

Classification of Errors in Text

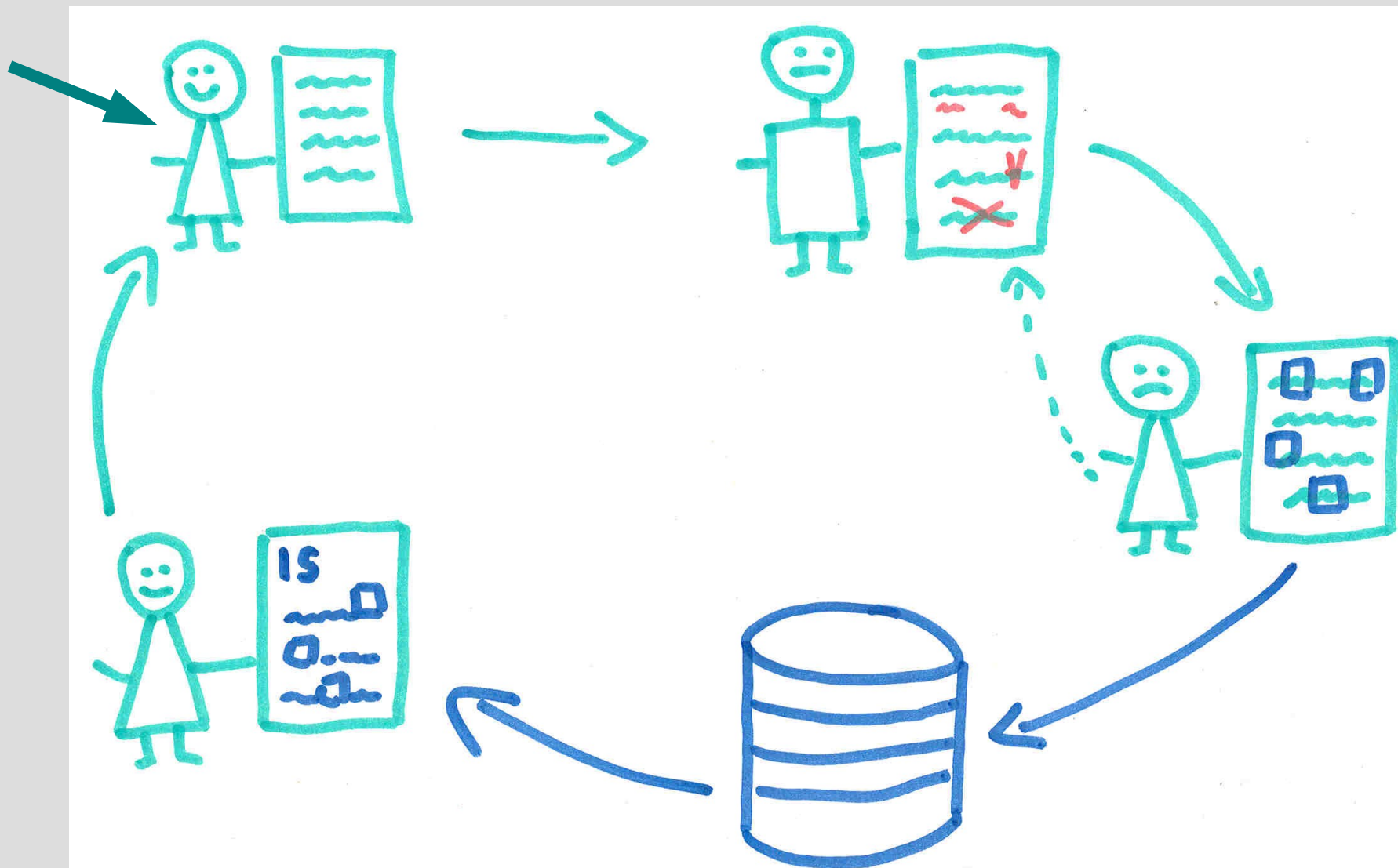
Jan Bušta, Dana Hlaváčková, Miloš
Jakubíček, Karel Pala

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)

Modern computer technologies make it possible to approach studying language in novel, very different ways, that interestingly complement traditional linguistic methods. <corr corrtype="change" errtype="slovo sled" old="The main idea is that the computer " learns" language in a way analogical to a small child - by searching parallels in utterances of people in its environment. In this respect, a computer can take advantage of a large amount of texts and the searching for parallels can be implemented for instance by machine learning algorithms.">The main idea is that the computer "learns" language in <corr corrtype="change" errtype="typo" old="away">a way</corr> analogical to a small child - by searching parallels in utterances of people in its environment. <corr corrtype="change" errtype="interp" old="In this respect, a computer can take advantage of a large amount of texts">In this respect, a computer can take advantage of a <corr corrtype="change" errtype="vyraz" old="large">huge</corr> amount of texts </corr> and the searching for parallels can be implemented for instance by machine learning algorithms. </corr> The aim is that the computer, based on large amount of data, infers meaning and usage of most words and expressions itself, without any human hard-coding them. ¶

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.