CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES

Automatic Enrichment of Croatian Morphological Lexicon Using Large Corpora and Web Search

Danijela Merkler, Željko Agić, Marko Tadić

Faculty of Humanities and Social Sciences
University of Zagreb

FASSBL-8, Dubrovnik, 2012-09-20

Motivation

- manual enlargment of inflectional lexica is a time-consuming task requiring expertise
  - assigning inflectional paradigms to potential entries
  - ca 20 lemmas per hour
  - 10.000 lemmas equals 500 hours or ca 60 days of work
- Croatian Morphological Lexicon v 4.6
  - 110.000 lemmas, 4.000.000 entries (wordform, lemma, MSD)
  - measured coverage: 96% on HNK and 91% on hrWaC
  - lemmas added manually on daily basis
  - "remaining" lemmas are expectedly infrequent
- linguistically motivated rules for automatic enrichment
  - derive female nouns (Ncf.*) from animate male nouns (Ncm.*y)
  - derive possessive adjectives from male and female nouns (N.m.*y, subset of N.f.*)
  - validation using hrWaC (and Google search index?)

Preliminary experiment

- used as a proof of concept
- manually selected 100 animate male nouns
- designed a set of derivation rules to produce female counterparts
  - 41 rule
  - 210 candidate lemmas generated
- queried hrWaC and Google for frequency evidence
- frequency ≥ 10

	Female		Male-female pairs	
hrWaC	95.24%	/	93.00%	/
Gold	92.38%	91.43%	94.00%	86.00%
	Google	hrWaC	Google	hrWaC

Table: Preliminary experiment results

Extending experiment scope

- preliminary conclusions
  - lemmas identified with high accuracy
  - inflectional patterns assigned with derivation rules
  - hrWaC and Google scores comparable
  - Google Search API limited to 100 queries per day
- experiment extension
  - replace 100 manually selected nouns with real entries from the morphological lexicon
    - consider entries from other lexical sources
    - dictionaries, Croatian WordNet
  - include the possessive adjectives test case
  - use only hrWaC
  - data sparseness, frequency ≥ 2

Experiment setup

- male-to-female test case
  - input nouns selected from the lexicon
    - filter Ncm.*y lemmas: 2.937 nouns
    - wordnet filtering not feasible (domain: person, SUMO: human, male)
    - dictionary does not denote animateness
  - 41 derivational rule
  - generated 6.810 candidate female nouns
    - 5.904 not covered by the lexicon
    - 1.713 confirmed by hrWaC, 985 not covered by the lexicon
  - evaluated both candidate lists (all not covered vs. confirmed not covered)

Experiment setup

- noun-to-adjective test case
  - input nouns selected from the lexicon
    - male filter: N.m.*y
    - female filter: N.f.* with specific inflectional patterns
    - included previously generated female nouns
    - input size: 12.950 candidate lemmas
  - 66 derivational rules
  - generated 6.583 candidate possessive adjectives
    - 6.486 not covered by the lexicon
    - 777 confirmed by hrWaC, 746 not covered by the lexicon
  - evaluated both candidate lists (all not covered vs. confirmed not covered)

Results

- substantial difference between accuracy on confirmed and unconfirmed female nouns
  - ambiguous suffixes in derivational rules
  - e.g. *ribič* → *ribička* (adjective lemma *ribički*)
- high accuracy for both possessive adjective cases

Test case	Count	Accuracy	New lemmas	New wordforms
Female conf'd	985	76%	750	10.500
Female all	5.904	27%	1.594	22.316
Adjective conf'd	746	98%	731	10.234
Adjective all	6.486	89%	5.773	80.822
Total	/	/	8.848	123.872

Table: Experiment results

Conclusions

- introduced 8.848 new lemmas to Croatian morphological lexicon
    - cleaning entries much faster than creating ones
    - saved ca 55 days of manual work
    - new version of the lexicon being prepared
        - includes these results and results of manual enlargement
        - expected ca 130.000 lemmas, more than 5.500.000 entries
- future work directions
    - guessing inflectional patterns for lemmas
    - including verb patterns

Thank you for your attention.