Rekonstrukce standardizovaného textu z mluvené řeči v Pražském závislostním korpusu mluvené češtiny

Manuál pro anotátory

Zpracovala: Marie Mikulová

Abstrakt

Dokument obsahuje pravidla pro manuální anotaci, kterou je třeba provést při budování závislostního korpusu mluveného jazyka. Tato anotace spočívá v tzv. rekonstrukci standardizovaného textu z mluvené řeči, tj. původní segmenty mluvené řeči, mnohdy velmi vzdálené gramaticky správným větám, se zde popsaným způsobem převádí do takové „standardizované" podoby, na kterou již je možné uplatnit další anotační pravidla (přidávající zejména informaci o syntaktické struktuře věty).

Anotační manuál je určen anotátorům Pražského závislostního korpusu mluvené češtiny, ale lze jej chápat jako obecný návod pro podobně pojatou anotaci mluveného korpusu kteréhokoli jazyka.

speech_reconstruction_guide
31. 3. 2008

Obsah

Poznámka:
V textu je několikrát zmíněn tektogramatický manuál. Jedná se o dokument: Mikulová a kol.: Anotace na tektogramatické rovině pražského závislostního korpusu. Anotační manuál. TR-2005-28. ÚFAL MFF UK Praha, 2005.
www: http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/cz/t-layer/html/index.html

1 Základní principy anotace

Práci anotátora při rekonstrukci standardizovaného textu z mluvené řeči lze přirovnat k redaktorovi, který zpracovává nahraný rozhovor k otištění v časopise: rozhovor dostává psanou podobu (tj. dodržuje pravidla psané řeči) a jeho výsledná podoba musí být potenciálnímu čtenáři nejen srozumitelná, ale musí se tomuto čtenáři i dobře číst.

Výstupem anotace je tzv. **standardizovaný text**, který vymezujeme na základě následujících podmínek:

- text neobsahuje neřečové události,
- 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.

1.1 Reprezentace anotace

Podrobný popis reprezentace anotace je k dispozici v technické zprávě TR-2006-33 (ÚFAL MFF UK Praha, 2006) a na www-stránkách projektu. Zde uvádíme jen základní principy s ohledem na potřeby anotačního manuálu.

1.1.1 Roviny anotace

Při anotaci rekonstrukce standardizovaného textu z mluvené řeči pracujeme s korpusem minimálně o dvou anotačních rovinách, v Pražském závislostním korpusu mluvené češtiny počítáme ale s tím, že korpus má tři hierarchicky uspořádané roviny.

1.1.1.1 Z-rovina

Z-rovina je nejnižší rovina korpusu. Obsahuje automaticky rozpoznané a automaticky segmentované promluvy.

1.1.1.2 W-rovina

W-rovina zachycuje manuálně transkribovaný text promluvy, tj. to, co mluvčí řekl, včetně všech přeřeknutí, zakašlání, pauz apod.
Základními jednotkami w-roviny jsou tzv. události, z nichž (nejen) pro rekonstrukci standardizovaného textu jsou nejdůležitější tzv. **obsahové události**, kterými jsou zachyceny:
- rozpoznané slovní tvary (tokeny, w-uzly typu **`w`**)
- rozpoznané neřečové události (w-uzly typu **`nonspeech`**)
- rozpoznané hluky na pozadí (w-uzly typu **`background`**)

Události (w-uzly) jsou na w-rovině segmentovány do **replik** (turn). Replika je primárně vymezena jedním mluvčím (při překrývání mluvčích však může mít replika mluvčích více).

1.1.1.3 M-rovina

M-rovina obsahuje standardizovaný text, na kterém se následně provede morfologická anotace (text pak může být anotován na vyšších syntaktických rovinách).

Základními jednotkami m-roviny jsou slovní jednotky (slovní tvary, čísla, interpunkce) reprezentované **m-uzly typu `m`**. Speciálními **m-uzly typu `nontext`** jsou pak zachyceny další obsahově relevantní jevy mluvené řeči (zejména neřečové události).
M-uzly jsou segmentovány do vět reprezentovaných tzv. **s-elementy**.

Anotátor tvoří standardizovanou podobu promluvy na m-rovině korpusu různými úpravami manuální transkripce zachycené na w-rovině. V některých případech, kdy manuální transkripce na w-rovině není k dispozici, je třeba nejprve takovou transkripci vytvořit, tj. manuálně opravit automaticky rozpoznané a segmentované promluvy, zachycené na z-rovině korpusu. Pravidla manuální transkripce segmentů mluvené řeči na w-rovině nejsou součástí tohoto manuálu; jsou částečně popsána v TR-2006-33. ÚFAL MFF UK Praha, 2006 a kompletně budou zpracována v samostatném manuálu.

1.1.2 Vztahy mezi jednotkami m-roviny a w-roviny

Rozdíly, kterými se vstupní segmenty manuálně transkribované mluvené řeči (zachycené na w-rovině) liší od svých standardizovaných podob na m-rovině, tj. provedené modifikace, jsou zachyceny ve vztazích mezi oběma rovinami, ve vztazích mezi jednotkami m-roviny (m-uzly) a jednotkami w-roviny (obsahovými událostmi, w-uzly).

Z m-uzlu, kterému odpovídá nějaký w-uzel na w-rovině, vede na tento w-uzel odkaz.

Jádro odkazů mezi m-rovinou a w-rovinou tvoří odkazy mezi m-uzly typu `m` (reprezentujícími tokeny na m-rovině) a w-uzly typu `w` (reprezentujícími tokeny na w-rovině).
O vztazích mezi m-uzly typu `m` a w-uzly typu `w` platí následující tvrzení.

Z m-uzlu typu `m` nemusí vést žádný odkaz do w-roviny.

M-uzel typu `m`, ze kterého nevede žádný odkaz do w-roviny, nazýváme **vložený m-uzel**; reprezentuje vložené gramaticky a obsahově nezbytné slovní jednotky, kterým neodpovídá žádný w-uzel typu `w` (token) na w-rovině (viz *4.2.1.2 Vkládání*).

Na w-uzel typu `w` nemusí vést žádný odkaz z m-roviny.

W-uzel typu `w`, na který nevede žádný odkaz z m-roviny, představuje vymazané obsahově nerelevantní slovní jednotky (viz *4.2.1.1 Mazání*). Hovoříme o **vymazaném w-uzlu**.

Pořadí m-uzlů typu `m` na m-rovině nemusí odpovídat pořadí w-uzlů typu `w` na w-rovině.

Změny ve slovosledu (viz *4.2.1.4 Změny ve slovosledu*) jsou zachyceny rozdílným uspořádáním uzlů na obou rovinách.

V případě domyšleného nesrozumitelného úseku textu (viz *5.3 Nesrozumitelný úsek textu*) vedou z m-uzlů typu `m` odkazy na w-uzel typu `nonspeech`, který má v atributu `desc` hodnotu `unintelligible`.

Jiným typem odkazů jsou odkazy z m-uzlů typu `nontext` (reprezentujících obsahově relevantní neřečové události) na w-uzly typu `nonspeech` (reprezentující neřečové události), případně na w-uzly typu `background` (reprezentující hluky na pozadí; viz *4.2.2 Zachycení obsahově relevantních neřečových událostí*).
O vztazích mezi m-uzly typu `nontext` a w-uzly typu `nonspeech`, případně `background` platí následující tvrzení.

Z m-uzlu typu `nontext` nemusí vést žádný odkaz do w-roviny.

Pokud z m-uzlu typu `nontext` nevede žádný odkaz do w-roviny, pak zachycuje obsahově relevantní neřečovou událost, která nebyla zachycena w-rovině (např. důraz na slově, šepot).

Na w-uzel typu `nonspeech` a typu `background` nemusí vést žádný odkaz z m-roviny.

Pokud na w-uzel typu `nonspeech` a typu `background_begin` nevede žádný odkaz z m-roviny, pak neřečová událost reprezentovaná tímto w-uzlem je z hlediska m-roviny obsahově nerelevantní, případně byl její význam zachycen prostředky psaného textu.

Přehled odkazů z m-uzlů na w-uzly

Typ m-uzlu	**Odkazované typy w-uzlů**
m-uzel typu `m`	w-uzel typu `w`
	w-uzel typu `nonspeech` (`desc` = `unintelligible`)
	Ø
m-uzel typu `nontext`	w-uzel typu `nonspeech`
	m-uzel typu `background_begin`
	Ø

Poznámka k typování odkazů do w-roviny:
Odkazy z m-uzlů na w-uzly nejsou při manuální anotaci typovány. Typy odkazů budou do anotovaných dat doplněny automaticky po skončení anotace.

Zatím počítáme s následujícími typy odkazů mezi m-uzly a w-uzly:
A. Typy odkazů z m-uzlů typu `m` na w-uzly typu `w`:
- **`basic`**: forma m-uzlu se rovná tokenu w-uzlu nebo dochází pouze k tzv. ortografickým modifikacím (viz *4.1 Ortografické modifikace*)
- **`num`**: ortografické modifikace čísel (viz 4.1.3 *Standardizace čísel*)
- **`substitution`**: forma nebo lema m-uzlu bylo vůči odpovídajícímu w-uzlu upraveno (byla provedena substituce – viz *4.2.1.3 Substituce*)

B. Typy odkazů z m-uzlů typu `nontext` na w-uzly typu `nonspeech` a `background`:
- **`nonspeech`**

1.1.3 Atributy věty

M-uzly jsou na m-rovině segmentovány do vět reprezentovaných tzv. **s-elementy**.

Každému s-elementu náleží atributy:
- **`w-speaker.rf`**: identifikace mluvčího, který dané obsahové sdělení pronesl. Atribut bude vyplněn automaticky po skončení anotace.
- **`is_modified`**: atribut určuje, zda věta reprezentovaná s-elementem byla nebo nebyla (musela nebo nemusela být) vůči odpovídajícímu segmentu na w-rovině modifikována. Atribut bude vyplněn automaticky po skončení anotace.
- **`stype`**: druh obsahu dané věty. Atribut je manuálně anotován (viz *3 Typy vět podle obsahu*).

Z každého s-elementu vedou dva (netypované) odkazy do w-roviny: na první a poslední obsahovou událost patřící do rekonstruované věty.

Odkazy s-elementu (`w-begin.rf`, `w-end.rf`) určují, jaký úsek w-roviny byl použit jako vstup pro rekonstruovanou větu reprezentovanou s-elementem (viz *2.1 Vyznačení hranic vět v proudu mluvené řeči*).

