Babel Dependency Treebank of Public Messages in Croatian

Danijela Merkler, Željko Agić, Ana Agić
University of Zagreb, Croatia
{dmerkler, zagic, aagic}@ffzg.hr

Introduction

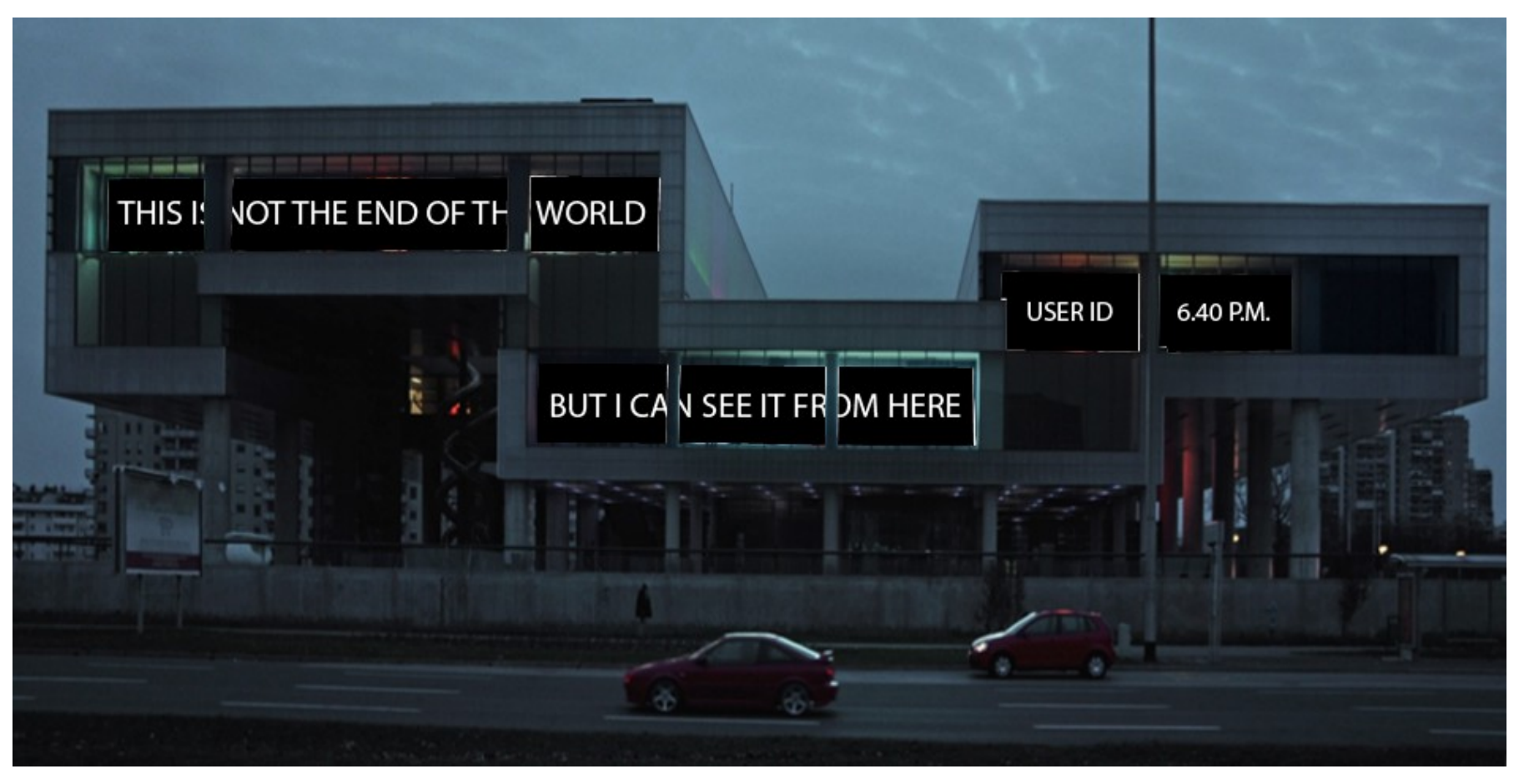
Croatian is an under-resourced language for which, up to this point, only one publicly available treebank exists – Croatian Dependency Treebank (HOBS, http://hobs.ffzg.hr), consisting of 118 kw of sentence-delimited, tokenized, lemmatized, morphosyntactically and syntactically annotated newspaper text.

We present the process of constructing a publicly available dependency treebank of public messages written in Croatian – the Babel dependency treebank – to complement HOBS in terms of genre, domain and temporal spread of the texts. Babel treebank serves as a test case and showcase for introducing a new standard for syntactic annotation of Croatian texts. The corpus is based on the Babel contemporary art project by Francisco Jodice (http://www.francescojodice.com).

Treebank construction

In the Babel art project, messages were collected from various electronic sources – e-mail, blog, Facebook, SMS – and published on the Zagreb Museum of Contemporary Art LED facade, aiming to use the facade as an open-space blog, enabling the citizens to publicly express their views. Three types of messages were displayed, limited by the LED facade to 150, 300 and 450 characters.

A sample of approximately 800 messages was made available for Babel treebank construction. They were manually classified and transcribed into standard Croatian. The original and the transcription were then manually sentence-split, tokenized, POS and MSD-annotated and also syntactically annotated. There is approximately 10.000 tokens in 1.100 sentences in both the original and the transcription.

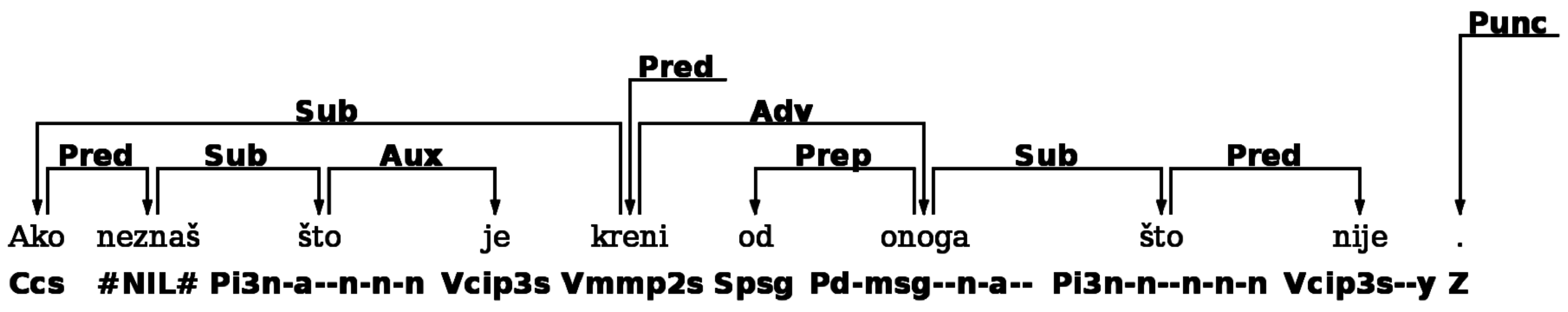


Morphology

Lemmatization and morphosnytactic tagging of original and transcribed messages was done semi-automatically by obtaining ambiguous unigram tagging by the Croatian Lemmatization Server (http://hml.ffzg.hr) and manual disambiguation. Multext East v4 tags were used.

466 different morphosyntactic tags were used for annotating the original text and 460 for the transcription. Token to type and token to lemma ratio for the original and trascribed text was observed at 2.38 and 3.42, 2.81 and 4.19, respectively. In comparison, HOBS has a token to type ratio of 4.25 and 8.40 tokens per lema.

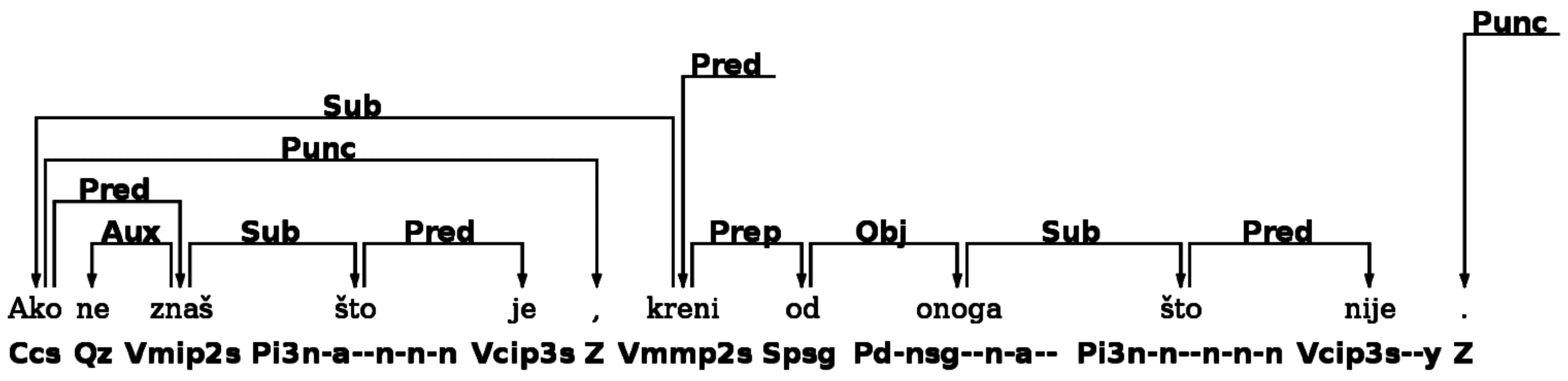
The distribution of parts of speech is virtually identical in both parts of the treebank: nouns (20%) are followed by verbs (15%), pronouns (9%) and adjectives (7%), with minor differences.


Punc
Pred
Sub
Adv
Pred
Sub
Aux
Prep
Sub
Pred
Ako neznaš što je kreni od onoga što nije .
Ccs #NIL# Pi3n-a--n-n-n Vcip3s Vmmp2s Spsg Pd-msg--n-a-- Pi3n-n--n-n-n Vcip3s--y Z

Syntax

Being based entirely on the Prague Dependency Treebank, HOBS uses 28 main and 41 etxtended, i.e., a total of 69 different syntactic tags. It makes the annotation process difficult in terms of IAA and the adaptation of PDT rules to Croatian is not always straightforward.

We propose a new annotation model. It has 15 syntactic tags introduced by closely observing grammatical rules of Croatian. The tagset reduction is motivated by the 100 kw JOS treebank of Slovene, which served to replace the PDT-motivated Slovene Dependency Treebank formalism with a model using only 10 syntactic tags. We introduce explicit encoding of coordination and subordination of clauses and a consistent model of complex predicate annotation. The distribution is governed by predicates (12%), attributes (12%), objects (10%) and ellipses (7%).


Punc
Pred
Sub
Punc
Pred
Aux
Sub
Pred
Prep
Obj
Sub
Pred
Ako ne znaš što je , kreni od onoga što nije .
Ccs Qz Vmip2s Pi3n-a--n-n-n Vcip3s Z Vmmp2s Spsg Pd-nsg--n-a-- Pi3n-n--n-n-n Vcip3s--y Z

Future work

As a treebank of non-standard Croatian text, Babel treebank introduces various courses of future work. Transcribed and non-transcribed portion of the treebank could be merged into a parallel treebank to conduct experiments on their differing properties.

Additional texts could be introduced, e.g. from the Croatian SMS corpus, forums, blogs and news commentary.

Once the syntactic tags from HOBS are converted in accordance with the new annotation guidelines, an experiment dependency parsing of sentences from the Babel treebank will be conducted to establish parsing baselines for non-standard Croatian text.

Experiments with lemmatization and morphosyntactic tagging of non-standard text are also supported by the treebank.

Acknowledgements

We would like to thank Francesco Jodice for envisioning the Babel project and Iva Radmila Janković from the Museum of Contemporary Art for implementing it in Zagreb and making a sample of the texts available for constructing the Babel treebank. We hope to get approved by the artist for making the treebank publicly available via permissive Creative Commons licensing.

Many thanks to all the unknown authors who participated in the project by publishing their texts on the LED facade at MSU.

Please feel free to conatct us regarding the treebank! ☺