Instrukce pro přepisování mluvených projevů pro korpusy řady ORTOFON

0 Úvod

V programu ELAN je projev každého mluvčího přepisován do zvláštních stop (tzv. partiturový přepis). Pro každého mluvčího jsou vytvořeny 3 stopy: **ort**, **fon** a **meta**. Stopy ort a fon si navzájem odpovídají, jsou stejně členěny a liší se pouze druhem přepisu. Ve stopě ort jde o přepis převážně ortografický (viz instrukce pro stopu ort), ve stopě fon o přepis převážně fonetický (zaznamenává, jak byla slova skutečně vyslovena, viz instrukce pro stopu fon). Ve stopě meta (viz instrukce stopy meta) jsou zaznamenány ostatní neverbální zvuky a komentáře ke skutečnostem, které jsou ze zvukové nahrávky zřejmé a jsou důležité pro to, co bylo řečeno (např. to, že někdo mluví ke zvířeti).

Při přepisování nahrávky se nejprve pořizuje ortografický přepis ve stopě ort (viz instrukce pro stopu ort) a v ní se také provádí základní rozčlenění zvuku (tzv. segmentace). Tato segmentace je stejná i pro stopu fon, stopa meta je segmentována jinak. Při ortografickém přepisu je možné stopy fon skrýt (viz Skrytí a zobrazení přepisovacích stop), umožní to lepší orientaci mezi zápisem projevu jednotlivých mluvčích.

K celé nahrávce patří ještě další dvě stopy: **META** a **anom** (viz instrukce pro stopy meta). Do stopy META jsou zapisovány neverbální zvuky a jiné metatextové informace, které nelze přiřadit k žádnému mluvčímu (např. štěkot psa). Stopa **anom** slouží k označení úseků, které mají být z nahrávky odstraněny. Jsou to úseky obsahující osobní data, která nemohou být zveřejněna (jako je např. příjmení, telefonní číslo atp.), protože by mohla vést k určení totožnosti nahrávaného. Takové údaje jsou v přepise anonymizovány (kódovány) a ze zvuku budou následně vymazány.

Pokud se v nahrávce objeví ojedinělá promluva jiné osoby či jiných osob, vytvoříme pro ni zvláštní stopu **JO** (viz instrukce pro stopy meta, odd. 2.3 Jiná osoba nebo hlas v nahrávce).

STOPA	VZTAH	PŘEPIS
ort	mluvčí	ortografický
fon	mluvčí	fonetický
meta	mluvčí	neverbální zvuky a komentáře k příslušnému mluvčímu
META	nahrávka	neverbální zvuky a komentáře k celé nahrávce
JO	nahrávka	ortografický přepis replik jiných mluvčích
anom	nahrávka	anonymizace osobních dat

tab. 1 Přehled přepisovacích stop.

Postup přepisu

Přepis probíhá ve dvou fázích. V první fázi nasegmentujete a přepíšete stopy ort, meta a META (a případně též stopy anom a JO). Stopu fon v této fázi nepřepisujte, i když je součástí šablony. Soubor s přepisem a zvukový soubor vložíte do databáze a předáte koordinátorovi ke kontrole (viz Vkládání sond). Koordinátor předá soubor administrátorovi, a pokud je vše v pořádku, sonda je uzavřena a předána specialistovi na přepis stopy fon. Pokud je editor zároveň i tímto specialistou, uzavřená sonda

se mu objeví znovu v databázi s předpřipraveným přepisem stopy fon: do stopy fon je zkopírován přepis stopy ort a jsou v něm provedeny některé automatické opravy, které mají tento přepis usnadnit: jsou změněna písmena nepoužívaná ve stopě fon: *ě*, *y*, *ý*, *ů*, *q*, *x*; slabiky di, ti, ni jsou převedeny na *ďi*, *ťi*, *ňi*, velká písmena (kromě *N*, které je součástí anonymizačních kódů) jsou změněna na malá. Tyto změny nejsou spolehlivé a je třeba je při přepisu stopy fon zkontrolovat a případně opravit (zejména v cizích slovech, např *politik* bude převedeno na *poliťik*).

Pokud jste tedy specialistou na přepis fon (což předpokládá předchozí fonetické vzdělání či dlouhou praxi v přepisu dat do mluvených korpusů), ve druhé fázi si stáhnete soubor s přepisem z databáze a věnujete se přepisu stopy fon. Po dokončení přepisu vložíte do databáze už pouze soubor s přepisem a předáte ho opět koordinátorovi. Přepis stopy fon kontroluje jiný koordinátor, proto vám sonda během této kontroly načas zmizí z databáze. Pokud specialistou nejste, už se o sondu nemusíte dále starat.

Následující pokyny se týkají způsobu zápisu do stop ort a fon.

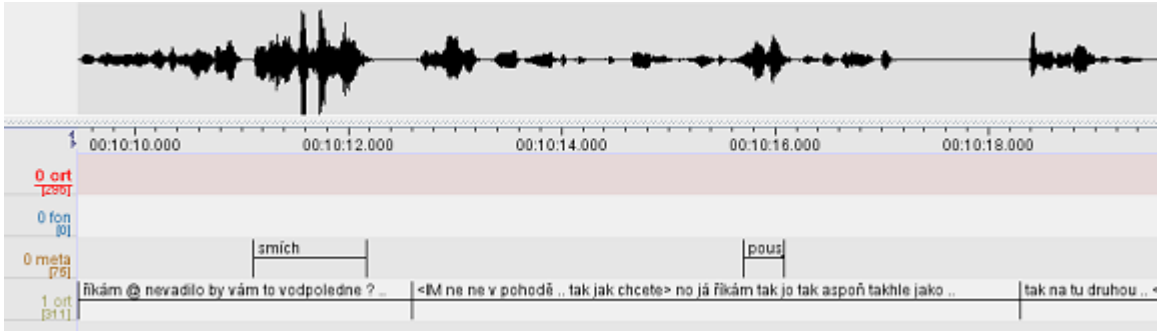
---

1 Segmentace

Při přepisu používáme pojmy replika a segment. **Replika** je úsek, ve kterém mluví jeden mluvčí, dokud ho v hovoru nevystřídá mluvčí jiný nebo dokud nenastane dlouhá pauza. V přepisu se repliky člení na **segmenty**: některé repliky jsou tvořeny pouze jedním, jiné mohou být poměrně dlouhé, proto je rozdělujeme na menší části (viz dále).