1.2 Anotační postup

Pro rekonstrukci standardizovaného textu z mluvené řeči je stanoven následující **anotační postup**:

1. Přečíst manuální transkripci mluvené řeči zachycenou na w-rovině.
2. Pokud je význam textu nejasný či nejednoznačný, poslechnout si odpovídající zvukový záznam textu.
3. Provést segmentaci textu do vět (viz *2 Větná segmentace*).
4. Pomocí modifikací mazání, vkládání, substituování a přesouvání slovních jednotek vytvořit větu splňující podmínky standardizovaného textu a zachovávající principy anotace (viz *4 Modifikace textu*).
5. Zkontrolovat odkazy do w-roviny (od m-uzlů i od s-elementu).
6. Označit typ věty (viz *3 Typy vět podle obsahu*).
7. Po dokončení anotace souboru – přečíst výsledný standardizovaný text a provést případné další úpravy.

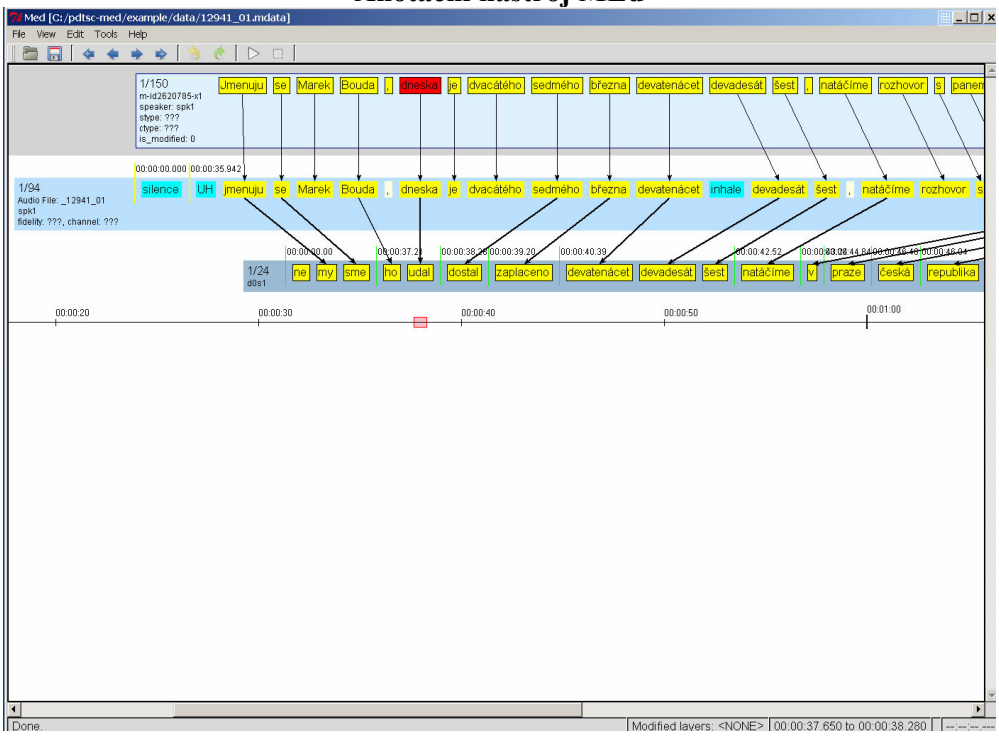
1.3 Anotační nástroj MEd

Anotace se provádí ve speciálně vyvinutém anotačním nástroji MEd. V anotačním nástroji je v hlavním anotačním okně zobrazena z-rovina, w-rovina a m-rovina a jejich vzájemné propojení. Pod jednotlivými rovinami korpusu je znázorněna časová linka reprezentující audio nahrávku (viz obr. *Anotační nástroj MEd*).

Anotační nástroj umožňuje:

- segmentovat proud řeči (manuální transkripci) do větných celků, přiřazovat atributy větným celkům;
- přesouvat libovolně slovní jednotky na m-rovině z hlediska jejich pořadí ve větě;
- slovní jednotky vymazat, vložit, spojit, jinak modifikovat, včetně změny formy nebo lematu;
- propojit m-uzel na m-rovině s odpovídajícími w-uzly na w-rovině tak, aby bylo zřejmé, se kterými jednotkami na w-rovině daný m-uzel souvisí (ze kterých "vznikl"), a případně určit typ propojení;
- poslech původní audio nahrávky, který je často nutný v případech, kdy ani původní transkripce (například vzhledem k absenci prozodické informace, informace o délce pauz a vzhledem k další "ztrátě informace"), ani její kontext neumožňuje anotátorovi rozhodnout o vhodné modifikaci.

Anotační nástroj MEd



Podrobný popis anotačního nástroje MEd není součástí tohoto manuálu.

2 Větná segmentace

Segmentace mluvené řeči na z-rovině je vždy výsledkem automatické procedury v rámci použitého rozpoznávače mluvené řeči, primárně je (automatickou procedurou) provedena podle výskytu (delšího) úseku nějaké neřečové události. Výsledné segmenty zhruba odpovídají větám, ne však nutně. W-uzly jsou na w-rovině segmentovány pouze do replik. V rámci repliky není žádná další segmentace provedena. Skutečná segmentace do vět tedy nastává až v rámci rekonstrukce standardizovaného textu na m-rovině.

Při rekonstrukci standardizovaného textu jsou vytvářeny větné celky, které odpovídají obvyklým pravidlům pro psaný text. Výsledná (rekonstruovaná) věta, která může být i neúplná (jde-li například o nedokončenou myšlenku), musí odpovídat jednomu ze čtyř typů klauzí popsaných v tektogramatickém manuálu (v sekci *Slovesné a neslovesné klauze*), tj. musí jít o:

- **slovesnou klauzi** (i elidovanou),
- **nominativní klauzi**,
- **citoslovečnou klauzi**,
- **vokativní klauzi**,

nebo o spojení jedné nebo více těchto klauzí.

Příklady:

řekla to dobře ➞ *Řekla to dobře.*

{kdy přijdeš} v pátek odpoledne ➞ *V pátek odpoledne.*

pane Barňák ➞ *Pane Barňák!*

pryč s fašisty ➞ *Pryč s fašisty!*

ach ano ➞ *Ach, ano.*

Věta na m-rovině je obsahové sdělení (tj. má nějaký obsah), bezobsažné úseky textu (obsažené v proudu mluvené řeči a zachycené na w-rovině) nemají na m-rovině svůj protějšek. K tomu viz více v *2.2.5 Bezobsažný úsek textu*.

2.1 Vyznačení hranic vět v proudu mluvené řeči

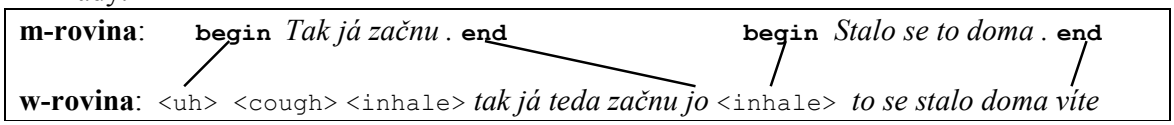
Na m-rovině je větou posloupnost m-uzlů, která je identifikovaná tzv. s-elementem. Tato posloupnost vždy odpovídá nějakému úseku (případně i úsekům) rozpoznaných obsahových událostí na w-rovině. Tento úsek obsahových událostí, který byl použit jako vstup pro výstupní rekonstruovanou větu reprezentovanou s-elementem, je určen pomocí dvou odkazů do w-roviny.

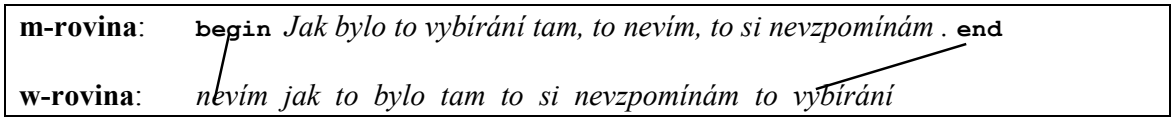
Z každého s-elementu vedou dva odkazy do w-roviny: odkaz na první a poslední obsahovou událost, která byla použita pro rekonstruovanou větu.

Odkazy ze dvou různých s-elementů se mohou křížit (v případě překrývání mluvčích, viz *2.2.4 Vzájemné přerušování mluvčích*), na w-rovině mohou být obsahové události, které

nebyly použity jako vstup pro žádnou rekonstruovanou větu (viz *2.2.5 Bezobsažný úsek textu*).

Příklady:





2.2 Určování hranic klauzí a vět

2.2.1 Hranice klauzí

Při určování hranic klauzí se 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.

Příklady:

sešli jsme se <noise> *v Praze* <noise>

➜ *Sešli jsme se v Praze.*

sešli jsme se <noise> *v Praze* <noise> *na Vyšehradě* <noise>

➜ *Sešli jsme se v Praze na Vyšehradě.*

sešli jsme se <noise> *v Praze* <noise> *já a Pavel* <noise>

➜ *Já a Pavel jsme se sešli v Praze.*

2.2.2 Hranice vět (spojování vět v souvětí)

Při spojování vět v (souřadná) souvětí platí, že nevytváříme příliš dlouhá souvětí. Standardem jsou **souvětí o dvou až třech větách hlavních**. Pokud mluvčí překotně, bez přerušení, dlouho mluví (neklesá hlasem, nedává signál o konci věty, stále používá spojku *a*), rozčleníme takový proud mluvené řeči na několik kratších souvětí.

Opakovaně používaná spojka *a* (případně *pak* aj.) je ze standardizovaného textu odstraněna (viz i *4.2.1.1.4 Nadbytečné konektory*).
Poznámka: Odstraňovaná spojka *a* patří do věty následující.

Příklad:

a pak jsme šli | a já už nevím jak dlouho | a jak sme tam došli tak se to stalo a to byl konec všech nadějí

➜ *Pak jsme šli. Už nevím jak dlouho. Jak jsme tam došli, tak se to stalo a to byl konec všech nadějí.*

Další příklady:

<uh> *na dětství si určitě stěžovat nemohu moji rodičové byli* <inhale> *velmi hodní a tolerantní já sem na straně druhé tak tolerantní k nim nebyl a* <inhale> *patřil sem k těm dětem který jim* <noise> *nadělaly dost starostí se domnívám* <inhale> *zejména potom v pozdějších létech kdy sem* <inhale> *byl již to čemu se dá říci politicky činný* <inhale> *těch starostí u rodičů přibývalo zejména po okupaci* <inhale> *Československa*

➜ *Na dětství si určitě stěžovat nemohu, moji rodičové byli velmi hodní a tolerantní.*
Já jsem na straně druhé k nim tak tolerantní nebyl a domnívám se, že jsem patřil k těm dětem, které rodičům nadělají dost starostí.
Zejména potom v pozdějších létech, kdy jsem byl již to, čemu se dá říci politicky činný, těch starostí u rodičů přibývalo, zejména po okupaci Československa.

myslím že to odhodlání Čechů nebo tohoto národa <inhale> *které vedlo až k dvěma mobilizacím byl takový že* <uh> *ten optimismus* <inhale> *a to* <noise> *spoléhání* </noise> *na pomoc tehdejších spojenců Anglie Francie* <inhale> *i Sovětského svazu bylo tak veliké že sme se cítili jaksi přece jenom bezpeční za tou za* <noise> *tou českou Mažinotovou linií že jo za těmi* <uh> *za těmi* <noise> *pevnostmi* <inhale> <noise> *s tou prakticky dobře vycvičenou armádou* <inhale> <noise> *a jak pozdější historické výzkumy ukázaly tak to tento optimismus byl oprávněný*

➜ *Myslím, že odhodlání Čechů nebo tohoto národa, které vedlo až k dvěma mobilizacím, bylo takové, že optimismus a spoléhání na pomoc tehdejších spojenců, Anglie, Francie i Sovětského svazu, byly tak veliké, že jsme se cítili jaksi přece jenom bezpeční za tou českou Mažinotovou linií, za těmi pevnostmi s prakticky dobře vycvičenou armádou, a jak pozdější historické výzkumy ukázaly, tak tento optimismus byl oprávněný.*

2.2.3 Nedokončené výpovědi

Pokud výpověď mluvčího evidentně nebyla dokončena, například proto, že jej druhý mluvčí přerušil, ale může se tak stát i z vlastní vůle mluvčího, pak nedokončené výpovědi necháváme nedokončené. Nedokončení výpovědi naznačíme **třemi tečkami na konci výpovědi**.

