M A T E M A T I C K O - F Y Z I K Á L N Í F A K U L T A

P R A H A

REKONSTRUKCE STANDARDIZOVANÉHO TEXTU Z MLUVENÉ ŘEČI V PRAŽSKÉM ZÁVISLOSTNÍM KORPUSU MLUVENÉ ČEŠTINY

MANUÁL PRO ANOTÁTORY

Marie Mikulová

úfal/ckl technical report

TR-2008-38

U N I V E R S I T A S C A R O L I N A P R A G E N S I S

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

Manuál pro anotátory

Marie Mikulová

ÚFAL/CKL Technická zpráva TR-2008-38
ISSN 1214-5521

Prosinec 2008

Copies of ÚFAL/CKL Technical Reports can be ordered from:

Institute of Formal and Applied Linguistics (ÚFAL MFF UK)
Faculty of Mathematics and Physics, Charles University
Malostranské nám. 25, CZ-11800 Prague 1
Czech Republic

or can be obtained via the Web: http://ufal.mff.cuni.cz

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 kteréhokoli jazyka.

Acknowledgement
Příspěvek vznikl za finanční podpory projektů: LC 536 a GA 405/06/0589.
This work was funded in part by the Companions project (www.companions-project.org) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-034434.

Obsah

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 mu 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). 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 rozpoznaný a automaticky segmentovaný přepis audio nahrávky.

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 nejdůležitější (nejen) pro rekonstrukci standardizovaného textu jsou 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 zachyceny další obsahově relevantní jevy mluvené řeči (zejména neřečové události).
M-uzly jsou segmentovány do tzv. **s-elementů**. S-elementy reprezentují jednotlivé věty standardizovaného textu.

Anotátor tvoří standardizovanou podobu promluvy na m-rovině korpusu různými úpravami manuální transkripce zachycené na w-rovině. V 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 na w-rovině neodpovídá žádný w-uzel typu `w` (token) (viz 4.2.2 *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 vymazanou obsahově nerelevantní slovní jednotku (viz 4.2.1 *Mazání*). Hovoříme o **smazaném w-uzlu**.

Pořadí m-uzlů typu `m` na m-rovině nemusí odpovídat pořadí w-uzlů typu `w` na w-rovině.
Změny ve slovosledu (viz 4.2.4 *Změny ve slovosledu*) jsou zachyceny rozdílným uspořádáním uzlů na obou rovinách.

V případě domyšleného nesrozumitelného textu (viz 4.2.3.5 *Nesrozumitelný úsek textu*) vedou z m-uzlů typu `m` odkazy na w-uzel typu `nonspeech`, který má 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 4.2.5 *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 na 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` nevede žádný odkaz z m-roviny, pak je neřečová událost reprezentovaná tímto w-uzlem 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` (`unintelligible`)
	Ø
m-uzel typu `nontext`	w-uzel typu `nonspeech`
	m-uzel typu `background`
	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 4.1.3.1 *Číslice*),
- **`substitution`**: forma nebo lema m-uzlu byly vůči odpovídajícímu w-uzlu upraveny (byla provedena substituce – viz 4.2.3 *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 tzv. **s-elementů**, které reprezentují jednotlivé věty standardizovaného textu.

Každému s-elementu náleží atributy:

- **`w-speaker.rf`**: identifikace mluvčího, který dané obsahové sdělení pronesl. Atribut se vyplňuje automaticky. Pouze v případech, kdy automatický nástroj selže, vyplňuje atribut anotátor.
- **`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.
4. Pomocí modifikací mazání, vkládání, substituování a přesouvání slovních jednotek vytvořit větu splňující podmínky standardizovaného textu a zachovávající principy anotace.
5. Zkontrolovat odkazy do w-roviny (od m-uzlů i od s-elementu).
6. Označit typ věty.
7. Po dokončení anotace souboru – přečíst celý výsledný standardizovaný text a provést případné další úpravy.

1.3 Anotační nástroj MEd

Anotuje se pomocí speciálně vyvinutého anotačního nástroje 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.

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 nutný zejména 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 neumožňují anotátorovi rozhodnout o vhodné modifikaci.

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 2.2.5 *Bezobsažný úsek textu*.

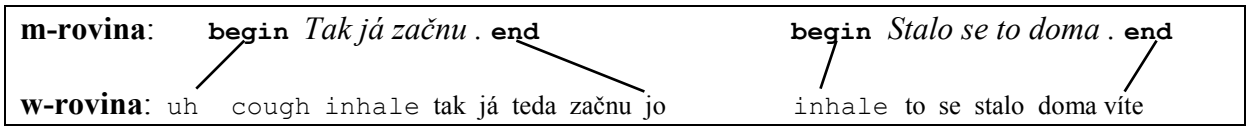
2.1 Vyznačení hranic vět v proudu mluvené řeči

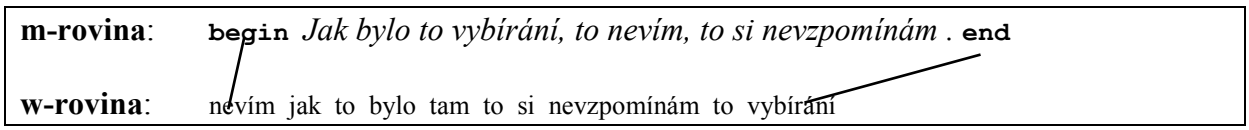
Na m-rovině je větou posloupnost m-uzlů, která je identifikovaná tzv. s-elementem. Tato posloupnost vždy odpovídá nějakému úseku (případně i úsekům) rozpoznaných obsahových událostí na w-rovině. Tento úsek obsahových událostí, který byl použit jako vstup pro výstupní rekonstruovanou větu reprezentovanou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

Neřečové události, různá `ehm`, `breath`, `cough`, `mouth`, které bezprostředně předcházejí vlastním slovům mluvčího (nebo za nimi bezprostředně následují), vždy zahrnujeme do nějaké věty na m-rovině. Mluvčí často, než opravdu něco řekne, nejprve kašle, řekne „ehm“, nadechne se. To všechno je zahrnuto odkazem `w-begin.rf` (a `w-end.rf`) do věty na m-rovině. Podobně zahrnujeme do věty na m-rovině i všechny z ní vymazané úseky textu – vymazané nadbytečné konektory, deiktická slova, vycpávková slova a fráze, falešné začátky, opakující se úseky textu, fragmenty atp. (viz 4.2.1 *Mazání*).
Většinou nezbude na w-rovině žádná obsahová událost, které by na m-rovině neodpovídala žádná věta. To nastane pouze v případě tzv. bezobsažného úseku textu.





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 maximálně o dvou až třech větách hlavních**. Dáváme přednost kratším větám. Pokud mluvčí překotně, bez přerušení, dlouho mluví (neklesá hlasem, nedává signál o konci věty, stále používá nějakou spojku), rozčleníme takový proud mluvené řeči na několik kratších vět.

Nespojujeme nesourodé obsahy.
Příklad:
to jsme byla v sedumapadesátym roce na rekreaci v begin Jeseníkách end vlastně prvně na horách breath a to byla turistická turistycká rekreace , každej den jsme chodili na různý túry breath a bydleli jsme na begin Červenohorskym sedle
V roce 1957 jsem byla na rekreaci v Jeseníkách, prvně na horách.
Byla to turistická rekreace.
Každý den jsme chodili na různé túry.
Bydleli jsme na Červenohorském sedle.

K odpovědi na otázku nepřipojujeme žádnou novou informaci.
Příklad:
Stýkáte se s ní dodnes?
ne nestýkám já jsem se potom v devětapadesátym roce odstěhovala do begin Plzně end a potom nějak už jsme breath vlastně ona se vdala taky jinam z begin Roztok end , takže už jsme se potom neviděly , ale máme teď mít setkání , breath tak se snad uvidíme
Ne, nestýkám.
Odstěhovala jsem se potom v roce 1959 do Plzně a ona se vdala taky jinam z Roztok, takže jsme se už potom neviděly.
Teď ale máme mít setkání, tak se snad uvidíme.

Signálem začátku nového segmentu jsou nadbytečné konektory (viz 4.2.1.4 *Nadbytečné konektory*).
Příklad:
moji rodiče brzo zemřeli , takže v našem baráku bydleli třicet let nějací Hermanové , měli jediného syna , tak když jsme neměli žádnou babičku , tak to byla teta
Moji rodiče brzo zemřeli.
V našem baráku bydleli třicet let nějací Hermanové, měli jediného syna.
Když jsme neměli žádnou babičku, tak to byla teta.

Další příklady:
breath ale teď se nějak prostě dozvěděli zase od~ od jiný holky , která breath taky chodila s náma do třídy a s kterou jsem se potom stýkala , breath protože manžel byl vlastně z vedlejší vesnice , co byla , ta moje breath bývalá spolužačka , takže přes ní nějak se o mně dozvěděli
Teď se ale o mně nějak dozvěděli zase od jiné holky, která taky chodila s námi do třídy.
S tou jsem se potom stýkala, protože manžel byl z vedlejší vesnice, odkud byla i tato moje bývalá spolužačka.

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.
Byl to konec všech nadějí.

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 nadělaly dost starostí.
Zejména potom v pozdějších letech, po okupaci Československa, kdy jsem byl již to, čemu se dá říci politicky činný, těch starostí u rodičů přibývalo.

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 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 ke 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.
Jak pozdější historické výzkumy ukázaly, tento optimismus byl oprávněný.

Znal jste ho osobně i vy?
znal jsem ho osobně , znal jsem ho velice dobře osobně , protože si mě vybral
Znal jsem ho osobně.
Znal jsem ho velice dobře, protože si mě vybral.

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 se i ve standardizovaném textu mohou objevit nedokončené výpovědi.
Ve standardizovaném textu zachycujeme zejména ty nedokončené výpovědi, které již mají nějaký obsah, „nic neříkající“ nedokončené výpovědi (jako například: *A bylo...; To...; Protože potom, jak...*) bez náhrady mažeme a ve standardizovaném textu je nezachycujeme.
Nedokončení výpovědi naznačíme **třemi tečkami na konci výpovědi**.
Pokud je evidentní, co chtěl mluvčí říci, pak takovou nedokončenou výpověď změníme při rekonstrukci na výpověď dokončenou, domyslíme ji (viz 4.2.2.2 *Nevyřčené úseky textu*).

Pozor! Rozlišujeme mezi nedokončenými výpověďmi a fragmenty (viz 4.2.1.8 *Fragmenty*). Na rozdíl od fragmentu nedokončená výpověď je výpověď, kterou mluvčí chtěl pronést (sdělit její obsah), ale z nějakého důvodu ji nedokončil. Fragment naproti tomu je kus výpovědi, od jejíhož dokončení (od sdělení daného obsahu) mluvčí vědomě upustil.

Příklady:
spk1
v období když ste byl v Palestině měl ste nějakou korespondenci s Československem s rodičema nebo
V období, když jste byl v Palestině, měl jste nějakou korespondenci s Československem, s rodiči nebo...
spk2
s Československem ne ale s rodiči jo pomocí červeného kříže kde sem se také dozvěděl že byli deportováni
S Československem ne, ale s rodiči ano, pomocí Červeného kříže, kde jsem se také dozvěděl, že byli deportováni.

breath tak na téhle fotce je moje nejstarší dcera , já EHM vnuk , vnučka a já jsem si měla vzít brejle jó a dcera moje druhá
Na téhle fotce je moje nejstarší dcera, já, vnuk, vnučka...
Měla jsem si vzít brýle.
... a moje druhá dcera.

spk2
mně přes- vždycky říkali přesný termín kdy budu propuštěn nikdy sem v tom daném termínu propuštěn nebyl
Mně vždycky říkali přesný termín, kdy budu propuštěn. Nikdy jsem v tom daném termínu propuštěn nebyl.
spk1
jak ste se dostal teda nakonec
Jak jste 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
Nakonec přece jenom nadešel onen den, kdy jsem byl propuštěn, a to někdy v říjnu.
Snažím se vzpomenout si, kdy to bylo.

spk2
když jsme jeli zpátky , tak ve všech autobusech bylo ticho jako v kostele , kdyby upadl špendlík , tak to byla rána , bohužel to bylo velké zklamání , ale zase nám to přineslo
Když jsme jeli zpátky, ve všech autobusech bylo ticho jako v kostele.
Kdyby upadl špendlík, tak by to byla rána.
Bohužel to bylo velké zklamání, ale zase nám to přineslo...
spk1
takže sestoupili
Takže sestoupili?

spk2
a ta se kontrolovala ta výzdoba , a kdo jí neměl , tak měl průšvih okna musely být vyzdobený , naše praporky , breath sovětské praporky a běda , když domovní důvěrník
Výzdoba se kontrolovala a kdo ji neměl, měl průšvih.
Okna musela být vyzdobená - naše praporky, sovětské praporky.
A běda, když domovní důvěrník...
spk1
jaký průšvih
Jaký průšvih?

spk2
no bylo to tám moc
Bylo to tam moc...
spk1
Bohužel nám vypršel čas.