Při přepisu nahrávky do stopy ort je nutné nejprve provést **segmentaci** (viz Segmentace a anotace). **Segmentací** se rozumí rozdělení zvukové stopy na jednotlivé úseky a vytvoření jim odpovídajících částí (segmentů) ve stopě ort. Do těchto segmentů se pak přepisují příslušné zvukové úseky. Každý segment může obsahovat maximálně 20 slov (včetně symbolu pro hezitační zvuky a citoslovce). Symbol pro počet nesrozumitelných slov, např. (3), považujeme za jedno slovo.

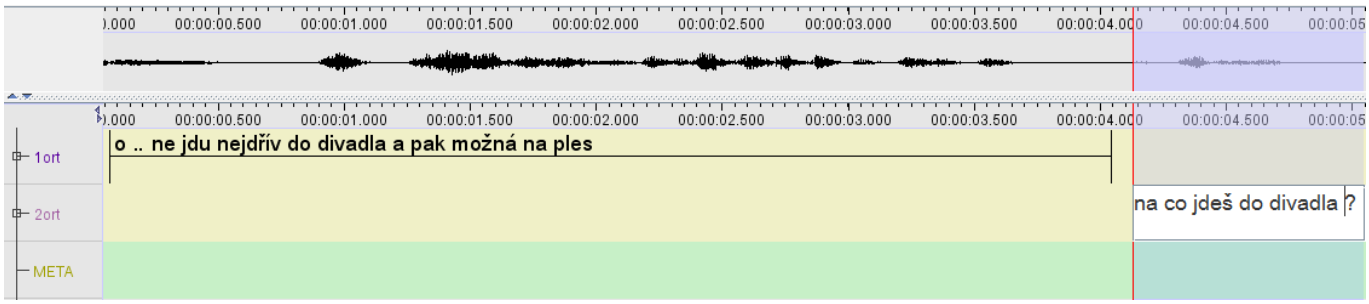
Je-li to možné, volíme hranici segmentů v pauze, příp. předělu. V případě mluvy, ve které se nevyskytují ani pauzy, ani předěly, lze v segmentu ponechat maximálně 25 slov. Jednotlivé segmenty o délce 20 slov (resp. 25 slov), které dohromady tvoří dlouhou nepřerušovanou repliku jednoho mluvčího, na sebe těsně navazují, tzn. sousední segmenty mají společnou hranici. Nevydělujeme tedy pauzy ani předěly jako samostatný segment. Budou také označeny v textu anotace – to znamená, že když vytvoříme hranici segmentu v místě pauzy, na konci anotace bude symbol pro pauzu (..) (viz obrázek č. 1)



obr. 1 Segmentace jedné nepřerušované repliky.

Segmenty ve stopách ort a fon si odpovídají. Stopa fon je závislá na segmentaci provedené ve stopě ort, a proto v ní nelze velikost ani počet segmentů měnit. Pokud se při přepisu ve stopě fon zjistí, že je třeba posunout hranice segmentu, musí se to udělat ve stopě ort.

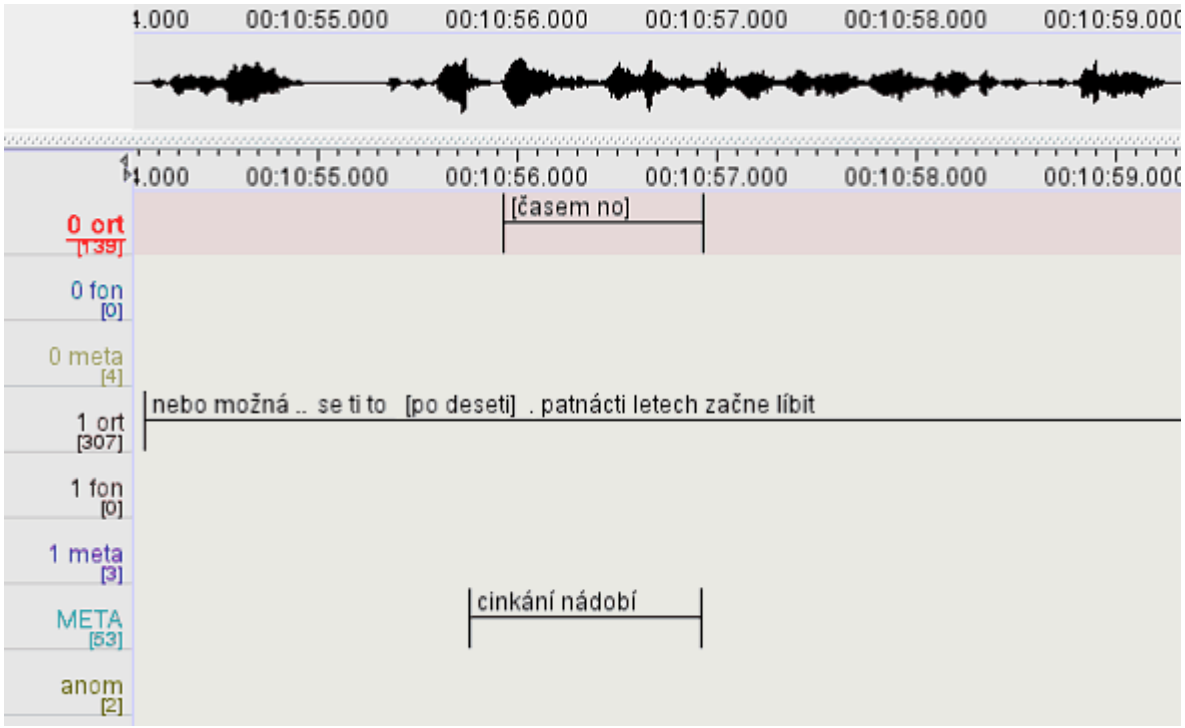
Pokud mluvčímu „skočí do řeči“ jiný mluvčí (nebo jiní mluvčí) a hovoří oba (nebo všichni) současně, také tuto část promluvy, která se překrývá s hovorem jiných mluvčích, přepisujeme stále do téhož segmentu, patří totiž do jedné repliky (více viz Překryvy). Je třeba bedlivě sledovat, aby byla ve zvukové stopě vybrána skutečně celá replika a aby segment nebyl časově posunutý. V přepisu nesmí chybět žádné slovo ani jeho část. Pokud poslední slovo segmentu končí na souhlásku nebo skupinu souhlásek, bývá předchozí samohláska vyslovena s delším trváním nebo koncové souhlásce předchází nepatrná pauza; zejména v těchto případech dávejte pozor na to, aby byla do segmentu zahrnuta i tato souhláska na konci slova. Také dbejte na to, aby příslušný úsek byl přepsán u toho mluvčího, který ho skutečně vyslovil.



obr. 2 Dvě na sebe navazující repliky (stopy fon a meta jsou skryté).