Příklady:

[spk1] *v období když ste byl v Palestině měl ste nějakou korespondenci s Československem s rodičema nebo*
[spk2] *s Československem ne ale s rodiči jo pomocí červeného kříže kde sem se také dozvěděl že byli deportováni*

➜

[spk1] *V období, když jste byl v Palestině, měl jste nějakou korespondenci s Československem, s rodiči nebo...*
[spk2] *S Československem ne, ale s rodiči ano, pomocí Červeného kříže, kde jsem se také dozvěděl, že byli deportováni.*

[spk2] *mně přes- vždycky říkali přesný termín kdy budu propuštěn nikdy sem v tom daném termínu propuštěn nebyl*
[spk1] *jak ste se dostal teda nakonec*
[spk2] *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*
[spk1] *štyrycetjedna to bylo*
[spk2] *byl to rok štyrycetjedna musel to být*
[spk1] *v říjnu* [spk2] *říjen*
[spk1] *dobře co ste potom dělal*

➜

[spk2] *Mně vždycky říkali přesný termín, kdy budu propuštěn. Nikdy jsem v tom daném termínu propuštěn nebyl.*
[spk1] *Jak jste se dostal teda nakonec...*
[spk2] *Nakonec přece jenom nadešel onen den, kdy jsem byl propuštěn, a to někdy v říjnu. Snažím se vzpomenout si, kdy to bylo.*
[spk1] *Bylo to 1941.*
[spk2] *Byl to rok 1941.*
[spk1] *V říjnu.*
[spk2] *Musel to být říjen.*
[spk1] *Dobře, co jste potom dělal?*

2.2.4 Vzájemné přerušování mluvčích

Pokud se mluvčí vzájemně přerušují (skáčou si do řeči, mluví přes sebe), pospojujeme výroky obou vzájemně se přerušujících mluvčích **do ucelených výpovědí**.
V krajních případech (je-li to vhodné) výpovědi do ucelených vět nespojujeme, ale naznačíme vzájemné přerušování mluvčích: nedokončení výpovědi zachytíme třemi tečkami na konci, navázání na dříve přerušenou výpověď naznačíme třemi tečkami na začátku výpovědi.

Příklady:

[spk1] `<inhale>` *jak se k vám chovali spolužáci jako když kteří věděli o vás že jste Žid setkal* `<inhale>` *jste se s*
[spk1] *projevy nesnášenlivosti v dětství* [spk2] *neměl sem*
[spk2] *neměl sem v tom směru problémy*
[spk1] *v žádném směru*
[spk2] *ne*
[spk1] *ani nadávky nějak nesetkal jste se*
[spk2] *ne no to spíš při dětských hrách v parku* `<inhale>` *když sem byl příliš úspěšný ve hře na kuličky tak prohrávající považoval za nutné mi* `<inhale>` *nadat do Židů*

➜

[spk1] *Jak se k vám chovali spolužáci, kteří o vás věděli, že jste Žid? Setkal jste se s projevy nesnášenlivosti v dětství?*
[spk2] *Neměl jsem v tom směru problémy.*
[spk1] *V žádném směru?*
[spk2] *Ne.*
[spk1] *Ani s nadávkami jste se nesetkal?*
[spk2] *Ne, to spíš při dětských hrách v parku, když sem byl příliš úspěšný ve hře na kuličky, tak prohrávající považoval za nutné mi nadat do Židů.*

[spk1] <inhale> *po devítiletce po obecné škole jste začal studovat na gymnáziu*
[spk2] *na reálném*
[spk1] *jaké to tam bylo* [spk2] *gymnáziu ano*
[spk1] <inhale> *co jste tam dělal*
[spk2] <inhale> *na tom reálném gymnáziu který patřilo ke klasickým německým gymnáziím v Praze* <inhale> *byl jak sem již* <uh> *předeslal* <inhale> *velmi silný levý proud*

➜

[spk1] *Po devítiletce, po obecné škole jste začal studovat na gymnáziu.*
[spk2] *Na reálném gymnáziu, ano.*
[spk1] *Jaké to tam bylo, co jste tam dělal?*
[spk2] *Na tom reálném gymnáziu, které patřilo ke klasickým německým gymnáziím v Praze, byl, jak sem již předeslal, velmi silný levý proud.*

[spk1] *jak vono to je s tim počasim teďka má bejt*
[spk2] *já sem se nedívala, na počasí se dívá táta a řikal že*
[spk1] *teďka má bejt teplo a pak už zas zima*
[spk2] *že tři dni teplo a ale neska se můžeš podivat*
[spk1] *no, jestli to stihnu*

➜

[spk1] *Jak je to teď s počasím? Teď má být teplo a pak už zase zima?*
[spk2] *Já jsem se nedívala, na počasí se dívá táta a říkal, že tři dny teplo. Dneska se ale můžeš podívat.*
[spk1] *No, jestli to stihnu.*

Varianta se zachyceným přerušováním mluvčích (přednost má varianta první):
[spk1] *Jak je to teď s počasím? Teď má být...*
[spk2] *Já jsem se nedívala, na počasí se dívá táta a říkal, že...*
[spk1] *...teplo a pak už zase zima?*
[spk2] *...tři dny teplo. Dneska se ale můžeš podívat.*
[spk1] *No, jestli to stihnu.*

2.2.5 Bezobsažný úsek textu

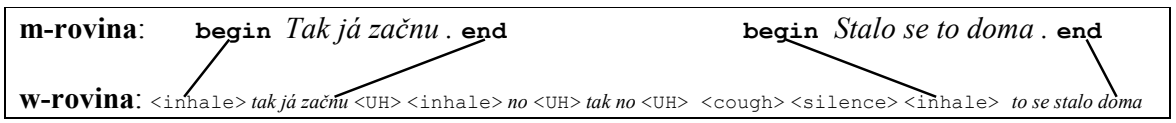
V proudu mluvené řeči se mohou někdy objevit i delší úseky, které nemají žádný obsah. Jde zpravidla o posloupnosti neřečových událostí a neplnovýznamových slov, které při větné segmentaci na w-rovině evidentně nelze zahrnout do žádného z úseků odpovídajícího na m-rovině větě (například jako váhání na začátku věty nebo na jejím konci).

Bezobsažným úsekům neodpovídá na m-rovině žádný s-element.

Příklady:

<UH> <inhale> *no* <UH> *tak to* <cough> <noise>

➜ Ø



3 Typy vět podle obsahu

Každá věta je ohodnocena z hlediska obsahové důležitosti v kontextu celého textu. Určuje se, jakého druhu je obsah dané věty, tj. zda daná věta přináší novou informaci, nebo je otázkou po takové informaci, příkazem, přitakáním mluvčího apod.

Z tohoto hlediska rozlišujeme osm typů vět a informaci o typu věty ukládáme v atributu `stype`, který náleží s-elementu identifikujícímu hranice vět. Přehled hodnot atributu `stype` viz tab. *Hodnoty atributu* `stype`.

Hodnoty atributu `stype`

`information`	informace, obsahově relevantní věta
`instruction`	příkaz, žádost, aby druhý mluvčí něco vykonal
`question`	otázka po informaci
`confirmation`	kladné přitakání druhého mluvčího (posluchače) k obsahu projevu prvního mluvčího
`surprise`	překvapení posluchače nad novou informací, kterou mu mluvčí sděluje
`disbelief`	věta, která signalizuje, že posluchač není přesvědčen o tom, co mluvčí sděluje
`repetition`	zopakovaná myšlenka
`other`	jiný typ

Hodnota `information` náleží větám, které do výsledného rekonstruovaného textu přináší podstatné nové informace. Věty s hodnotou `information` za žádných okolností nelze ze standardizovaného textu vypustit.
Z formálního hlediska jde primárně o věty oznamovací (případně věty přací, zvolací, řečnické otázky).

Příklady:

Je mi osmdesát let.
V Praze.
Ach, to byla hrůza.
Kéž by se to nikdy nestalo.
Ano. (odpověď na zjišťovací otázku)

Hodnota `question` náleží primárně zjišťovacím a doplňovacím otázkám, tedy otázkám po informaci, nikoli otázkám zvolacím, řečnickým a otázkám, které jsou ve skutečnosti žádostí.
Z formálního hlediska jde o věty tázací.

Příklady:

Kolik je vám let?
Jak jste strávil dětství?

Hodnota `instruction` náleží větám, které vyjadřují příkaz, žádost, přání jednoho mluvčího, aby druhý mluvčí něco vykonal, řídil se jeho pokyny.
Z formálního hlediska jde primárně o věty rozkazovací a některé otázky vyjadřující žádost.

Příklady:

Držme se ještě vašeho dětství. *Řekněte, jak jste strávil dětství.* *Povězte nám něco o té době.* *Můžete zavřít okno?*

Hodnota `confirmation` náleží větám, které vyjadřují kladné přitakání druhého mluvčího (posluchače) k obsahu projevu prvního mluvčího, aniž by tento projev prvního mluvčího byl danou větou nějak přerušován; první mluvčí na základě přitakání nemění směr hovoru. V konverzaci jsou tyto věty naprosto běžné, nenesou žádnou informaci, nepřispívají k obsahu konverzace, mohou být z textu i vypuštěny a jeho informační hodnota se tím neztratí.
Pozor! Hodnota `confirmation` nenáleží odpovědím na zjišťovací otázky!

Příklady:

To je pravda. *Jo.* *Souhlas.* *Souhlasím.* *Aha.* *Ano.* *Jasně.*

Specifické případy přitakání posluchače k obsahu projevu mluvčího jsou označovány hodnotami `surprise` a `disbelief`.

Hodnota `surprise` náleží větám, které v širokém smyslu vyjadřují překvapení posluchače nad novou informací, kterou mu mluvčí sděluje.