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 výpovědi, 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 spk2 neměl sem spk1 s projevy nesnášenlivosti v dětství
Jak se k vám chovali spolužáci, kteří o vás věděli, že jste Žid?
Setkal jste se v dětství s projevy nesnášenlivosti?
spk2
neměl sem spk1 s projevy nesnášenlivosti v dětství spk2 neměl sem v tom směru problémy
Neměl jsem v tom směru problémy.

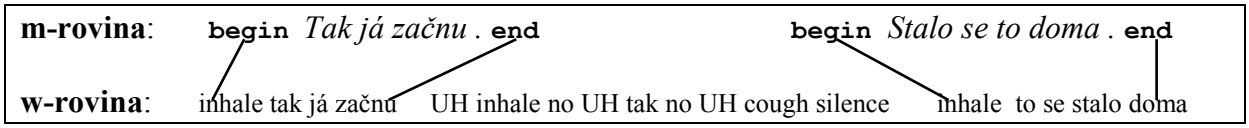
spk1
inhale po devítiletce po obecné škole jste začal studovat na gymnáziu
Po devítiletce, po obecné škole jste začal studovat na gymnáziu.
spk2
na reálném spk1 jaké to tam bylo spk2 gymnáziu ano
Na reálném gymnáziu, ano.
spk1
jaké to tam bylo spk2 gymnáziu ano spk1 inhale co jste tam dělal
Jaké to tam bylo?
Co jste tam dělal?

spk1
jak vono to je s tim počasim
Jak je to s počasím?
spk1
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
Teď má být teplo a pak už zase zima?
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
Já jsem se nedívala, na počasí se dívá táta a říkal, že tři dny teplo.
Dneska se ale můžeš podívat.

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/nebo neplnovýznamových slov, které při větné segmentaci 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.



Jako bezobsažný úsek textu interpretujeme:

A. posloupnost neřečových událostí a/nebo neplnovýznamových slov, která je zřetelně oddělena od ostatních obsahově relevantních výpovědí například delší pauzou, hlukem, kašláním.
Příklad (bezobsažný úsek textu je podtržen):
odešel sem UH inhale no UH tak to cough noise nikdy sem se tam nevrátil
Odešel jsem.
Nikdy jsem se tam nevrátil.

B. obsahově nerelevantní neřečové události `noise`, `cough`, `laugh` realizované na pozadí dialogu: jejich původcem je účastník dialogu, který zrovna nemluví (ale například kašle, protože má kašel), nebo je jejich původcem osoba, která se dialogu vůbec neúčastní.
Příklad:
spk1
byla tam lepší místa na focení než toto , my jsme to ale dělali digitálním fotoaparátem spk2 cough spk1 takže si to doma promítáme
Byla tam lepší místa na focení než toto.
My jsme to ale dělali digitálním fotoaparátem, takže si to doma promítáme.

C. obsahově nerelevantní kladné přitakání druhého mluvčího (posluchače) k obsahu projevu prvního mluvčího (*ano, rozumím*, neřečová událost `EHM` atp.). Projev prvního mluvčího není tímto přitakáním nijak přerušen (viz i 3 *Typy vět podle obsahu*) .
Příklad:
spk1
byla tam lepší místa na focení než toto , my jsme to ale dělali digitálním fotoaparátem spk2 ano spk1 takže si to doma promítáme
Byla tam lepší místa na focení než toto.
My jsme to ale dělali digitálním fotoaparátem, takže si to doma promítáme.

D. „nic neříkající“ nedokončené výpovědi, které nebyly zahrnuty do standardizovaného textu (viz i 2.2.3 *Nedokončené výpovědi*).
Příklad:
spk2
a ta se kontrolovala ta výzdoba , a kdo jí neměl , tak měl průšvih okna musely být vyzdobený , naše praporky , breath sovětské praporky a bylo to
Výzdoba se kontrolovala a kdo ji neměl, měl průšvih.
Okna musela být vyzdobená - naše praporky, sovětské praporky.
spk1
jaký průšvih
Jaký průšvih?

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.

Rozlišujeme pět typů vět a informaci o typu věty ukládáme v atributu `stype`, který náleží s-elementu identifikujícímu hranice vět.

Hodnoty atributu `stype`

`information`	informace, obsahově relevantní věta
`instruction`	příkaz, žádost, aby druhý mluvčí něco vykonal
`guestion`	otázka po informaci
`confirmation`	kladné přitakání druhého mluvčího (posluchače) k obsahu projevu prvního mluvčího
`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).
Například:
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 (a dáváme jim podle smyslu buď `information` nebo `instruction`).
Z formálního hlediska jde o věty tázací.
Například:
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.
Například:
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 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.
Poznámka: Ve většině případů tato „přitakání“ při rekonstrukci vypouštíme jako bezobsažný úsek textu (viz 2.2.5 *Bezobsažný úsek textu*), někdy však může být vhodné takové přitakání ve standardizovaném textu zachytit, pak má hodnotu `confirmation`.
Pozor! Hodnota `confirmation` nenáleží odpovědím na zjišťovací otázky!
Například:
To je pravda.
Souhlas.
Souhlasím.
Aha.
Ano.
Jasně.
Rozumím.

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

Pokud se vyskytne souvětí, které je spojením dvou vět, z nichž každá by měla dostat jinou hodnotu atributu `stype`, dostane celé souvětí hodnotu podle poslední klauze souvětí. Avšak, je-li to vhodné (u delších klauzí), rozdělíme daný úsek do dvou vět a každé přiřadíme vlastní hodnotu atributu `stype`.

Příklady:
`instruction`
To je veselé, pojďme na další fotku.

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:

- 4.1 **Ortografické modifikace**,
- 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:

- 4.1.1 **Odstranění obsahově nerelevantních neřečových událostí**,
- 4.1.2 **Pravopisné náležitosti psaného textu**,
- 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. 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 4.2.5 *Zachycení obsahově relevantních neřečových událostí*).

Neřečové události, které nemají žádný podstatný význam pro obsah sdělení (většina), jsou na m-rovině bez náhrady odstraněny. Odstraněné neřečové události zpravidla vždy zahrnujeme odkazy `w-begin.rf` a `w-end.rf` do nějakého segmentu na m-rovině (viz 2.1 *Vyznačení hranic v proudu v mluvené řeči*). Do segmentů na m-rovině zůstávají nezahrnuté pouze neřečové události, které jsou součástí bezobsažných úseků textu (viz 2.2.5 *Bezobsažný úsek textu*).

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, ehm`	uh, um, uh-huh, uh-hum, hm, ehm
`noise`	hluk v pozadí
`unintelligible`	nesrozumitelný úsek

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

Příklady:
silence mouth inhale tak možná že bych ještě něco řek breath uh silence
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:

- 4.1.2.1 **Vložení interpunkčních znamének**,
- 4.1.2.2 **Náhrada malých písmen za velká**.

4.1.2.1 Vložení interpunkčních znamének

Ve standardizovaném textu jsou správně doplněna veškerá **interpunkční znaménka** (čárky, tečky, pomlčky, uvozovky, dvojtečky, závorky). Jsou reprezentována samostatným m-uzlem.

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

4.1.2.1.1 Čárka, tečka, vykřičník, otazník, uvozovky

Čárky, koncová interpunkční znaménka (tečka, vykřičník, otazník) a interpunkční znaménka pro zápis přímé řeči (uvozovky a dvojtečka) používáme v souladu s pravidly českého pravopisu.
Pokud přímou řeč tvoří více vět, pak jsou to jediné uvozovky (počáteční a koncové), ale je to tolik segmentů, kolik je vět. Případná uvozovací věta je v jednom segmentu společně s první větou přímé řeči.
Do uvozovek dáváme i „neslovníková“ slova (viz 5.2 *Standardizace „neslovníkových“ slov*) a též silné expresivní, vulgární, metaforické a jiné podobné výrazy.

Příklady:
on řekl byl sem tam ale nikdo mu nevěřil
Řekl: „Byl jsem tam,“ ale nikdo mu nevěřil.

inhale my si řikali , no jo , ale dyť to to neni daleko , click inhale jestli nás vemou , tak s nima jedem , to se dostaneme už sem bl - do blízkosti Prahy
Říkali jsme si: „Vždyť to není daleko.
Jestli nás vezmou, tak s nimi pojedeme.
To už se dostaneme do blízkosti Prahy.“

já můžu říct to , inhale se do dneška pamatuju , že inhale EHM jako malý holce inhale EHM moje maminka mi napsala třeba do památníku inhale když bolí něco nechtěj soucitu , neb aspoň u lidí ho nehledej inhale zamkni se před světem a hrdě trp , inhale by slz tvých nezřeli raději se směj
Já můžu říct, to se do dneška pamatuju, že jako malé holce mi moje maminka napsala do památníku: „Když bolí něco, nechtěj soucitu, neb aspoň u lidí ho nehledej.
Zamkni se před světem a hrdě trp, by slz tvých nezřeli, raději se směj.“

4.1.2.1.2 Závorky

Ve standardizovaném textu je též možné s výhodou využívat závorek pro ohraničení vsuvky. Avšak závorkami (vsuvkami) šetříme, vždy zvážíme nutnost jejich použití. Vsuvky a odbočky jsou v proudu mluvené řeči naprosto běžné, v psaném textu je však spíše než vsuvkou (ohraničenou závorkami) nahradíme samostatným, do textu začleněným větným segmentem. Jako vsuvku zachycujeme zejména syntakticky nezačleněné výrazy.

Příklady:
a když se mi podařilo dceru dostat do Anglie na oper ópér , to jest pojem známý , nemusím vysvětlit , je to jakási pomoc v domácnosti , člen rodiny , tak sme se s manželkou dohodli , že to zkusíme
Když se mi podařilo dostat dceru do Anglie na au pair (to je známý pojem, ten nemusím vysvětlovat, je to jakási pomoc v domácnosti, člen rodiny), dohodli jsme se s manželkou, že to zkusíme.

konečně se mi podařilo najít zaměstnání v Ó PÉ BÉ HÁ Praha jedna , obvodní podnik bytového hospodářství.
Konečně se mi podařilo najít zaměstnání v OPBH Praha 1 (Obvodní podnik bytového hospodářství).

breath a druhej den jsme jeli do begin Sokolova end do begin Falknova end , breath dnešní begin Sokolov end a tám jsme byli do pátnáctýho dubna tisíc devětset štyricet šest taky u muziky
Druhý den jsme jeli do Falknova (dnešní Sokolov) a tam jsme byli do 15. dubna 1946 taky u muziky.

4.1.2.1.3 Pomlčka

Ve standardizovaném textu lze též užívat pomlčky, zejména pro oddělení dodatečně připojených členů. Podobně jako v případě závorek i užití pomlčky vždy pečlivě zvážíme. Pomlčky nepoužíváme pro ohraničení vsuvky (tj. namísto závorek).

Příklady:
hrály jsme jako děti kuličky , vždyť to znáte ty dětský hry
Jako děti jsme hráli kuličky - vždyť znáte ty dětské hry.

ale všichni jako bydleli tady okolo begin Plzně end , begin Dobřany end , a begin Litice end , a ty begin Robčiče end , no a mami byla v těch begin Útušicích
Všichni bydleli tady okolo Plzně - Dobřany, Litice, Robčiče.
Maminka byla v Útušicích.

v begin Rokycanech end jsem chodil tři roky do měšťanky breath první , druhý a do třetí
V Rokycanech jsem chodil tři roky do měšťanky – do první, druhé a třetí.

breath no tak učila nás ty všechny , ty předměty , ketrý byly v první třídě jako počty , čtení , prvouku
Učila nás všechny předměty, které byly v první třídě - počty, čtení, prvouku.