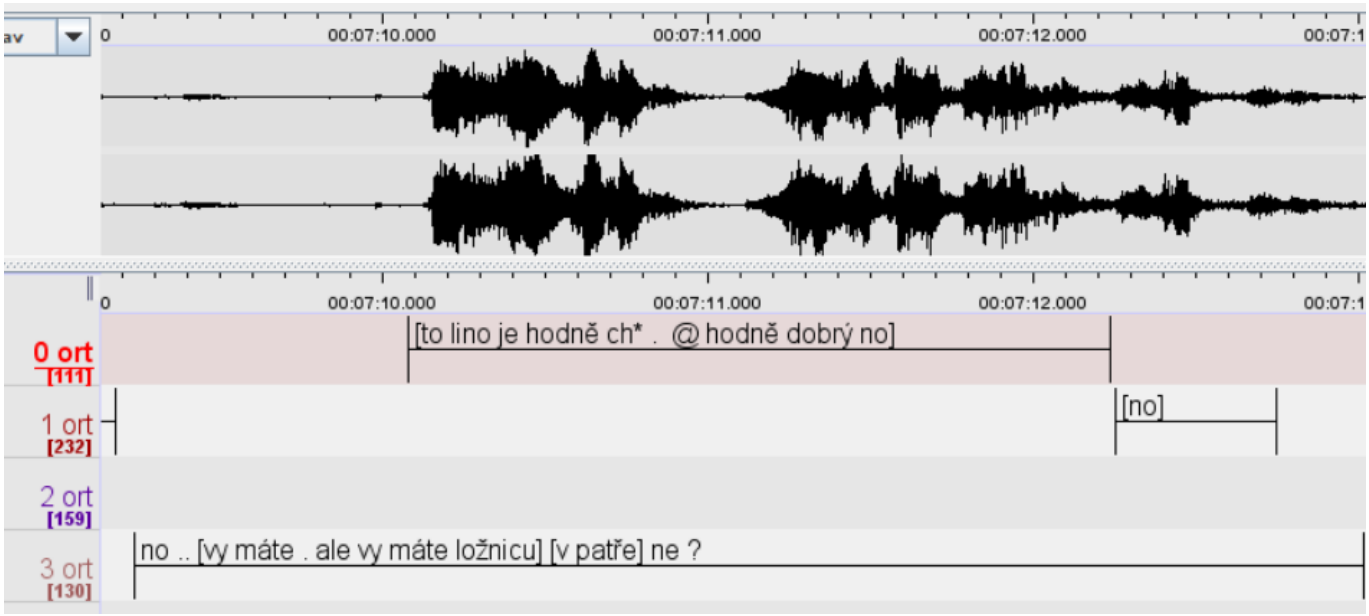
1.1 Překryvy

Překryvy jsou úseky, ve kterých mluví dva anebo i více mluvčích současně. Tyto úseky se nevydělují do zvláštního segmentu, jejich přepis je součástí repliky konkrétního mluvčího. Překrývající se slova je třeba v každé replice v přepise ort i fon přesně označit pomocí hranatých závorek [], které zapisujeme bez mezery. Označujeme vždy celá slova nikoliv slabiky. Značení překryvů se týká i hezitací @ a citoslovcí zapsaných kódem &. Nezapomeňte označit překryvy ve všech replikách, kterých se to týká.

