

Anotace orálních korpusů češtiny

Olomoucký mluvený korpus jako model

Petr Pořízka

Tato studie stručně předkládá v základních bodech anotační koncepci *Olomouckého korpusu mluvené češtiny* (budovaného autorem tohoto článku od r. 2002 na Katedře bohemistiky Filozofické fakulty Univerzity Palackého), jež je – srovnáme-li ji s koncepcí, dle které vznikly *Pražský* a *Brněnský mluvený korpus* (PMK a BMK), – v mnoha aspektech nová a zřetelně odlišná. Ve snaze o maximální kompatibilitu se zmíněnými mluvenými korpusy Českého národního korpusu (ČNK) jsme sice jejich metodologii (včetně pravidel přepisů) převzali, postupem času však bylo i přes veškerou snahu o zachování koncepce stále zřetelnější, že je nadále neudržitelná. Blíže se této problematice věnujeme v dřívějších studiích (viz Pořízka 2004a, 2004b a 2005).

Zásady sběru a anotace mluvených korpusů ČNK

Sběr nahrávek Olomouckého mluveného korpusu (OMK) byl a je založen (ve shodě s PMK i BMK) na kombinaci čtyř sociolingvistických proměnných u nahrávaných mluvčích: na kombinaci (1) pohlaví (muž–žena); (2) věku (vyšší–nižší, s hranicí 35 let); (3) vzdělání (nižší–vyšší) a (4) jazykových dat získaných řízeným a neřízeným způsobem nahrávky: tj. (4a) nahrávkou formální, podle předem připravených tematicky širokých otázek, a (4b) neformální, neřízeným dialogem dvou (tří) mluvčích (kteří se znají), tematicky neomezeným a neurčeným; s délkou nahrávky kolem 20 minut (cca 2 300 slov). Modifikovali jsme zásady sběru v několika aspektech: Předně jsme rozšířili kritérium výběru respondentů: Původně (PMK a BMK) museli být nahrávanými pouze rodilí mluvčí dané lokace (města), jedinou výjimkou byli respondenti, kteří sice pocházeli odjinud, ale v daném městě žili nejméně dvě generace, tj. 20 let. U OMK pro nás byl stejně zásadní dennodenní kontakt s olomouckou mluvou. Proto jsme zrušili (poměrně přísnou) hranici 20 let pobytu; podstatné je, zda mluvčí žije nebo žil v Olomouci (doba pobytu se uvádí v identifikačním štítku každé nahrávky), nebo zde má zaměstnání apod. Držíme se jen vyloučení jazyka mládeže dané lokality.

Informace o mluvčích jsou u OMK doplněny o zaměstnání, konkrétní věk nahrávaných, oblastní (nářeční) původ mluvčího, není-li původem z Olomouce, a dobu pobytu v Olomouci. Pro korpusové vyhledávání počítáme s detailnější segmentací věku (alespoň po dekáдах), neboť stávající dělení věku na vyšší–nižší s hranicí 35 let je příliš „hrubé“, neumožňující sociolingvistické analýzy prostřednictvím korpusového manažeru.

Pravidla přepisů mluvených korpusů ČNK

Pravidlům přepisů PMK a BMK jsme se detailněji věnovali ve studii *K možností vyhledávání dat a struktury atributů korpusových manažerů v mluvených korpusech ČNK* (Pořízka 2004a). Ukázali jsme, že transkripční pravidla PMK jsou nedostačující, prakticky zcela založená na ortografickém principu. Šla „proti“ mluvenému jazyku, zejm. u jevů jako opravy či doplnění slov (ať už nedopovězených či různých přechytlů), neznačení znělostních či artikulačních asimilací ad. Problematické je i užití syntaktické interpunkce (v transkriptu čárky) – a to i tam, kde není realizován zvukový předěl (pauza). Není tedy reflektována zvuková podoba komunikátů.

Transkripční pravidla BMK doznala oproti PMK jistých (a přínosných) změn. Užit byl tzv. hybridní princip přepisu – účelová kombinace fonetického a ortografického zápisu, stále ovšem akcentací standardních pravopisných norem. Syntaktické čárky byly nahrazeny tečkami značícími pauzu (reflektujícími i její relativní délku), tj. tradiční interpunkce u PMK byla nahrazena interpunkcí

„pauzovou“, čímž bylo upřednostněno hledisko zvukové realizace. Důležitá byla i snaha o zachycení simultánnosti dialogických promluv.

