Storing morphology information in a wiki

Radovan Garabík

Ľ. Štúr Institute of linguistics
Panská 26
813 64 Bratislava
Slovakia
e-mail: korpus@korpus.juls.savba.sk
www: http://korpus.juls.savba.sk

Morphology analysers

- different ways of describing morphology information
- Slavic languages – (prefix)+root+affix
- changes in the root, morphing of suffixes
- paradigm classes – common root (or lemma) modifications
- special treatment to either: reduce number of paradigms, allow guessing of unknown words or accommodate different linguistic premises
- partial paradigms
- our approach: no paradigms at all, for each word the paradigm is spelt out in full

Wiki

- to store all the information: wiki – easy collaborative editing, tracking of changes
- software of choice: MoinMoin http://moinmo.in/
- Python http://www.python.org/
- everything in UTF-8: minus one big problem
- plugins
- built in full text search engine or more efficient Xapian search engine bindings
- ~70 kwords (pages), ~$2.5 \cdot 10^6$ wordforms
- design: easily computer parseable, but also human readable

```
== Lema ==
ucho

== Paradigma ==
SSns1: ucho
SSns2: ucha
SSns3: uchu
SSns4: ucho
SSns5: ucho
SSns6: uchu
SSns7: uchom
SSnp1: uši, uchá
SSnp2: úch, ušú, uší
SSnp3: ušiam, uchám
SSnp4: uši, uchá
SSnp5: uši, uchá
SSnp6: uchách, ušiach
SSnp7: ušami, uchami
----
[[Kategória:Substantíva]]
```

- sections: Lema, Paradigma, kategórie

- homonymy: special page names: *mať (V)*, *mať (S)*
- disambiguation pages

```
== Lema ==
mať

== Pozri ==
[[mať_(S)]] [[mať_(V)]]
----
[[Kategória:Dezambiguácia]]
```

Quirks

- reflexive verbs: very efficient solution: we just ignore them :-)
- reflexive particle/pronoun tag **R**
- analytical forms: we ignore them too
- conditional particle tag **Y**
- analytical verbs: hey, it's just *byť* + infinitive or L-participle
- words cannot contain spaces/hyphens

28163	verbs
26061	substantives
13100	adjectives
5069	adverbs
1297	abbreviations
1104	participles
656	interjections
369	particles
369	pronouns
311	numerals
123	prepositions
110	conjunctions
72	citation elements
26	part of multiword expression
2	*sa/si*
1	*by*
716	disambiguation pages

Scalability

- each page in its own directory (several files)
- tens of thousands of directory entries in the main directory
- filesystem capable of efficiently handling such amount of data
- all the major contemporary Linux filesystems
- but the winner is....
- reiserfs (B-trees, tail packing)

Issues

- built in full text search engine cannot cope with such amount of data – multi minute long searches
- Xapian is fine
- category pages do not work conveniently – formatting of moderately long pages
- solution: hide category pages form the users
- otherwise everything works fine

To be continued...

- design interwiki links – easy
- design interwiki data transfer – tricky
- design data transfer to/from external data sources - ???
- XML-RPC?
- macros for easier editing (new entries)