4.1.2.1.4 Spojovník

Spojovník užíváme tam, kde je to pravopisně náležité.

Příklad:
na okrese , teď už vlastně okres neň~ není , begin Plzni end jihu
Na okrese, teď už vlastně okres není, Plzeň-jih.

Poznámka: Slova spojená spojovníkem/pomlčkou (*Plzeň-jih, je-li*) jsou na m-rovině reprezentována třemi m-uzly.

4.1.2.1.5 Dvojtečka

Po pečlivém uvážení používáme i dvojtečku, zejména pro uvození výčtu.

Příklady:
a taková zajímavost jednou jsme byli breath EHM v cha~ EHM chata begin Vrátná end se to EHM jmenovalo byli jsme tam na obědě breath a najednou jsme koukali breath a von tam byl herec begin Vojta
Taková zajímavost: Jednou jsme byli v chatě Vrátná.
Byli jsme tam na obědě a najednou jsme koukali a on tam byl herec Vojta.

a když jsem rozbalil , EHM EHM rozbalil obálku , breath tak tam byla cedulka breath begin František Slivoně end breath begin Křimická end sto šest begin Plzeň
Když jsem rozbalil obálku, byla tam cedulka: František Slivoně, Křimická 106, Plzeň.

inhale toto sou moji rodiče , click vlevo maminka Hermína , inhale vpravo otec Eduard
Toto jsou moji rodiče: vlevo maminka Hermína, vpravo otec Eduard.

máme tám jabka , hrušky , třešně breath a manželka pěstuje zeleninu celer , breath kedluben , karfiol všechno možný
Máme tam jablka, hrušky, třešně a manželka pěstuje zeleninu: celer, kedluben, karfiol, ředkvičky, všechno možné.

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:

- **zvětšení písmena na začátku vět**,
- **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).

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

jedna plus dvě rovná se tři
1 + 2 = 3

dvacátý šestý
26.

osmkrát
osmkrát

dvacet pětkrát
25krát

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 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 nutné, nealfabetické znaky nepoužíváme, tj. dáváme přednost zápisu slovy. Nealfabetický znak použijeme jen tam, kde je naprosto běžný.

Příklady:

dvě procenta
dvě procenta

dvacet tři procent
23 procent

jedenadvacet dolarů
21 dolarů

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. Na rozdíl od ortografických modifikací představují podstatný zásah do podoby vstupního textu.

K dispozici jsou následující typy vlastních modifikací:

- 4.2.1 **Mazání**,
- 4.2.2 **Vkládání**,
- 4.2.2 **Substituce**,
- 4.2.4 **Změny ve slovosledu**,
- 4.2.5 **Zachycení obsahově relevantních neřečových událostí**.

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.

Při rekonstrukci je třeba **hodně přemýšlet, jak z toho, co mluvčí řekl, při co nejmenším počtu úprav postavit hezké věty**.
Jednotlivé modifikace mají různou váhu. Vždy se snažíme dosáhnout standardizovaného textu použitím modifikací, které jsou menším zásahem vzhledem k původnímu přepisu mluvené řeči.

Slovosledné změny jsou nejmenším možným zásahem do původního textu. Změny slovosledného uspořádání provedeme vždy, dosáhneme-li jimi „hezčího“ standardizovaného textu.

Modifikace neplnovýznamových slov a tvarů slov je menší úprava než modifikace plnovýznamových slov. Modifikace neplnovýznamových slov a tvarů slov (substituce, mazání, vkládání) jsou běžné úpravy.

Modifikace plnovýznamových slov vždy pečlivě zvažujeme. Přistupujeme k nim jen v nejnutnějších případech, pokud standardizovaného textu nelze dosáhnout jinými úpravami.

Modifikace mazání je menší úprava než modifikace vkládání.
Je lepší některé vyslovené úseky nepoužít než nové úseky vymýšlet (zejména pokud jde o plnovýznamová slova). Mažeme zejména ty úseky, které mluvčí v proudu svého mluveného projevu nějakým způsobem opravil (nahradil jinými, zpřesnil, doplnil, změnil atp.). Domýšlení textu (vkládání plnovýznamových slov) ale použijeme vždy, když je chybějící text evidentní a bez jeho doplnění by výsledný segment byl nedokončený nebo jinak syntakticky neúplný.

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

4.2.1 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:

- 4.2.1.1 **Výplňková slova**,
- 4.2.1.2 **Výplňkové fráze**,
- 4.2.1.3 **Nadbytečná deiktická slova**,
- 4.2.1.4 **Nadbytečné konektory**,
- 4.2.1.5 **Nadbytečná nebo nesprávně užitá gramatická slova**,
- 4.2.1.6 **Opravené úseky textu (restarty)**,
- 4.2.1.7 **Opakující se úseky textu**,
- 4.2.1.8 **Fragmenty**.

4.2.1.1 Výplňková slova

jako
vlastně
tedy
prostě
no
jo
žejo

Výplňková slova (též vycpávková) jsou slovní jednotky, které 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.

Příklady:
hledali nějakýho ubožáčka že jo
Hledali nějakého ubožáčka.

no tam jsme byli dva roky
Tam jsme byli dva roky.

tam je jako poutní místo
Je tam poutní místo.

tam sa jde kus jako pěšky
Jde se tam kus pěšky.

breath no já jsem tam byla myslim dvakrát jako za svobodna a potom už jsem tam nebyla
Byla jsem tam za svobodna myslím dvakrát a potom jsem tam už nebyla.

breath a potom , když jsme se přestěhovali do EHM jiného domu , tak jsme vlastně byli sousedi , ona bydlela vedle nás
Potom, když jsme se přestěhovali do jiného domu, jsme byli sousedi, ona bydlela vedle nás.

pro mě to bude první setkání vlastně po těch asi víc , než padesáti letech , breath kdy se uvidim s tima spolužákama , co jsem s nima chodila do třídy
Pro mě to bude první setkání po víc než padesáti letech, kdy se uvidím s těmi spolužáky, co jsem s nimi chodila do třídy.

Proč se vám tam nelíbilo?
breath no za prvé tedy moc mi nešla matematika nikdy a potom jako k tomu zemědělství jsem neměla nějaký vztah breath i když moje rodiče oba byli ze zemědělství , ale breath mě se to nelíbilo , neměla jsem k tomu prostě vztah
Nikdy mi moc nešla matematika a k zemědělství jsem neměla žádný vztah, i když oba moji rodiče byli ze zemědělství.
Mně se to ale nelíbilo, neměla jsem k tomu vztah.

4.2.1.2 Výplňkové fráze

to víte
myslím
jak vidíš
vždyť víš
nedej bůh
páne jo
jéžišmarjá

K **výplňkovým frázím** řádíme většinou ustrnulé slovesné konstrukce, které klesají v pouhé částice. Z věty je odstraňujeme tehdy, jestliže narušují 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).

Příklady:
to bylo v Praze na myslím na Vánoce
To bylo v Praze na Vánoce.

Co dělá dnes váš syn?
breath no dnes jéžiš noise marjá laugh já si nemůžu zapamatovat , co je to za podnik
Já si nemůžu zapamatovat, co je to za podnik.

4.2.1.3 Nadbytečná deiktická slova

ten, ta, to
takový
tam

Ukazovací zájmena *ten, ta, to* a další deiktická slova (*takový, tam, tady*) jsou v mluveném projevu užívána mnohem častěji než v psaném textu. Mluvený projev probíhá v konkrétním

prostoru a čase, v aktuální komunikační situaci a mluvčí k této situaci neustále odkazuje. Deiktická slova jsou v mluveném projevu často používána jako vycpávková slova pro zdůraznění pojmenovaných slov jako "toho, jak jsme o tom už mluvili" a z jiných důvodů, které nejsou v psaném textu relevantní.
Při rekonstrukci **tato deiktická slova mažeme**.

Za **nadbytečné deiktické slovo** (jde o nadbytečnost z hlediska psaného textu) považujeme zejména:

A. deiktické slovo, které **vyplňuje 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. deiktické slovo užité jako **vycpávkové slovo** (mluvčí si není jist, přemýšlí, proto než jméno uvede, nahradí ho nejprve zájmenem).
Příklad:
byl jsem v tom eh táboře
Byl jsem v táboře.

C. deiktické slovo užité pro identifikaci pojmenovaných věcí jako **"toho, jak jsme o tom už mluvili"**. Jako nadbytečná deiktická slova tohoto typu chápeme především ukazovací zájmena před vlastními jmény a názvy, jež jsou samy dostatečnými identifikátory pojmenované entity.
Příklad:
jel sem do té Prahy
Jel jsem do Prahy.

D. neutrální **zájmeno *to* ve funkci počátečního výrazu nebo ve funkci připojovací**.
Příklad:
to jsme jeli na motorkách tehdy b breath tam to , jak bylo také dobrodružná cesta
Jeli jsme tam tehdy na motorkách, byla to také dobrodružná cesta.

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.*)

Další příklady**:**
a noise za ní stojí můj dědeček tedy otec , který b~ byl , tuším v tom ten čtvrtý , EHM na té fotografii předešlé
Za ní stojí můj otec, který byl na předešlé fotografii, tuším, čtvrtý.

tak to bylo nějak vždycky spojeno s tou přírodou pro ty vesnické lidi
Pro vesnické lidi to bylo vždycky nějak spojeno s přírodou.

takže děda byl v tom begin Rumburku end , a byl tam také nějak do té vzpoury zapleten , ale breath protože neorganizoval přímo breath nebyl ten hlavní organizátor breath nestalo se mu tedy dohromady nic , ale mnoho nechybělo aby skončil zrovna tak , jako skončil ten jeho známý begin Noha end
Děda byl v Rumburku a byl tam také nějak do té vzpoury zapleten, ale protože nebyl hlavní organizátor, nestalo se mu dohromady nic.
Ale mnoho nechybělo, aby skončil zrovna tak, jako skončil jeho známý Noha.

breath no a tam jsme si sedli , tak do stínu , aby jsme tak pozorovali tu krajinu a to všecko okolo to dění , breath tak jsme si tam sedli a ten muj nastávající , tak si lehl na to moje koleno , aby jsme si odpočinuli
Sedli jsme si do stínu, abychom pozorovali krajinu a všecko dění okolo.
Sedli jsme si a můj nastávající si lehl na moje koleno, abychom si odpočinuli.

ona bydlela tam dole v begin Tichém údolí end
Ona bydlela dole v Tichém údolí.

4.2.1.4 Nadbytečné konektory

a
tak
takže
že
pak
no

Spojky *a, tak, takže*, ale i podřadicí spojka *že* (někdy též *protože*), příslovečný výraz *pak*, nespisovná částice *no* často nemusí v proudu mluvené řeči nést své významy - slučovací, důsledkové, časové, připojení obsahové věty vedlejší. Velmi často jsou používány jen jako prostředek připojení nové informace. Plní funkci prostě připojovací, navazovací, segmentovací (mají funkci „tečky a velkého písmena“).
Při rekonstrukci jsou tyto výrazy **signálem pro rozdělení segmentu a začátku nové věty**.

Poznámka: Odstraněný nadbytečný konektor je odkazem `w-begin.rf` zahrnut do věty následující.

Příklady:
a tam to trvalo roky a spolu začali dělat todleto
Tam to trvalo roky.
Spolu začali dělat tohleto.

ona bydlela tam dole v begin Tichém údolí end breath no já jsem často k nim chodila , takže jsme tak jako v těch letech , co děvčata dělají no breath povídaly jsme si , hrály takový různý, ty věci z tý doby no
Ona bydlela dole v Tichém údolí.
Často jsem k nim chodila.
Povídaly jsme si, hrály takové různé věci z té doby tak, jak to v těch letech děvčata dělají.

breath její syn potom byl v begin Indii end indyji , tak a v tu dobu , když byl v té begin Indii end indyji , tak jí zemřel manžel , no tak jsme samozřejmě byli breath všechno na nás , protože oni příbuzný breath měli v begin Praze end , nikoho jiného taky neměli a to byly neteře breath no tak jsme s~ jako o tu tetu se starali
Její syn byl potom v Indii a v tu dobu jí zemřel manžel.
Všechno to bylo samozřejmě na nás, protože oni měli příbuzné v Praze.
Byly to neteře, nikoho jiného neměli.
Tak jsme se o tetu starali.

tam choďijá hodně mladí na ty , na tu pouť , že sa tam jako seznamujú a tak
Na pouť tam chodí hodně mladí.
Seznamují se tam a tak.