Foneticky jsou v BMK transkribovány pouze výrazy realizované v rozporu s ortoepickou normou, tj. odchylky od spisovné realizace výslovnosti. Šlo především o vybrané jevy moravské nespisovné sonorizace neznělých konsonantů-obstruentů ve spojení se sonorami (typ *sme* × *zme*) apod.¹ Neřešeny tak ovšem zůstávají spisovné ortoepické dublety typu [*hled'me* i *hlet'me*] či [*renesance* i *renezance*], není značena spojitá výslovnost a ráz. Zaznamenána je segmentální rovina, suprasegmentální nikoli, až na pauzy a signalizaci otázkové intonace pomocí otazníku.

Důvodem pro volbu transkripčních pravidel PMK i BMK založených výrazným dílem na ortografii byla dle tvůrců byla potřeba morfologické analýzy. Doposud ovšem tyto korpusy lingvisticky označovány nejsou, což je poněkud paradoxní situace, neboť jsou-li ponechány tyto korpusy bez lingvistické anotace, pak by bývalo nesporně vhodnější transkribovat nahrávky foneticky.²

U korpusu ORAL2006, jenž je prozatím nejnovějším mluveným korpusem ČNK, nalezneme další pozitivní změny metodologického charakteru. Korpus je složen pouze z neformálních nahrávek a u mluvčích již lze zjistit jejich přesný věk. V rámci pravidel přepisu nahrávek je pamatováno citlivějším způsobem na nedořečené výrazy a přerušování – jsou značeny hvězdičkou³ (*to bylo vše**, *především*, *neda* nedořečené*). Třemi tečkami se signalizuje neukončenost vět nebo přerušení jednoho mluvčího druhým. Krom zmíněných modifikací je přepis bohužel i nadále založen na původních pravidlech PMK – viz syntaktická interpunkce i tam, kde není realizován předěl (pauza) apod.

Výsledky analýzy z r. 2003 a změna pravidel transkripce

Pro potřeby tohoto textu uvádíme jen stručně výsledky analýzy, k níž lze nalézt zevrubné informace v již zmiňované publikaci (viz Pořízka 2004a). Podstatou a východiskem provedené analýzy bylo prověřit reliabilitu a validitu zvolených přepisů mluvených korpusů ČNK (tj. PMK a BMK). Jako zdroj pro vyhledávání dat nám sloužil BMK, a to pro možnosti vyhledávat částečně i podle výslovnosti (atributy *Word* a *Pron* v korpusovém manažeru). PMK byl ponechán stranou – důvodem byla pravidla přepisu, a tedy absence fonetických aspektů transkriptů (PMK má pro vyhledávání jediný atribut *Word*).

Zvolenými jazykovými jevy (daty) byly takové slovní tvary, které jsou pro mluvenou formu češtiny v dané oblasti typické a hojně užívané: znělostní asimilace, zejm. typická „moravská“ tendence k markantnější sonorizaci neznělých konsonantů-obstruentů, než je tomu na území Čech. Předpokládali jsme, že budou v přepisu reflektovány. Omezení jsme byli absencí zvukových nahrávek, tj. neměli jsme možnost zpětné kontroly a srovnání získaných statistických výsledků. Zvolili jsme proto případy téměř „kanonizované“, jež se realizují v mluvě Moravanů tak často, že znělá realizace případů s možnou dvojí výslovností převládá nad neznělou. Zvolené jevy jsou nadto explicitně zmíněny i v zásadách přepisu BMK, tzn. přepis si všímá spojení neznělých konsonantů se sonorami (*r*, *l*, *m*, *n*, *ň*, *j*, resp. *v*) a hláskové skupiny *sh*.

Konkrétně jsme zkoumali spojení sonor s párovými konsonanty *s* a *k* ve dvou kontextech: (a) spojení neslabičných neznělých předložek *s* a *k* se slovem začínajícím sonorou (např. *k ruce*, *k městu*, *s mícem*) a (b) sandhiové jevy ve spojení dvou slov, kdy první končí na konsonanty *s* a *k*, druhé slovo opět začíná sonorou (resp. konsonantem *v*), např. *jak rybka*, *vlak jel*. Mimoto jsme vybrali další slova typická pro moravskou výslovnost a srovnávali tvary *sme*/[*zme*], *bysme*/[*bizme*], *kvůli*/[*gvůli*] a konsonantické spojení *sh* na počátku slov (př. *shon*, *shazovat*). Pro vyhledávání jsme využili formalizovaný tvar dotazů prostřednictvím definovaných regulárních výrazů. Výsledky a statistická zjištění ukazuje následující Tabulka.

¹ Viz pravidla přepisu BMK na webových stránkách ÚČNK (<http://ucnk.ff.cuni.cz>).

² Je ovšem třeba brát v úvahu i kvalitu nahrávek. Podmínky, možnosti i kvalita nahrávání se od r. 1988, kdy bylo započato s budováním PMK, nesrovnatelně změnily.

³ Viz opět pravidla přepisu ORAL2006 na webových stránkách ÚČNK.

Výsledky analýzy (Tab)⁴

vyhledané tvary	celkový výskyt	neznělá varianta	sonorizovaná podoba
<i>sme</i> /[<i>zme</i>]	1851	1716	135
<i>bysme</i> /[<i>bizme</i>]	181	175	6
<i>kvůli</i> /[<i>gvůli</i>]	73	66	7
<i>skupina</i> [sx...]/[zh...]	72 (70)	33 (31)	39
spojení <i>k_r...</i>	23	21	2
spojení <i>s_r...</i>	41	41	0
spojení <i>k_l...</i>	14	14	0
spojení <i>s_l...</i>	26	26	0
spojení <i>k_m...</i>	28	28	0
spojení <i>s_m...</i>	97	96	1
spojení <i>k_n(ň)...</i>	182	180	2
spojení <i>s_n(ň)...</i>	430	429	1
spojení <i>k_ň...</i>	12	12	0
spojení <i>s_ň...</i>	42	42	0
spojení <i>k_j...</i>	16	16	0
spojení <i>s_j...</i>	85	85	0
spojení <i>..k_j...</i>	2077	2056	21
spojení <i>...s_j...</i>	304	301	3
spojení <i>..k_r...</i>	123	121	2
spojení <i>...s_r...</i>	52	52	0
spojení <i>..k_l...</i>	68	68	0
spojení <i>...s_l...</i>	46	44	2
spojení <i>..k_m...</i>	875	868	7
spojení <i>...s_m...</i>	295	291	4
spojení <i>..k_n(ň)...</i>	1485	1479	6
spojení <i>...s_n(ň)...</i>	815	811	4
spojení <i>..k_ň...</i>	73	73	0
spojení <i>...s_ň...</i>	48	48	0
sandhi <i>..k_v...</i>	899	892	7
sandhi <i>...s_v...</i>	275	273	2

Koncepční změny na základě provedených analýz

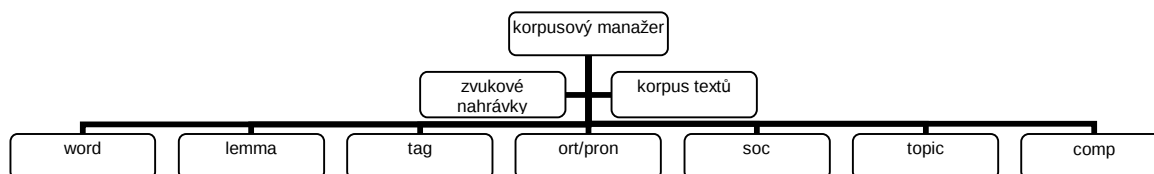
Ze statistických údajů vyplynulo, že přepisy s největší pravděpodobností neodpovídají skutečné realizaci, a tedy poměru zkoumaných hláskových realizací v daných kontextech. Podle přepisů se sonorizovaná podoba až na „zažité“ výjimky lišení *sme* × *zme* a zvukové realizace spojení konsonantů *sh* v korpusu BMK prakticky nevyskytuje! Existuje tedy zřejmá diskrepance mezi přepisem a zvukovou nahrávkou komunikátů. Obzvlášť zajímavé jsou všechny případy spojení neslabičné předložky *s* (v 7. pádě) se sonorami. Dostáváme chybné statistické údaje na základě zvolených pravidel přepisu. Vzhledem k tomu, že dubletní výslovnost *s* i *z* je spisovná, v přepisu se realizace *s* vyslovenou předložkou *z* vlastně neznačí (jak jsme uvedli výše, dublety nejsou v přepisu vyřešeny). Protože *s* je tvar ortografický, je mu dána přednost.

Ukázalo se jako nezbytné přikročit ke změně transkripčních pravidel směrem k vizualizaci fonetické realizace řečového kontinua. Rozhodli jsme se zavést fonetický přepis, který nebude uložen metatextově, ale jako „plný“ paralelní text vedle přepisu založeného na ortografii (sloužícího potřebám morfologické analýzy). U OMK počítáme i s propojením zvukových záznamů komunikátů a přepisů (*s* možností zobrazit texty v úplnosti). Velmi žádoucí by bylo i rozšíření funkcí korpusových

⁴ Autorova studie *K možností vyhledávání dat a struktury atributů korpusových manažerů v mluvených korpusech ČNK* obsahuje obsáhlý komentář ke statistickým výsledkům.

manažerů zavedením nových atributů/funkcí umožňujících širše využít mluvené komunikáty korpusových souborů (viz schéma), aby bylo možno vyhledávat i prostřednictvím sociolingvistických proměnných.⁵

Návrh struktury atributů korpusového manažeru



Vysvětlivky atributů:

- Ø word – vyhledávání podle klíčového slova nebo slovního spojení
- Ø lemma – vyhledávání podle základního slovníkového tvaru
- Ø tag – vyhledávání podle morfologické značky
- Ø ort/pron (orthographic / pronunciation) – přepínání verze komunikátu, možnost vyhledávání jak podle ortografie, tak podle výslovnosti
- Ø soc (sociolinguistic) – vyhledávání podle sociolingvistických proměnných
- Ø topic – vyhledávání podle témat komunikátů či jejich tematických částí
- Ø comp (complete) – zobrazení komunikátu v jeho úplnosti

Fáze budování OMK – sběr a přepis komunikátů (2002–2007)

Na třech fázích přepisu nahrávek OMK lze vidět rodící se anotační koncepci od prvotního převzetí pravidel ČNK až po duální paralelní transkripci.

První fáze (r. 2002–2003): tzv. hybridní (ortograficko-fonetický) přepis na základě pravidel BMK (cca 45 nahrávek).

OMK: Přepis formálního rozhovoru dle pravidel BMK (formální typ komunikátu)

(1)

no myslim si že **ted'ka** je to náročnější těši . dřív **chodily děti** .. osum let do š teda **my sme** [zme] **chodili osum let do školy** pak to se **změnilo** na **devět** [devjed] let . **chodilo** se do **školy** . pak se **chodilo** . pak to **bylo dobrovolny** a **ted'** se **chodi ani nevím** . taky **devět** [devjed] let ... to je všechno

(2)

no **tak** [tag] mysim si že **by měli** ve školstvi byt muži . protože **děti** só čim dál živjější a .. žensky na to nestačijó . chachacha

Vysvětlivky:

Tučné písmo – ortografický princip (vybrány jsou jen typické jevy)

Kurzíva – fonetický princip

Podtržené písmo – hybridní ortograficko-fonetický tvar

Ve druhé fázi (r. 2004) jsme přistoupili ke změně v zavedení dvojího přepisu – (a) zjednodušením stávajícího hybridního přepisu (byla ponechána malá písmena, pauzová interpunkce, ale odstranili jsme např. systém hranatých závorek...) a (b) zavedením paralelního fonetického přepisu komunikátů – vzhledem k potřebě mít k dispozici font nezávislý na konkrétní počítačové platformě jsme zkoušeli systém SAMPA (českou verzi) (cca 45 nahrávek).

⁵ Pokoušíme se o to ve spolupráci s kolegy z NLP Fakulty informatiky Masarykovy univerzity v Brně.

OMK: Fonetický přepis – SAMPA

<ZVBF.1> Skolska: so_ustava |a vubet_s|
<T> |no prosim| ce norma:lJe mluf jag dibi
<1> Skolstvi: jako je f so_ut_Sasne: dobje pro generat_si pQ\eceZuji:t_si: . a zaceZuji:t_si: . protoZe diS ji srovna:m se svim
Jet_stvim . tak na:roki ktere: bili na na:s kladeni bili sit_se jako . na:m pQ\ipadali visoke: . ale: mJeli zme prosce prostor pro
vlastJi rozvoj ne: pro lumpa:rni ale pro vlastJi rozvoj . vi:t_s volne:h\o t_Sasu . Ze se h\odJe smJerovala vi:uka do Skoli .. a ne:
mimo .. a tec sou J\eci pQ\eceZovani: --- nebo viJ\ela sem uZ v dobje pQ\et cema dvat_seci leti
<T> jako mis pot_Skej jag Zes to viJ\ela uZ v dobje pQ\ed dvat_seci leti ?
<1> no gdiZ vlasJe tis zat_Sala xoJ\it do Skoli . tak
<T> jako mysi:S uS na me: za:klatt_se
<1> uS na tve: za:klatt_se bilo toh\o ut_SeJi: daleko vi:t_s neZ zme mJeli mi . t_Sili bibi prosce jako ti: osnovi: a tak da:l
prosce uspu:sobovani: .. prosce bili sou na ti J\eci kladeni t_Sim da:l viSi: na:roki na to zvla:da:Ji: toh\o ut_Siva na xa:pa:Ji: a
tag da:le . a nemaj prostor vlastJe . pro . klasit_ski: J\et_stvi . po h\ri . |pro kamara:t_stvi:|
<T> |Ze otupuji:| misiS |jako|
<1> |ano|
<T> no . to jako celkem zJ\i:li:m tvu:j na:zor .. no ---

Nespornou devizou transkripčního systému SAMPA je jeho univerzálnost a nezávislost na konkrétním kódování, fontu či programu, ale – jak vidno z ukázky i níže uvedených příkladů – je uživatelsky „nepřátelská“: je mnohem těžší čitelná pro ty, kteří ji pravidelně neužívají. Čtení SAMPY vyžaduje určitou dávku cviku.

Czech SAMPA

?eSce je tQ\i t_Stvr=ce na ?osm=	(eště je tři čtvrtě na osm)
sot_si_a:lJi: pra:t_se	(sociální práce)
pQ\et_sci:ra: Ze teJ\ u:pl=Je nepQ\et_Ste t_Si:slo	(předstírá, že teď úplně nepřečte číslo)

Až ve třetí fázi (r. 2004–2007) (cca 200 nahrávek) jsme vyřešili ambivalentnost situace, kdy je třeba mít pokud možno co nejvěrnější písemný zápis zvukového záznamu, ale na druhé straně přepis takového druhu, aby umožňoval i počítačové zpracování. Zůstali jsme u paralelní verze přepisu, ale zavedli jsme transkript (a) důsledně ortografický (pro potřeby morfologické analýzy) a (b) fonetický. Ad (b): Upustili jsme od systému SAMPA a užili zjednodušený systém české fonetické transkripce (zjednodušený ve srovnání s IPA)⁶. Používáme font FONETIC-X (© Kiril Ribarov).

OMK: Fonetický přepis – font FONETIC-X

M1: / tag mňe na to ta bo~~ž~~ka napašela / tagže teď je to _ upčne mokří /
M2: / bila ses tam poďivat / jo ÷
M1: / bila / já +sem tam ?ešt’e p~~ř~~vješovala to co +sem doprala viš / to tmavi / a tak +sem s poťešením z+ist’ila / že dneska to
zbirad nebudu muset / protože to neusxne /
M2: / no / +dibi fo~~ž~~kalo / tak si mislim / že bi t’i to ?ešt’e usxlo /
M1: / ?ale já na tom net~~ř~~vám / mňe to jako / [1 f poctaťe gdištak to seberu zitra / to]1
M2: [1 ?ale tak co tebe nehoň / že jo]1
M2: [2 ?ale je to počasi fakt pošahani / protože +neska se zatahlo dvakrát / ?a vipadalo to / že bude takovej hroznej děšť]2 [3
h è / že +ste dali jako na tu p~~ř~~etpovjet’ / že ÷]3
M1: [2 ?ale prosim t’e ráno / gdiž +zme +stávali / to bilo tak krásně / že muj muž velice jako se roščiloval / že +zme nejeli na
moráfku /mi +zme původně xt’eli / ale]2
M1:[3 ?ale gdiž řikali potom / budo~~ž~~kako prutáže mračen šileni / že +sme si řikali / budem stejně sed’et tam doma ?a nič s toho
nebude / jo]3

⁶ IPA (International Phonetic Association): mezinárodně platná a definovaná fonetická transkripční soustava. Je relativně obsáhlá díky množství speciálně definovaných symbolů.

Vysvětlivky: Hranaté závorky (s číslicemi) slouží k jednoznačnému, přesnému a přehlednému značení překrývání replik (viz níže v textu).

Korpusy mluvené češtiny ČNK (stav: září 2007)

mluvený korpus	dobu budování korpusu	počet nahrávek	počet mluvčích	velikost korpusu (počet slov)
PMK	1988–1996	304	?	674 992
BMK	1994–1999	250	294	596 009
ORAL2006	2002–2006	221	754	1 000 798
OMK	2002–dosud	289 (578) ⁷	658 ⁸	* ⁹

Zásady anotace orálních korpusů

Při vytváření nové anotační koncepce OMK jsme si vytyčili několik zásad, jež jsme u mluvených korpusů považovali za relevantní:

- zvolit adekvátní způsob vizualizace zvukových komunikátů a rozhodnout o detailnosti přepisu (OMK: metoda dvou paralelních verzí transkriptu)
- vyřešit tzv. překrývání replik jednotlivých mluvčích a vnést do přepisu komunikační aspekty textu (komentáře a metatextové informace)
- odlišit roviny přepisu, aby bylo jasné a zřejmé, co je vlastní transkript textu a co metatext
- komunikačně relevantní kontextové a komentářové informace zaznačit takovým způsobem, aby příliš nenarušovaly přepis
- vytvořit pokud možno přehledný a jasné strukturovaný systém anotace s obecnou platností
- pečlivě zvážit volbu extralingvistického i intralingvistického anotačního systému (viz typologie transkripčních principů)

Typologie transkripčních a anotačních principů

Možnosti i způsoby lingvistické anotace jsme pečlivě zvažovali. Vyhledávat, pokud jde o jazyková data, lze totiž v korpusu jen to, co bylo anotováno (velmi záleží i na tom, jakým způsobem). Lingvistická anotace (a její záběr i šíře) tedy určuje způsob práce a možnosti využití korpusových/jazykových dat (srov. Pořízka 2004, s. 81). Ve výčtu uvádíme základní typy transkripčních a anotačních principů (v závorkách jsou uvedeny korpusy, které danou anotaci obsahují):

- ortografická (OMK)
- fonetická (fonologická) (OMK)
- ortoepická, tzv. foneticky kanonická (Pražský fonetický korpus)

⁷ Počítáme set FOR+NEFOR rozhovorů jako jednu nahrávku, jinak by byl počet prakticky dvojnásobný, tj. 578! (V PMK i BMK jsou počítány FOR a NEFOR nahrávky odděleně.) Konkrétní počet nahrávek, respektive transkriptů, jež budou oficiálně zveřejněny, se ještě může změnit v závislosti na konečném a pečlivém výběru. Zvažována bude technická kvalita nahrávky, autentičnost komunikátu, zda mluvčí odpovídá kritériím výběru apod. První předvýběr byl již proveden, neboť jsme měli k dispozici 330 (resp. 660) nahrávek.

⁸ Konečný počet mluvčích v oficiálně zveřejněném OMK může být ještě mírně modifikován finálním výběrem nahrávek, viz pozn. 7.

⁹ Na půlmilionový korpus je třeba cca 50 hodin nahrávek; jeden set F+NF nahrávek OMK trvá v průměru 30 minut (celkově: cca 145 hodin, tj. kolem 1,5 milionu slov).

