CzAccent – Simple Tool for Restoring Accents in Czech Texts

Pavel Rychlý

pary@fi.muni.cz

CzAccent

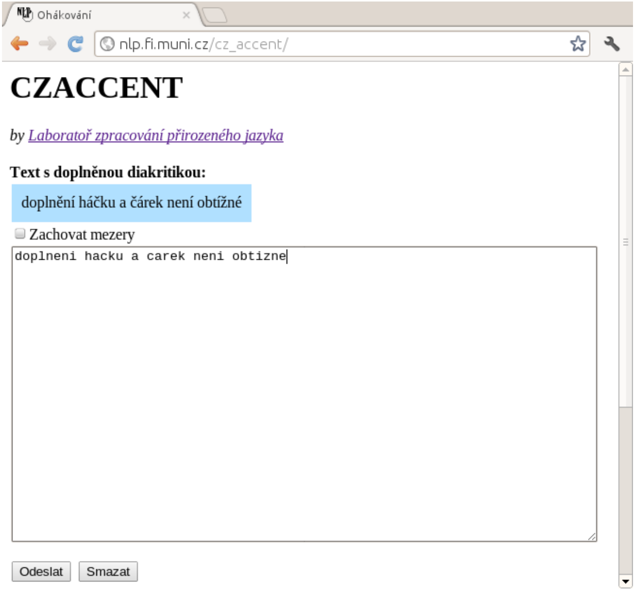
- tool for restoring accents in Czech texts
- input: plain text without accents
- output: same plain text with added accents

Example

```
 $ echo realny problem | czaccent

reálný problém
```

Web interface



CZACCENT

by Laboratoř zpracování přirozeného jazyka

Text s doplněnou diakritikou:

doplnění háčku a čárek není obtížné

☐ Zachovat mezery

```
doplneni hacku a carek neni obtizne
```

Algorithm

- use most frequent variant

Algorithm

- use most frequent variant
- big lexicon (ajka)
- corpus frequency
- finite state automata
- Daciuk – Finite State Utilities

Never *simple* in a paper

- never ever use word *simple* in a research paper

Never *simple* in a paper

- never ever use word *simple* in a research paper
- especially in title

Never *simple* in a paper

- never ever use word *simple* in a research paper
- especially in title
- but czaccent *IS* simple
- simple idea, simple code, simple usage
- simplicity is the core feature of czaccent

Conclusion

- simplicity rules
- use simple methods, algoriths, languages, interaface, . . .