Příklady:

Opravdu? *A helemese.* *To jsou věci!* *Vážně?*

Hodnota `disbelief` náleží větám, které signalizují, že posluchač nevěděl o tom, co mluvčí říká, nebo si myslel opak a nové informaci příliš nevěří, není o ní ještě přesvědčen.

Příklady:

To není možný. *To nemůže být tak.* *Opravdu?* (tónem, který říká: “Nepřesvědčil jste mě.”)

I pro přiřazení hodnoty `surprise` a `disbelief` (podobně jako pro hodnotu `confirmation`) platí, že náleží větám, které nepřerušují projev hlavního mluvčího; mluvčí na základě těchto vět nemění směr hovoru.

Hodnota `repetition` náleží větám, které znova opakují celou předcházející myšlenku, například proto, aby byl podtrhnut její význam nebo aby byla potvrzena její platnost. (Nemusí jít o doslovné opakování.)

Příklad:

{<silence> *dostali sme pět korun* <cough>} *pet korun sme dostali* <noise>

➜ {*Dostali jsme pět korun.*} *Pět korun jsme dostali.*

Pro ostatní nedefinované případy (pro případy, kterým nevyhovuje žádná ze zde uvedených hodnot) je zavedena **hodnota `other`**.

4 Modifikace textu

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

- **ortografické modifikace** (viz *4.1 Ortografické modifikace*),
- **vlastní modifikace** (viz *4.2 Vlastní modifikace*).

4.1 Ortografické modifikace

Ortografické modifikace představují pravidelné úpravy vstupního textu, vyplývající ze základní podmínky na standardizovaný text, totiž že standardizovaný text splňuje obecné charakteristiky psaného textu a jsou v něm dodržena pravidla českého pravopisu.

K ortografickým modifikacím patří:

- odstranění obsahově nerelevantních neřečových událostí (viz *4.1.1 Odstranění obsahově nerelevantních neřečových událostí*),
- pravopisné úpravy (viz *4.1.2 Pravopisné náležitosti psaného textu*)
- přepis slov pomocí nealfabetických znaků (viz *4.1.3 Přepis slov pomocí nealfabetických znaků*)

4.1.1 Odstranění obsahově nerelevantních neřečových událostí

Neřečové události (jako nádechy, zakašlání) jsou důsledně zachycovány na w-rovině korpusu (seznam typů neřečových událostí rozlišovaných na w-rovině korpusu viz tab. *Přehled značek pro neřečové události*). Ve výsledném standardizovaném textu na m-rovině jsou neřečové události zaznamenávány jen tehdy, pokud nesou nějaký význam, pokud přispívají k obsahu sdělení (k tomu viz *4.2.2 Zachycení obsahově relevantních neřečových událostí*).
Neřečové události, které nemají žádný důležitý význam pro obsah sdělení (většina), jsou na m-rovině bez náhrady odstraněny.

Obsahově nerelevantní neřečové události neodpovídá na m-rovině žádný uzel.
Na obsahově nerelevantní neřečovou událost nevede z m-roviny žádný odkaz.

Přehled typů neřečových událostí rozlišovaných na w-rovině

`click`	mlaskání jazykem
`mouth`	mlaskání rty
`cough`	kašlání
`laugh`	smích
`breath`	zvuk dechu
`inhale`	nádech
`silence`	ticho, pauza
`uh`	uh, um, uh-huh, uh-hum, hm, ehm
`noise`	hluk v pozadí
`unintelligible`	nesrozumitelný úsek

Příklady:

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

➜ *Tak možná, že bych ještě něco řekl.*

<silence> <inhale> *někteří lidé mně* <uh> *utkvěli* <inhale> *velmi v paměti* <inhale> *z toho koncentračního tábora* <silence>

➜ *Někteří lidé z koncentračního tábora mně velmi utkvěli v paměti.*

4.1.2 Pravopisné náležitosti psaného textu

Ve standardizovaném textu jsou dodržována všechna pravopisná pravidla pro psaný text (přijaté transkripční zásady pro zápis segmentů mluvené řeči na w-rovině přitom tato pravidla dodržovat nemusí).

K úpravám tohoto typu patří zejména dvě následující:

- vložení interpunkčních znamének (viz *4.1.2.1 Interpunkce*)
- náhrada malých písmen za velká (viz *4.1.2.2 Velká písmena*)

4.1.2.1 Interpunkce

Ve standardizovaném textu jsou správně doplněna veškerá **interpunkční znaménka** (čárky, tečky, pomlčky, uvozovky, dvojtečky).

Z vloženého m-uzlu reprezentujícího interpunkční znaménko nevede žádný odkaz do w-roviny.

Příklad:

on řekl byl sem tam ale nikdo mu nevěřil

➜ *On řekl: „Byl jsem tam," ale nikdo mu nevěřil.*

4.1.2.2 Velká písmena

Velká a malá písmena jsou ve standardizovaném textu psána v souladu s pravidly českého pravopisu.

Při rekonstrukci jde zejména o následující změny:

a. **zvětšení písmena na začátku vět**.
b. **zvětšení písmena na začátku vlastních jmen a názvů**.

Příklad:

ať žije havel

➜ *Ať žije Havel.*

4.1.3 Přepis slov pomocí nealfabetických znaků

Na w-rovině korpusu je zpravidla vše, co bylo řečeno, zaznamenáváno slovně (pomocí písmen). V psaném textu však často s výhodou užíváme k zápisu některých slov nealfabetických znaků (číslic a jiných symbolů). V následujících sekcích popisujeme, kdy je možné použít tento způsob zápisu také na m-rovině, ve standardizovaném textu.

4.1.3.1 Číslice

Různé číselné údaje zaznamenané na w-rovině tak, jak byly vysloveny (tj. slovy), zapisujeme na m-rovině způsobem co nejobvyklejším pro psaný text (tj. buď slovy, nebo pomocí číslic). Obecně platí, že jednoslovná čísla se standardizují pomocí čísel zapsaných slovy, víceslovná čísla se standardizují pomocí čísel zapsaných číslicemi (číslicemi zapisujeme i jednoslovný složený typ *jedenadvacet*). V matematických kontextech píšeme čísla vždy číslicemi.

Příklady:

tři ➞ *tři*

dvacet tři ➞ *23*

jedenadvacet ➞ *21*

první ➞ *první*

dvacátý šestý ➞ *26.*

osmkrát ➞ *osmkrát*

dvacet pětkrát ➞ *25krát*

jedna plus dvě rovná se tři ➞ *1 + 2 = 3*

Úpravu čísla zapsaného na w-rovině slovy na číslo zapsané číslicemi považujeme za specifický typ ortografické modifikace (tj. ne za modifikaci vlastní), a to i v případě kdy je číslicemi nahrazeno číslo vyslovené s nespisovnými koncovkami (např. *čtyřicátej pátej* → *45.*).
Více viz sekce *5.1 Standardizace čísel*.

4.1.3.2 Ostatní nealfabetické značky a symboly

Vedle čísel lze nealfabetickými znaky přepsat i další slova. Zde se řídíme pravidlem: pokud to není násilné nealfabetické znaky nepoužíváme, tj. dáváme přednost zápisu slovy. Nealfabetická znak použijeme jen tam, kde je naprosto běžný.

Příklady:

dvě procenta ➞ *dvě procenta*

dvacet tři procent ➞ *23 procent*

jedenadvacet dolarů ➞ *21 dolarů*

třicetiprocentní ➞ *třicetiprocentní*

dvacetipětiprocentní ➞ *25procentní*

dvě plus tři rovná se pět ➞ *2 + 3 = 5*

byt dvě plus jedna ➞ *byt 2+1*

4.2 Vlastní modifikace

Nejdůležitější částí anotace jsou tzv. **vlastní modifikace** vstupního transkribovaného textu, představují na rozdíl od ortografických modifikací podstatný zásah do podoby vstupního textu.

K dispozici jsou následující typy vlastních modifikací:

- **Modifikace slovních jednotek** (viz *4.2.1 Modifikace slovních jednotek*):
  - vymazání slovní jednotky (viz *4.2.1.1 Mazání*)
  - vložení nové slovní jednotky (viz *4.2.1.2 Vkládání*)
  - substituce slovní jednotky (viz *4.2.1.3 Substituce*)
  - změny ve slovosledu (viz *4.2.1.4 Změny ve slovosledu*)
- **Zachycení obsahově relevantních neřečových událostí** (viz *4.2.2 Zachycení obsahově relevantních neřečových událostí*)

Pozor! Jednotlivé typy modifikací jsou v této příručce ilustrovány na příkladech izolovaných větných segmentů; vhodnost uplatnění jakékoli popisované modifikace je však třeba vždy posuzovat vzhledem ke kontextu celého rekonstruovaného textu.

Terminologie modifikací (názvy mazání, vkládání atp.) je odvozena od procesu rekonstrukce jdoucí od vstupního transkribovaného textu na w-rovině k výstupnímu standardizovanému textu na m-rovině.

4.2.1 Modifikace slovních jednotek

V této sekci popisujeme úpravy týkající se primárně slovních jednotek, tj. jde zejména o vztahy mezi tokeny na w-rovině (w-uzly typu `w`) a slovními jednotkami na m-rovině (m-uzly typu `m`).

4.2.1.1 Mazání

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 jsou při rekonstrukci standardizovaného textu ze vstupní transkripce odstraňovány. Jde o slovní jednotky obsahově nerelevantní.

W-uzlu (typu `w`) reprezentujícímu obsahově nerelevantní slovní jednotku neodpovídá na m-rovině žádný m-uzel.
Na w-uzel (typu `w`) reprezentující obsahově nerelevantní slovní jednotku nevede z m-roviny žádný odkaz.

K obsahově nerelevantním slovním jednotkám řadíme zejména:

- výplňková slova (viz *4.2.1.1.1 Výplňková slova*)
- výplňkové fráze (viz *4.2.1.1.2 Výplňkové fráze*)
- nadbytečná deiktická slova (viz *4.2.1.1.3 Nadbytečná deiktická slova*)
- nadbytečné konektory (viz *4.2.1.1.4 Nadbytečné konektory*)

- nadbytečná a nesprávně užitá gramatická slova (viz *4.2.1.1.5 Nadbytečná nebo nesprávně užitá gramatická slova*)
- restarty (viz *4.2.1.1.6 Restarty*)
- opakující se úseky textu (viz *4.2.1.1.7 Opakující se úseky textu*)
- fragmenty (viz *4.2.1.1.8 Fragmenty*)

4.2.1.1.1 Výplňková slova

Výplňková slova (též vycpávková) jsou slovní jednotky, která nenesou žádný význam, mluvčí je používá tehdy, když se rozmýšlí, co říci, když hledá správná slova pro to, co chce říci. K výplňkovým slovům řadíme i slova, která někteří mluvčí ve svém mluveném projevu často opakují, vkládají je bez zjevného důvodu na různá místa ve větě.