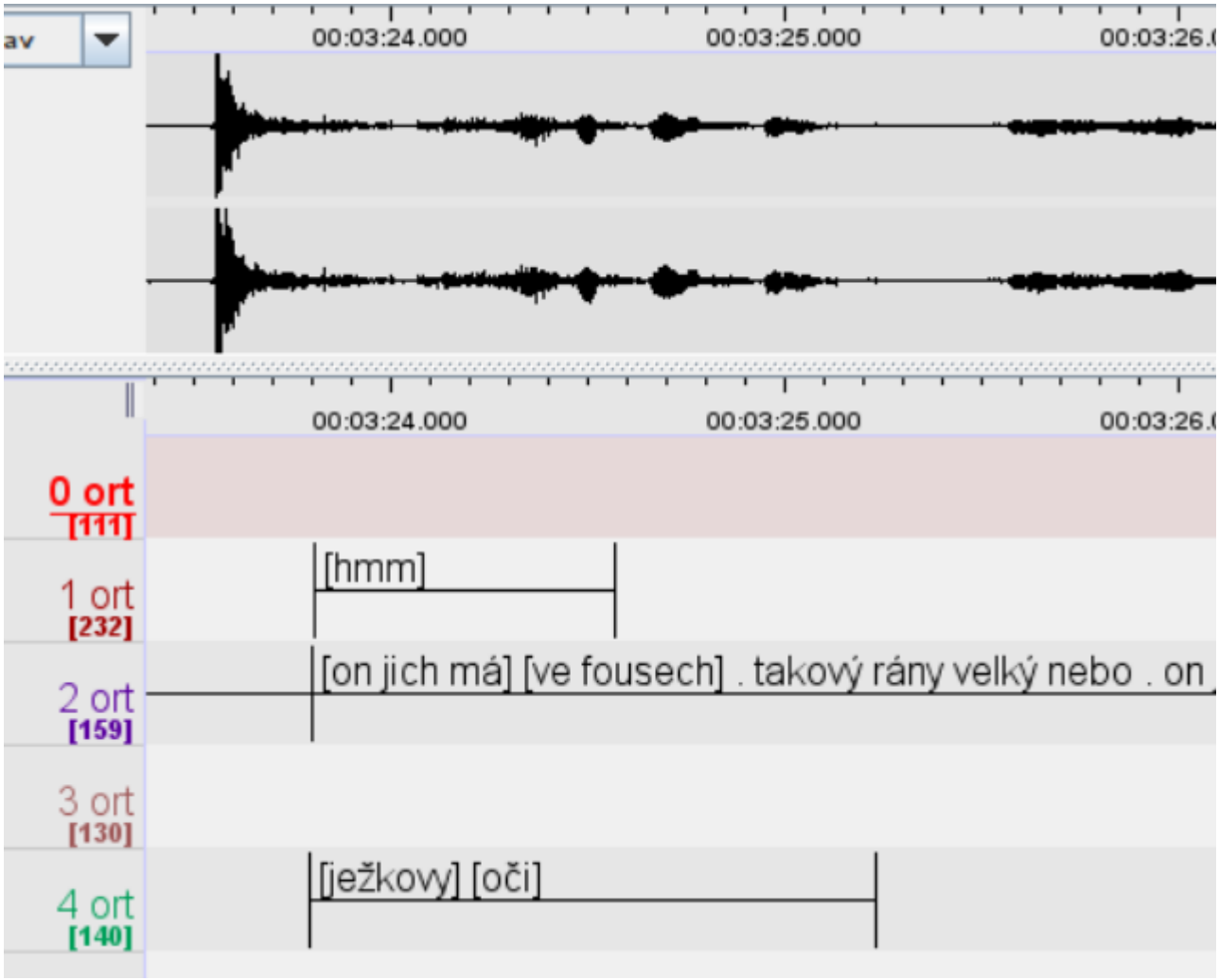


obr. 3 Vyznačení překryvů ve stopě ort u jednotlivých mluvčích.

Pokud v nahrávce vystupuje tři a více mluvčích, může dojít k překryvu nejen mezi dvěma, ale i mezi třemi či více mluvčími. I v takových případech se snažíme zahrnout do hranatých závorek celá slova u všech překrývajících se mluvčích. Snažíme se oddělit překryvy podle toho, který mluvčí s kterým se překrývá (viz obr. 4 a 5).



obr. 4 Vyznačení překryvů mezi více než dvěma mluvčími ve stopě ort.



obr. 5 Vyznačení překryvů mezi více než dvěma mluvčími ve stopě ort.

Přerušení a nedokončení repliky (znaménka + a -) zapisujeme vždy až za hranatou závorkou. V zachyceném překryvu, tj. uvnitř hranatých závorek, se mohou objevit znaky jiných závorek (jako např. kulaté závorky značící nesrozumitelnou výpověď, špičaté závorky signalizující metainformaci), znaky pro hezitaci, citoslovce.

Překryvy se stopami meta pomocí hranatých závorek nevyznačujeme: ve stopě meta vytvoříme příslušný segment a ve stopě ort ani fon nic neznačíme.

2 Anonymizace

Některé části nahrávky je třeba anonymizovat. To znamená, že příslušný anonymizovaný údaj se nesmí objevit v žádném přepise a ze zvukové stopy musí být odstraněn. V ortografickém i fonetickém přepise se vyslovený údaj nahradí tzv. anonymizačním kódem.

Povinně jsou kódována příjmení (NP) s výjimkou známých osobností (politiků, herců, sportovců a podobně). Podle rozhodnutí přepisujícího či na přání nahrávaných osob mohou být anonymizovány i další údaje: vlastní jména, přezdívky, názvy míst, telefonního čísla (kód vkládáme místo posledního dvoučíslí). Pokud je nutné anonymizovat jiné jméno (např. název firmy, občanského sdružení) nebo jiný údaj (mailovou adresu, webové stránky, adresu), použijeme kód NO.

V přepisech se podle typu údaje používají následující typy anonymizačních kódů:

Anonymizační kód	Anonymizovaný údaj
NP	příjmení
NJ	osobní jméno
NN	přezdívka
NM	název místa
NO	ostatní vlastní jména
NT	telefonní číslo

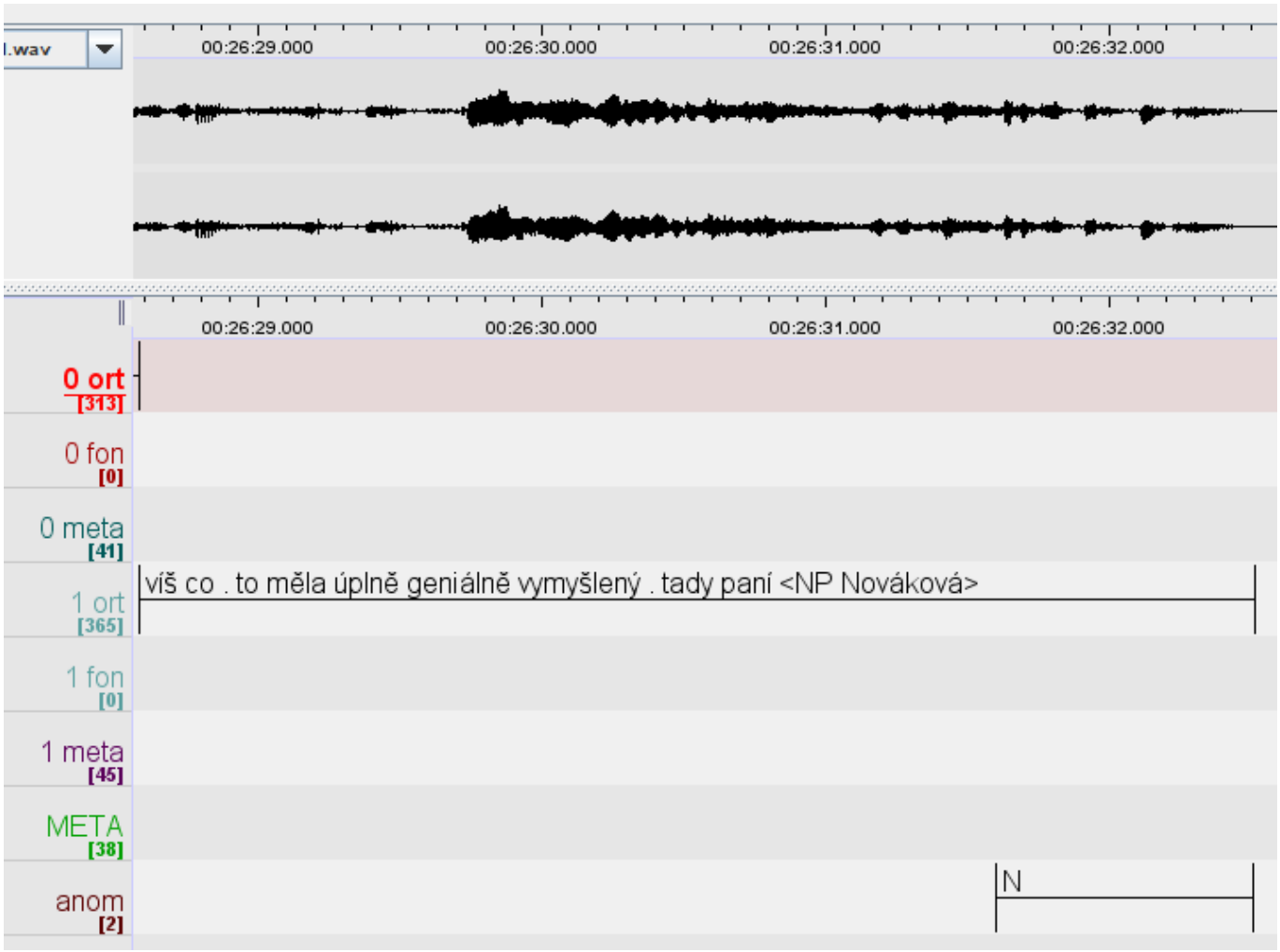
tab. 2 Seznam anonymizačních kódů.

