Abstract

The paper presents the process of constructing a publicly available treebank of public messages written in Croatian. The messages were collected from various electronic sources – e-mail, blog, Facebook and SMS – and published on the Zagreb Museum of Contemporary Art LED facade within the Babel art project. The project aimed to use the facade as an open-space blog or social interface for enabling citizens to publicly express their views. Construction and current state of the treebank is presented along with future work plans. A comparison of Babel Treebank with Croatian Dependency Treebank and SETimes.HR treebank regarding differing domains and annotation schemes is briefly sketched. The treebank is used as a test platform for introducing a new standard for syntactic annotation of Croatian texts. An experiment with morphosyntactic tagging and dependency parsing of the treebank is conducted, providing first insight to computational processing of non-standard text in Croatian.

© 2013 The Authors. Published by Elsevier Ltd.
Selection and peer-review under responsibility of CILC2013.

Keywords: dependency treebank; dependency parsing; public messages; non-standard text; Croatian language

1. Introduction

At the time of conducting the work presented in this paper, Croatian Dependency Treebank (HOBS) (Tadić, 2007) was the only existing and publicly available syntactically annotated corpus of Croatian texts. It consists exclusively of texts from the newspaper domain assembled from a Croatian daily newspaper published in 2000 (Agić, 2012; Berović et al., 2012). Its current version contains 4,626 sentences. It is annotated by adapting the
syntactic formalism used in the Prague Dependency Treebank (PDT) (Hajč et al., 2000) to differing properties of Croatian syntax. This version is available through META-SHARE (Tadić and Varadi, 2012.), but syntactic tags have been stripped from it, i.e., only the morphosyntactic layer and tree structure of sentences is maintained, while actual syntactic roles of word forms are unavailable. Another treebank of Croatian is currently in prototype stage. The SETimes.HR treebank (Agić and Merkler, 2013) is based on the morphosyntactically annotated SETimes.HR corpus (Agić et al., 2013) of newspaper text from the publicly available SETimes parallel corpus (http://www.nljubesic.net/resources/corpora/setimes/). This treebank will be made available under the CC-BY-SA-3.0 license once it leaves prototype stage. As treebanks are generally considered to be a sparse linguistic resource (Ambati and Chen, 2007), any effort in enlarging and enhancing the Croatian Dependency Treebank and in developing complementary – in terms of genre and domain selection or temporal spread of the texts – syntactically annotated corpora of Croatian can be considered desirable, especially being that Croatian is an under-resourced language in general (Tadić et al., 2011).

The construction of Babel Treebank of Croatian is based on the Babel project. Babel is an artistic project led by an Italian contemporary artist Francesco Jodice (http://www.francescojodice.com/). It was basically envisioned as an attempt at providing a social interface for citizens of an urban environment to interact and "publicly express their intentions, strategies, views and disagreements" (http://www.msu.hr/#/en/19359/). More specifically, several electronic communication interfaces – the project blog and Facebook page, e-mail and SMS – were used to collect messages from users and the messages were published in real time on the 400 m² LED facade, i.e., video wall of the Zagreb Museum of Contemporary Art to enable a channel for users to express their viewpoints. The messages were collected and manually transferred to the video wall by museum staff in the period of 11 days, from 2011-11-23 to 2012-12-03, with the latter specifically targeted as being the day before the Croatian parliamentary elections of 2012. The messages were initially envisioned with a limitation to 150 characters but the limit was finally defined by specific video wall requirements. Namely, the video wall consists of three panels – small, medium and large – which are in turn limited to 150, 300 and 450 characters. A sample of approximately 800 collected messages was made available by the Babel project team to build a prototype Babel Treebank. As it was the intention of the project team and the artist to make the messages published on the video wall publicly available upon project completion, the Babel Treebank is also to be made available for research purposes under the CC-BY-NC-SA-3.0 license.

In the following section, Babel Treebank construction is described, its basic properties are sketched and compared with the properties of HOBS and SETimes.HR treebank. A new standard for syntactic annotation of Croatian texts is introduced and applied in manual syntactic annotation of Babel Treebank. Results of an experiment with morphosyntactic tagging and dependency parsing of Babel treebank using SETimes.HR corpus and treebank as language models are provided and discussed. Future work directions are indicated.

2. Treebank construction

Babel text sample was delivered from the Zagreb Museum of Contemporary Art as a spreadsheet document containing approximately 800 messages published on the video wall. The metadata was sparse and unstructured – other than the message text itself, it contained the information on author and creation date, these two being included within message text itself, i.e., without any explicit annotation. The data was thus manually validated and structured. Invalid and meaningless entries were deleted, leaving a total of 742 messages. Each message was assigned with structured metadata on creation date, author, source, language and category. Six general categories were defined by observing the data: personal (ca 52.29% of the messages), political (34.23%), social (6.47%), advertisement (3.37%), religion (2.29%) and sports (1.35%). A total of 42 messages (5.66%) were not written in Croatian, 32 of which were in English and the remaining 10 in Serbian, Slovene, Italian, Latin, Spanish and German. 59 messages were annotated as potential spam but were still held in the corpus. More than 65% of the messages were anonymous.

All non-foreign messages were manually transcribed to standard Croatian. Both the original and the transcribed messages were manually sentence delimited and tokenized. As indicated in Table 1, each message contained 1.46 sentences and 13.08 word forms on average in the original corpus and 1.39 sentences and 13.06 word forms in the transcribed corpus. Stats in Table 1 are given in comparison with HOBS and SETimes.HR treebank. The newspaper treebanks have at least 10 more word forms in a sentence on average when compared to Babel, as expected given the different nature of the texts. The word form to lemma ratio is maintained across corpora. All corpora use the
Multext East v4 tagset (Erjavec, 2010) and Babel and SETimes.HR also share the syntactic formalism, while HOBS uses the PDT syntactic model, which is reflected in the number of different syntactic tags used for syntactic annotation.

Table 1. Babel Treebank stats compared to HOBS and SETimes.HR

| Feature      | Babel original | Babel transcribed | HOBS      | SETimes.HR  |
|--------------|----------------|-------------------|-----------|-------------|
| Sentences    | 1 084          | 1 035             | 4 626     | 2 488       |
| Tokens       | 9 710          | 9 672             | 117 369   | 56 334      |
| Word forms   | 4 076          | 3 446             | 25 038    | 13 409      |
| Lemmas       | 2 841          | 2 307             | 12 388    | 6 901       |
| MSD tags     | 466            | 460               | 914       | 804         |
| Syntactic tags | 15             | 15                | 70        | 15          |

2.1. Morphology

Once both original and transcribed Babel corpus were manually verticalized, the word forms were linguistically annotated, i.e., lemmatized and morphosyntactically tagged ambiguously at unigram level by using Croatian Lemmatization Server (Tadić, 2005). Ambiguous annotations, i.e., all possible lemma-tag pairs assigned by the service were manually disambiguated by linguists. Multext East v4 morphosyntactic tags for Croatian (Erjavec, 2010) were used in the annotation process. Basic stats regarding this procedure are given in Table 1 and Table 2.

Table 2. Distribution of morphosyntactic tags in Babel Treebank, HOBS and SETimes.HR

| MSD tag     | Babel original (%) | Babel transcribed (%) | HOBS (%) | SETimes.HR (%) |
|-------------|---------------------|-----------------------|----------|----------------|
| Abbreviation | 1.04                | 1.73                  | 1.01     | 1.41           |
| Adjective   | 6.54                | 7.09                  | 12.22    | 11.42          |
| Adverb      | 4.33                | 4.70                  | 3.92     | 3.30           |
| Conjunction | 5.54                | 6.27                  | 7.05     | 5.24           |
| Interjection| 0.13                | 0.30                  | 0.01     | 0.00           |
| Noun        | 19.55               | 20.90                 | 29.89    | 32.24          |
| Numeral     | 1.73                | 1.47                  | 1.83     | 2.40           |
| Particle    | 1.41                | 1.80                  | 0.47     | 0.40           |
| Preposition | 6.48                | 5.98                  | 9.64     | 9.73           |
| Pronoun     | 10.14               | 8.86                  | 6.22     | 5.08           |
| Punctuation | 22.64               | 21.85                 | 12.56    | 13.70          |
| Residual    | 4.66                | 1.38                  | 0.51     | 0.38           |
| Verb        | 15.81               | 17.67                 | 14.67    | 14.72          |

Specifically, Table 2 displays certain morphosyntactic tagging differences between the original and the transcribed Babel corpus and compares them to HOBS and SETimes.HR. