Pravidla pro manuální transkripci zvukového signálu na w-rovině Pražského závislostního korpusu mluvené češtiny

Zpracovala: Marie Mikulová

transcription_guidelines
28. 2. 2008

Obsah

Úvod

transcriber, med

Zvukové nahrávky byly transkribovány za použití speciálního anotačního nástroje Transcriber 1.4.1, který je volně k dispozici na htttp://www.ldc.upenn.edu/. Pro transkripci platila následující anotační pravidla:

reprezentace anotace v PDTSC, atributy

Reprezentace anotace

Segment obsahuje sled událostí (`events`).
První událostí každého segmentu je událost `speaker`, za kterou následuje událost `sync`.
Přepis promluvy je zachycen posloupností obsahových událostí `w`, `nonspeech`, `background_begin` a `background_end`, případně `comment`.

Případ, kdy mluvčí mluví přes sebe, je zachycen tak, že v rámci jednoho segmentu je nejprve zachyceno, co řekl jeden mluvčí (sledem obsahových událostí následujících za událostmi `speaker` a `sync`), pak následuje událost `speaker` identifikující druhého mluvčího (a povinně událost sync) a sled obsahových událostí zachycujících souběžnou řeč druhého mluvčího.

Událostí `speaker` je identifikován mluvčí promluvy.
Událost `sync` obsahuje časový údaj, který je odkazem do audio nahrávky. V rámci segmentu (repliky) je umísťována povinně po každé události `speaker` a pak přibližně po každých deseti rozpoznaných obsahových událostí.
Událostí `w` jsou zachyceny rozpoznané slovní tvary.
Událostí `nonspeech` jsou zachyceny rozpoznané neřečové události, které nepřekrývají promluvu (které nejsou hlukem na pozadí).
Událostmi `background_begin` a `background_end` jsou zachyceny rozpoznané hluky na pozadí.
Událost `comment` obsahuje případný komentář k anotaci.

I Segmentace nahrávky

Každá nahrávka je rozdělena na segmenty tak, že **jeden segment odpovídá jedné replice** (`turn`).
Replika je primárně vymezena mluvčím. Změna řečníka znamená vždy začátek nového segmentu.
Jako samostatný segment je označen také každý úsek, kdy řečníci mluví přes sebe. Pak má jeden segment dva a více řečníků.
Ke každému segmentu je přiřazen přepis promluvy, která byla v segmentu vyslovena.
Pokud je v segmentech, kdy mluví řečníci přes sebe, promluva srozumitelná, mělo by být u každého řečníka přepsáno, co řekl.

Příklady:
Přepis plynulého dialogu:
[spk1 *nakonec přece jenom nadešel onen den kdy sem byl propuštěn a to někdy v říjnu já se snažim si o na zapamatovat kdy to bylo*]
[spk2 *štyrycetjedna to bylo*]
[spk1 *byl to rok štyrycetjedna musel to být říjen*]
[spk2 *dobře co ste potom dělal*]

Přepisy dialogu, ve kterém mluvčí mluví přes sebe (najednou bylo vysloveno *v říjnu* a *říjen*):
[spk1 *nakonec přece jenom nadešel onen den kdy sem byl propuštěn a to někdy v říjnu já se snažim si o na zapamatovat kdy to bylo*]
[spk2 *štyrycetjedna to bylo*]
[spk1 *byl to rok štyrycetjedna musel to být*]
[spk1 *říjen* spk2 *v říjnu*]
[spk2 *dobře co ste potom dělal*]

[spk1 *jak se k vám chovali spolužáci jako když kteří věděli o vás že jste žid setkal jste se s projevy*]
[spk1 *nesnášenlivosti v dětství* spk2 *neměl sem*]
[spk2 *v tom směru problémy*]

Přepis dialogu, ve kterém mluvčí mluví přes sebe a jednomu mluvčímu není rozumět:
[spk1 *jak se k vám chovali spolužáci jako když kteří věděli o vás že jste žid setkal jste se s projevy*]
[spk1 *nesnášenlivosti v dětství* spk2 `<unintelligible>`]
[spk2 *v tom směru problémy*]

II Přepis promluvy

V přepisu promluvy mluvčího zachycujeme **slovní jednotky**, které mluvčí vyslovil, a **neřečové události**, které se staly během doby, co mluvčí promlouval.

1 Přepis rozpoznaných slovních jednotek

Při přepisu slovních jednotek používáme tzv. **ortografickou transkripci**, kdy se získané nahrávky přepisují na základě standardních pravopisných zvyklostí daného jazyka (např. se nepoužívají žádné speciální znaky, jen písmena běžné české abecedy). Protože však mluvená řeč má oproti psané formě své zvláštnosti, které nelze vyjádřit pouze pomocí pravopisných pravidel, je třeba k ortografické transkripci přidat ještě další pravidla.

Všechno, co bylo řečeno, se transkribuje slovy, nejsou používány žádné číslice a jiné znaky (například: *%*, *$*).

V transkripci se neužívají žádná interpunkční znaménka.