Patří sem: *eh*; *hm*, *to*, *no*, *jo*, *že jo*; *vlastně*, *prostě*, *jako*, *teda* aj.

Příklady:

hledali nějakýho ubožáčka že jo

➜ *Hledali nějakého ubožáčka.*

no tam jsme byli dva roky

➜ *Tam jsme byli dva roky.*

4.2.1.1.2 Výplňkové fráze

K **výplňkovým frázím** řádíme většinou ustrnulé slovesné konstrukce, které klesají v pouhé částice. Z věty je odstraňujeme tehdy, narušují-li její strukturu a nemají podstatný význam. Ve větě však, pokud nenarušují její plynulost, mohou takové fráze i zůstat (jde o typy popsané v tektogramatickém manuálu jako: kleslá parenteze, ustrnulé infinitivní a participiální konstrukce).

Patří sem: *to víte; myslím; jak vidíš; vždyť víš; nedej bůh.*

Příklad:

to bylo v Praze na já myslím na Vánoce

➜ *To bylo v Praze na Vánoce.*

4.2.1.1.3 Nadbytečná deiktická slova

Za **nadbytečná deiktická slova** (jde o nadbytečnost z hlediska psaného textu) považujeme zejména:

a. deiktická slova, která vyplňují pozici členu umístěného jinde ve větě (například kvůli zdůraznění).

Příklad:

ono snad všechny vagóny tam nebyly jenom děti

➜ *Snad ve všech vagónech nebyly jenom děti.*

b. ukazovací zájmena před jmény, kdy zájmeno nenese význam, tj. nemá význam identifikace entity pojmenované jménem. Je užito proto, že:
    - mluvčí si není jist, přemýšlí, proto nejprve nahradí jméno zájmenem, než jej uvede.

Příklad:

byl jsem v <u>tom</u> eh táboře ➜ *Byl jsem v táboře.*

- pojmenovaná entita je zdůrazněna jako "to, jak jsme o tom už mluvili".

Příklady:

jel sem do <u>té</u> Prahy ➜ *Jel jsem do Prahy.* *šla jsem do <u>té</u> školy* ➜ *Šla jsem do školy.*

Jako nadbytečná deiktická slova chápeme zejména ukazovací zájmena před vlastními jmény a názvy, jež samy jsou dostatečnými identifikátory pojmenované entity.

Pozor! Nadbytečná deiktická slova je třeba odlišit od ukazovacích zájmen jednoznačně určujících (identifikujících) objekt (mezi jinými podobnými objekty), o kterém se mluví (například: *Přišel jsem do toho tábora a ne do tamtoho.*)

4.2.1.1.4 Nadbytečné konektory

Za **nadbytečné konektory** (jde o nadbytečnost z hlediska psaného textu) považujeme zejména spojky na začátcích větných celků, které nevyjadřují žádný významový vztah připojované věty k větě předcházející, tj. pouze navazují text na předešlou promluvu.

Nejčastějším nadbytečným konektorem je spojka *a*. Jiné spojky jako *ale*, *nebo*, *tak*, jsou odstraněny jen tehdy, nenesou-li svůj vlastní specifický připojovací význam (kontrast, disjunkci, důsledek).

Příklady:

<u>a</u> spolu začali dělat tohleto ➜ *Spolu začali dělat tohleto.* *<u>a</u> tam to trvalo dva roky.* ➜ *Tam to trvalo dva roky.*

4.2.1.1.5 Nadbytečná nebo nesprávně užitá gramatická slova

K **nadbytečným nebo nesprávně užitým gramatickým slovům** patří z hlediska psaného textu a jeho obvyklých stylistických vlastností nadbytečně nebo nesprávně užitá:

a. pomocná slovesa:

Příklad:

pak byl přišel ➜ *Pak přišel.*

b. předložky:

Příklad:

to bylo v Praze na já myslím o Vánocich ➜ *To bylo v Praze o Vánocích.*

c. spojky:

Příklad:

to se stalo mně a Járovi a Pavlovi ➜ *To se stalo mně, Járovi a Pavlovi.*

d. osobní zájmena v pozici subjektu:

Příklad:

já se menuju Marek ➜ *Jmenuju se Marek.*

4.2.1.1.6 Restarty

Restartem označujeme úseky textu, ve kterých tzv. falešný začátek je nahrazen novým začátkem. Restarty mají zpravidla následující strukturu:

- **falešný začátek - (korektor) - nový začátek**.

Falešný začátek je úsek textu, který mluvčí posléze buď opustí a již na něj nenavazuje, nebo jej nahradí jiným úsekem textu - novým začátkem. **Korektorem** (v angl. interregnum) rozumíme výraz (nebo výrazy), kterým mluvčí uvozuje následující nový začátek toho, co předtím nepřesně vyjádřil. Korektor může ve struktuře restartu chybět. **Nový začátek** je pak oprava původního falešného začátku.

Při rekonstrukci jsou ze vstupního textu odstraněny jak úseky představující falešné začátky, tak případné korektory. Ve standardizovaném textu se objeví jen opravené nové začátky.

Příklad:

v pátek teda vlastně v sobotu sme tam šli **Falešný začátek**: *v pátek* **Korektor**: *teda vlastně* **Nový začátek**: *v sobotu* ➜ *V sobotu jsme tam šli.*

Restartů je celá řada různých druhů:

a. **zakoktání**.

Příklad:

v tomto z- zd- zděném baráku byly betonové kobky
➜ *V tomto zděném baráku byly betonové kobky.*

b. **opakování** stejných slov (hned za sebou).

Příklad:

a odvedli nás do do do toho karanténního bloku
➜ *Odvedli nás do karanténního bloku.*

c. Za restart považujeme i případy, kdy má opakování slov význam zdůraznění.

Příklad:

to bylo poslední poslední jídlo
➜ *To bylo poslední jídlo.*

d. **oprava**.

Příklad:

syn můj syn už se nevrátil
➜ *Můj syn už se nevrátil.*

f. **zadrhnutí**: úseky, kde se mluvčí zadrhnul, zakoktal, hledal správná slova. Jedná se o úsek textu, který je posléze (většinou) přeformulován a nahrazen jiným (například kvůli změně vazby).

Příklad:

a to byli většinou to byl většinou ten personál
➜ *To byl většinou personál.*

h. **samotný falešný začátek** (bez opravy, bez nového začátku); blíží se fragmentu (viz *4.2.1.8 Fragmenty*).

Příklad:

já to byl většinou ten personál
➜ *To byl většinou personál.*

4.2.1.1.7 Opakující se úseky textu

Z původního transkribovaného textu jsou na m-rovině odstraněny **úseky textu, které se opakují** v případě, že opakování nemá žádný podstatný význam pro obsah sdělení (srov. k tomu možnost hodnoty `repetition` v atributu `stype` popsanou v *3 Typy vět podle obsahu*).

Příklad:

my sme tam dostávali v Bratislavě podporu že jo asi deset korun denně sme dostávali že
➜ *V Bratislavě jsme dostávali podporu asi deset korun denně.*

4.2.1.1.8 Fragmenty

Fragmentem rozumíme úsek textu (jedno nebo několik plnovýznamových slov), který zůstal nedokončený a nikde dále v textu se na něj nenavazuje, ani nepřímo (tj. pro obsah textu nemá žádný podstatný význam). Fragment je třeba odlišit od nedokončené výpovědi - srov. k tomu *2.2.3 Nedokončené výpovědi*.

Příklad:

v pátek sem `<cough>` *Barňák pak odešel*
➜ *Barňák pak odešel.*

4.2.1.2 Vkládání

Standardizovaný text může obsahovat i slovní jednotky, které nebyly vyřčeny, ale které jsou nezbytné pro vytvoření gramaticky i lexikálně správné věty (standardizovaného textu).

Při rekonstrukci je pro takovou slovní jednotku vytvořen na m-rovině nový, vložený m-uzel.

Na m-rovině mohou být (vložené) m-uzly (typu `m`) reprezentující slovní jednotky, které nejsou přítomné na w-rovině. **Z m-uzlu (typu `m`) reprezentujícího slovní jednotku nepřítomnou na w-rovině nevede žádný odkaz do w-roviny.**

Vložené m-uzly reprezentují zejména:

- chybějící gramatická slova (viz *4.2.1.2.1 Chybějící gramatického slova*)
- nevyjádřená plnovýznamová slova (viz *4.2.1.2.2 Nevyjádřená plnovýznamová slova*).

Pozor! Vloženým m-uzlem je reprezentována také doplněná interpunkce; k tomu viz *4.1.2.1 Interpunkce*.

4.2.1.2.1 Chybějící gramatická slova

Do vstupního textu jsou na m-rovině vkládána gramatická slova na pozice, kde chybí a jsou nezbytná pro vytvoření gramaticky správné věty. K **chybějícím gramatickým slovům** patří:

a. pomocná a modální slovesa.

Příklad:

on v té válce zabit
➜ *On <u>byl</u> v té válce zabit.*

b. předložky.

Příklad:

bratrem sme byli v těch vybraných ➜ *<u>S</u> bratrem jsme byli v těch vybraných.*

c. spojky.

Příklad:

přines chleba čaj ➜ *Přinesl chleba <u>a</u> čaj.*

d. zájmena. Zájmena doplňujeme tam, kde je to nezbytné z důvodu koherence textu.

Příklad:

{*přišla Hana a Pavel*} *přines chleba* ➜ {*Přišla Hana a Pavel.*} *<u>On</u> přinesl chleba.*

4.2.1.2.2 Nevyjádřená plnovýznamová slova

Do vstupního textu jsou na m-rovině vkládána i **plnovýznamová slova**, ale jen v těch případech, kdy jsou tato slova jednoznačně z kontextu odvoditelná a jejich doplnění je nezbytně nutné k vytvoření plynulého standardizovaného textu.

Příklad:

`<silence>` `<inhale>` *tak* `<uh>` *to bylo* `<breath>` *tušim že to bylo na Silvestra toho roku třicet devět gestapáci* `<inhale>` *se p- potřebovali pobavit tak najednou* `<inhale>` *prostě* `<inhale>` *vnikli do našeho tábora takže ihned* `<inhale>` *ihned alarm že jo ihned do pozoru* `<silence>`
➜ *Tuším, že to bylo na Silvestra roku 1939, gestapáci se potřebovali pobavit, tak najednou vnikli do našeho tábora, takže <u>byl</u> ihned alarm, ihned jsme museli do pozoru.*
 	
`<silence>` `<inhale>` *revolverem mu takle začali před nos* `<inhale>` *a chtěli aby řekl sieg heil* `<cough>` *jo* `<silence>`
➜ *Revolverem mu takhle začali <u>dělat</u> před nosem a chtěli, aby řekl: "sieg heil."*

4.2.1.3 Substituce

Ve standardizovaném textu jsou užívána jen slova spisovná a též jen správně utvořené tvary slov. Lema slovní jednotky odpovídá vyjadřovanému významu.
Při rekonstrukci jsou proto měněny vstupní nespisovné a nesprávně utvořené formy slov a v případě slov užitých nesprávně z hlediska vyjadřovaného významu nebo i z jiných důvodů (např. z důvodu koherence textu) jsou měněna též i celá slova.