4.2.1.5 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
Bylo to v Praze o Vánocích.

C. spojky.
Příklad:
to se stalo mně a Járovi a Pavlovi
Stalo se to mně, Járovi a Pavlovi.

D. osobní zájmena v pozici subjektu.
Příklad:
já se menuju Marek
Jmenuju se Marek.

4.2.1.6 Opravené úseky textu (restarty)

Mluvený projev probíhá v čase, lineárně. Na rozdíl od psaného textu se mluvčí při výstavbě mluveného projevu nemůže vrátit, aby vymazal a opravil, co řekl nesprávně, zlepšil výběr slov, změnil použité gramatické kategorie. Takové opravy může mluvčí dělat jen tak, že daný úsek projevu řekne znovu. Výsledkem je velmi rozvolněná syntax s mnoha odbočkami, anakoluty, nedokončenými a znovu započatými větami, doplňujícími sděleními.

Prototypickým příkladem „opravy“ v proudu mluvené řeči je **restart** ve své základní formě:
`falešný začátek – (korektor) – oprava.`
Falešný začátek je úsek textu, který mluvčí posléze 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. **Oprava** je pak opravený původní falešný začátek.

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.

Restart může mít ale i daleko složitější podobu. Falešných začátků (než se mluvčí dobere k tomu, co chtěl říci) může být i několik.

Z tohoto pohledu lze proud mluvené řeči dělit spíše na opravené a neopravené úseky textu: `úsek1 – (korektor) – oprava(úsek1)`, přičemž oprava nějakého úseku nemusí následovat bezprostředně za tímto úsekem.

Při rekonstrukci následně **opravené úseky textu a korektory mažeme**.

Existuje celá řada různých druhů restartů:

A. vlastní oprava.
Příklad:
v pátek teda vlastně v sobotu sme tam šli
V sobotu jsme tam šli.

B. zakoktání.
Příklad:
v tomto z- zd- zděném baráku byly betonové kobky
V tomto zděném baráku byly betonové kobky.

C. upřesnění.
Příklad:
syn můj syn už se nevrátil
Můj syn už se nevrátil.

D. 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
Byl to většinou personál.

Další příklady:
a přes EHM breath přes léto vlastně od jara do podzimu
od jara do podzimu

breath to jsem oslavovali moje EHM dědečkovi šedesátiny manželovi
Oslavovali jsme manželovy šedesátiny.

co sedí breath vedle je dědeček jako pradědeček , no manže~ manželůj otec
Ten, co sedí vedle, je manželův otec.

Kdy se vaše dcera odstěhovala do Plzně?
breath no , myslim roku devatenáscet , devatenáscet šedes~ unintelligible šedesát nemyslim , ňák devatenáscet semdesát nebo tak nejak
Myslím, že roku 1970 nebo tak nějak.

breath no , to bylo tak ňak breath padesátém pátém roku nebo tak nějak devatenáscet padesát pět
Bylo to tak v roce 1955.

a to je nejmladší je sestra jako no mladší sestra , to je begin Anička end sa jmenuje
To je mladší sestra, jmenuje se Anička.

Jaké bylo tenkrát počasí?
nádherné, tam bylo krásně , sluněčko tam krásně svítilo , nádherně tam bylo , nádherné počasí aj ta voda to bylo tam nádherné
Bylo tam krásně, sluníčko krásně svítilo.
Bylo tam nádherné počasí i voda.

ten je jako inženýr a dělá ještě dálkově dalej školu breath študuje ještě dál
Je inženýr a dálkově ještě studuje dál.

Kdy chodil do tanečních?
breath jéžiš čtyry , šest breath asi v osumdesátym roce , protože mu bylo šestnáct a narodil se breath v roce šedesát čtyři no , breath takže v tom osumdesátym až noise na přelomu osumdesátýho a jednaosumdesátýho roku
Bylo mu šestnáct a narodil se v roce 1964, takže až na přelomu roku 1980 a 1981.

Kolik máte celkem dětí?
breath no měli jsme dvě , breath ale starší syn před osmi zemřel , takže teď mam jenom toho syna a manžel před třema lety taky zemřel , no tak breath mam jenom toho syna a vnoučata , takže jedno dítě no
Měli jsme dvě, ale starší syn před osmi lety zemřel.
Manžel před třemi lety taky zemřel, tak mám jenom syna a vnoučata, takže jedno dítě.

4.2.1.7 Opakující se slova i celé úseky textu

Z původního transkribovaného textu jsou na m-rovině odstraněny **slova i celé úseky textu, které se opakují** v případě, že opakování nemá žádný podstatný význam pro obsah sdělení. V případě, že se opakuje doslovně tentýž úsek textu, mažeme první úsek. Při nedoslovném a jinak složitě komplikovaném opakování je možné do standardizovaného textu promítnout různé části opakujících se úseků textu.

Příklady:
a odvedli nás do do do toho karanténního bloku
Odvedli nás do karanténního bloku.

to bylo poslední poslední jídlo
Bylo to poslední jídlo.

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.

a byla tam vždycky veliká slávn~ slavná mša tam byla
Byla tam vždycky velká slavná mše.

dobrý , dobrý žák breath von se učil velmi dobře
Dobrý žák, učil se velmi dobře.

Kolik tehdy bylo těm dětem na fotce?
těm dětem na fotce , já si myslím , breath že to nebylo dlouho noise předtím než došlo k té tragédii breath protože mluví se o tom , že ta mladší ta begin Maruška end , že jí bylo pět let když zemřela no
Myslím si, že to nebylo dlouho předtím, než došlo k té tragédii, protože se mluví o tom, že mladší, Marušce, bylo pět let, když zemřela.

Takže po horách jenom chodíte?
jenom chodim jenom chodim
Jenom chodím.

Teď už to tam vypadá jinak?
úplně jinak , úplně jinak
Úplně jinak.

breath dědeček , ten jako rybu nejedl , ten vždycky říkal , že sa ryb přejedl v begin Itálii end , když byl na vojne , tak ten rybu nejedl
Dědeček rybu nejedl.
Vždycky říkal, že se ryb přejedl v Itálii, když byl na vojně.

Co je to ta závěrečná?
no jako breath jako prodloužená , ale zavěreč~ úplně poslední breath na ukončení těch tanečních breath a tam vlastně jsou pozvaný i rodiče a je to prostě na konec tanečních no
Je jako prodloužená, ale závěrečná je úplně poslední na ukončení tanečních.
Jsou tam pozvaní i rodiče.

4.2.1.8 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). Jde vlastně o **samotný falešný začátek**, od kterého bylo úplně upuštěno, nový začátek, oprava se týká úplně něčeho jiného.
Fragment je třeba odlišit od nedokončené výpovědi (viz 2.2.3 *Nedokončené výpovědi*).

Příklady:
v pátek sem `cough` Barňák pak odešel
Barňák pak odešel.

já to byl většinou ten personál
Byl to většinou personál.

breath to jsem byla , já strašně ráda cestuju a manžel , ten hrozně nerad , takže kamarádka mě vždycky s manželem , když jezdili breath no my jsme měli takový party , breath že jsme , byli jsme tři rodiny , tři dvojice , děti už jsme měli odrostlý , breath tak jsme dost jezdili
Já strašně ráda cestuju a manžel hrozně nerad.
Měli jsme takové party, byli jsme tři rodiny, tři dvojice, děti už jsme měli odrostlé, tak jsme dost jezdili.

breath no a ta díčka , slavila osumnácté narozeniny breath a chlapcí , já jsem prvně byla v nemocnici
Ta dívka slavila osmnácté narozeniny.
Já jsem byla prvně v nemocnici.

4.2.2 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:
- 4.2.2.1 **Chybějící gramatická slova**,
- 4.2.2.2 **Nevyřčené úseky textu**.

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

4.2.2.1 Chybějící gramatická slova

Do vstupního textu jsou na m-rovině vkládána gramatická slova na pozice, kde chybí a kde jsou nezbytná pro vytvoření gramaticky správné věty.

K **chybějícím gramatickým slovům** patří:

A. pomocná a modální slovesa.
Příklad:
on v té válce zabit
On byl v té válce zabit.

B. předložky.
Příklad:
bratrem sme byli v těch vybraných
S bratrem jsme byli v těch vybraných.

C. spojky.
Příklad:
přines chleba čaj
Přinesl chleba a čaj.

D. zájmena. Zájmena doplňujeme tam, kde je to nezbytné z důvodu koherence textu.
Příklad:
přišla Hana a Pavel přines chleba přines chleba čaj
Přišla Hana a Pavel.
On přinesl chleba a čaj.

4.2.2.2 Nevyřčené úseky textu

Mluvčí může některé části svých výpovědí z nejrůznějších důvodů neříct (je jasné, co chce říci; naznačí to, co chce říci, gestem; je přerušen).
Při rekonstrukci domýšlíme nevyřčené úseky textu jen v těch případech, kdy je **nevyřčený text jednoznačně odvoditelný z kontextu** a jeho doplnění přispívá k vytvoření plynulého standardizovaného textu. Domýšlení textu je ta nejposlednější úprava, kterou děláme, ale provedeme ji vždy, dosáhneme-li touto úpravou plynulého „hezčího" standardizovaného textu. V případech, které jsou evidentní, je lepší věta domyšlená než nedokončená.
Typickou úpravou tohoto typu je doplnění chybějícího slovesa.

Příklady (domyšlené úseky jsou podtržené):
gestapáci najednou inhale prostě inhale vnikli do našeho tábora ihned inhale ihned alarm že jo ihned do pozoru
Gestapáci najednou vnikli do našeho tábora, ihned byl alarm, ihned jsme museli do pozoru.

nám pro pořád jenom připomínali jak~ ~ké je nebezpečí , breath že může každou chvíli ta válka nastat , takže breath nebyli to také docela , i k~ i když v míru , breath tak přece jenom to nebylo tak mírové
Pořád nám jenom připomínali, jaké je nebezpečí, že může každou chvíli nastat válka.
I když jsme byli v míru, přece jenom to nebylo tak mírové.

inhale no to víte velké uvítání no inhale spousta slz a radosti
Bylo velké uvítání, spousta slz a radosti.

povídám proboha , kdo to zvoní a on to syn , že usnul a přejel do Ždírce
Povídám: „Proboha, kdo to zvoní"?
Byl to syn, volal, že usnul a přejel do Ždírce.

silence inhale revolverem mu takle začali před nos inhale a chtěli aby řekl sieg heil cough jo silence
Revolverem mu takhle začali dělat před nosem a chtěli, aby řekl: "Sieg heil".

Těsnopis jsem se učila trochu v Terezíně, ale nebyla jsem schopna se tím živit.
ale inhale no řekla sem si no EHM musim jít dělat a budu se snažit si nějaký
Řekla jsem si ale: "Musím jít dělat a budu se snažit si nějaké peníze vydělat."

a takže si koupila inhale EHM právě tady EHM v tý obci Labuť EHM pod tou Přimdou inhale EHM
Koupila si domek právě tady v obci Labuť pod Přimdou.

m - manželka byla mladá inhale vona nechtěla bez dětí
Manželka byla mladá, nechtěla jít bez dětí.

Nebyl žádný předpoklad, že bych mohla jít studovat.
inhale manžel nechtěl abych dojížděla do Prahy a a taky prostě při malých dětech to
Manžel nechtěl, abych dojížděla do Prahy a taky při malých dětech to nešlo.

Musela si obstarat křestní listy svých předků až do třetí generace, to nikde jinde nebylo.
tak to EHM to s - nevim , nesetkala sem se s nikým komu by to bývali takovymto způsobem
Nesetkala jsem se s nikým, komu by to takovýmto způsobem ztrpčovali.

Vzpomínky vždycky zůstanou na věky věčné a nikdo je až do konce mého života nesmaže.
no inhale todleto inhale nemůžu
Nic jiného k tomu nemůžu říct.

