* jsou v češtině několikanásobně méně četné než *EE*.

3.2. Lexikální členící signály (discourse markers)

„Discourse markers“ jsou slova nebo krátké fráze, které primárně slouží k signalizování struktury mluveného projevu. DM nenesou samostatný význam, ale indikují posluchači takové události, jako je změna řečníka, snaha udržet si slovo nebo změna tématu. V rámci MDE nás zajímají pouze ty DM, které mají funkci výplně (tedy jakési „slovní vaty“) a jejichž vymazáním neztratíme žádnou zásadní informaci.

Pro jakýkoliv jazyk je prakticky nemožné sestavit vyčerpávající seznam DM, protože jejich používání je silně ovlivněno individuálním řečnickým stylem. Příklady častých českých DM mohou být následující: *dobře, jako, jaksi, no, podívejte se, prostě, tak, takže, tedy, víte, víte co, vlastně, v podstatě.*

Stejné výrazy se mohou objevovat jako DM i jako „obsahová slova“. Proto je třeba vždy posoudit daný případ z hlediska kontextu celého dialogu. Např. často problematický je výraz *takže,* u kterého je třeba vzít v potaz, zda řečník chce vyjádřit vztah k předchozí větě, nebo zda chce označit předěl v rozhovoru. U jiných výrazů je rozlišení funkce jednodušší (*Běžel jako vítr. x To jako není nic neobvyklého.*).

3.3. Explicitní editační výrazy

EEV jsou výplně, které explicitně indikují skutečnost, že se řečník při formulaci promluvy spletl. EEV se vyskytují poměrně řídce (jako EEV bylo anotováno jen cca 0,1 % slov), jejich funkci ve spontánních projevech obvykle plní VP. V MDE se ale VP jako EEV neoznačují. Jednoznačně nejčastějším českým EEV je *nebo.*

3.4. Vsuvky a odbočky

Striktně vzato, vsuvky a odbočky nejsou pravé výplně, ale pro zjednodušení celého systému se kvůli podobnosti jejich anotace (označení určitého rozsahu textu) v rámci MDE mezi ně zařazují. Charakteristika vsuvek je zřejmá, odbočky[4] jsou v původním MDE chápány jako krátké komentáře k jinému než hlavnímu tématu, po kterých se mluvčí vrátí opět k původnímu tématu. Odbočka může být adresována někomu, kdo není účastníkem hlavní konverzace (*„A pak se rozsvítil takovej {pardon, musím si vypnout telefon} takovej obrovskej nápis.“)*. Tyto odbočky jsou oproti vsuvkám obvykle provázeny výraznějšími prozodickými příznaky. Pro účely anotace se mezi odbočkami a vsuvkami nedělá rozdíl, používá se pouze jeden společný symbol. Jako O/V se neoznačují lexikalizované (pokleslé) vsuvky jako např. *řekněme* nebo *myslím*. Tato spojení obvykle nejsou provázena prozodickými změnami, které většinou doprovázejí vložení O/V. Kvůli podpoře anotační shody byl připraven ilustrativní seznam těchto vsuvek. Jejich maximální délka je v našem systému omezena na dvě slova.

4. Anotace editačních neplynulostí

Neplynulosti (angl. disfluencies) jsou nechtěná selhání řečníka při formulování a produkci promluvy. Editační neplynulosti jsou úseky řeči, ve kterých řečník koriguje nebo zcela mění obsah části svého mluveného projevu. Může také započatou promluvu zcela opustit a začít formulovat novou myšlenku (tzv. restart). Charakter neplynulostí je v češtině velmi podobný angličtině, proto anotační pravidla pro tyto jevy nebylo třeba nijak zvlášť upravovat. Struktura neplynulosti (Shriberg 1994) je v MDE následující:

- *Odstranitelná oblast* – (deletable region, DelReg, v literatuře též *reparandum*) – obsahuje původní úsek řeči, který řečník následně opravuje.
- *Bod přerušení* – (BP) – je moment, ve kterém mluvčí přeruší plynulost své promluvy. Všechny neplynulosti mají alespoň jeden BP na pravém konci DelRegu. Existují i tzv. složité neplynulosti, které mají více než jeden BP.
- *Editační fáze* – (v lit. též *interregnum*) – v tomto úseku si řečník uvědomuje, že se dopustil neplynulosti. Může obsahovat EEV, DM či VP.
- *Oprava* (korekce) – obsahuje plynulý úsek promluvy, ve kterém řečník opravuje neplynulost. Restarty korekci neobsahují. Přestože opravy nejsou v anglickém MDE explicitně označovány, rozhodli jsme se je označovat, abychom získali relevantní data pro jejich studium.