Forma a lema m-uzlu (typu `m`) nemusí odpovídat tokenu odpovídajícího w-uzlu (typu `w`).

4.2.1.3.1 Změna formy slovní jednotky

Na m-rovině jsou nespisovné a nesprávné tvary slov nahrazeny spisovnými.

M-uzly typu `m` reprezentují jen správně utvořené a spisovné tvary slov.

Forma slovních jednotek se mění z následujících důvodů:

A. forma slova je **nespisovná**.

Jde o případy užití slova s nespisovnou koncovkou nebo s nespisovnou (obecně českou) hláskovou změnou uvnitř slova:

Příklad:

to musi bejt vo vozejk

➜ *To musí být o vozík.*

Spisovné tvary slov lze ze stylistického hlediska rozdělit na spisovné tvary knižní, neutrální a hovorové. Stylistické změny při rekonstrukci neprovádíme. Pokud mluvčí například použil spisovný tvar hovorový, neměníme ho na spisovný tvar neutrální.

Za spisovné tvary hovorové jsou dnes již považovány například tvary:
sousedi, komunisti vedle *sousedé*, *komunisté*;
nesem, žijem, kupujem, můžem vedle *neseme, žijeme, kupujeme, můžeme*;
moct vedle *moci*;
myju, *žiju, kupuju, lyžuju* vedle *myji*, *žiji, kupuji, lyžuji*;
myjou, *žijou*, *kupujou*, *lyžujou* vedle *myjí*, *žijí*, *kupují*, *lyžují*;
oni sází, *se vrací, chybějí* vedle *sázejí*, *se vracejí*, *chybí*;
komunizmus vedle *komunismus.*
Stylově neutrální dvojtvary: *mohu* i *můžu*; *mažu* i *maži*, *mažou* i *maží*, *kopu* i *kopám*, *řežu* i *řezám*.
Nespisovné naproti tomu je například: *bysme*, *začnul, načnul, začla*.

B. forma slova je **nesprávně utvořená**; vyjadřuje nesprávně hodnotu nějaké gramatické kategorie.

Příklady:

nechtělo se mu tam jet samotnýho

➜ *Nechtělo se mu tam jet samotnému.*

revolverem mu začali takhle dělat před nos

➜ *Revolverem mu začali takhle dělat před nosem.*

tyto auta se vracely prázdné

➜ *Tato auta se vracela prázdná.*

Pozor! Expresivní slova, slova vulgární se neutrálními spisovnými protějšky nenahrazují.

4.2.1.3.2 Změna lematu slovní jednotky

Lema slovních jednotek se mění z následujících důvodů:

A. lema slova je zvoleno nesprávně **z hlediska vyjadřovaného významu**.

Jde zejména o případy, kdy mluvčí užije zvukově podobné, avšak významem zcela odlišné slovo (tzv. paronymum), nebo o případy, kdy mluvčí užije slovo významově velmi blízké, avšak v daném kontextu nevhodné.

Příklady:

tak jsem začal mluvit jaký má krásný obrazy

➜ *Tak jsem začal říkat, jaké má krásné obrazy.*

architekt zelenka má velikou zálohu o tuto činnost

➜ *Architekt Zelenka má velikou zásluhu na této činnosti.*

B. lema slova je zvoleno nesprávně **z hlediska vyjadřované vazby**.

Jde zejména o případy nesprávně užitých předložkových vazeb.

Příklad:

architekt Zelenka má velikou zásluhu o tuto činnost

➜ *Architekt Zelenka má velikou zásluhu na této činnosti.*

C. lema slova je zvoleno nesprávně **z hlediska zachování koherence textu**.

Standardizovaný text na m-rovině dodržuje pravidla koherence textu. Z důvodu plynulé návaznosti textu a udržení správných koreferenčních vztahů mezi jednotlivými referenčně totožnými větnými členy, je někdy žádoucí nahradit pronesený deiktický výraz (zaznamenaný na w-rovině) plným lexikálním pojmenováním, někdy je naproti tomu vhodná opačná úprava.

Příklad:

m-rovina:	*z*	*těch*	*domů*	*pak*	*vyšli*
w-rovina:		*z*	*nich*	*pak*	*vyšli*

m-rovina:	*z*	*nich*	*pak*	*vyšli*	
w-rovina:	*z*	*těch*	*domů*	*pak*	*vyšli*

{Petr dobíhal na poslední chvíli Honza taky} on to už ale pak nestihnul

➜ *Honza už to pak ale nestihnul.*

nalil mi kávu do hrnku pak si nabral omáčku a podal mi ji

➜ *Nalil mi kávu do hrnku, pak si nabral omáčku a podal mi tu kávu.*

a s <u>tou</u> paní sme na tý lavičce seděli až do oběda ➜ *S paní <u>Novákovou</u> jsme na té lavičce seděli až do oběda.*

4.2.1.3.3 Náhrada nesrozumitelného úseku textu domyšleným

Nesrozumitelné úseky textu (reprezentované na w-rovině w-uzly typu `nonspeech` označené jako `unintelligible`) se, pokud to na základě kontextu jde, pokusíme při rekonstrukci domyslet (třeba jen pomocí obecných, ne příliš významových slov). Od všech doplněných m-uzlů (typu `m`; s výjimkou interpunkce), které představují domyšlený text, vedou na w-uzel typu `nonspeech` s hodnotou `unintelligible` v atributu `desc` odkazy.

Příklad:

m-rovina:	*Setkal jste se s takovými projevy v dětství ?*
w-rovina:	*setkal jste se s* `<unintelligible>` *projevy v dětství*

V případě, že text domyslet nelze, řídíme se pravidly uvedenými v *4.2.2 Zachycení obsahově relevantních neřečových událostí*. Souhrnně též *5.3 Nesrozumitelný úsek textu*.

4.2.1.4 Změny ve slovosledu

Na m-rovině mají rekonstruované věty gramatický slovosled, který nenarušuje plynulost textu.

Pořadí uzlů na m-rovině nemusí odpovídat pořadí uzlů na w-rovině.

Příklady:

po pěti sme leželi	*prosté měření terénu sme dělali*
➜ *Leželi jsme po pěti.*	➜ *Dělali jsme prosté měření terénu.*
tam my sme autem jeli	*sem jel s ním do Zvolena*
➜ *My jsme tam jeli autem.*	➜ *Jel jsem s ním do Zvolena.*

4.2.2 Zachycení obsahově relevantních neřečových událostí

Standardizovaný text, ve kterém se řídíme pravidly psaného textu, primárně neobsahuje značky pro neřečové události. Obsahově nerelevantní neřečové události se při rekonstrukci bez náhrady odstraňují (viz *4.1.1 Odstranění obsahově nerelevantních neřečových událostí*).

Obsahově relevantní neřečové události zachycujeme na m-rovině primárně prostředky psaného textu.

Obsahově relevantní neřečové události, tj. takové, které nesou nějaký význam, kterým přispívají k obsahu sdělení, zachycujeme ve standardizovaném textu primárně prostředky textu psaného, tj. zejména pomocí interpunkčních znamének, slovosledu. Takto zaznamenáváme například:
- věty pronesené s důrazem (vykřičník),
- delší pauzy (pomlčka),
- ironicky pronesené slovo (uvozovky)
- důraz na slově (slovosled, aktuální členění).

Význam pro obsah sdělení může mít ale celá řada neřečových událostí, které jen pomocí běžných prostředků psaného textu nezachytíme (ironický smích, šeptání, náhlé zvýšení hlasu aj.).

Obsahově relevantní neřečové události mohou být na m-rovině zachyceny i speciálním typem m-uzlu (m-uzlem typu `nontext`).

M-uzlu typu `nontext` náleží atribut `type`, ve kterém anotátor (vlastními slovy) uvede popis neřečové události, kterou považuje za obsahově relevantní.

Příklady popisů:

smích
váhá (ticho)
nejspíš kývnul na souhlas
ztišil hlas
předchozí slovo vysloveno hodně nahlas
hvízdnul
pochichtává se

M-uzel typu `nontext` je vždy součástí nějaké věty (větu, s-element může tvořit i jen tato speciální značka).

Pokud jsou na w-rovině zachyceny odpovídající neřečové události (w-uzly typu `nonspeech` a `background_begin`) vedou na tyto odpovídající w-uzly z m-uzlu typu `nontext` odkazy.

Příklady:

m-rovina: [spk1] *Je to tak ?* [spk2] `<nejspíš kývnul>`

w-rovina: *no a je to tak* `<silence>`

m-rovina: [spk1] *Je to tak?* [spk2] `<souhlasí>`

w-rovina: [spk1] *no a je to tak* [spk2] `<uh>`

m-rovina: [spk1] *Odjeli jsme dvanáctého.* [spk2] `<přitakává>`

w-rovina: [spk1] *odjeli sme dvanáctýho* [spk2] `<uh>`

m-rovina:	<pochichtává se> *To nemyslíš vážně?*
w-rovina:	<background_begin laugh> *to* <uh> *nemyslíš vážně co* <background_end>

m-rovina:	*Byl jsem velký* <předchozí slovo důrazně> *pán.*
w-rovina:	*jo já sem byl velký pán*

Obsahově relevantní bývá často i **nesrozumitelný úsek textu**, zachycený na w-rovině taktéž w-uzlem typu `nonspeech` (s hodnotou `unintelligible` v atributu `desc`).
Obsahově relevantní nesrozumitelný úsek textu nahrazujeme na m-rovině primárně textem domyšleným (viz *4.2.1.3.3 Náhrada nesrozumitelného úseku textu domyšleným*). Pokud však taková náhrada není možná (text si domyslet nelze), reprezentujeme na m-rovině nesrozumitelný úsek textu m-uzlem typu `nontext` s hodnotou `unintelligible` v atributu `type`. Vzájemně si odpovídající uzly jsou opět propojeny odkazem.

Příklad:

m-rovina:	*Setkal jste se s* <unintelligible> *projevy v dětství ?*
w-rovina:	*setkal jste se s* <unintelligible> *projevy v dětství*

Viz též *5.3 Nesrozumitelný úsek textu.*

5 Specifické případy

5.1 Standardizace čísel

Různé číselné údaje zaznamenané na w-rovině tak, jak byly vysloveny (tj. slovy), jsou na m-rovině zapsány obvyklým způsobem pro psaný text (tj. slovy nebo pomocí číslic).
Změnu čísla zapsaného na w-rovině slovy na číslo psané číslicemi považujeme za ortografickou modifikaci (viz *4.1.3.1 Číslice*).

Jednoslovná čísla se standardizují pomocí čísel zapsaných slovy, víceslovná čísla se standardizují pomocí čísel zapsaných číslicemi.

Číslicemi zapisujeme i jednoslovný složený typ *jedenadvacet*.

Příklady:

tři ➞ *tři*	*první* ➞ *první*
šedesát ➞ *šedesát*	*dvacátý šestý* ➞ *26.*
dvacet tři ➞ *23*	*šestadvacátý* ➞ *26.*
třiadvacet ➞ *23*	*osmkrát* ➞ *osmkrát*
sto jedna ➞ *101*	*dvacet pětkrát* ➞ *25krát*

m-rovina: *4321*

w-rovina: *čtyři tisíce tři sta dvacet jedna*

m-rovina: *dvě*

w-rovina: *dvě*

Pozor! Odkaz (na odpovídající řadovou číslovku) vede i z m-uzlu reprezentujícího tečku za řadovou číslovkou.

Příklad:

m-rovina: *21* .

w-rovina: *dvacátej první*

V matematických (případně fyzikálních aj.) kontextech píšeme čísla vždy číslicemi.

Příklady:

jedna plus dvě rovná se tři ➞ *1 + 2 = 3*

jedna plus dvě je tři ➞ *1 + 2 je tři.*

jedna a dvě je tři ➞ *Jedna a dvě je tři.*

v poměru jedna ku třem ➞ *v poměru 1 : 3*

je to třicet km [*k m*] ➞ *Je to 30 km.* Ale: *je to třicet kilometrů* ➞ *Je to třicet kilometrů.*

5.1.1 Časové údaje

5.1.1.1 Letopočet

Letopočty jsou primárně psány číslicemi.

Příklad:

m-rovina: *1945*

w-rovina: *devatenáct set štyřicet pět*

Způsob standardizace nejrůznějších variant odvozených od modelového příkladu ukazují následující příklady.

Další příklady:

bylo to čtyřicet pět	*v osmašedesátým se to stalo*
➞ *Bylo to 1945.*	➞ *Stalo se to 1968.*
bylo to v roce čtyřicet pět	*to se stalo v šedesátém osmém*
➞ *Bylo to v roce 1945.*	➞ *To se stalo 1968.*

5.1.1.2 Desetiletí

Desetiletí (léta) jsou primárně zachycována řadovými číslovkami psanými slovy.

Příklad:

m-rovina: *v šedesátých letech*

w-rovina: *v šedesátých letech*

5.1.1.3 Datum

Označení dne v datu je primárně standardizováno číslicí, označení měsíce je standardizováno pomocí názvu měsíce nebo čísla měsíce (zapsaného číslicemi) podle toho, jak bylo datum vysloveno; srov. dva následující příklady.

m-rovina: *22 . září*

w-rovina: *dvacátého druhého září*

m-rovina: *22 . 9 .*

w-rovina: *dvacátého druhého devátý*

Ve „vyprávěcím“ kontextu může však i označení dne být standardizováno slovem, zejména jde-li o jednoslovnou číslovku.

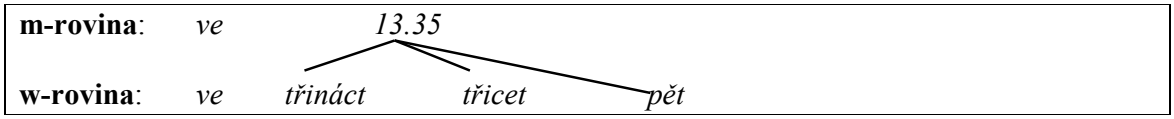
prvního ledna sme odjeli dvanáctého sme zastavovali a dorazili sme tam až dvacátého osmého
➜*Prvního ledna jsme odjeli, dvanáctého jsme zastavovali a dorazili jsme tam až 28.*

5.1.1.4 Čas

Hodinový časový údaj standardizujeme podle následujících dvou příkladů (digitální čas číslicemi, ostatní typy slovy).

Příklady:

m-rovina:	*v*	*půl*	*druhé*	*a*	*pět*	*minut*
	|	|	|	|	|	|
w-rovina:	*v*	*půl*	*druhé*	*a*	*pět*	*minut*



Poznámka: V časovém údaji se mezi hodinami a minutami píše tečka (bez mezery), tedy: *13.35*. Digitální čas ve formě číslo+tečka+číslo je jeden m-uzel.

Další příklady:

přišel o třetí hodině odpoledne	*přišel ve čtrnáct hodin dvacet pět minut*
➜ *Přišel o třetí hodině odpoledne.*	➜ *Přišel ve 14.25.*
už byla jedna pryč	*bylo dvacet jedna hodin*
➜ *Už byla jedna pryč.*	➜ *Bylo 21 hodin.*
přišel ve čtrnáct hodin a dvacet pět minut	*přijďte v patnáct nula nula*
➜ *Přišel ve čtrnáct hodin a 25 minut.*	➜ *Přijďte v 15.00.*

5.1.2 Vyjadřování množství

Při standardizaci čísel vyjadřujících množství počítaného předmětu se řídíme základním pravidlem uvedeným v úvodu této sekce: jednoslovná čísla se standardizují pomocí čísel zapsaných slovy, víceslovná čísla se standardizují pomocí čísel zapsaných číslicemi.

Příklady:

našel dvě koruny	*každý třetí v řadě si vystoupil*
➜ *Našel dvě koruny.*	➜ *Každý třetí v řadě si vystoupil.*
bylo mu třicet šest let	*skončil dvacátý pátý*
➜ *Bylo mu 36 let.*	➜ *Skončil 25.*

5.2 Standardizace „neslovníkových“ slov

Tato sekce popisuje pravidla, jak při standardizaci nakládat s tzv. „neslovníkovými“ slovy. Za „neslovníková“ slova považujeme slova, která běžně nepatří do slovní zásoby českého jazyka – jde o slova cizí (včetně cizojazyčných jmen a názvů), dále o slova neznámá, nově utvořená a různá přeřeknutí a zkomoleniny slov známých.

Úkolem anotátora je i těmto neslovníkovým „slovům“ na základě následujících pravidel přidělit nějakou podobu lematu (formy). Případy, kdy anotátor není s to konečnou podobu slova vyřešit, označuje v anotátorské poznámce typu `form` (viz i 5.6 *Anotátorská poznámka*).

5.2.1 Cizojazyčné výrazy

Řečník může během výpovědi vyslovit některá slova v jiném jazyce, než je původní jazyk výpovědi (tj. v našem případě v jiném jazyce než českém). Řekne například pár slov anglicky nebo v jazyce jidiš. Cizojazyčné výrazy může mluvčí vyslovit nejrůznějším způsobem, často původní cizí výraz nějak počešťuje (přidává českou flexivní koncovku).

Při zápisu cizojazyčných výrazů na m-rovině se řídíme následujícími pravidly:

- vyslovené nepočeštěné podoby slov zapisujeme tak, jak se správně píší v daném cizím jazyce (tj. nikoli například foneticky)
- vyslovené různě počeštěné podoby cizích slov zapisujeme:
  - v kodifikované počeštěné podobě (u slov přejatých, u kterých počeštěná podoba existuje)
  - tak, jak se správně píší v daném cizím jazyce (tj. například bez počeštěné koncovky),

Je-li více možností zápisu, volíme tu podobu, která je nejbližší tomu, co mluvčí skutečně vyslovil.

Příklady:

říkali sme jim agrutke a to znamená vedení

➜ *Říkali jsme jim agrutke a to znamená vedení.*

anglicky se to řekne identity card [*ajdentyty kárt*]

➜ *Anglicky se to řekne identity card.*

Citační kontext. Je-li však žádoucí zachytit skutečně to, co mluvčí vyslovil (například proto, že na danou špatnou výslovnost/tvar se v další části dialogu reaguje, mluvčí chce zdůraznit právě onu neobvyklou výslovnost/tvar), píšeme takový výraz foneticky a pak jej dáváme do uvozovek.
Tyto případy tzv. citačních kontextů (kdy je slovo užito nikoli kvůli tomu, co označuje, ale kvůli tomu, jak se vyslovuje, jaký má tvar) označujeme v anotátorské poznámce typu `metalanguage` (viz i *5.6 Anotátorská poznámka*).

Příklad:

a on to vyslovoval identity [*ídentyty*] *místo identity* [*ajdentyty*]

➜ *A on to vyslovoval „ídentity“ místo „ajdentity“.*

Způsob zápisu v uvozovkách použijeme i v případech, kdy je spojením cizojazyčného základu a české koncovky vytvořeno nové slovo, které nelze do standardizovaného textu jednoduše převést ani v původní cizojazyčné podobě, ani v nějaké správné české podobě.

Příklad:

talkovali [*tolkovali*] *sme celé dvě hodiny*

➜ *"Talkovali" jsme celé dvě hodiny.*

Poznámka: Podle pravidel anotace w-roviny by měly cizojazyčné výrazy na w-rovině být zapsány v zásadě tak, jak zde uvádíme pro m-rovinu, tj. tak, jak se správně píší v daném cizím jazyce nebo v přejaté počeštěné podobě a jejich skutečná výslovnost by měla být uložena ve speciálním atributu w-uzlu (zde ji uvádíme v hranatých závorkách). Pouze v případech, kdy anotátor nebyl schopen zjistit správný zápis cizích slov, jsou cizojazyčné výrazy zapsány foneticky přímo. Ve většině případů by tudíž cizojazyčné výrazy měly na m-rovinu být z w-roviny přejaty beze změn. Chyby na w-rovině poznamenáváme v anotátorské poznámce typu `w-token` (viz i *5.5 Chyby v manuální transkripci* a *5.6 Anotátorská poznámka*).

5.2.2 Cizojazyčná vlastní jména a názvy

Při standardizaci cizojazyčných jmen a názvů postupujeme podobně jako u obecných cizojazyčných výrazů (viz *5.2.1 Cizojazyčné výrazy*). Platí, že cizojazyčné jméno či název zapisujeme na m-rovině v té podobě, v jaké se v česky psaném textu obvykle vyskytuje (která je kodifikovaná). Je-li více možností, volíme tu podobu, která je nejbližší tomu, co mluvčí skutečně vyslovil.

Obecně známá jména a názvy, které mají českou (počeštěnou) podobu, skloňujeme.
U jmen a názvů, kdy neznáme žádnou „správnou" českou podobu názvu, tj. v češtině se používá jako „správná" domovská podoba (německá, polská, anglická aj.), použijeme tuto podobu názvu. Domovskou podobu názvu obvykle neskloňujeme.

Je-li žádoucí zachytit skutečně to, co mluvčí vyslovil, píšeme takový výraz do uvozovek (a označujeme jej v anotátorské poznámce `metalanguage`; viz i *5.4 Citační kontexty* a *5.6 Anotátorská poznámka*).

Příklady:

bydlely jsme v Maiselově ulici [*majslově*]

➜ *Bydleli jsme v Maiselově ulici.*

firma Franc Cimermann z Freudentálu dnešním Bruntále

➜ *firma Franc Cimermann z Freudentálu, dnešního Bruntálu*

odvezli nás do Osvětimy [*osvěčimi*]

➜ *Odvezli nás do Osvětimi.*

5.2.3 Nová slova a slova neznámá

Nejrůznější nově vytvořená slova, neobvyklé vulgarismy, méně známé (neznámé) nářeční výrazy zapisujme na m-rovině v uvozovkách.