4.2.3 Substituce

Některé úseky textu, přestože nebyly vyřčeny úplně přesně a mají některé nedostatky zejména ve tvarech použitých slov, ponechává mluvčí z nejrůznějších důvodů neopravené, tj. neříká je znovu správně (chyba není na závadu srozumitelnosti, mluvčí si chybu neuvědomuje).
Při rekonstrukci **opravujeme syntakticky neúplné a porušené věty.**
Použitá slova by ve standardizovaném textu měla odpovídat vyjadřovanému významu. Mluvčí však může z nejrůznějších důvodů užít slovo z významového hlediska nevhodné, špatné (z neznalosti, z přeřeknutí). **Slova užitá nesprávně z hlediska vyjadřovaného významu jsou při rekonstrukci nahrazována významově vhodnějšími protějšky.**
Ve standardizovaném textu jsou navíc užívány jen spisovné a též jen správně utvořené tvary slov. Při rekonstrukci **jsou měněny vstupní nespisovné a nesprávně utvořené formy slov.**

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

Při rekonstrukci substituujeme:

- 4.2.3.1 **Nespisovné a nesprávně utvořené tvary slov**,
- 4.2.3.2 **Slova užitá nesprávně z hlediska vyjadřovaného významu**,
- 4.2.3.3 **„Neslovníková“ slova**,
- 4.2.3.4 **Syntakticky neúplné a nesprávné konstrukce**,
- 4.2.3.5 **Nesrozumitelný úseku textu**.

4.2.3.1 Nespisovné a nesprávně utvořené tvary slov

Na m-rovině jsou nespisovné a nesprávně utvořené formy slov nahrazeny spisovnými tak, aby standardizovaný text byl „spisovná čeština“.
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.

Mezi **spisovné** tvary (hovorové) patří například tvary:
mohu i *můžu*; *mažu* i *maži*, *mažou* i *maží*, *kopu* i *kopám*, *řežu* i *řezám*.
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.*

tadyhleto, tuhleto, tenhle, tamhlety, tyhlety, tohleto, těhle
teďka, dneska
taky

Nespisovné naproti tomu je například:
bysme, *by jsme, by jste*
začnul, načnul, začla,
todleto, tudlety, tamdlety, todle, tendle,
kór, kórd

no, *jo* ve významu „ano“ nahrazujeme spisovným *ano*. V jiných funkcích tato slůvka mažeme.

B. forma slova je nesprávně utvořená.
Forma slova vyjadřuje nesprávně hodnotu některé 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, nářeční slova, slova vulgární se neutrálními spisovnými protějšky nenahrazují. Nezměněna tedy zůstanou slova jako:
barák, lágr, cimra, čuně, holt, furt, špitál, sanitárka, děcka, pracka, mamina, mami

Silně expresivní, vulgární nebo nářeční slova dáváme do uvozovek.

4.2.3.2 Slova užitá nesprávně z hlediska vyjadřovaného významu

Slova užitá nesprávně z hlediska vyjadřovaného významu jsou nahrazována vhodnějšími významovými protějšky.
Substituci plnovýznamových slov, tj. náhradu jednoho plnovýznamového slova jiným plnovýznamovým slovem, vždy pečlivě zvažujeme. Provádíme ji jen ve významově nesourodých případech: užité plnovýznamové slovo se do kontextu opravdu nehodí, je evidentní, že bylo použito omylem, z neznalosti.

Substituci slov užitých nesprávně z hlediska vyjadřovaného významu provádíme zejména tehdy, když namísto „správného“ slova mluvčí užil:

A. slovo zvukově podobné, avšak významem zcela odlišné (tzv. paronymum).
Mluvčí si „správné“ slovo spletl s jiným slovem, přeřekl se.

Příklad:
nastává hysterický okamžik
Nastává historický okamžik.

B. slovo významově velmi blízké, avšak v daném kontextu nevhodné.
Jde o záměnu „správného“ slova slovem příbuzným, významově blízkým, nicméně v daném kontextu nevhodným, nebo nepříliš vhodným. „Nesprávné“ slovo užil mluvčí například z toho důvodu, že už je starší a někdy špatně hledá správná slova, nebo proto, že byl delší dobu v cizině a některá slova se mu pletou, nebo prostě jen z neznalosti.
Může se jednat nejen o nesprávně užité slovo plnovýznamové, ale i o nevhodně užité slovo pomocné (například *takže* místo náležitějšího *tak*).
Takové případy opravujeme, ale jen tehdy, když mluvčím užité slovo by se v daném kontextu v psaném textu vyskytlo jen velmi nepravděpodobně, neprovedení substituce by znamenalo, že výsledný standardizovaný text nebude „rozumná“ čeština.

Příklady:
když má někdo narozeniny a svátek takže si sejd~ sedneme, sejdeme se, popovídáme no, oslavíme to
Když má někdo narozeniny a svátek, tak se sejdeme, sedneme si, popovídáme a oslavíme to.

jedině když se sedí večer v hospodě , takže se něco vypije
Jedině když se sedí večer v hospodě, tak se něco vypije.

tak jsem začal mluvit jaký má krásný obrazy
Začal jsem říkat, jaké má krásné obrazy.

architekt zelenka má velikou zálohu o tuto činnost
Architekt Zelenka má velikou zásluhu na této činnosti.

celá řada inhale obchodních přátel se bála mě zaměstnat , inhale protože se báli , že by jim klesly kontakty s Československem
Celá řada obchodních přátel se mě bála zaměstnat, protože se báli, že by jim ubyly kontakty s Československem.

takže sem potom nakonec EHM přijal nabídku jet do Salzburgu a pokusit se vytáhnout z bláta a louže inhale jednu menší firmu , která se dostala do potíží , protože tu organizaci vůbec neuměli
Nakonec jsem přijal nabídku jet do Salzburgu a pokusit se vytáhnout z bláta a louže jednu menší firmu, která se dostala do potíží, protože vůbec neuměli organizovat.

takže t~ tohle si vždycky vzpomenu o co ta práce je dneska daleko snažší , a co už jí ubylo
Vždycky si uvědomím, o co je ta práce dneska snazší a co už jí ubylo.

4.2.3.3 „Neslovníková“ slova

Některé typy „neslovníkových“ slov (nesprávně užitá cizí slova, přeřeknutí, nedokončená slova aj.) nahrazujeme náležitými protějšky. Standardizace „neslovníkových“ slov je podrobně popsána v samostatné sekci – viz 5.2 *Standardizace „neslovníkových“ slov*.

Příklady:
a myslim si , že jsme by~ breath v begin Praze end v těch zahradách nějakejch
Myslím si, že jsme byli v Praze v nějakých zahradách.

dal sem si dobrou klábosu
Dal jsem si dobrou klobásu.

4.2.3.4 Syntakticky neúplné a nesprávné konstrukce

Výstavba segmentů mluveného projevu probíhá lineárně, mluvčí se nemůže vrátit, aby vymazal a opravil, co řekl nesprávně, taktéž zcela jistě nemá vždy přesně dopředu rozmyšlené, jak to, co chce říci, řekne. Výsledkem je, že syntaktická stavba mluveného projevu je často velmi nedokonalá, nejrůznějším způsobem porušená. Mluvčí, přestože si může některé nedostatky svého projevu uvědomovat, je však nemusí opravit.
Věty ve standardizovaném textu jsou však ze syntaktického hlediska správně utvořené, nejsou v nich nesprávné vazby, anakoluty, kontaminace a jiné nevhodné formy výpovědí. Nesprávné syntaktické konstrukce nahrazujeme náležitými syntaktickými protějšky.

K nesprávným syntaktickým konstrukcím patří:

A. kontaminace a jiné nesprávné vazby slov.
Kontaminace je zkřížení vazeb náležejících k různým (významově, případně i zvukově podobným) slovesům nebo dějovým jménům. Nově vzniklá vazba je hodnocena jako nenáležitá. Nahrazujeme veškeré nesprávně užité předložkové vazby a pádové tvary.

Příklady:
pokoušela sem se rodiče přesvědčit k odchodu
Pokoušela jsem se rodiče přesvědčit o odchodu.
nebo: *Pokoušela jsem se rodiče přimět k odchodu.*

architekt Zelenka má velikou zásluhu o tuto činnost
Architekt Zelenka má velikou zásluhu na této činnosti.

B. atrakce.
Atrakce, neboli syntaktická spodoba, je nevhodné přizpůsobení, připodobnění jednoho výrazu předchozímu nebo následujícímu výrazu. Příslušný výraz obdrží vlivem sousedního nebo vzdálenějšího výrazu jiný tvar, než jaký mu podle gramatické závislosti přísluší.

Příklady:
dbali na stav nám svěřeným nástrojům
Dbali na stav nám svěřených nástrojů.

nebyli schopni většině postiženým pomáhat
Nebyli schopni většině postižených pomáhat.

C. zeugma.
Zeugma je zanedbání odlišné vazby dvou různovazebných sloves, případně jiných dějových slov. K oběma slovesům je připojen jediný společný závislý člen, zatímco náležitě by měl ke každému slovesu být připojen samostatný výraz v náležitém tvaru.

Příklady:
v tanečních se poprvé setkal a oslovil svou příští dívku
V tanečních se poprvé setkal se svou příští dívkou a oslovil ji.

překvapovaly mě a nesouhlasila jsem s jeho názory
Překvapovaly mě jeho názory a nesouhlasila jsem s nimi.

D. anakolut.

Anakolut je vybočení, vyšinutí z náležité konstrukce výpovědi. K začátku výpovědi, který je dále nějak komplikovaněji rozvíjen a zpřesňován, případně přerušen vsuvkou, se mluvnicky nenáležitě připojí zbytek textu.

Příklady:

všelijakým zklamáním , která jsme nečekali a s jejichž dopadem jsme předem nemohli počítat , i když jsme se snažili o prozíravost a nadhled , nám život také uštědřil

Všelijaká zklamání, která jsme nečekali a s jejichž dopadem jsme předem nemohli počítat, i když jsme se snažili o prozíravost a nadhled, nám život také uštědřil.

se jménem tvůrce , který žil ve Spojených státech, kde vydal většinu svých odborných pojednání, jistě neslyšíte poprvé

Jméno tvůrce, který žil nejprve ve Spojených státech, kde vydal většinu svých odborných pojednání, jistě neslyšíte poprvé.

a EHM k tomu chlapci , který tam tu fotku tedy údajně zkazil dle jeho otce breath to je dnes velice známý malíř , který EHM učil na filozofické fakultě breath v Praze

Ten chlapec, který tu fotku dle jeho otce údajně zkazil, je dnes velice známý malíř, který učil na Filozofické fakultě v Praze.

protože begin Pavel Nedvěd end , breath ten EHM pochází ze begin Skalný u Chebu end a přišel do~ EHM do mladšího dorostu , myslim , že přišel až breath do begin Viktorky end spk2 a když neměl kde spk1 EHM spk2 kde~ kde bydlet , tak bydlel EHM u begin Pepíka Žaloudka end breath v begin Nejřanech

Pavel Nedvěd pochází ze Skalné u Chebu.
Do mladšího dorostu, myslím, přišel až do Viktorky.
A protože neměl kde bydlet, tak bydlel u Pepíka Žaloudka v Nýřanech.

tady možná , že je jich o něco méně nevím , nepočítal jsem je , ale breath většinou nás bylo tak šedesát , kteří zpívali , breath a řídil nás , je tam jako dirigent dyrygent také profesor begin Bohumír Liška end , breath který později působil jako dirigent dyrygent v begin Národním divadle

Tady jich je možná o něco méně, nevím, nepočítal jsem je, ale většinou nás bylo tak šedesát, kteří jsme zpívali.
Řídil nás jako dirigent profesor Bohumír Liška, který později působil jako dirigent v Národním divadle.

Jaký byl ten Jarda Kracík žák?

dobrý , dobrý žák breath von se učil velmi dobře a ono vždycky já nepamatuju ani , že by ti mí známí žáci , kteří jsou dnes z~ třeba známější než begin Jarda Kracík end , třeba dneska breath begin Jirka Kučera end , který tam dělá trenéra prvnímu mu~ mužstvu nebo , který byl v reprezentaci a tam dělal kapitána , breath tak bych mohl jmenovat jich mn~ mnoho a mnoho tak oni se všichni učili dobře , velmi dobře , nebo alespoň dobře

Dobrý žák, učil se velmi dobře.
Bylo to tak vždycky, všichni mí známí žáci, kteří jsou dnes třeba známější než Jarda Kracík, se učili velmi dobře, nebo alespoň dobře.
Třeba dneska Jirka Kučera, který dělá trenéra prvnímu mužstvu nebo který byl v reprezentaci a dělal tam kapitána.
Mohl bych jich jmenovat mnoho a mnoho.

