Development and Applications of the Croatian *1984* Corpus for the MULTEXT-East Resources

Željko Agić**, Danijela Merkler*, Daša Berović*, Marko Tadić*

* Department of Linguistics

** Department of Information Sciences

Faculty of Humanities and Social Sciences, University of Zagreb

{zagic, dmerkler, dberovic, marko.tadic}@ffzg.hr

SlaviCorp 2011
Dubrovnik
2011-09-14

MULTEXT-East project

- MULTEXT-East resources
  - multilingual dataset for LT research and development
- covering today
  - Bulgarian, Croatian, Czech, English, Estonian, Hungarian, Lithuanian, Macedonian, Persian, Polish, Resian, Romanian, Russian, Serbian, Slovak, Slovene, and Ukrainian
- language resources
  - morphosyntactic specifications
    - Croatian included since 1998
  - lexica
  - annotated *1984* corpus
  - MULTEXT-East parallel and comparable text and speech corpora
  - associated documentation

The *1984* corpus

- the central component of the MULTEXT-East corpus
- XML marked-up in accordance with the TEI reccomendations
- annotation with hand validated MSDs and lemmas
  - suitable for MSD tagging and lemmatisation experiments
- separate alignment files
  - hand-validated pair-wise sentence alignments between English and the translations
- version 4 adds pair-wise alignments between all the languages
  - automatically induced from the alignments with English

The *1984* corpus

- the Croatian translation of *1984* semi-manually annotated for lemmas and morphosyntactic tags
  - first step
    - text of *1984* and Croatian Morphological lexicon matched yielding all lemma and MSD interpretations for tokens
    - manual selection of proper interpretation by 20 students
  - second step
    - two expert annotators
    - checking the corpus for errors in the previous round
      - sentence segmentation, tokenization, lemmatization, assignment of MSD tags
    - overlap of 1/4 of the overall corpus size in sentences
      - in order to calculate the inter-annotator agreement on tokens, lemmas and morphosyntactic tags, including PoS and other MSD categories

Problems

- manual verification of the semi-manually lemmatized and MSD-tagged corpus
- two kinds of problems
  - problems of processing
    - incorrect lemmatization and MSD tagging
    - such errors were corrected
  - problems of the text
    - errors in lemmatization and MSD tagging because of the errors in the text itself
    - we did not intervene, errors were marked

Problems of processing

- correcting problems of processing by consulting two Croatian grammars and a Croatian dictionary
  - nouns with numerals 2, 3 and 4
  - adverbialised nouns and pronouns
  - conjunctions
  - modal particles

Nouns with numerals 2, 3 and 4

- nouns that occur with the numbers 2, 3 and 4 always appear in a special morphosyntactic category
- masculine, neuter and feminine nouns corrected to genitive case singular
    - *tri (tri, Mc-p-l) čovjeka (čovjek, Ncmsa--y) → tri (tri, Mc-p-l) čovjeka (čovjek, Ncmsg)*
    - *dva (dva, Mc-p-l) mišljenja (mišljenje, Ncnpa) → dva (dva, Mc-p-l) mišljenja (mišljenje, Ncnsg)*

Adverbialised nouns and pronouns

- adverbialization of nouns
  - result of lexicalization of certian word-forms that become a new lexical unit and move to new PoS

    Mjesecima prije toga nije se dijelila čokolada.
    mjesec, Ncmpd
    mjesecima, Rt
- adverbialised pronouns
  - occur as intensifiers of an adverb

    Onda se udalji od mene što brže možeš.
    što, Pi3n-n--n-n-n
    što, Rn

Conjunctions

- different types of words can have function of the conjunction in the sentence
  - adverbs (pronominal adverbs with interrogative semantics)

    U tom trenu, iznenada mu je sinulo kako su potpuno sami.

    kako, Rn

    kako, Css
  - pronouns (relative – *što*, *koji* → conjunctions introduce causal, temporal and comparative clauses)

    Ubrzo su bili u neprilici što tu sjede tako šutke.

