

***Klasifikace a typologie chyb ve vstupních textech a
koncepte značkování chybných textů***

Milena Hnátková – Tomáš Jelínek – Vladimír Petkevič
Ústav teoretické a počítačové lingvistiky
Filozofické fakulty Karlovy Univerzity

1. Klasifikace a typologie chyb ve vstupních textech

Na úvod jako příklad chybných textů dva cizinecké texty:

Legenda:

pravopis

vazby

shoda

styl

liberecký Vietnamec:

Esej: Nákup Na Párty

Dnes je hezká sobota. Máme malý párty, že můj přítel a já, jedeme do Tesca.

Potřebujeme koupit nápoje, jídlo a ovoce. Můj přítel je Luat, má rad koupit jídlo a

protože koupím nápoje a ovoce. Máme moc rád pivo [,] že koupím hodně pivo a také

dobré víno. Po tom koupím ovoce, jablko, pomeranč a jahoda. Luat koupí hovězí, kuře, housky a chléb. Potom jedeme do koleji a máme znamenitý párty.

Legenda:

pravopis

vazby

slovník

špatný slovosled

Polák začátečník:

Nekolil tydnu spatki byl jsem s moi přítelkyni v Řecku na dovolene. Počasi bylo užasne a sloničko svetlo poržat. Markéta, moja holka, opalila se krasně. Pro mně to ne bylo velmi jednoduhe. Byl jsem spaleny na zdech a morská sůl sposubila mi beďary na čele. Ne vadlo mi to vůbec. Cele dni jsmy travili na plazi, jenom jednou udělali jsmy vylet. Rozhodli jsmy se [,] že bude to „Olimp“. Je to nejvyšší hora Řecku a hloubi [chlubí] se **ekstraordinární historii**.

Klasifikace chyb a vhodný standard

Otázka, co je chyba – s čím srovnávat, s jakým standardem

- spisovná čeština
- hovorová čeština
- obecná čeština

Míra obecnosti klasifikace chyb:

A. Hrubší členění chyb

B. Jemnější členění chyb

A. Hrubší členění chyb

Chyby

- **pravopisné**
- **hláskoslovné**
- **lexikální**
- **morfologické**
- **syntaktické**
- **ve slovosledu**
- **v úzu (nevhodné vyjadřování) / stylistické**
- **makarónský text**
- **zcela rozvrácený text**
- **jiná chyba**

Příklady chyb:

- pravopisné/hláskoslovné

V taškentu

spivani

Nepálsky

Kathmandu je Hlavní město z nepalu

*středniho, penež, manaželka, poržat (pořád), dlouhe, skříňi, kouzelné
(kouzelně/é), aktivny, [j]ím (jím), sest[r]a (sestra), nakupojou (nakupujou), prsší*

- **lexikální**

očné (oční) víčka

*A tam **takže** moc krásné metro.*

***Po** zase jdu do školy (potom)*

*Přesto dopadlo to velmi **přírodně***

*Chtěl bych hodně cestovat, dělat sport, **expecialuje** na fotbal*

*Když jsem byl v české republice, našel jsem hodně **jinaků** mezi životy ve Vietnamu a v České republice*

*Ona se mnou studovala ve stejné třídě. **Obvykle** jsme spolu chodili do školy.*

- morfologické

monumentov (monumentů), *plaváme* (plaveme), *řekě* (řece)

Jsem z Nepalu a bydlím v *kathmandu*

Ahoj *kamarády* (kamarádi)

Mám ráda *zima*

Se *myje* a *snidám*

- syntaktické

Mám ráda zima

plaváme v bázeň (plaveme v bazénu)

Máme historická místo taky

V Taškentě jsou hodně monumentov

moje rodiče

Tam je velki horka a kopec a řekě

Tam je doprava problém a mnoho dopravní zácpa.

stále vzpomínám vás

pro moje přitele

Mám rádosť [že] bydlím tady

Centrum Prahy [je] pořád plné lidí hlavně v turistické sezóně.

- **ve slovosledu**

Se myje a snidám

já musel jsem jít do práce

Rádio je taky na skříni (Na skříni je taky rádio.)

- v úzu (nevhodné vyjadřování) / stylistické

Tam je hodně cizinec od rozdílny země.

Potom ja spim v noci

Začal jsem umět milovat holčičku

hezky poloha a výhled umí být založit.

Tehdy jsem byl chlapec šestnáctého (tehdy jsem byl šestnáctiletý chlapec)

nekdy ale většinou ja obědvám v domě

- makarónský text

*Policie zavírá **the way***

*I cheng my dress a **again***

*I **chek my mail** a si čistíme zuby*

*vídím moje kamarádka a **her** přítele*

- zcela rozvrácený text

Den místo.

Obvykle pracuje na počítači proto období jako Mongolsko.