E. korefenční důvody.

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.

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

Další příklady:
Petr dobíhal na poslední chvíli Honza taky on to už ale pak nestihnul
Petr dobíhal na poslední chvíli.
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 tou paní sme na tý lavičce seděli až do oběda
S paní Novákovou jsme na lavičce seděli až do oběda.

4.2.3.5 Nesrozumitelný úseku textu

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

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

V případě, že text domyslet nelze, řídíme se pravidly uvedenými v 4.2.5 *Zachycení obsahově relevantních neřečových událostí*. Souhrnně též 5.3 *Nesrozumitelný úsek textu*.

4.2.4 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
Leželi jsme po pěti.

tam my sme autem jeli
Jeli jsme tam autem.

to bylo loni v červnu
Bylo to loni v červnu.

prosté měření terénu sme dělali
Dělali jsme prosté měření terénu.

sem jel s ním do Zvolena
Jel jsem s ním do Zvolena.

4.2.5 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*).

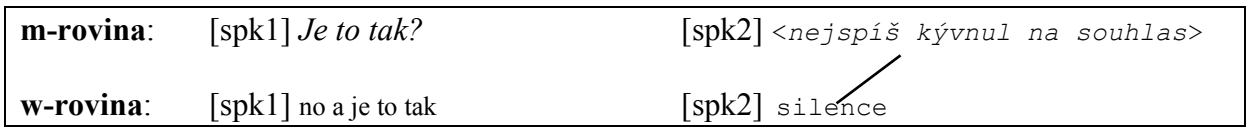
Obsahově relevantní neřečové události, tj. takové, které nesou nějaký význam, jímž přispívají k obsahu sdělení, **zachycujeme ve standardizovaném textu primárně prostředky psaného textu**, tj. zejména pomocí interpunkčních znamének, slovosledu.
Takto zaznamenáváme například:
- věty pronesené s důrazem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

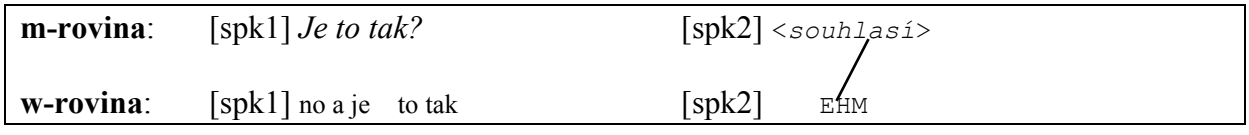
Význam pro obsah sdělení může mít ale celá řada neřečových událostí, které jen pomocí běžných prostředků psaného textu nezachytíme (ironický smích, šeptání, náhlé zvýšení hlasu aj.). Takové **obsahově relevantní neřečové události** zachycujeme na m-rovině **speciálním typem m-uzlu, m-uzlem typu `nontext`.** M-uzlu typu `nontext` náleží atribut `type`, ve kterém anotátor (vlastními slovy) uvede popis, interpretaci významu neřečové události.

Příklady popisů:
smích
pochichtává se
hlasitě zakašlal
váhá (ticho)
nejspíš kývnul na souhlas
souhlasí
přitakává na souhlas
ztišil hlas
předchozí slovo vysloveno hodně nahlas
hvízdnul

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

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





m-rovina:	[spk1] *Odjeli jsme dvanáctého.*	[spk2] `<přitakává na souhlas>`
w-rovina:	[spk1] odjeli sme dvanáctýho	[spk2] `EHM`

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`).
Obsahově relevantní nesrozumitelný úsek textu nahrazujeme na m-rovině primárně textem domyšleným (viz 4.2.3.5 *Nesrozumitelný úsek textu*). Pokud však taková náhrada není možná (text si domyslet nelze), reprezentujeme na m-rovině nesrozumitelný úsek textu m-uzlem typu `nontext` s textem *unintelligible* v atributu `type`. Vzájemně si odpovídající uzly jsou opět propojeny odkazem.

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 text.*

5 Další pravidla, konvence a příklady

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*).

Základní pravidlo: **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

dvacet tři
23

třiadvacet
23

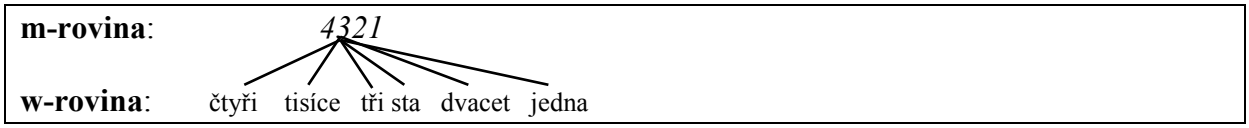
sto jedna
101

dvacátý šestý
26.

šestadvacátý
26.

osmkrát
osmkrát

dvacet pětkrát
25krát



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

5.1.1 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
Našel dvě koruny.

bylo mu čtyřicet pět let
Bylo mu 45 let.

každý třetí v řadě si vystoupil
Každý třetí v řadě si vystoupil.

skončil dvacátý pátý
Skončil 25.

tisíc tři sta jedenáct korun
1 311 korun

milion lidí
milion lidí

jeden milion lidí
1 000 000 lidí

milion dvě stě tisíc lidí
1 200 000 lidí

možná ňákejch pětadvacet , třicet breath dětí chodilo do tý obecní školy v těch begin Litohlavech
Možná 25, třicet dětí chodilo do obecné školy v Litohlavech.

Poznámka: Při zápisu čísla číslicemi oddělujeme mezerou řády, celé číslo je vždy reprezentováno jedním m-uzlem.

Desetinná čísla standardizujeme podobně jako čísla celá: jednoslovná desetinná čísla slovy, víceslovná číslicemi. Výjimku tvoří číslovka typu „*jednoslovná a půl*“, kterou v nematematických kontextech standardizujme slovy.

Příklady:

tři celé pět desetin hodiny
3,5 hodiny

dvě desetiny hodiny
0,2 hodiny

půl koláče
půl koláče

bylo mu dva a půl roku
Bylo mu dva a půl roku.

tři a půl hodiny pracoval
Pracoval tři a půl hodiny.

třicet osm a půl hodiny pracoval
Pracoval 38,5 hodiny.

za poukázky jste mohl dostat zrovna tak peníze , dejme tomu tři koruny padesát a sto krát tři koruny padesát je tři sta padesát korun, což v té době bylo dost značný obnos
Za poukázky jste mohl dostat zrovna tak peníze, dejme tomu 3,50 Kč, a 100 x 3,50 Kč je 350 korun, což v té době byl dost značný obnos.

Poznámka: Desetinné číslo zapsané číslicemi je jeden m-uzel.

V matematických, fyzikálních a jiných kontextech, ve kterých je v psaném textu obvyklý jiný způsob zápisu čísla, než který stanoví základní pravidlo standardizace čísel, použijeme ve standardizovaném textu tento obvyklý způsob zápisu daného čísla. Váháme-li, standardizujeme podle základního pravidla standardizace čísel.

Příklady:

jedna plus dvě rovná se tři
1 + 2 = 3

v poměru jedna ku třem
v poměru 1 : 3

jedna a dvě jsou tři
Jedna a dvě jsou tři.

Číslo před fyzikální značkou zapsanou zkratkou je vždy standardizováno číslicemi. Jednoslovná čísla před fyzikální značkou zapsanou slovem jsou standardizována slovy.

Příklady:

je to třicet kilometrů
Je to třicet kilometrů.

je to třicet k m
Je to 30 km.

je to třicet jedna kilometrů
Je to 31 kilometrů.

žárovka na čtyři a půl voltu
žárovka na 4,5 V

to jsem chytil štiku, ta měla šest kilo dvacet.
To jsem chytil štiku , měla 6,20 kg.

to je naše chalupa v Železnej Rudě, ale máme jenom půlku půlku ty chalupy, protože je to veliký, je to třinát krát třinát metrů
To je naše chalupa v Železné Rudě.
Máme té chalupy ale jenom půlku, protože je to veliké.
Je to 13 x 13 m.

Přívlastkové spojení čísla a slova standardizujeme analogicky podle základního pravidla: je-li ve složení číslo v normálním kontextu víceslovné, standardizujeme ve formě „číslo zapsané číslicemi+slovo“, je-li ve složení číslo v normálním kontextu jednoslovné, standardizujeme ve formě „číslo zapsané slovy+slovo“.

Příklady:

dvacetiletý muž
dvacetiletý muž

desetiprocentní roztok
desetiprocentní roztok

pětadvacetiletý muž
25letý muž

šedesátivoltová žárovka
60V žárovka

sto čtyřiceti osmimetrová dráha
148metrová dráha

Poznámka: Spojení číslo+slovo(písmeno) je jeden m-uzel. Spojení číslo+znak (např. *10%*) jsou dva m-uzly.

5.1.2 Čísla „nálepky“

Čísly-nálepkami rozumíme použití číslovky/čísla ve významu očíslování, označení určitých objektů (telefonní číslo, fax, číslo domu, poštovní směrovací číslo, výrobní číslo, čísla dokumentů, číslo jako součást názvu výrobku, IP adresa, rodné číslo).

Protože jedno a totéž číslo-nálepka (zejména je-li víceciferné) může být mluvčím řečeno různě, platí pro přepis čísel-nálepek do standardizovaného textu následující pravidla:
Číslo-nálepku zapisujme podle běžných konvencí pro zápis daného čísla-nálepky v psaném textu (ať už dané číslo bylo vysloveno jakkoli). **Čísla-nálepky se píší číslicemi**.
Pokud není zřejmé, zda vyslovovaná posloupnost čísel je jedno číslo, či posloupnost několika čísel, zapíšeme (číslicemi) posloupnost jednotlivých vyslovených čísel a jednotlivá čísla oddělíme čárkou (a to i kdyby všechna čísla v dané posloupnosti byla jednociferná).

Příklady:

měla sem číslo šest osm sedm
Měla jsem číslo 687.

bydlí Purkyňova osm
bydlí Purkyňova 8

telefon čtyři sta šedesát pět sto osm nula nula tři
telefon 465 108 003

zvítězil závodník s číslem pět
Zvítězil závodník s číslem 5.

výrobek měl označení třicet dva jedna nula dvacet pět
Výrobek měl označení 32, 1, 0, 25.

Poznámka: Celé jedno číslo nálepka je vždy reprezentováno jedním m-uzlem; například telefonní číslo *465 108 003* bude jeden m-uzel.

5.1.3 Časové údaje

5.1.3.1 Letopočet

Letopočty jsou primárně psány číslicemi.

Způsob standardizace nejrůznějších variant vyjadřování letopočtu ukazují následující příklady. Před vlastní letopočet vždy vkládáme slovo *rok* v příslušném tvaru.

Příklady:

bylo to čtyřicet pět
Bylo to roku 1945.

v osmašedesátým se to stalo
Stalo se to v roce 1968.

bylo to v roce čtyřicet pět
Bylo to v roce 1945.

to se stalo v šedesátém osmém
Stalo se to v roce 1968.

manželovi zemřela maminka ve dvaapadesátém roce
Manželovi zemřela maminka v roce 1952.

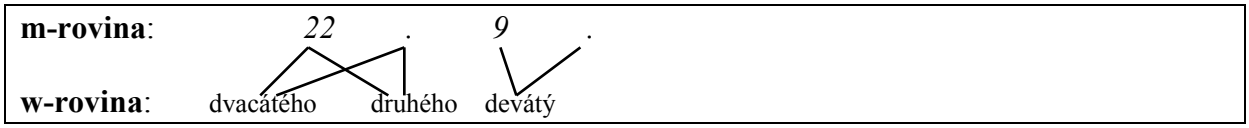
takže prakticky od toho padesátého pátého roku jsem v Domažlicích
Prakticky od roku 1955 jsem v Domažlicích.

5.1.3.2 Desetiletí

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

5.1.3.3 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.