Pro každou neplynulost se identifikuje plný rozsah DelRegu a označí se všechna slova, která do něj patří. BP na koncích DelRegů jsou automaticky identifikovány a označovány anotačním programem, případné vícenásobné BP v rámci jednoho DelRegu již musí být anotátorem označeny manuálně. Pokud v neplynulosti existuje EEV, označí se podle výše uvedených pravidel pro výplně. Do opravy se zahrnou právě ta slova, kterými řečník nahrazuje slova a slovní fragmenty v DelRegu. Ukázka anotace je uvedena v následujícím příkladu. DelReg je označen hranatými závorkami, BP hvězdičkou, EEV tučným řezem písma a oprava podtržením.

*Chtěl bych vědět odjezdy [z Prahy v * v * v] * EE* ***teda*** *<u>z Plzně v</u> úterý večer /.*

[4] V angličtině se tato kategorie výplní označuje jako Asides/Parentheticals, proto i pro češtinu používáme termín „odbočka“, který není v zdejším mluvnickém názvosloví příliš užíván.

5. Segmentace přepisu do syntakticko-sémantických jednotek (SU)

Klíčovým úkolem MDE anotace je rozdělení promluv do jednotek, které zhruba odpovídají větám v psaném projevu. Protože mnoho řečníků má ve spontánním projevu tendenci používat velmi dlouhá a nejasně prozodicky strukturovaná souvětí, je téměř nemožné konzistentně určovat hranici „věty“ pouze podle zvukového vyznačení. Možným řešením je rozdělovat spojité projevy do úseků minimální délky, které splňují podmínku syntaktické a sémantické uzavřenosti.[5] Tyto jednotky se v MDE označují zkratkou SU.[6] SU typicky obsahuje jednu hlavní (nezávislou) klauzi. Jednotlivé SU jsou v přepisech identifikovány na základě symbolu své hranice. Rozlišujeme dva druhy označovaných hranic – interní (uvnitř SU) a externí (hranice SU). Zatímco externí hranice jsou pro MDE fundamentální, interní symboly byly zavedeny hlavně pro zvýšení přehlednosti zápisu a lepší orientaci anotátorů. Zjednodušeně lze anotační proces popsat jako segmentaci přepisu do klauzí a následné přiřazení odpovídajících SU symbolů jejich hranicím. Všechny SU symboly začínají lomítkem.

Jako interní SU symboly používáme:

- /, – hranice klauze (*Mám velký hlad* /, *protože jsem ráno nic nejedl* /.)
- /& – souřadné spojení vedlejších klauzí nebo hlavních klauzí, které nemohou stát samostatně („*On mi tvrdil /, že tam nepůjde /& a že Petra taky zůstane doma /.“, „Buď mi ty peníze dáš hned /& nebo s tebou už nikdy nebudu obchodovat /.“*)

Externí SU symboly jsou následující:

- /. – oznámení (*Dnes je pěkné počasí* /.)
- /? – otázka (*Máš večer čas* /?)
- /~ – nekompletní SU přerušená jiným řečníkem
  (*A: Mluvili jsme o tom člověku* /, *který včera /~ B: Počkejte /.)*
- /- – nekompletní SU svévolně opuštěná řečníkem
  (Určitě to bylo už včera /, protože bylo /- Prostě on lže /.)

Ve srovnání s anglickým systémem, který používá jen jeden symbol pro nekompletní SU, je v českém MDE přidáno rozlišení jejich dvou typů. Tato změna byla motivována tím, že svévolně nedokončené SU se od SU přerušených „skočením do řeči“ liší v prozódii i syntaxi. Další jazykově nezávislou změnou v anotaci SU bylo přidání symbolů „//.“ a „//?“ pro výrazné předěly. Symboly s dvojitými lomítky jsou použity v případě, že je označovaný předěl doprovázen silným prozodickým vyznačením, tj. pauzou, závěrečným prodloužením nebo výrazným snížením nebo zvýšením výšky hlasu. Důležitou otázkou před jejich zavedením je, jaká je na síle prozodického označení shoda mezi anotátory, protože konzistentní ruční značkování prozódie je často velmi obtížné. Naše pozorování ukázala, že pokud se dva anotátoři shodnou na tom, že na určitém místě je konec určitého typu SU, pak se přibližně v 85 % případů shodnou i na tom, zda mají použít symbol s jedním, nebo dvěma lomítky. Tuto míru shody považujeme za velmi dobře akceptovatelnou.

Další modifikace byly motivovány specifiky české syntaxe. Nejvýraznější změnu jsme museli provést s ohledem na skladbu českých souřadných souvětí. Při naší práci jsme zejména vycházeli z popisů v (Svoboda 1970, Grepl a Karlík 1998). Z pohledu MDE je nejzásadnějším rozdílem mezi stavbou věty v angličtině a češtině možnost zamlčeného podmětu. V angličtině je vynechání explicitního uvedení podmětu povoleno pouze v souřadných souvětích

[5] Tedy platí pravidlo “děl, kde jen můžeš!”.
[6] Zkratka SU má v angličtině zároveň více významů: Sentential/Semantic/Syntactic/Slash Unit.

v případě, že dvě klauze sdílejí společný podmět. Naopak v češtině může být podmět ve větě vynechán vždy, když je zřejmý z okolního kontextu a použitého slovesného tvaru. Tento fakt ovlivňuje rozhodování mezi použitím /. a /& v souřadných souvětích. V anglických pravidlech zamlčení podmětu v souřadně připojené hlavní klauzi automaticky implikuje použití /&, protože taková klauze nemůže stát samostatně. Protože ale v češtině samostatně stát může, použijeme místo /& SU-externí symbol (viz srovnání anotace v angličtině a češtině v následujícím příkladu).