    što, Pi3n-a--n-n-n

    što, Css

Conjunctions

- combination of multiple words
  - conjunction + conjunction or
  - conjunction + non-conjunction word (adverb, present participle, etc.)

Ali budući da u zbilji Veliki Brat nije svemoćan, a Partija nije nepogrešiva, postoji stalna potreba za neumornom i uvijek budnom elastičnošću u postupanju s činjenicama.

budući biti1 Vcpp, da Css

budući Csc, da Csc

Modal particles

- many words in a sentence can have a function of modal particles
- in the most cases in *1984* these are conjunctions *i* and *ni*, that have function of intensifiers

...presjeći će se i posljednja karika koja vezuje s prošlošću.

i, Css

i, Qo

Ni tisuću je raketnih bombi ne bi razorilo.

ni, Css

ni, Qo

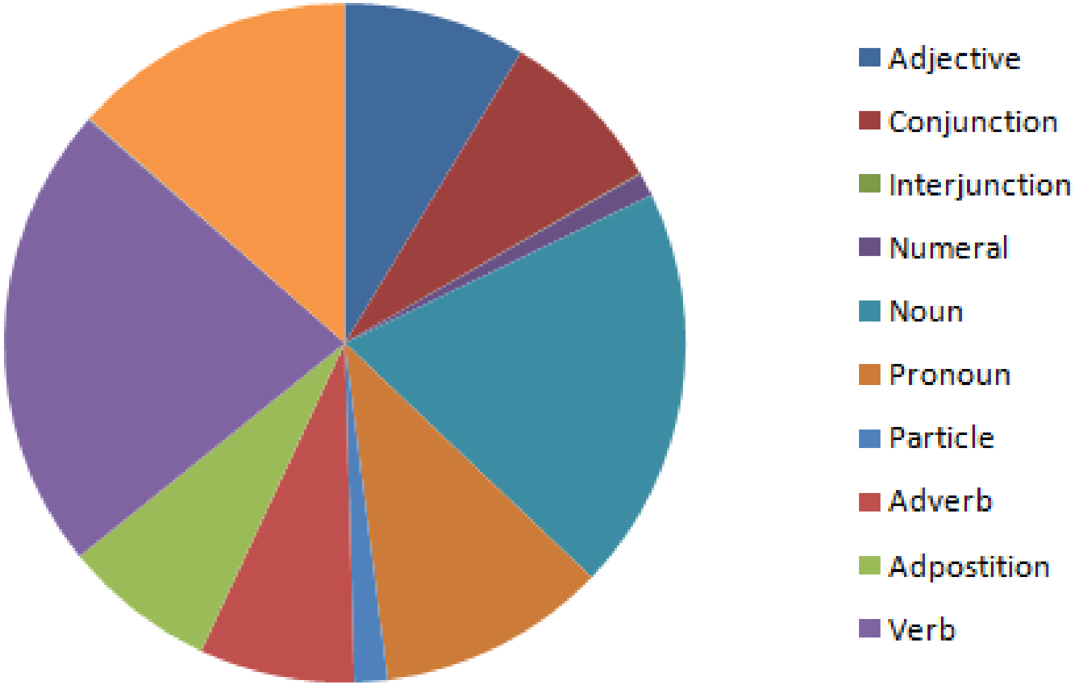
Problems of the text

- problems of the text imply errors in lemmatization and MSD tagging because of the errors in the text itself:
  - spelling errors: *čeovjek* (*čovjek*), *šte* (*što*), *ćene* (*neće*), *veli* (*voli*), *i i*...
  - grammar and spelling errors: *zaspe* < *zasuti* (*zaspi* < *zaspati*)...
  - disagreement: *ovaj* (Pd-msn--n-a--) *puta* (Ncmsg), *u razini oko* (Spsg) *metar* (Ncmsa--n), *poslije* (Spsg) *podne* (Ncnsn)...
- not corrected, but marked for possible later correction

Corpus stats

- 6625 sentences, 106632 tokens
  - 18846 different wordforms, 8671 different lemmas
- annotated by 802 different MSD-tags
- distribution of word types in the corpus