- gramatická (lexikálně-morfologická) (BMK, OMK)¹⁰
- pragmatická (komunikační – kontextové komentáře) (OMK)
- hybridní (kombinující v rámci jediné transkripce dva či více principů, nejčastěji ortograficko-fonetická) (PMK, BMK, ORAL2006)

Z uvedeného přehledu je vidět anotační šíře záběru OMK. Otázkou zůstává potenciální zařazení ortoepického přepisu. Zvažujeme i zavedení této transkripce, ovšem je závislá na vytvoření speciálního softwaru – konvertoru, který by z ortografické verze „generoval“ text na základě definovaných pravidel založených na platné ortoepické normě češtiny.

Anotace OMK a její struktura

záhlaví

v komplexní informace v záhlaví dokumentu:

- informace o délce nahrávky, její datace
- sociolingvistické proměnné – mluvčí a jejich charakteristika
- žánr – např. interview, přednáška ap.
- sekce: vypsána jednotlivá témata komunikace – slouží i k rychlé orientaci o obsahu celého komunikátu

transkript

v vrstvy textu: (1) dva módy přepisu: ORT a FON

Příklad:

ORT protože on si si fakt myslel , že musím mít důvod

FON p+že von s+ si fag+ mis+e:l / že musím _ míd _ důvot

(2) strukturální metatextové informace:¹¹

- komentáře, komunikační labelizace
- tematická segmentace sekcí
- časové údaje (zejm. u sekcí)
- ostatní informace

Strukturální (meta)znaky anotace¹²

- signalizují typ (meta)textu: zda jde o vlastní text přepisu, nebo o novou sekci, komentář, časový údaj aj.
- vždy předcházejí každý samostatný řádek přepisu (oddělené klávesou *enter*), tj. stojí na počátku každé repliky, ale i metatextových řádků
- význam metaznaků:

\$ vlastní text komunikátu (předchází značce mluvčího), tj. „\$A:“

komentář, časový údaj, jiná metatextová informace, př. # < zakašlání >

¹⁰ U BMK není doposud zveřejněna, a není možno s ní tedy pracovat. Pro OMK počítáme s morfologickou anotací prostřednictvím softwaru AJKA (FI MU Brno, © R. Sedláček).

¹¹ Komentářové, sekční a časové metatextové značky jsou součástí foneticky transkribovaného textu; díky strukturaci tematických sekcí je vytvořena možnost vyhledávat podle témat komunikátů či jejich tematických částí.

¹² Inspirací nám v některých aspektech byl tzv. *Göteborg transcription standard*. Globálně jde o rozsáhlý anotační systém, místy ovšem až příliš (a zbytečně) komplikovaný, u některých jevů se způsoby značení dokonce dublují. Pro běžnou praxi prakticky příliš „zatěžuje“ samotný text přepisu.

§ sekce (téma) komunikátu, změna tématu během konverzace, př. § < koupě nového automobilu >

- transkripční metaznaky:

/ // ///	pauza (odstupňovaná dle relativní délky)
()	slova či pasáže, u nichž si přepisovatel není jist zněním (vliv např. kvality nahrávky či deformace slov mluvčím)
< >	komentářové závorky (je-li více komentářů k jediné replice, posloupně se číslovají)
+	nedořečené slovo, elidovaná hláska či skupina hlásek
---	nesrozumitelný úsek
[]	simultánnost replik (posloupně číslovaných)
:	nenáležité (či emfatické) dloužení vokálů
_	spojitá, geminovaná výslovnost hlásek (př. do _ uxa, musím _ míd _ dúvot)
@	hezitace (je-li hezitace delší, je užito sekvence dvou či tří „zavináčů“, tj. @ @ @)
:-)	smích; odstupňované dle relativní „intenzity“ či délky, např. :-) :-)) :-)))
÷	signalizace otázkové intonace

Obecné schéma strukturálních metatextových značek anotace

§ Téma 1

< čas: HH::MM::SS >

\$A: replika prvního mluvčího

\$B: replika druhého mluvčího

\$A: replika prvního mluvčího < uvnitř je pasáž označená počátkem a koncem, k níž je vztažen komentář > replika prvního mluvčího replika prvního mluvčího ...

< komentář k předchozí replice >

\$B: replika druhého mluvčího

... ..