Robert walked to work /& but took the train home /.
Robert šel do práce pěšky /. ale domů jel vlakem /.

Naopak pokud souřadně spojené klauze sdílejí pomocné sloveso, druhá klauze nemůže existovat samostatně ani v češtině, a proto použijeme /& (*Zítra budu odpočívat /& a pak venku hrát fotbal /.*).

Z důvodu omezeného prostoru bohužel není možné zde diskutovat všechna pravidla pro identifikaci SU. Detailní anotační pravidla (Kolář 2005b) dále popisují, jak pracovat s několikanásobnými přísudky, vedlejšími větami závislými na více větách hlavních, různými druhy elips, volně připojenými větnými členy, anakoluty, osamostatnělými vedlejšími větami, idiomatickými spojeními, přímou a nepřímou řečí, parcelací, apozicemi a jinými charakteristickými jevy.

6. Závěr

V tomto článku byl představen anotační systém MDE použitelný pro zápis strukturálních metadat v mluvené češtině. Tyto anotace zahrnují rozdělení doslovného přepisu řeči do syntakticko-sémantických jednotek (SU) a identifikaci různých druhů výplní a neplynulostí. Jednotlivé MDE symboly plní v řečových transkripcích funkci, která je podobná funkci interpunkce v psaném textu. Původní anotační pravidla, která byla vytvořena pro angličtinu, byla upravena tak, aby zohledňovala specifika češtiny. Vedle úprav motivovaných rozdílnou syntaxí byl také navržen způsob zápisu vyplněných pauz typických pro češtinu. Mimo převedení pravidel do češtiny bylo také provedeno několik jazykově nezávislých změn, které podle našeho názoru zpřehledňují původní anotační standard.

S použitím těchto pravidel byl anotován spontánní řečový korpus obsahující nahrávky rozhlasového diskusního pořadu. Tento korpus plánujeme v brzké době zpřístupnit široké vědecké komunitě prostřednictvím jeho publikace v LDC. S využitím zkušeností nabytých při anotaci tohoto korpusu chceme dále zpřesňovat anotační pravidla pro sporné případy tak, aby co nejlépe podporovala shodu mezi anotátory. Plánujeme také rozšíření celé databáze, abychom získali více dat pro dostatečně robustní odhad parametrů automatických systémů.

Poděkování

Autor děkuje Dagmar Kozlíkové, Stephanii Strasselové a Christopheru Walkerovi za cenné konzultace při vytváření anotačních pravidel a Janu Švecovi za naprogramování anotačního nástroje QAn. Tato práce byla podporována MŠMT ČR v rámci projektů 1M0567 a ME909.

LITERATURA

Clark H.H., Fox Tree J., 2002, Using uh and um in spontaneous speaking, *Cognition,* 84, 73-111.

O'Connel D., Kowal S., 2004, The history of research on the filled pause as evidence of the written language bias in linguistics, *Journal of Psycholinguistic Research,* 33, č. 6, 459-474.
Grepl M., Karlík P., 1998, *Skladba češtiny,* Votobia, Olomouc.
Kolář J., Švec J., Strassel S., Walker Ch., Kozlíková D., Psutka J., 2005a, Czech spontaneous speech corpus with structural metadata, In *Proc. INTERSPEECH 2005,* Lisbon, Portugal.
Kolář J., 2005b, *Jednoduchá anotace strukturálních metadat pro češtinu,* Dostupné z WWW: http://www.mde.zcu.cz.
Müllerová O., 1994, *Mluvený text a jeho syntaktická výstavba*, Academia, Praha.
Psutka J., Hajič J., Byrne W., 2004, The development of ASR for Slavic languages in the MALACH project, In *Proc. IEEE ICASSP,* Montreal, Canada.
Shriberg E., 1994, *Preliminaries to a theory of speech disfluencies,* PhD thesis, University of California at Berkeley.
Strassel S., 2004, Simple metadata annotation specification, Dostupné z WWW: http://www.ldc.upenn.edu/Projects/MDE/.
Strassel S., Kolář J., Song Z., Barclay L., Glenn M., 2005, Structural metadata annotation: Moving beyond English, In *Proc. INTERSPEECH 2005,* Lisbon, Portugal.
Svoboda K., 1970, *Souvětí spisovné češtiny*, SPN, Praha.