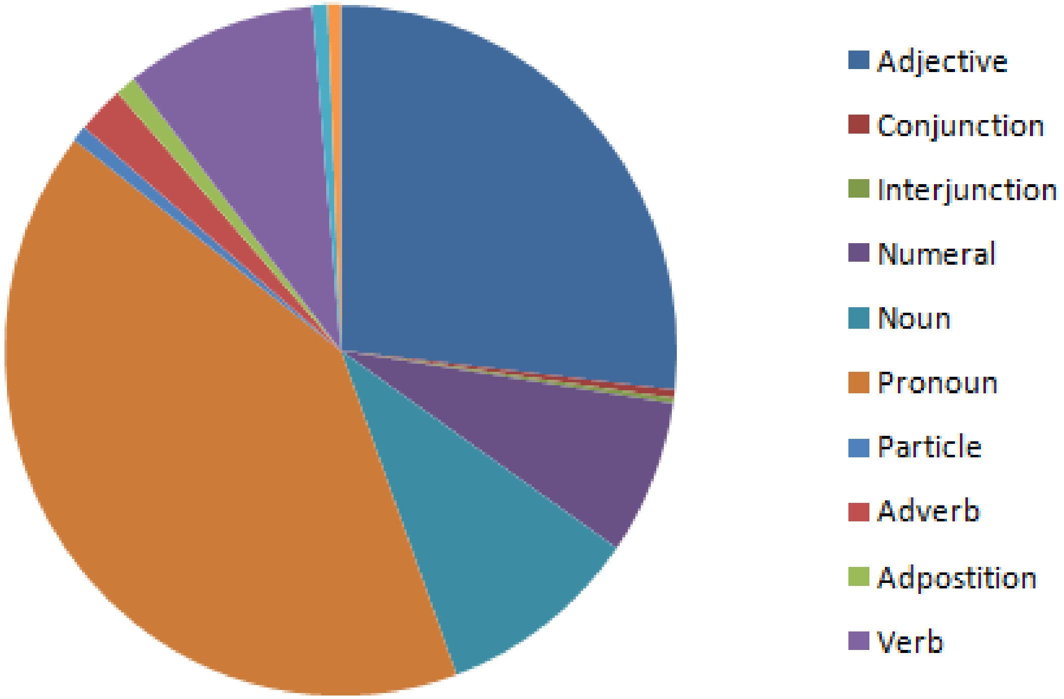
adjective	9244
conjunction	8553
interjunction	45
numeral	1177
noun	20524
pronoun	11627
particle	1680
adverb	7881
adpostition	7760
verb	23659
abbreviation	56
other	14426



Corpus stats

- distribution of different MSD-tags on parts of speech
- adjectives, nouns, pronouns and verbs expected to be the most difficult to annotate

adjective	215
conjunction	3
interjunction	2
numeral	58
noun	78
pronoun	329
particle	6
adverb	18
adpostition	8
verb	74
abbreviation	6
other	5

