Typos in Czech corpora

Marek Grác

Motivation

- Attempt to find “better” corpus
  - Existing tools are not robust enough
  - Solve simple problems first
- Comparision of corpora
  - Is CBB.blog good enough?
- In semantic network Sholva – we annotate also typos and we want to identify them

Rapid Development

- We have a known goal
- Annotation task can be reduced to confirm/deny
- Expectation of high IAA
- Ability to use existing annotating tools

the-score

Czech corpus	The-score	Frequency
DESAM	9,200	12
CBB.blog	**1,418**	48
Syn2K12	7,897	18,847
czes2	56	530,289
czTenTen12	1,331	346,706
czTenTen12.clean	2,087	216,940

typo-ratio

- First 100K tokens from czTenTen which were not identified by morph. analyzer
- Should we annotate lemma or token?
- 32,766 tokens are not valid Czech words (agreement 2 from 2)
- Typical cases: missing diacritics, english words, texts in non-latin alphabets

typo-ratio

Czech corpus	Typos in thousands	Size in millions	Typo-ratio
DESAM	1.65	1.04	0.16 %
CBB.blog	2.32	0.81	**0.29 %**
SYN2K12	15,710	1,294	1.21 %
czes2	7,053	465	1.52 %
czTenTen12	55,650	5,436	1.02 %
czTenTen12.clean	28,482	5,214	0.55 %

The difference between czTenTen12 and czTenTen12.clean has typo-ratio 12.24%