2.1 Anonymizovaný údaj v překryvu

Pokud se anonymizovaný údaj nachází v překryvu, zapíšeme odpovídající část vyslovenou druhým mluvčím do kulatých závorek jako špatně srozumitelný úsek, protože bude také odstraněna ze zvuku.

2.2 Segment anonymizovaného údaje

K tomu, aby byl údaj odstraněn i ze zvukové stopy, slouží pomocná stopa **anom**. V té je třeba vytvořit segment, který co nejpřesněji odpovídá části zvuku, v níž bylo vysloveno jméno. Do tohoto segmentu vepíšeme N, tato část bude z nahrávky odstraněna spolu s celou pomocnou stopou. Obr. 6 zachycuje anonymizaci příjmení, které vyslovil první mluvčí. Příjmení je ve stopách 1 ort a 1 fon přepsáno a opatřeno anonymizačním kódem NP (např. <NP Nováková>) a ve stopě anom kódem N. Ve stopě anom se segmentuje pouze kódované jméno, proto je segment kratší než ve stopách ort a fon, kde je kód pro příjmení součástí repliky.

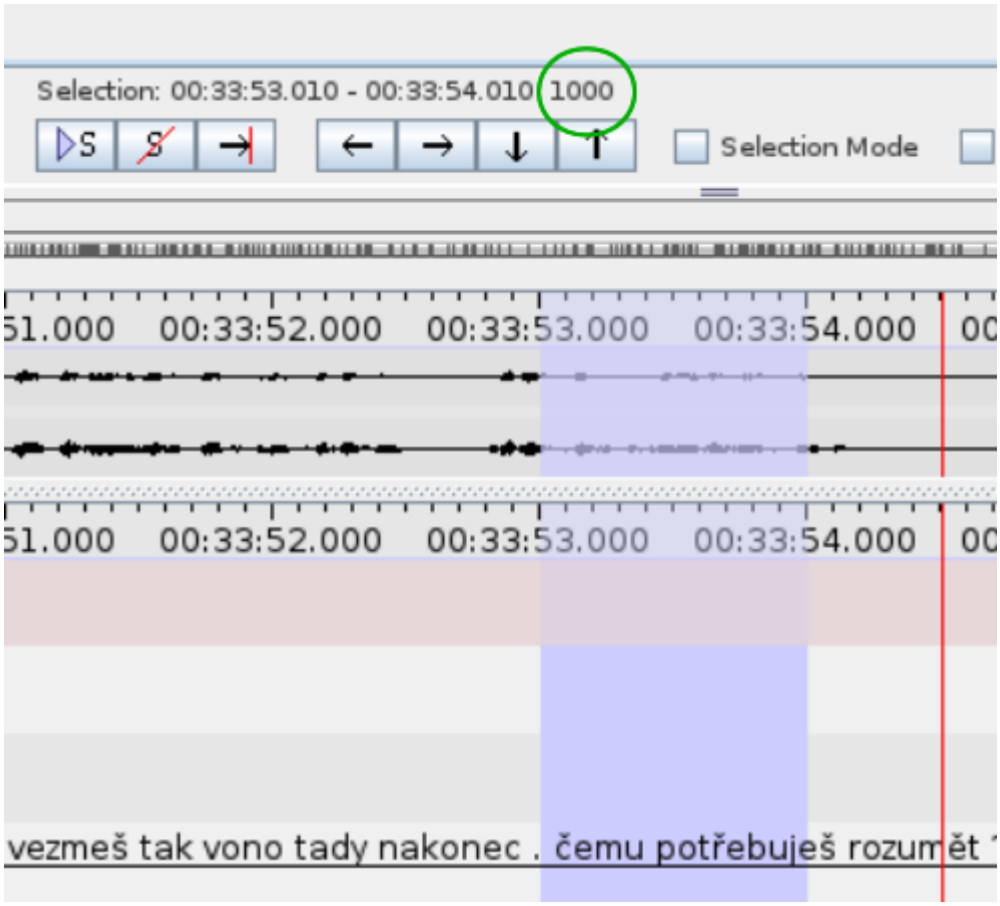


obr. 6 Anonymizace jména ve stopě META a kódování příjmení ve stopách ort a fon.

3 Pauzy a předěly

3.1 Rozdíl mezi pauzou a předělem

Pauzou rozumíme alespoň **120 ms ticha** (obr. 7 popisuje, kde v ELANu tuto informaci zjistit), případně neřečových zvuků jako je **nádech** apod., v rámci projevu jednoho mluvčího (jiní mluvčí po tuto dobu samozřejmě mluvit mohou), během nichž promluvu přeruší a následně naváže (nedojde tedy k vystřídání mluvčích). **Předěl** nastává tehdy, kdy mezi dvěma částmi promluvy pociťujeme hranici (z důvodů intonace apod.), aniž by mezi nimi reálně nastala pauza (okamžik ticha/neřečových zvuků). Pauzy jsou tedy dány objektivně, naopak značení předělů je do jisté míry subjektivní. Předěly proto aktivně nevyhledávejte: značte je tehdy, kdy se vám na první poslech zdá, že jde o pauzu, ale při bližším ohledání zjistíte, že v nahrávce není potřebný úsek ticha.



obr. 7 Když v ELANu v Annotation Mode dáte část nahrávky do výběru (modře podbarvený úsek), její trvání se zobrazí v milisekundách na místě označeném zelenou elipsou (zde tedy 1000 ms, tj. 1 s).