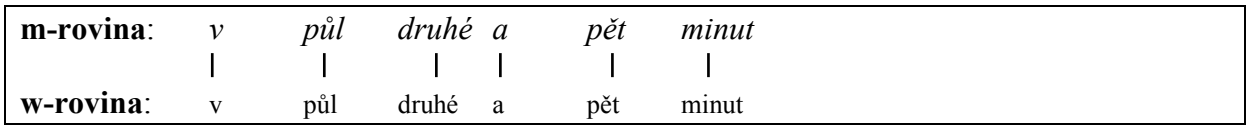


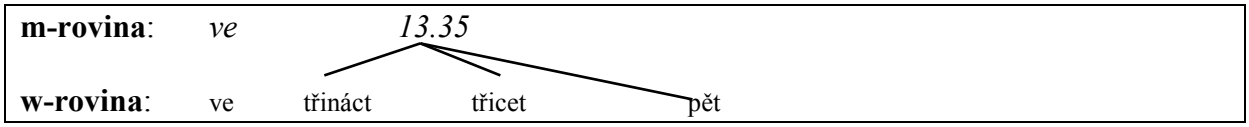
Ve „vyprávěcím“ kontextu může však i označení dne být standardizováno slovem, zejména jde-li o jednoslovnou číslovku.

Příklad:
v 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.

5.1.3.4 Čas

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





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.

Příklady:
přišel o třetí hodině odpoledne
Přišel o třetí hodině odpoledne.

už byla jedna pryč
Už byla jedna pryč.

přišel ve čtrnáct hodin a dvacet pět minut
Přišel ve čtrnáct hodin a 25 minut.

přišel ve čtrnáct hodin dvacet pět minut
Přišel ve 14.25.

bylo dvacet jedna hodin
Bylo 21 hodin.

přišel v patnáct nula nula
Přišel v 15.00

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 5.5 *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, německy, 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), nebo 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 dáváme jej 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 5.4 *Citační kontexty*).

Příklad:
a on to vyslovoval identity [ídentyty] místo identity [ajdentyty]
Vyslovoval to „ídentity“ místo „ajdentity“.

německy neuměl , ale vím , že tam skomolil to jméno ur~ urláb na to ulráb , breath že tam píše , že by rád breath na ten ulráb EHM se dostal domů
Německy neuměl, ale vím, že tam zkomolil jméno urlaub na „ulráb“.
Píše tam, že by se rád na ten „ulráb“ dostal domů.

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 (viz 5.5.1 *Anotátorské poznámky pro zaznamenání chyb na w-rovině*).

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, kde 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.

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.

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

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.

Příklad:
talkovali sme celé dvě hodiny
„Talkovali“ jsme celé dvě hodiny.

5.2.4 Přeřeknutí

Přeřeknutí nahrazujeme nepřeřeknutými tvary slov. Jde o zvláštní typ substituce.

Příklady:
pak přijela lokotomiva
Pak přijela lokomotiva.

jo holokost to bylo něco
Holocaust to bylo něco.

Citační kontext. Jen v těch případech, kdy přeřeknutí má nějaký význam pro další vývoj dialogu – mluvčí na něj nějak reaguje, 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*).

5.2.5 Nedokončená slova

Nedokončená (nedovyslovená) slova nahrazujeme slovy úplnými. Jde o zvláštní typ substituce.

Příklad:
pak přijela lokomo~
Pak přijela lokomotiva.

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

5.2.6 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.
Poznámka: Každé písmeno je samostatný m-uzel.

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.

5.2.7 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.

Poznámka: Zkratka je vždy reprezentována jedním m-uzlem.

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.

5.3 Nesrozumitelný text

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

Obsahově relevantní nesrozumitelný úsek textu nahrazujeme na m-rovině primárně textem domyšleným (viz 4.2.3.5 *Nesrozumitelný úsek textu*).
Pokud však taková náhrada není možná (text si na základě kontextu domyslet nelze), reprezentujeme jej na m-rovině m-uzlem typu `nontext` s textem *unintelligible* (viz 4.2.5 *Zachycení obsahově relevantních neřečových událostí*).

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

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`), nicméně při poslechu nahrávky anotátor nyní mluvčímu dobře rozumí, slyší, co říká. V takovém případě zapíše anotátor v anotátorské poznámce typu `w-recognize`, 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`.

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* aj. Význam meta-užití je obvyklý také u sloves: *znamenat*, *značit*, *označovat*, *psát*, *vyslovovat* aj.

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

5.5 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.

5.5.1 Anotátorské poznámky pro zaznamenání chyb na w-rovině

Při rekonstrukci standardizovaného textu z mluvené řeči jsou důsledně odlišovány „nedostatky“, které způsobil mluvčí, 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 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.

5.5.1.1 w-token

Poznámka **`w-token`** slouží pro zaznamenání chybně rozpoznaných w-uzlů, tj. pro případy, kdy slovo na w-rovině je přepsáno špatně. Chyby u w-uzlů, které nemají na m-rovině protějšek, zaznamenáváme v anotátorské poznámce nějakého (nejbližšího) m-uzlu.
Pozor! Zapsané přeřeknutí (když se mluvčí skutečně přeřekl) není chyba v přepisu na w-rovině.

Příklady:
a bačkora pak vstala a odešla
Babička{`w-token`} *pak vstala a odešla.*

ženy byli smutný
Ženy byly{`w-token`} *smutné.*

5.5.1.2 w-missing

Poznámka **`w-missing`** 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, tj. vypíše se, jak má být w-rovina opravena.
Rozpoznaný chybějící úsek nemusí být totožný s odpovídajícím rekonstruovaným úsekem na m-rovině, text poznámky je proto povinný.

Příklad:
a potom přišel inhale muž a jeho žena cough
Potom přišel muž a také{`w-missing`: *také ehm*} *jeho žena*.

5.5.1.3 w-extraneous

Poznámka **`w-extraneous`** slouží pro případy, kdy na w-rovině je zaznamenáno slovo nebo i celý úsek textu, který nikdo neříká. Do textu poznámky nějakého nejbližšího jednoho m-uzlu se vypíše celý nadbytečný úsek, tj. vypíší se všechna slova, která mají být z w-roviny vymazána. Text poznámky je povinný.

Příklad:
a potom přišel inhale muž a také jeho žena cough
Potom přišel muž a{`w-extraneous`: *také*} *jeho žena*.

5.5.1.4 w-recognize

Poznámka **`w-recognize`** slouží pro případy, kdy je na w-rovině 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, tj. vypíše se, jak se má w-rovina opravit.
Rozpoznaný nesrozumitelný úsek nemusí být totožný s odpovídajícím rekonstruovaným úsekem na m-rovině, text poznámky je proto povinný.

Příklad:
a potom přišel inhale muž a unintelligible
Potom přišel muž a také{`w-recognize`: *také ehm jeho žena cough*} *jeho žena*.

Pozor! Něco jiného je, když nesrozumitelný úsek textu domýšlíme (textu opravdu není rozumět), pak se poznámka `w-recognize` nepoužije.

5.5.1.5 w-speaker

Poznámka **`w-speaker`** slouží pro případy, kdy na w-rovině chybí označení změny mluvčího. Poznámku dáváme tomu nejbližšímu m-uzlu, před jehož protějškem na w-rovině chybí označení změny mluvčího („speaker“).
Zároveň je třeba manuálně změnit hodnotu atributu `w-speaker.rf` v hlavičce segmentu.
Text poznámky není povinný. Pokud nebude v poznámce nic uvedeno, bude se poznámka chápat takto: před w-uzlem, na který odkazuje m-uzel s poznámkou, má být doplněna změna mluvčího („speaker“), který bude stejný, jako mluvčí daného segmentu. U složitějších případů uvedeme v poznámce stručně popis problému.

Příklad:
spk1 Je to tak správně ? Je no .
spk1 *Je to tak správně?*
spk2 *Ano je*{`w-speaker`} .

5.5.1.6 other

Pro jiné typy chyb na w-rovině použijeme poznámku **`other`**.

5.5.2 Ostatní anotátorské poznámky

5.5.2.1 metalanguage

Poznámku **`metalanguage`** používáme pro označení citačního kontextu (viz 5.4 *Citační kontexty*). Slova v citačním kontextu dáváme zpravidla do uvozovek a označujeme je anotátorskou poznámkou `metalanguage`. V případě, že je v citačním kontextu celé slovní spojení nebo i celá věta, stačí anotátorskou poznámkou označit pouze řídící člen spojení v citačním kontextu.
Text poznámky není povinný.

Příklad:
cedule s nápisem Romy neobsluhujeme
cedule s nápisem „Romy neobsluhujeme{`metalanguage`}*“*

5.5.2.2 form

Poznámku **`form`** používáme pro zaznamenání nejistoty v lematu nebo 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 o nejistotu v psaní velkých a malých písmen aj.
Text poznámky není povinný.

Příklad:
menuju se Jiří [Bém]
Jmenuji se Jiří Boehm{`form`}

5.5.2.3 other

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

6 Šablony

V této kapitole uvádíme návody a vzorové šablony pro rekonstrukci často se opakujících vět a jevů.

6.1 Šablony pro mluvící hlavu

Úvod.
Dobrý den.
Jmenuji se Petra a budu si tu s vámi teď chvíli povídat.
Ještě než začneme, je mě dobře slyšet? (`stype`: `question`)
Ráda bych si s vámi popovídala o vašich fotkách.
Začněme třeba s touhle. (`stype`: `instruction`)
Začneme třeba s touhle. (`stype`: `information`)
Co je na ní vidět?

V průběhu rozhovoru.
Dobře. (nepřipojujeme k následující otázce nebo instrukci)
Řeknete mi ještě něco o téhle fotce?
Chcete mi říct ještě něco k téhle fotce?
Povíte mi k tomu ještě něco?
Chtěla byste k tomu ještě něco dodat?
Co byste mi ještě řekl k této fotce?
Podíváme se na další fotku, nebo chcete ještě něco dodat? (`stype`: `question`)

Dobře. (nepřipojujeme k následující otázce nebo instrukci)
Podíváme se na další fotku. (nikoli: *Tak se podíváme na další fotku.)*
Přejdeme na další fotku. (`stype`: `information`)
Jdeme na další.
Jdeme dál.
Vrátíme se k fotce.

Co je na ní vidět?
Copak to máme tady?
Co je tohle za fotku?
Co tady můžeme vidět?
Kdo je na téhle fotce?

Ukončení.
Bohužel nám vypršel čas.
Vypršel čas.

Dobře. (nepřipojujeme k následující otázce nebo instrukci)
Tohle byla poslední fotka.

Děkujeme vám za váš čas.
Moc hezky se mi s vámi povídalo.
Počkejte chvilku, kluci vás přijdou vysvobodit. (`stype`: `information`)

Literatura

Allwood, J.; Grnqvist, L.; Ahlsn, E.; Gunnarsson, M. (2002): *Annotations and Tools for an Activity Based Spoken Language Corpus*. Proc. of 2nd SIGdial Workshop on Discourse and Dialogue, Aalborg, Denmark.

Barras, C.; Geoffrois, E.; Wu, Z.; Liberman, M. (2001): *Transcriber: development and use of a tool for assisting speech corpora production*. Speech Communication special issue on Speech Annotation and Corpus Tools, vol. 33, no. 1–2, pp. 5–22.

Bradley, J.; Mival, O.; Benyon, D. (2008): *A Novel Architecture for Designing by Wizard of Oz*. In proceedings of CREATE08, pp. 1–4.

Byrne, W.; Doermann, D.; Franz, M.; Gustman, S.; Hajič, J.; Oard, D. Picheny, M.; Psutka, J.; Ramabhadran, B.; Soergel, D.; Ward, T.; Zhu, W. (2004): *Automatic Recognition of spontaneous Speech for Access to Multilingual Oral History Archives.* IEEE Transactions on Speech and Audio Processing, vol. 12, no. 4, pp. 420–435.

Čmejrková, S. (1993): *Slovo psané a mluvené*. Slovo a slovesnost, 54, s.51-58.

Čmejrková, S.; Daneš, F.; Havlová, E. (eds.) (1994): *Writing vs. Speaking: Language*. Text, Discourse, Communication. Tübingen: Gunter Narr Verlag.

Fitzgerald, E.; Jelinek, F. (2008): *Linguistic resources for reconstructing spontaneous speech text.* In LREC Proceedings, Marrakesh, Morocco, ELRA, pp. 1–8.

Glücksmannová, H. (2008): *Spontaneous Speech Reconstruction.* In Proceedings of WDS2008, Prague, Czech Republic

Graff, D.; Bird, S. (2000): *Many uses, many annotations for large speech corpora: Switchboard and TDT as case studies*. Proceedings of the Second International Conference on Language Resources and Evaluation, pp 427-433.

Godfrey, J.; Holliman, E.; McDaniel, J. (1992): *SWITCHBOARD: Telephone speech corpus for research and development*. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Hajič, J.; Mikulová, M.; Otradovcová, M.; Pajas, P.; Podveský, P.; Urešová, Z. (2006): *Pražský závislostní korpus mluvené češtiny. Rekonstrukce standardizovaného textu z mluvené češtiny.* Technical Report, UFAL MFF UK, Praha.

Hajič, J.; Psutka, J.; Ircing, P.; Ramabhadran, B.; Gustman, S.; Byrne, W. J.; Psutka, J. V.; Radová, V. (2002): *Automatic Transcription of Czech Language Oral History in the MALACH Project: Resources and Initial Experiments*. In Text, Speech and Dialogue. 5th International Conference, TSD 2002, pp. 253-260. Springer.

Heeman, P.; Allen, J. (1994): *Tagging Speech Repairs*. In ARPA Workshop on Human Language Technology, Princeton, NJ, pp. 187–192.

Heeman, P.; Allen, J. (1994): *Detecting and Correcting Speech Repairs*. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94), Las Cruces, pp. 295–302.

Kořenský, J. a kol. (1999): *Komplexní analýza komunikačního procesu a textu*. Č.Budějovice 1991.

Leech, G. (2000): *Anotační systémy pro značkování korpusů*. In: F. Čermák – J. Klímová – V. Petkevič (eds.). Studie z korpusové lingvistiky. Praha: Univerzita Karlova v Praze – Nakladatelství Karolinum.

Kolář, J.; Švec, J.; Strassel, S.; Walker, C.; Kozlikova, D.; Psutka, J. (2005): *Czech Spontaneous Speech Corpus with Structural Metadata.* In Proceedings of the 9th European Conference on Speech Communication and Technology, INTERSPEECH 2005, Lisboa, Portugal, pp. 1165-1168.

Mikulová a kol. (2005): *Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka*. Technická zpráva ÚFAL TR-2005-28. MFF UK, Praha. (http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/cz/t-layer/html/index.html)

Mikulová, M.; Urešová, Z. (2008): *Rekonstrukce standardizovaného textu z mluvené řeči*. In Kopřivová, M.; Waclawičová, M.: *Čeština v mluveném korpusu*. Lidové noviny, Praha.

Miller, J.; Weinert, R. (1998): *Spontaneous Spoken Language*. Syntax and Discourse. Clarendon Press, Oxford.

Müllerová, O. (1994): *Mluvený text a jeho syntaktická výstavba*. Praha.

Pajas, P.; Mareček, D. (2007): *MEd - an editor of interlinked multi-layered linearly-structured linguistic nnotations* (http://ufal.mff.cuni.cz/˜pajas/med).

Psutka, J.; Ircing, P.; Psutka, J. V.; Radová, V.; Byrne, W.; Hajič, J.; Mírovský, J.; Gustman, S. (2003): *Large Vocabulary ASR for Spontaneous Czech in the MALACH Project*. In EUROSPEECH 2003 Proceedings (8th European Conference on Speech Communication and Technology), pp. 1821-1824. ISCA.

Sgall, P.; Hajičová, E.; Buráňová, E. (1980): *Aktuální členění věty v češtině*. Academia, Praha.

Sgall, P. (2000): *Jan Firbas, Functional sentence perspective in written and spoken communication*. Review of Jan Firbas, Functional sentence perspective in written and spoken communication. In Journal of Pragmatics, pp. 639-644.

Sgall, P. (2002): *Spoken Czech revisited*. In Where One's Tongue Rules Well. A Festschrift for Charles E. Townsend, pp. 299--309. Slavica Publishers. Supported by LN00A063.

The ÚFAL/CKL Technical Report Series

ÚFAL

ÚFAL (Ústav formální a aplikované lingvistiky; http://ufal.mff.cuni.cz) is the Institute of Formal and Applied linguistics, at the Faculty of Mathematics and Physics of Charles University, Prague, Czech Republic. The Institute was established in 1990 after the political changes as a continuation of the research work and teaching carried out by the former Laboratory of Algebraic Linguistics since the early 60s at the Faculty of Philosophy and later the Faculty of Mathematics and Physics. Together with the "sister" Institute of Theoretical and Computational Linguistics (Faculty of Arts) we aim at the development of teaching programs and research in the domain of theoretical and computational linguistics at the respective Faculties, collaborating closely with other departments such as the Institute of the Czech National Corpus at the Faculty of Philosophy and the Department of Computer Science at the Faculty of Mathematics and Physics.

CKL

As of 1 June 2000 the Center for Computational Linguistics (Centrum komputační lingvistiky; http://ckl.mff.cuni.cz) was established as one of the centers of excellence within the governmental program for support of research in the Czech Republic. The center is attached to the Faculty of Mathematics and Physics of Charles University in Prague.

Technical Reports

The ÚFAL/CKL technical report series has been established with the aim of disseminate topical results of research currently pursued by members, cooperators, or visitors of the Institute. The technical reports published in this Series are results of the research carried out in the research projects supported by the Grant Agency of the Czech Republic, GAČR 405/96/K214 ("Komplexní program"), GAČR 405/96/0198 (Treebank project), grant of the Ministry of Education of the Czech Republic VS 96151, and project of the Ministry of Education of the Czech Republic LN00A063 (Center for Computational Linguistics). Since November 1996, the following reports have been published.

ÚFAL TR-1996-01 Eva Hajičová, *The Past and Present of Computational Linguistics at Charles University*
Jan Hajič and Barbora Hladká, *Probabilistic and Rule-Based Tagging of an Inflective Language – A Comparison*

ÚFAL TR-1997-02 Vladislav Kuboň, Tomáš Holan and Martin Plátek, *A Grammar-Checker for Czech*

ÚFAL TR-1997-03 Alla Bémová at al., Anotace na analytické rovině, *Návod pro anotátory (in Czech)*

ÚFAL TR-1997-04 Jan Hajič and Barbora Hladká, *Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structural Tagset*

ÚFAL TR-1998-05 Geert-Jan M. Kruijff, *Basic Dependency-Based Logical Grammar*

ÚFAL TR-1999-06 Vladislav Kuboň, *A Robust Parser for Czech*

ÚFAL TR-1999-07 Eva Hajičová, Jarmila Panevová and Petr Sgall, *Manuál pro tektogramatické značkování (in Czech)*

ÚFAL TR-2000-08 Tomáš Holan, Vladislav Kuboň, Karel Oliva, Martin Plátek, *On Complexity of Word Order*

ÚFAL/CKL TR-2000-09 Eva Hajičová, Jarmila Panevová and Petr Sgall, *A Manual for Tectogrammatical Tagging of the Prague Dependency Treebank*

ÚFAL/CKL TR-2001-10 Zdeněk Žabokrtský, *Automatic Functor Assignment in the Prague Dependency Treebank*

ÚFAL/CKL TR-2001-11 Markéta Straňáková, *Homonymie předložkových skupin v češtině a možnost jejich automatického zpracování*

ÚFAL/CKL TR-2001-12 Eva Hajičová, Jarmila Panevová and Petr Sgall, *Manuál pro tektogramatické značkování (III. verze)*

ÚFAL/CKL TR-2002-13 Pavel Pecina and Martin Holub, *Sémanticky signifikantní kolokace*

ÚFAL/CKL TR-2002-14 Jiří Hana, Hana Hanová, *Manual for Morphological Annotation*

ÚFAL/CKL TR-2002-15 Markéta Lopatková, Zdeněk Žabokrtský, Karolína Skwarská and Vendula Benešová, *Tektogramaticky anotovaný valenční slovník českých sloves*

ÚFAL/CKL TR-2002-16 Radu Gramatovici and Martin Plátek, *D-trivial Dependency Grammars with Global Word-Order Restrictions*

ÚFAL/CKL TR-2003-17 Pavel Květoň, *Language for Grammatical Rules*

ÚFAL/CKL TR-2003-18 Markéta Lopatková, Zdeněk Žabokrtský, Karolina Skwarska, Václava Benešová, *Valency Lexicon of Czech Verbs VALLEX 1.0*

ÚFAL/CKL TR-2003-19 Lucie Kučová, Veronika Kolářová, Zdeněk Žabokrtský, Petr Pajas, Oliver Čulo, *Anotování koreference v Pražském závislostním korpusu*

ÚFAL/CKL TR-2003-20 Kateřina Veselá, Jiří Havelka, *Anotování aktuálního členění věty v Pražském závislostním korpusu*

ÚFAL/CKL TR-2004-21 Silvie Cinková, *Manuál pro tektogramatickou anotaci angličtiny*

ÚFAL/CKL TR-2004-22 Daniel Zeman, *Neprojektivity v Pražském závislostním korpusu (PDT)*

ÚFAL/CKL TR-2004-23 Jan Hajič a kol., *Anotace na analytické rovině, návod pro anotátory*

ÚFAL/CKL TR-2004-24 Jan Hajič, Zdeňka Urešová, Alevtina Bémová, Marie Kaplanová, *Anotace na tektogramatické rovině (úroveň 3)*

ÚFAL/CKL TR-2004-25 Jan Hajič, Zdeňka Urešová, Alevtina Bémová, Marie Kaplanová, *The Prague Dependency Treebank, Annotation on tectogrammatical level*

ÚFAL/CKL TR-2004-26 Martin Holub, Jiří Diviš, Jan Pávek, Pavel Pecina, Jiří Semecký, *Topics of Texts. Annotation, Automatic Searching and Indexing*

ÚFAL/CKL TR-2005-27 Jiří Hana, Daniel Zeman, *Manual for Morphological Annotation (Revision for PDT 2.0)*

ÚFAL/CKL TR-2005-28 Marie Mikulová a kol., *Pražský závislostní korpus (The Prague Dependency Treebank) Anotace na tektogramatické rovině (úroveň 3)*

ÚFAL/CKL TR-2005-29 Petr Pajas, Jan Štěpánek, *A Generic XML-Based Format for Structured Linguistic Annotation and Its application to the Prague Dependency Treebank 2.0*

ÚFAL/CKL TR-2006-30 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Annotation on the tectogrammatical level in the Prague Dependency Treebank (Annotation manual)*

ÚFAL/CKL TR-2006-31 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Petr Sgall, Magda Ševčíková, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Anotace na tektogramatické rovině Pražského závislostního korpusu (Referenční příručka)*

ÚFAL/CKL TR-2006-32 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Petr Sgall,Magda Ševčíková, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Annotation on the tectogrammatical level in the Prague Dependency Treebank (Reference book)*

ÚFAL/CKL TR-2006-33 Jan Hajič, Marie Mikulová, Martina Otradovcová, Petr Pajas, Petr Podveský, Zdeňka Urešová, *Pražský závislostní korpus mluvené češtiny. Rekonstrukce standardizovaného textu z mluvené řeči*

ÚFAL/CKL TR-2006-34 Markéta Lopatková, Zdeněk Žabokrtský, Václava Benešová (in cooperation with Karolína Skwarska, Klára Hrstková, Michaela Nová, Eduard Bejček, Miroslav Tichý) *Valency Lexicon of Czech Verbs. VALLEX 2.0*

ÚFAL/CKL TR-2006-35 Silvie Cinková, Jan Hajič, Marie Mikulová, Lucie Mladová, Anja Nedolužko, Petr Pajas, Jarmila Panevová, Jiří Semecký, Jana Šindlerová, Josef Toman, Zdeňka Urešová, Zdeněk Žabokrtský, *Annotation of English on the tectogrammatical level*

ÚFAL/CKL TR-2007-36 Magda Ševčíková, Zdeněk Žabokrtský, Oldřich Krůza, *Zpracování pojmenovaných entit v českých textech*

ÚFAL/CKL TR-2008-37 Silvie Cinková, Marie Mikulová, *Spontaneous speech reconstruction for the syntactic and semantic analysis of the NAP corpus*

ÚFAL/CKL TR-2008-38 Marie Mikulová, *Rekonstrukce standardizovaného textu z mluvené řeči v Pražském závislostním korpusu mluvené češtiny. Manuál pro anotátory*