„Věta“ začíná malým písmenem. Velká písmena používáme jen u vlastních jmen a zkratek (v souladu s pravidly českého pravopisu).
~~Konec věty je označen „.“, „?“ nebo „!“. Kromě těchto interpunkčních znamének je povoleno použít ještě znak „„“ uprostřed věty (v souladu s gramatickými pravidly), žádná další interpunkční znaménka se používat nesmějí.~~

2 Přepis slov, která se jinak vyslovují a jinak píší

2.1 Jediná spisovná varianta výslovnosti

plod [plot], *lev* [lef],
Zapisujeme tak, jak se dané slovo píše

2.2 Více spisovných variant výslovnosti

? *shoda* [schoda, zhoda]
Definujeme výslovnost základní, a sekundární.
Zapisujeme tak, jak se dané slovo píše, pokud bylo vysloveno sekundární variantou výslovností, píšeme tuto variantu do atributu `pronounce_as`.

Příklady:
[ke schodě nedošlo]
ke shodě nedošlo

[ke zhodě nedošlo]
ke shodě <pronounce_as: *zhodě*> *nedošlo*

2.3 Nespisovná výslovnost

vosum, kanička, eště, ňáký

3.1 Nespisovné tvary slov

Pokud řečník mluví nespisovně, přepíše se jeho řeč nespisovně. To znamená, že při přepisu je povoleno používat například slova, *ňákej, ženama* a podobně.
vosum, vozejk, ňákej – výslovnostní problémy!

3.2 Slova vyslovená v cizím jazyce a cizí jména

Slova vyslovená v cizím jazyce a cizí jména je třeba označit a připsat k nim, jak byla skutečně vyslovena. Pokud u korpusů spontánní řeči není možné zjistit správný zápis cizích slov. slova se přepíšou foneticky a výslovnost se k nim nepřipojuje.
Příklad:
anglicky se to řekne [identity card] {ajdentyty kárt}
říkali jsme jim [agrutke] a to znamená vedení

3.3 Zkratky

U korpusů spontánní řeči není povoleno zkracovat přepis promluvy pomocí zkratek typu atd. např. apod., pokud tyto zkratky byly vysloveny jako slova.
Zkratky typu IBM se přepisují tak, že se zapíše správný ortografický tvar (pokud je znám) a ta ním se uvede skutečná výslovnost. Pokud správný ortografický tvar znám není, přepíše se zkratka foneticky.
Příklad:
Vysloveno: týká se to firmy aj bí em
Přepis: týká se to firmy IBM {aj bí em}
Zkratky, které se vyslovují jako slova (například NATO, Benelux, Opec), se přepisují podle stejných zásad jako všechna ostatní slova.

3.4 Přeřeknutí

Pokud se řečník přeřekne tak, že například v nějakém slově přehodí písmena či slabiky a vznikne tak nesmyslné slovo (třeba místo *lokomotiva* řekne *lokotomiva*), je nutné toto slovo označit.
Příklad: jezdila tam parní *lokotomiva*

3.5 Koktání

Jestliže řečník při řeči zadrhává, je třeba slovo, které nebylo vysloveno celé, vyznačit. Stejné označení se využije i v případě, že například z důvodů poruchy na záznamovém zařízení nebylo vyřčené slovo zaznamenáno celé.
Příklad: bylo to během mého dět- dětství

III Záznam neřečových událostí

Všechny zvuky, které jsou zaznamenány v nahrávce a není to řeč, by měly být označeny. Těmto zvukům se říká neřečové události a mohou to být jednak zvuky vydávané samotným řečníkem (smích, kašlání, nádech, mlaskání apod.) a jednak zvuky z řečníkova okolí (skřípění židle, šustění papíru, bouchnutí dveří apod.).
Jestliže neřečová událost nastává současně s nějakým slovem, mělo by to být v anotaci názorně vyznačeno.
Příklad: Neřečová událost „nádech“ nenastává současně s žádným slovem:
Zatím není jasněé, [LOUD_BREATH] jestli budou následovat obvinění zodpovědných úředníků.
Příklad: Neřečová událost „bouchnutí dveří“ nastává současně se slovem „budou“:
Na hřebenech hor budou [<DOOR_SLAM] přeháňky sněhové.
Na hřebenech hor [DOOR_SLAM>] budou přeháňky sněhové.
(Oba zápisy jsou rovnocenné.)
Příklad: Přes část „mínus tři až mínus“ je slyšet blíže nespecifikovatelný šum:
Noční teploty [NOISE/] mínus tři až mínus [/NOISE] sedum stupňů.

Pouze vlastní jména a akronyma jako IBM, NATO jsou zapsána s velkými písmeny. Je-li nějaké slovo ve výpovědi spelováno, jsou jednotlivá spelovaná písmena zapsána velkými písmeny a oddělena mezerou.
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>. Viz Table 5.1, "Přehled značek pro neřečové události".
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 tabulce 5.1 Přehled značek pro neřečové události je uveden přehled všech značek užívaných při transkripci pro neřečové události.
Tabulka 5.1 Přehled značek pro neřečové události

<click>	mlaskání jazykem
<mouth>	mlaskání rty
<cough>	kašlání
<laugh>	smích
<breath>	zvuk dechu
<inhale>	nádech
<UH>	UH
<UM>	UM
<UH-HUH>	UH-HUH
<UH-HUM>	UH-HUM
<unintelligible>	nesrozumitelný úsek
<noise>	hluk v pozadí
<noise_begin>	začátek hluku v pozadí
<noise_end>	konec hluku v pozadí
<silence>	ticho, pauza

V tabulce 5.2 Ukázka transkripce je ukázka transkribovaného textu.

Tabulka 5.2 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