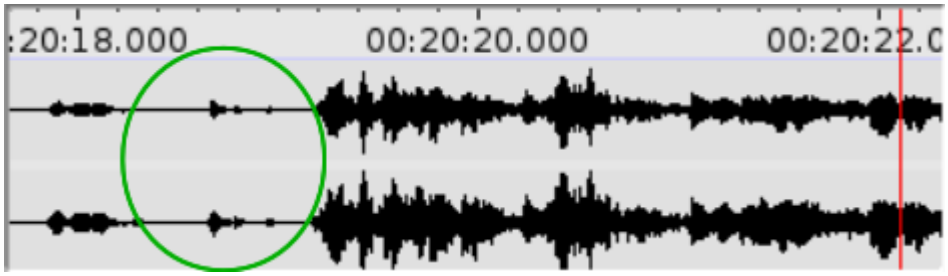
3.2 Zápis pauz a předělů ve stopě ORT

Pauzy v mluveném projevu značíme v přepise **dvěma tečkami a předěly tečkou jednou**, v obou případech oddělenými mezerou z obou stran (např. *takže když k sobě někoho seženu tak tak půjdu ale . kdybych nesehnal řekněme někdy .. kolem třeba šestý hodiny*). Pokud se během rozhovoru vyskytne **dlouhá pauza (více než 2 s)**, oddělíme ji do zvláštního segmentu v příslušné stopě meta a vybereme z nabídky dlouhá pauza (více viz instrukce pro přepisování stopy meta).[1] Pauzy zapisujeme skutečně tam, kde se vyskytnou. Nenechte se ovlivnit syntaxí psaného textu, tj. tím, kde by se podle pravidel pro psaný text měly psát čárky a tečky. Nenechte se ovlivnit ani intonací či prozodií obecně: když mluvčí jedno slovo zdůrazní, může nastat percepční dojem pauzy před slovem následujícím, aniž by se na onom místě mluvčí reálně odmlčel. Takový jev potenciálně představuje předěl, nikoli pauzu. Podmínkou pro zapsání pauzy je, že se mluvčí skutečně alespoň na chvilku odmlčí.

Příklady:

- audioukázka : *hmm .. nəa tobude ňáka schúze*

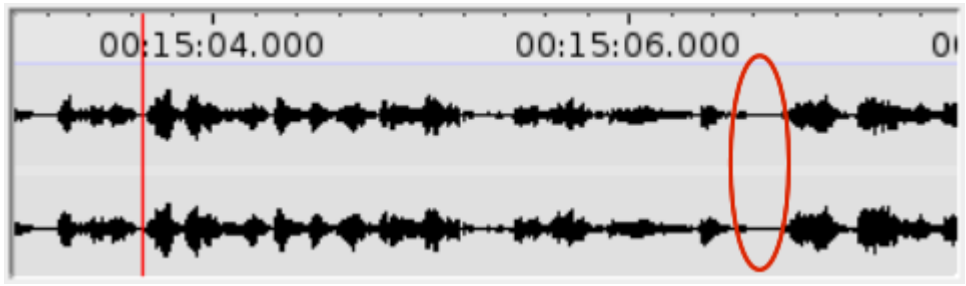
Tady pauza je, a je vidět i na zvukové vlně (viz část vlny na obr. 8 označená zelenou elipsou; drobné zvlnění uprostřed je způsobeno chrastěním klíčů); ovšem při vizuální kontrole je třeba dát pozor na to, aby úsek ticha nepředstavoval závěrovou fázi ražené hlásky ([p, b, t, d, ť, ď, k, g]) – ta pochopitelně pauzou není (viz další bod).



obr. 8 Příklad pauzy.

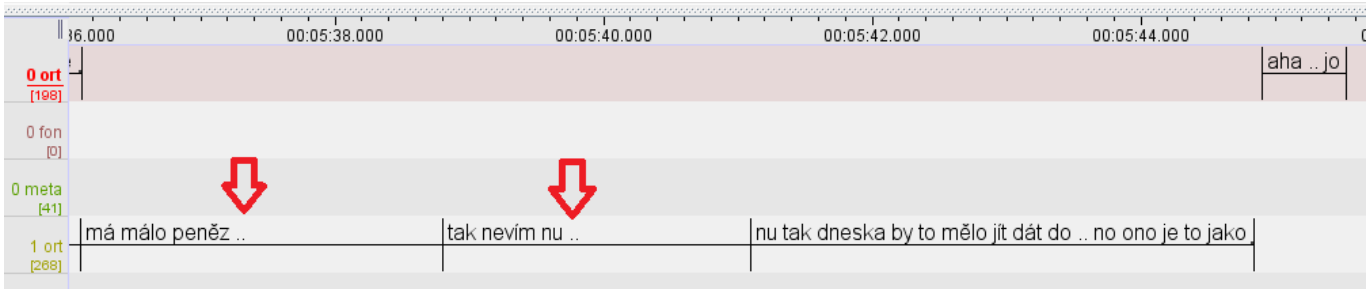
- audioukázka : *tuatam əcə zaplaťim aleten dlúch cotambil tenemúžu plaťit . noa toje jednou podruhí*

Zde není ani jedna pauza, ale jedno místo k jejímu zaznamenání obzvlášť svádí. Zaprvé, mezi “plaťit” a “noa” je něco, co je pauze alespoň vizuálně hodně podobné (viz část zvukové vlny na obr. 9 označená červenou elipsou), ale jedná se pouze o poněkud delší závěrovou (“tichou”) fázi při vyslovování ražené hlásky [t]. Zadruhé je na tom samém místě výrazný intonační pokles, který může vytvářet percepční dojem pauzy, byť tam pauza reálně není. V podobných případech výrazného dojmu pauzy bez opory v signálu je vhodné sáhnout k vyznačení předělu.



