Classification of Errors in Text

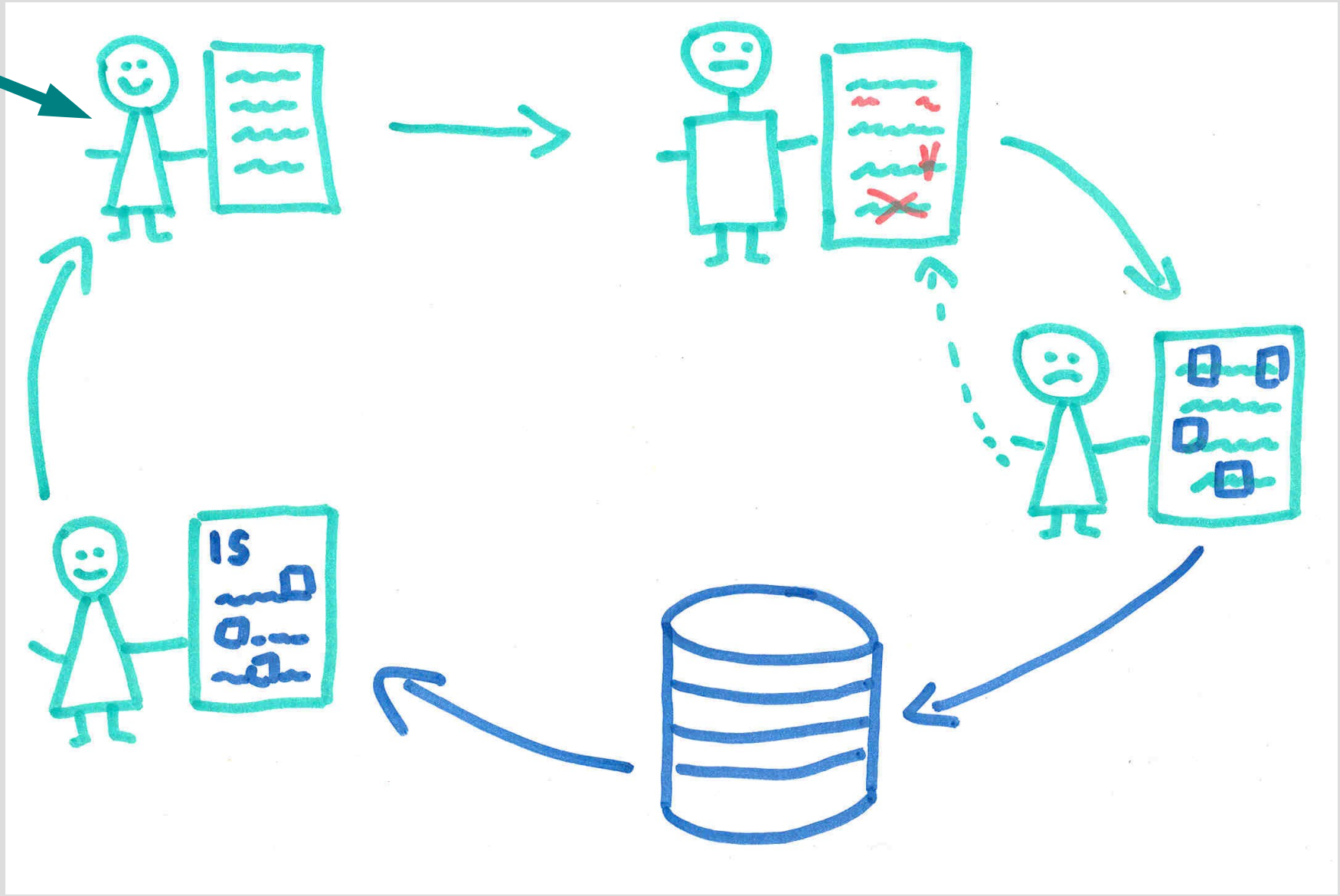
Jan Bušta, Dana Hlaváčková, Miloš Jakubíček, Karel Pala

RASLAN 2009

Korpus chyb (Chyby)

- Je z textů studentských prací s vyznačenými chybami.
- Obsahuje původní verzi chybného textu i jeho opravu.
- Má nyní 469 tis. tokenů.