B. Jemnější členění chyb

Chyby

- **pravopisné (klasické)**
 - nesprávné psaní **i/y**; **s/z** apod. (**s**pívání)
 - nesprávně psaná palatalizace (**d'**i)
 - neutralizace znělosti na konci slova
 - psaní velkých a malých písmen
 - chyba v interpunkci
 - jiná chyba

- **hláskoslovné**

- jiná hláska (délka vokálů, palatály/nepalatály) projevující se chybnou diakritikou (čárka, kroužek, háček)
- metateze grafémů
- chybějící nebo přebývající grafém
- jiná chyba ve slově

- **lexikální – chyby ve slovní zásobě**

- existující, ale nevhodně užitá slovo

- neexistující slovo

- **morfologické**

- ve flexi

- v derivační morfologii

- **obecněčeská morfologie**

- jiné

- **syntaktické**

- morfosyntaktické

- ve shodě

- ve vazbách (valenci)

- vynechané/nadbytečné slovo/slova rozvracející syntax

- jiné

- **ve slovosledu**

- negramatický slovosled
- gramatický, ale nevhodný slovosled
- aktuální členění

- v úzu (nevhodné vyjadřování)
- **stylistické / aktuální členění**
- **makarónské texty**
- **zcela rozvrácený text**
- jiná chyba

Příklady

- **pravopisné (klasické)**

- **nesprávné psaní i/y; s/z** apod.

spívání, strátit

- **nesprávně psaná palatalizace (ďi, ňi)**

radĭ, skřĭňi

- **neutralizace znělosti (na konci slova)**

let/led

- **psaní velkých a malých písmen**

Kathmandu je Hlavní město z nepalu

mám ráda Kučerí (kuřecí) řízek

- **chyba v interpunkci**

Zdejší všihní stavby (dům) je staré a historické [,] ale vypadá hezky

- hláskoslovné

- jiná hláska (délka vokálů, palatály/nepalatály) projevující se chybnou diakritikou (čárka, kroužek, háček)

dlouhe, *bazen*, *věsinou*, *nakupojou* (nakupujou), *notelu* (nutelu), *moj* (můj), *pochodě* (pohodě)

- metateze grafémů

hoidny (hodiny)

mám ráda Kučerí (kuřecí) řízek

- **chybějící nebo přebývající grafém**

[j]ím (jím), sestr[a] (sestra), jízd[e]nky, j[í]dлу, na kol[e]ji, ve[n]kovní
vestavam, prsší, ne-pracuju

- **jiná chyba ve slově**

- **lexikální – chyby ve slovní zásobě**

- **existující, ale nevhodně užitá slovo**

*Přesto dopadlo to velmi **přírodně**.*

*A tam **takže** moc krásné metro.*

- **neexistující slovo**

*Chtěl bych hodně cestovat, dělat sport, **expecialuje** na fotbal.*

- **morfologické**

- **ve flexi**

*Jsem z Nepalu a **bydlí**[m] v kathmandu*

- **v derivační morfologii**

Například užití nevhodné předpony

- **obecněčeská morfologie**

hnědýho, hnědej, novým, novom

- syntaktické

- morfosyntaktické

myslím se (myslím si)

Spolu jsme studovali a hrali [si].

byl jsem aktivní hodně jsem [si] *hrál*.

- ve shodě

Hned bude vánoce, které je velká svátek

Můj synovec je hloupá

○ **ve vazbách (valenci)**

Je konec prosinec

mám kurz cestinu (kurz češtiny)

s kamaradi

Potom ja jdu do dům

V Taškentě jsou hodně monumentov

- **vynechané/nadbytečné slovo/slova narušující/rozvracející syntax**

*V létě [je] ve městě hodně turistov
Mám radost [že] bydlím tady
Na mém životě se mi [toho] hodně líbí*

- **jiné**

*Protože ona je takova ja jsem motivovan **aby dělat** vic
jsou nakupjou jízdenky pro Kutna Hora*

- **ve slovosledu**

- **negramatický slovosled**

Se učím česky 2. hodiny

já musel jsem jít do práce

- **gramatický, ale nevhodný slovosled**

Rádio je taky na skříni (Na skříni je taky rádio.)

- v úzu (nevhodné vyjadřování)

Kathmandu je Hlavní město z nepalu

Ráda se dívá na film.

nekdy ale většinou ja obědvám v domě

Prah je hlavní město v české republice.

Chci mít dobrého manžela jako: hezký, intelligent, veselý, vysoký. a budeme mít 2 děti.

Můj přítel je Luat, má rad koupit jídlo a protože koupím nápoje a ovoce

- **stylistické / aktuální členění**

Rádio je taky na skříni

Vzadu jsou žlutá lampa a velká skříň

tady nemám hodně kamaradů protože neumím Český dobře

Nad dveřmi jsou hodiny a dveře jsou vedle skříně.