Příklady:

talkovali sme celé dvě hodiny
➜ *„Talkovali“ jsme celé dvě hodiny.*

5.2.4 Zkratky

Zkratky, které se při vyslovení hláskují by na w-rovině měly být přepsány tak, jak se skutečně píšou (skutečná výslovnost je zapsána ve speciálním atributu).
Na m-rovině píšeme zkratky tak, jak se správně píšou, včetně velikosti písmen.

Příklady:

byla to firma IBM [*aj bí em*] ➜ *Byla to firma IBM.*
byla to firma IBM [*í bé em*] ➜ *Byla to firma IBM.*
byla to firma IBM [*i b m*] ➜ *Byla to firma IBM.*
v SSSR [*es es es er*] ➜ *v SSSR*
tužka papír atd. [*a t d*] ➜ *tužka, papír atd.*
tužka papír a tak dále ➜ *tužka, papír a tak dále*
bylo to asi třicet km [*k m*] ➜ *Bylo to asi 30 km.*
bylo to asi třicet kilometrů ➜ *Bylo to asi třicet kilometrů.*

5.2.5 Hláskovaná slova

Hláskovaná slova jsou na w-rovině zapsána tak, jak byla hláskována. Na m-rovině je zapisujeme vždy jen (velkými) písmeny, které oddělujeme mezerou.

Příklady:

jmenuji se Dana DÉ Á EN Á ➜ *Jmenuji se Dana, D A N A.*
jmenuji se Dana D A N A ➜ *Jmenuji se Dana, D A N A.*

5.2.6 Přeřeknutí

Přeřeknutí nahrazujeme nepřeřeknutými tvary slov.

Příklady:

pak přijela <u>lokotomiva</u>	*jo <u>holokost</u> to bylo něco*
➜ *Pak přijela <u>lokomotiva</u>.*	➜ *<u>Holocaust</u> to bylo něco.*

Jen v těch případech, kdy přeřeknutí má nějaký význam pro další vývoj dialogu – mluvčí na něj nějak reaguje, pak uvedeme i na m-rovině „přeřeknutou“ podobu slova, kterou dáme do uvozovek a označíme ji anotátorskou poznámkou typu `metalanguage (`viz i *5.4 Citační kontexty* a *5.6 Anotátorská poznámka*).

5.3 Nesrozumitelný úsek textu

Nesrozumitelný úsek textu je na w-rovině zachycen w-uzlem typu `nonspeech` s hodnotou `unintelligible` v atributu `desc`.
Na m-rovině zachycujeme nesrozumitelný úsek textu jen tehdy, když je obsahově relevantní, tj. když je evidentní, že obsahuje nějakou důležitou informaci, a akorát není rozumět, jakou.

Obsahově relevantní nesrozumitelný úsek textu nahrazujeme na m-rovině primárně textem domyšleným (viz *4.2.1.3.3 Náhrada nesrozumitelného úseku textu domyšleným*).
Pokud však taková náhrada není možná (text si na základě kontextu domyslet nelze) reprezentujeme jej na m-rovině m-uzlem typu `nontext` s hodnotou `unintelligible` v atributu `type` (viz i *4.2.2 Zachycení obsahově relevantních neřečových událostí*).

Od „domyšlených“ m-uzlů nebo od m-uzlu typu `nontext` vede vždy odkaz (odkazy) na odpovídající w-uzel.

Příklady:

m-rovina: *Setkal jste se s takovými projevy v dětství ?*

w-rovina: *setkal jste se s* `<unintelligible>` *projevy v dětství*

m-rovina: *Prosím, podej mi tu* `<unintelligible>` .

w-rovina: *prosim podej mi tu* `<unintelligible>`

Pokud je patrné, že nesrozumitelný úsek textu obsahuje nějaké (nesrozumitelné) nesmyslné koktání, které je zjevně obsahově nerelevantní, pak takový nesrozumitelný úsek textu nemá na m-rovině žádný protějšek.

m-rovina: *Leželi jsme po pěti.*

w-rovina: `<unintelligible>` *leželi sme po pěti*

Poznámka: Může se stát, že na w-rovině je nějaký úsek projevu mluvčího označen jako nesrozumitelný (w-uzlem typu `nonspeech` s hodnotou `unintelligible` v atributu `desc`), nicméně při poslechu nahrávky nyní anotátor mluvčímu dobře rozumí, slyší, co říká. V takovém případě poznamená anotátor v anotátorské poznámce typu w-token, že „nesrozumitelnému“ úseku je rozumět (zapíše, i jak „nesrozumitelnému“ úseku rozumí, tj. jak se má w-rovina opravit). Při rekonstrukci pak přistupuje k tomuto „nesrozumitelnému“ úseku tak, jako by byl na w-rovině přepsán způsobem, který uvedl v anotátorské poznámce. Standardizuje jej povolenými modifikacemi. Případné odkazy vede všechny na ten jediný „nesprávný“ w-uzel s hodnotou `unintelligible` na w-rovině.

5.4 Citační kontexty

Citačním kontextem rozumíme výrazy, ve kterých nejde o běžné užití slov, ale o slova samotná, mluví se o jejich významu, zvukové nebo grafické podobě. Slovo (spojení, nebo i celé věty) v citačním kontextu bývá uvozeno substantivy, které signalizují, že nejde o běžný význam slova nebo slov: *nápis*, *slovo*, *text*, *otázka*, *označení*, *pojem*, *věta*, *výraz*, *výrok*, *význam* a jinými. Význam meta-užití je obvyklý také u sloves: *znamenat*, *značit*, *označovat*, *psát*, *vyslovovat* aj.

Slova v citačním kontextu dáváme zpravidla do uvozovek a označíme je anotátorskou poznámkou `metalanguage` (v textu poznámky nemusí anotátor uvést nic). V případě, že je v citačním kontextu celé slovní spojení, celá věta stačí označit anotátorskou poznámkou pouze řídící člen spojení v citačním kontextu.

Příklady (podtrženým slovům náleží anotátorská poznámka `metalanguage`):

Slovo „šebah“ znamená původně sedm.
V přídavném jménu „český“ se vyskytují dvě písmena mající dominantní význam, a to „č“ a „s“.
„Hvězdné nebe nade mnou a mravní zákon ve mně“ stojí rusky a německy na desce.
Germanismus „klika“ se užívá ve významu „štěstí“ a znamená také „držadlo k otvírání dveří“.
cedule s nápisem „Romy neobsluhujeme“
Vyznání „miluji tě“ i slovo „odchod“ lidé zprofanovali.
Za výchozí význam se považuje „hák“, „hákovitý předmět“.
Výrobky obsahující freony budou podle zákona zřetelně opatřeny textem „Výrobek obsahuje látky ničící ozónovou vrstvu Země“.

Viz i *5.2.1 Cizojazyčné výrazy* a *5.2.6 Přeřeknutí*.

5.5 Chyby v manuální transkripci

Při rekonstrukci standardizovaného textu z mluvené řeči jsou důsledně odlišovány „nedostatky“ způsobené mluvčím od chyb ve formách a lematech slovních jednotek, které jsou způsobeny automatickou transkripcí (nesprávným rozpoznáním slova). Zatímco „nedostatky“ způsobené mluvčím se odstraňují rekonstrukcí nového textu na m-rovině, chyby v transkripci by měly být odstraněny přímo na w-rovině, tj. nesprávně rozpoznané tokeny by se na základě poslechu audio nahrávky měly opravit na správné.

Anotátor při rekonstrukci standardizovaného textu nemá možnost měnit w-rovinu.
Pokud anotátor při rekonstrukci zjistí chybu v transkripci na w-rovině, poznamená tento fakt **do anotátorské poznámky**. Při anotaci pak postupuje tak, jako kdyby chyba na w-rovině byla odstraněna. (Chyby u w-uzlů, které nemají na m-rovině protějšek, zaznamenáváme v anotátorské poznámce nějakého (nejbližšího) m-uzlu.) Viz i *5.6 Anotátorská poznámka*.

5.6 Anotátorská poznámka

Pro potřeby anotace je zavedena tzv. anotátorská poznámka, atribut `comment`, který slouží pro zaznamenávání nejrůznějších komentářů anotátora k jím provedené anotaci.
Pro pozdější zpracování poznámek jsou anotátorské poznámky typovány.

Anotátorské poznámky pro zaznamenání chyb na w-rovině:
Anotátor při rekonstrukci standardizovaného textu nemůže zasahovat do anotace na w-rovině. Zjistí-li nějaké chyby na w-rovině, poznamená je v některé z následujících anotátorských poznámek a rekonstrukci provede tak, jako kdyby chyba na w-rovině nebyla. Text poznámky je ve všech typech povinný.

`w-token`: poznámka slouží pro zaznamenání chybně rozpoznaných w-uzlů, pro případy, kdy slovo je na w-rovině přepsáno špatně. (Chyby u w-uzlů, které nemají na m-rovině protějšek, zaznamenáváme v anotátorské poznámce nějakého (nejbližšího) m-uzlu.).
Příklad: místo *babička* je na w-rovině zapsáno *bačkora*.

`w-missing`: poznámka slouží pro případy, kdy na w-rovině chybí přepis nějakého slova nebo celého úseku textu. Do textu poznámky nějakého nejbližšího jednoho m-uzlu se vypíše celý rozpoznaný chybějící úsek.

`w-recognize`: poznámka slouží pro případy, kdy na w-rovině je w-uzel s hodnotou `unintelligible` značící nerozpoznaný úsek textu, ale anotátorovi se podařilo text rozpoznat. Do textu poznámky nějakého nejbližšího jednoho m-uzlu se vypíše celý rozpoznaný úsek.

Ostatní

`metalanguage`: označení citačního kontextu. Slova v citačním kontextu dáváme zpravidla do uvozovek a označíme je anotátorskou poznámkou `metalanguage` (v textu poznámky nemusí anotátor uvést nic). V případě, že je v citačním kontextu celé slovní spojení, celá věta, stačí označit anotátorskou poznámkou pouze řídící člen spojení v citačním kontextu. Viz i *5.4 Citační kontexty*.

`form`: nejistota v lematu, formě slova. Anotárskou poznámku `form` vybírá anotátor tehdy, když si není jistý výslednou podobou slova (zejména u slov cizích, neznámých, u tzv. „neslovníkových“ slov; viz *5.2 Standardizace „neslovníkových“ slov*), může však jít i nejistotu v psaní velkých a malých písmen aj. V textu poznámky může anotátor uvést vlastní komentář.

`other`: jiná poznámka. Jiné komentáře k anotaci poznamenává anotátor v poznámce typu `other`. Text poznámky je v tomto případě povinný.