Proces tvorby korpusu



Klasifikace chyb

- Pomocí XML tagu párového <corr>

  <corr errtype=’string’ corrtype=’string’ old=’old text’>
  new repaired text
  </corr>

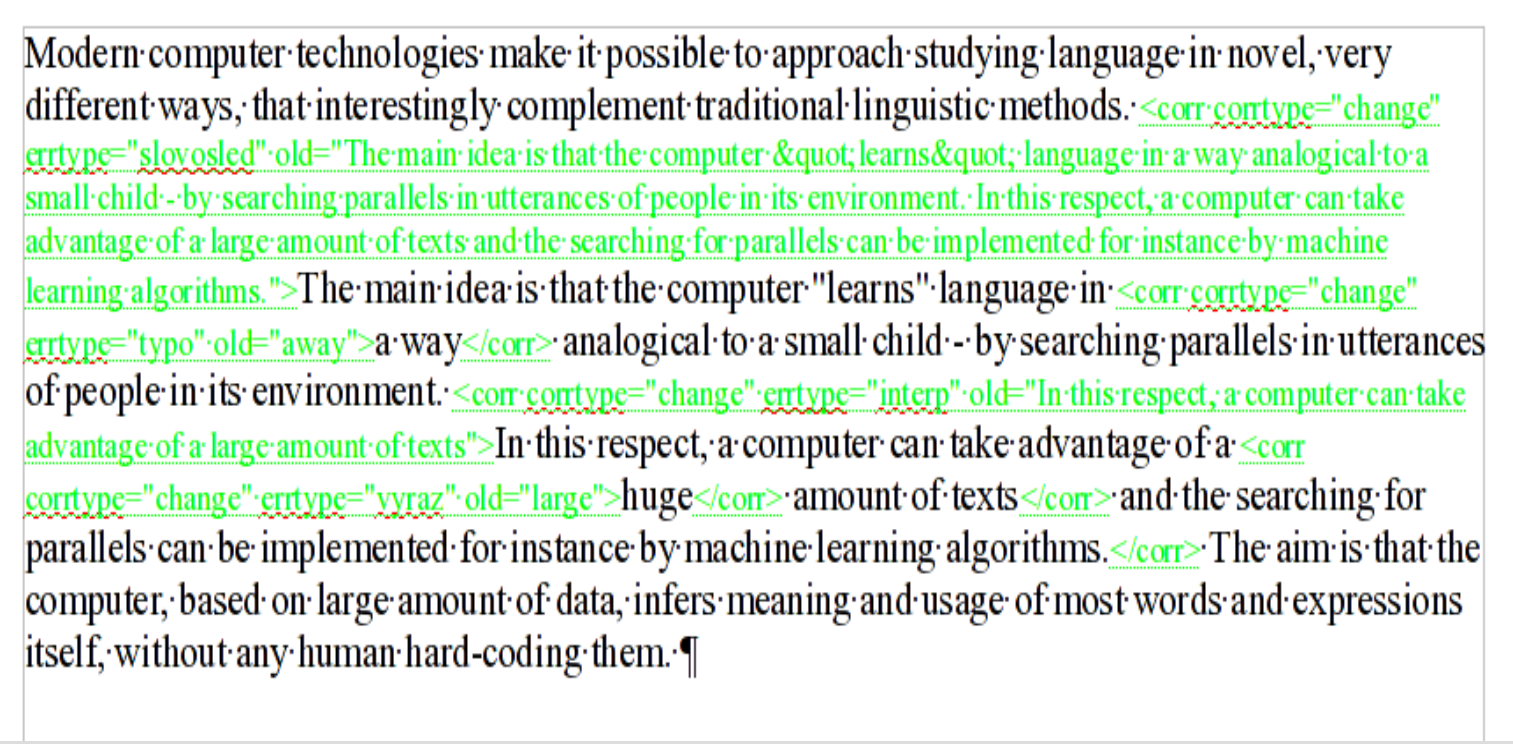
- Typy chyb:
  - překlepy
  - pravopisné chyby
  - typografické chyby
  - morfologicko-syntaktické chyby
  - lexikálně-sémantické chyby
  - stylistické chyby

Označení typů chyb

- prav-interp, prav-iy, prav-sz, prav-mneme, prav-malavelka, prav-sprezky, prav-prejata, prav-jine
- typo-pomlcka, typo-delenislov, typo-odstavce, typo-mezery, typo-predlozky, typo-jine
- synt-morf, synt-shoda, synt-vazba, synt-zajmena, synt-jine
- sem-vyraz, sem-nonsense, sem-jine

Nástroj pro označování chyb

- OOCorr – korekturní rozšíření do OpenOffice.org Writer (Jaroslav Moravec)



Rozvrstvení chyb v korpusu

Error Group	count	%
Spelling (simple)	2347	13.04
Morpho-syntactic	1689	9.39
Spelling (other)	867	4.82
Lexico-semantics	2536	14.09
Punctuation	3837	21.32
Stylistic	4184	23.25
Typography	2165	12.03
unsorted	371	2.06
Total	**17,996**	100.0

Závěr

- Získání reálných příkladů chyb.
- Distribuce výskytů různých druhů chyb v přirozeném jazyce.
- Základ pro interpunkční korektor.
- Mapování chyb v českém jazyce.
-
- E-learningový materiál.