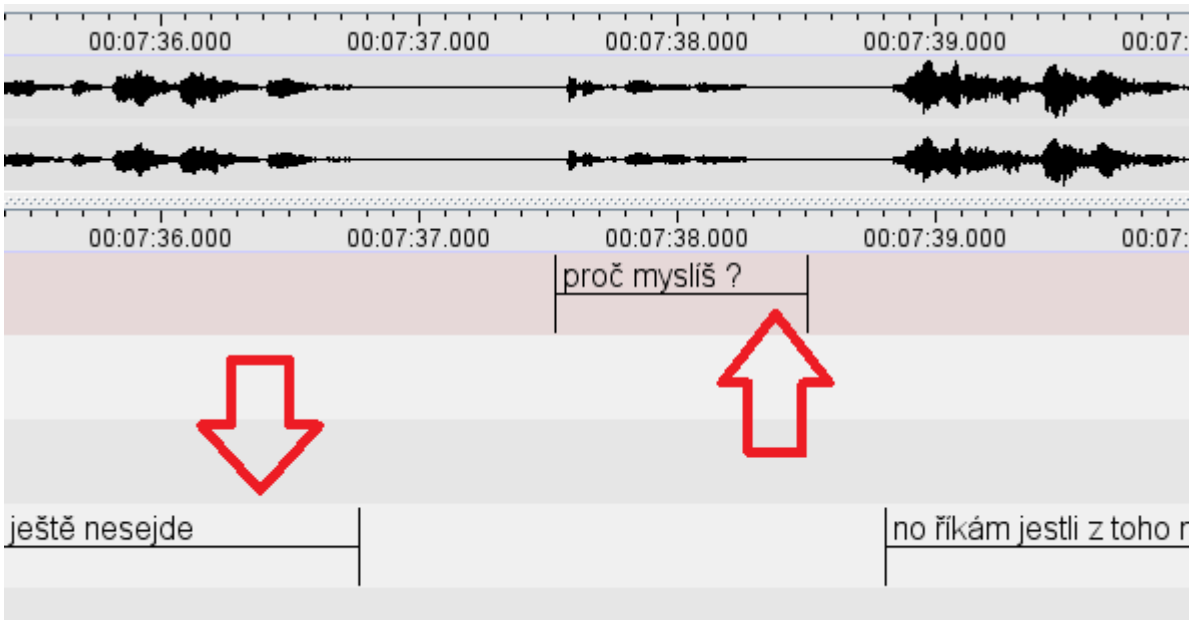
obr. 9 Příklad tiché pasáže ve zvukové vlně, která ovšem není pauzou, ale závěrovou fází ražené hlásky.

Pauzy a předěly zaznamenáváme pouze v rámci repliky (tj. nepřerušené řeči jednoho mluvčího). Repliky bývají často delší, rozdělují se na segmenty po max. 25 slovech. Hranice mezi segmenty by měla být právě v místech pauzy či předělu, protože tak hrozí nejmenší riziko, že segment nebude obsahovat celá slova nebo hlásky. Podle toho pak značíme pauzy a předěly, tj. na koncích segmentů, nikdy ne na začátku nového segmentu (viz šipky na obr. 10).



obr. 10 Zaznamenávání pauz v segmentech tvořících repliku jednoho mluvčího.

Při střídání mluvčích je mezi jejich replikami často pauza nebo předěl. Ty však do žádného segmentu nezaznamenáváme (viz šipky na obr. 11).



obr. 11 Zaznamenávání pauz při střídání mluvčích.

Technická poznámka: V případě potřeby (zejména u delších segmentů) je v ELANu i v Transcription Mode možné „zazoomovat“ na část zvukové vlny a zobrazit si ji tak ve větším detailu: stačí na zobrazení vlny najet myší, zmáčknout na klávesnici Ctrl a otočit kolečkem myši. Může se také stát, že je nahrávka celkově hodně potichu a celý segment na pohled vypadá jako pauza (rovná čára): pak je možné kliknout na zvukovou vlnu pravým tlačítkem myši a pomocí nabídky Vertical Zoom „roztáhnout“ vlnu i ve vertikálním směru.

4 Přerušení a nedokončení repliky

Replika mluvčího může být přerušena druhým mluvčím (“skočí mu do řeči”), současně se ale nejedná o překryv. To znamená, že oba mluvčí nemluví současně, první pouze neukončí svou repliku. Takové přerušení zaznamenáváme a rozlišujeme při tom, zda první mluvčí na svou promluvu znovu navázal nebo zda zůstala nedokončena. Pokud mluvčí během hovoru “přeruší sám sebe” a začne mluvit o jiném tématu, přerušení neznačíme, přepisujeme dál do jednoho segmentu jako jednu repliku.

4.1 Přerušení repliky

Přerušení repliky, po kterém mluvčí znovu pokračuje, označujeme znaménkem **plus** + (odděleným mezerou) za posledním slovem, které mluvčí pronesl. Pokud se jedná o nedořečené slovo, může + následovat i po signálu neukončeného slova, kterým je **hvězdička** * (*nevím co dě** +). Znaménko plus napíšeme i před slovo, které navazuje na přerušenou repliku (+ *teď dělá*).

Příklady:

M1: *sem taková prostší tak* +

M2: *no ne jako . že nejsou tak křupavý*

M1: + *nevím co se dělá*

M1: *ně* někdy do pěti to maj asi jako vyklidit jako +*

M2: *no no no .*

M1: + *definitivně . ale . jednou tady taky*

V hovoru může nastat situace, kdy první mluvčí udělá pauzu (zamyslí se, nemůže si vybavit potřebný výraz) a repliku za něj dokončí jiný mluvčí. I tento případ budeme označovat znaménkem plus za posledním slovem prvního mluvčího a před prvním slovem druhého mluvčího, který repliku dopoví. Pokud příslušné slovo první mluvčí zopakuje, napíšeme plus i před opakované slovo.

M1: *že pokud s ní ten chlap nechce mít žádnej +*

M2: + *vážnej poměr*

M1: + *vážnej poměr no prostě . tak nechce*

Pokud repliku dokončí oba mluvčí současně (tj. v překryvu), nezapomeneme překrývající se slova vyznačit do hranatých závorek. Znaménko + se v tomto případě u jednoho mluvčího objeví uprostřed repliky.

M1: *že pokud s ní ten chlap nechce mít žádnej .. + [vážnej poměr] no prostě . tak nechce*

M2: + [*vážnej poměr*]

4.2 Nedokončení repliky

V případě, že se přerušený mluvčí po replice druhého mluvčího už ke své promluvě nevrátí a začne jiné téma, mlčí nebo mluví další mluvčí, označíme přerušené místo spojovníkem (-) odděleným od posledního slova mezerou.

Příklad:

M1: *tam měli nějaký problémy . s tou vzduchotechnikou . a -*

M2: *aha*

M1: *hele a co to retro ?*

PLUS I MÍNUS ZAPISUJEME PO MEZEŘE, NIKDY NE PŘÍMO KE SLOVU!

Pokud se přerušení či nedokončení repliky objeví v nesrozumitelném úseku, zapisujte znaky plus a mínus až za mezerou následující po koncové kulaté závorce.