§ Téma 2

< čas: HH::MM::SS >

\$A: replika replika replika replika replika replika replika replika

\$B: replika replika replika < pasáž repliky, k níž je vtažen následující metatext > replika replika replika replika replika

< metatext k předchozí replice >

\$A: replika replika replika replika replika replika replika replika

\$B: replika replika replika replika replika replika replika replika

... ..

atp.

Modelová ukázka

§ Majitelův pes

< čas: 00:13:50 >

\$B: < něco proved jo ÷ >

< pes >

\$A: dobrí ÷ < ?ušs to zežral ÷ >

< táže se svého psa >

\$B: ?uš to zežral ?ale to nevaďí / tadi to je takový víno

\$A: <1 dobrí to bilo ÷ >1 / <2 d'ivej ted' bude somrovat / d'ivej >2

<1 táže se svého psa >1

<2 ukazuje na psa >2

\$B: nějaká reklama

\$A: no to bilo v novinách

... ..

§ Oslava u babičky – „ochutnávka“

< čas: 00:17:25 >

\$A: < tak počkej ?oxutnáme jo ÷ > / vždyť je to pro xlapi

< pikantní oříšky >

\$B: tadi máte /// no nevím /// na tom to +de bčbje vitáhnož

\$C: < ň+a:kí pomerančovní >

< konzumace čokolády různých chutí >

\$A: ti babičko je to zas to nejtrapnější co máž že ÷

\$B: ježišmarjá to fčera kolega / ?on si libuje f t'ex čokoládáj jag je to devadesád devjet procent

\$C: no to bíý nejedla

\$B: ?ale víš co / ta čokoláda

\$A: já +sem to ?oxutnala / to nepoznáš

\$B: ?ale jakožto má cenu gdiš to nepoznáš ÷

Závěr

Způsob anotace, anotační struktura OMK, umožňuje zaznačit v přepisu jak fonetické aspekty, tak komunikační rysy komunikátů, a disponuje i bohatou škálou externí lingvistické anotace. Díky ortografické standardizaci mluveného projevu lze snáze provést morfologické značkování, než by tomu bylo u hybridního přepisu. Prezentovaná anotace tedy řeší jak správní (vnější), tak strukturní a lingvistické (vnitřní) značkování, představujíc propracovaný a komplexně pojatý anotační systém, jenž by se mohl stát spolehlivou metodologickou bází českých mluvených korpusů.

LITERATURA A PRAMENY

Český národní korpus – PMK, 2001. Ústav Českého národního korpusu FF UK, Praha. Dostupný z World Wide Web: <<http://ucnk.ff.cuni.cz>>.

Český národní korpus – BMK, 2001. Ústav Českého národního korpusu FF UK, Praha. Dostupný z World Wide Web: <<http://ucnk.ff.cuni.cz>>.

Český národní korpus – ORAL, 2006. Ústav Českého národního korpusu FF UK, Praha. Dostupný z World Wide Web: <<http://ucnk.ff.cuni.cz>>.

Nivre J., Allwood J., Grönqvist L., Ahlsén E., Gunnarsson M., Hagman J., Larsson S., Sofkova, S., 2005, *Göteborg transcription standard. Semantics and Spoken Language, Version 6.3 – DRAFT*, Göteborg University. http://www.ling.gu.se/projekt/tal/doc/transcription_standard.html (červen 2005)

Olomoucký mluvený korpus (interní materiály Katedry bohemistiky FF UP Olomouc)

Pořízka P, 2005, Přepis(y) textů v korpusech mluvené češtiny. In *Jazyky v kontaktu/jazyky v konfliktu a evropský jazykový prostor*, Olomouc, s. 235–240.

Pořízka P, 2004a, K možnostem vyhledávání dat a struktury atributů korpusových manažerů v mluvených korpusech ČNK. In *AUPO Philologica 84, Bohemica IX*, Olomouc, s. 81–92.

Pořízka P, 2004b, Představení výzkumného projektu Olomoucký korpus mluvené češtiny. In: *AUPO Moravica I*, Olomouc, s. 125–131.

Pravidla přepisu PMK a BMK (interní materiály ÚČNK a KČJ FF MU).

SAMPA – <http://www.phon.ucl.ac.uk/home/sampa/home.htm> (červen 2005).