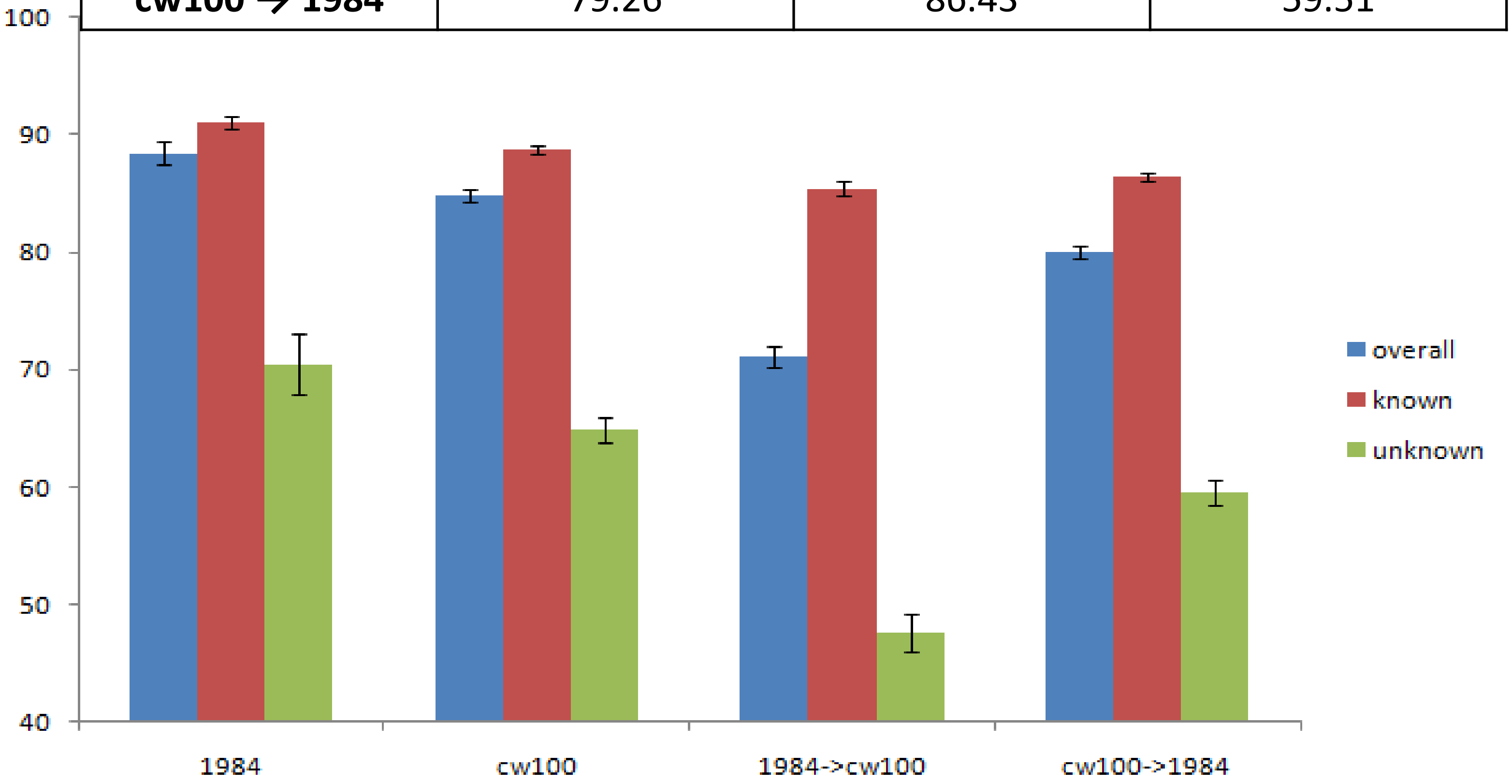


Experiment setup

- an illustrational experiment in morphosyntactic tagging of Croatian by using the *1984* corpus
  - CroTag HMM tagger (and lemmatizer)
  - tenfold cross-validated
  - cross-tagging by using the CW100 newspaper corpus of Croatian (differring in domain)
  - four different scenarios
    - train on 1984, test on 1984
    - train on CW100, test on CW100
    - train on 1984, test on CW100
    - train on CW100, test on 1984
- results predefined by the size of the model
  - CW100 ca 10% larger than *1984* in terms of tokens and different MSDs used in the annotation

Results

	overall	known	unknown
1984	88.46	91.09	70.48
cw100	84.80	88.70	64.94
1984 → cw100	71.11	85.44	47.67
cw100 → 1984	79.26	86.43	59.51



Future work

- completion of Croatian MULTEXT-East lexica and relevant documentation
- inclusion of Croatian *1984* corpus in the next version of MULTEXT-East Resources
- use also the Croatian *1984* for the experiments that all other MULTEXT-East resources are submitted to
- usage of Croatian translation of *1984* in experiments within the ACCURAT project
    - si, ro also included

ACCURAT

Thank you for your attention.

www.accurat-project.eu

The research within the project Accurat leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement nº 248347.

SEVENTH FRAMEWORK PROGRAMME