Můj přítel je Luat, má rad koupit jídlo a protože koupím nápoje a ovoce

2. Obecná koncepce jazykového značkování chybných textů

Specifické problémy jazykového značkování v našem projektu:

A. Různé typy cizojazyčných mluvčích

Oproti velké většině existujících žakovských korpusů popisovaných v literatuře máme v našem projektu dost **různé typy cizojazyčných mluvčích**:

a. mluvčí **různých jazyků L1** (L1 = mateřský/první jazyk) a jim odpovídající subkorporusy (kvazi)českých textů

- **romština**
- **vietnamština**
- **ruština**
- **směsný korpus**

b. různý typ jazyka mluvčích: **psaný** vs. **mluvený**

c. **různý věk** mluvčích, a tedy různá kvalita jejich jazykových projevů (od školních dětí po dospělé)

d. **různé vzdělání** mluvčích, a opět tedy různá kvalita jejich jazykových projevů

Tato velká pestrost klade velké nároky na značkování!

B. Detailnost značkování

a) Rozhodně je třeba počítat s **interpretačními alternativami**, nestačí proto *flat tagging*! Technicky je tedy asi nutné zvolit tzv. **standoff anotaci**.

Standoff anotace:

- skutečně standoff (tj. **interpretace** odkazující do zpracovávaného textu se nacházejí **v jiném souboru**)
- **interpretace** se připojují ke zpracovávanému textu **v témže souboru** (nejlépe vždy za příslušnou větu)

b) Jaká má být **detailnost značkování**?

- bude-li značkování **velmi jemné**, bude mít anotátor problémy s klasifikací chyby, navíc mu bude anotace dost dlouho trvat
- bude-li značkování **příliš hrubé**, bude mít značkování malou vypovídací hodnotu

Jaký kompromis zvolit?

C. Jak sofistikovaný software potřebujeme?

Tato otázka úzce souvisí s bodem b):

Navrhujeme tento postup práce:

fáze 1: tvorba surového psaného textu buď mluvčím samým, nebo přepisovačem (z mluveného jazyka mluvčího)

fáze 2: lingvistické značkování: aplikace **spelling checkeru** a **grammar checkeru** na surový text.

Na základě výsledků značkování anotátor (nebo spíše anotátoři, nejméně 2) opraví/interpretuje surový text, **načež se spustí fáze 2 na jeho opravu** a takto případně dále. Je navíc nutné, aby mohl anotovat alternativně (umožnit tedy více interpretací). Mimo opravy textu musí být každá oprava zdůvodněna, tj. nestačí jen opravit příslušnou pravopisnou, morfologickou či jinou chybu, ale také **zaznamenat druh chyby**.

Otázka **kvality anotátora**: anotátoři musí být bohemisticky velmi kvalitní, lze to však zaručit? Jde o to, že chyb se může dopustit anotátor (při přepisu gramaticky správných výroků mluvčích), nikoli cizinec, tak **abychom pak neanalyzovali anotátorské chyby!** Bylo by dobré, kdyby anotátoři byli nějak obeznámeni s jazykem L1, to však asi zaručit nelze.

Klíčová otázka: **jak podrobně se chyby mají anotovat**? Počítáme s těmito hlavními aspekty chyb:

- **Jazykové chyby** (včetně pravopisných) a jejich typy
- **Rozsah chyby**, tj. úsek ve větě, kde se chyba nachází
- **Náročnost opravy/emendace/interpretace chyby**

1. Jazykové chyby (včetně pravopisných) a jejich typy

o tom bylo již výše pojednáno

2. Rozsah chyby, tj. úsek ve větě, kde se chyba nachází

- oprava **jediného ortografického slova**
- oprava **spojité skupiny (shluku) slov**
- oprava **nespojité skupiny slov**
- oprava **celé klauze/věty**

oprava jediného ortografického slova

Mám ráda zima

Mám ráda zimu

manaželka

manažerka

manaželka

manželka

oprava spojité skupiny (shluku) slov

Potom ja jdu do dům

Potom ja jdu domů

oprava nespojité skupiny slov

já se divam na televiže nebo hraju pačitač

já se dívám na televizi nebo hraju na počítači

Snažím se, že se moc dobře naučím česky

Snažím se, abych se moc dobře naučil česky

oprava celé klauze/věty

Rádio je taky na skříni

Na skříni je taky rádio.

3. Náročnost a způsob opravy/emendace chyby

Poznámky

- někdy/často je obtížné vymezit chybný úsek a emendovat ho
- někdy je obtížné vůbec pochopit vyjádření mluvčího
- bylo by dobré, aby oprava byla co nejspecifičtější
- je nutné umožnit alternativní emendace

Metoda oprav – jak postupovat?

Ve fázích

Při opravách, je-li jich pro danou větu více, **je vhodné postupovat metodicky, v určitém stanoveném pořadí**, třeba tomto:

- **oprava pravopisu**
- **oprava hláskosloví**
- **oprava prohřešku proti slovní zásobě**
- **oprava morfologie**
- **oprava syntaxe**
- **oprava slovosledu**
- **oprava úzu**
- **oprava stylu**