5 Nesrozumitelné a špatně srozumitelné části

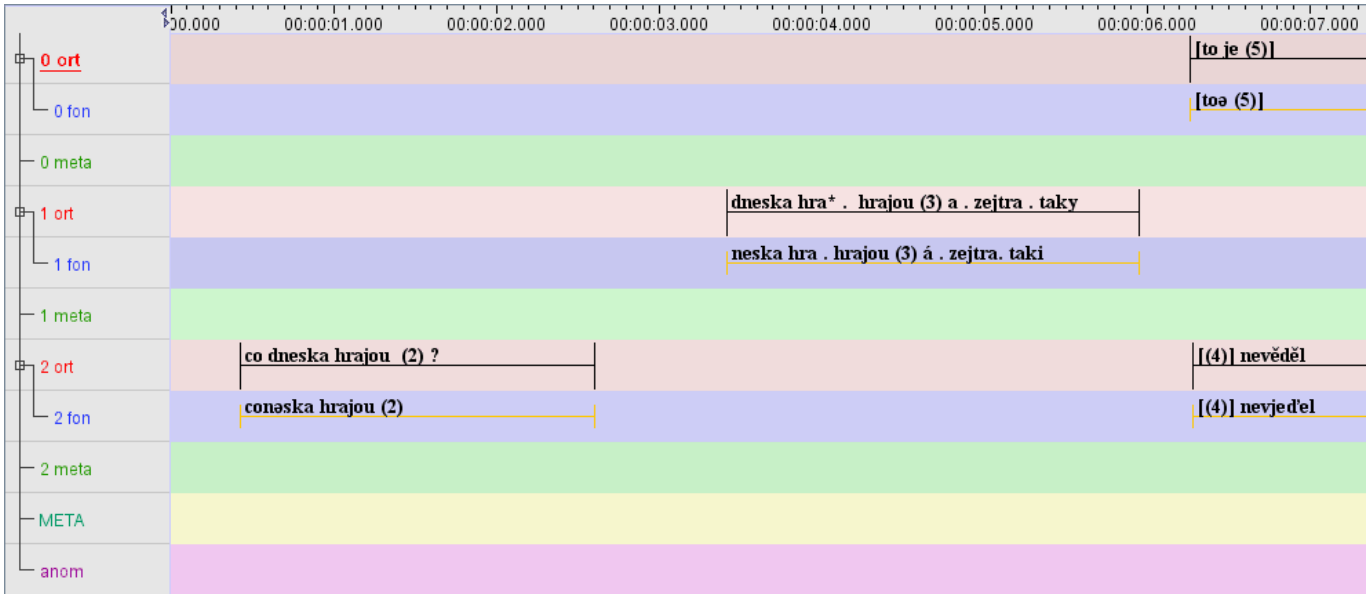
5.1 Nesrozumitelný úsek

Nesrozumitelný zvukový úsek označujeme v přepise kulatými závorkami a uvnitř nich uvádíme číslem počet slov, která byla v nesrozumitelném úseku vyřčena, např. (3). Počet slov musíme mnohdy pouze odhadnout. Snažíme se, aby takových úseků bylo v přepise nahrávky co nejméně; trpělivý, opakovaný poslech často umožní porozumění.

Příklady:

večer pošlu zprávu (2) protože já nevím je hroznej jako a nevím . (5) nějakej jinej

Pokud se takový úsek objeví v překryvech a není možné přepsat část repliky žádného z mluvčích, vepíšeme kulaté závorky s počtem nesrozumitelných slov ke všem mluvčím, kteří se účastní rozhovoru.



obr. 12 Zachycení nesrozumitelného úseku.

5.2 Špatně srozumitelný úsek

Podobně špatně srozumitelná slova, která částečně odhadujeme, ale přitom si nejsme úplně jisti jejich zněním, zapisujeme do kulatých závorek:

kecáš je to . (prej) dobrý

Jako špatně srozumitelný úsek označujeme rovněž části přepisu v překryvech, které se kryjí s anonymizovanými úseky (viz anonymizace).

6 Zápis závorek

V přepise užíváme následující závorky:

- **hranaté [ ]**, ty označují překryvy (viz Překryvy);

- **kulaté ( )**, ty označují nesrozumitelné úseky s odhadnutým zněním nebo počtem slov (viz Nesrozumitelné a špatně srozumitelné části);
- **špičaté <SM>** s příslušným dvojpísmenným kódem, ty ohraničují úseky vyslovené s určitým doprovodným rysem (viz Stopa meta)

Pouze ve stopě fon používáme pro zdůraznění slova nebo slabiky závorky složené {} (viz Stopa fon).

Často v přepise nastane situace, ve které se závorky překrývají nebo zanořují. Pokud na začátku stojí dvoje závorky, zapisujte jako druhé ty, které dříve skončí. Mezi jednotlivými závorkami nesmí být mezera. Dávejte pozor, aby u všech závorek byl označen začátek i konec.

Pro označení stejného úseku dvojími závorkami používejte následující zápis:

- špičaté závorky vždy uvnitř: *[<SM jo>]; (<SM jo>); [(<SM jo>)];*
- kulaté uvnitř hranatých: *[(jo)]*

Několik příkladů použití závorek ve stopě ort:

Vnořené závorky

1. špatně srozumitelný úsek v překryvu: *[do (toho) se mi nechce]; [(nevíš) v kolik to bylo] ?*
2. nesrozumitelný úsek v překryvu: *[(4) potom mu to došlo]*
3. překryv, nesrozumitelný a špatně srozumitelný úsek v části řečené se smíchem: *<SM (2) tomu (nebudeš [věřit])>*
4. všechna slova v překryvu řečená se smíchem: ani *[<SM tomu sám nechtěl věřit>] že -*

Závorky postupně začínající a končící

1. čtený text, jehož část je v překryvu špatně srozumitelná: *tady stojí <CT vytvořte účet . [stáhněte (program> . a kam) to jako uložím] ?*
2. zívání s překryvem: *<ZV a tak jsem se s ním [domluvil> na dnešek]*

Zdůraznění slova ve stopě fon

1. zdůraznění slabiky v překryvu *[{slá}va]*
2. zdůraznění slova v úseku, který je vysloven nahlas *<NH to nesmíš . {ne} >*

[1] Segment dlouhé pauzy na stopě meta **nesmí** probíhat zároveň se segmentem s promluvou jednoho z mluvčích: když k překryvu dojde, je potřeba segment s promluvou rozdělit.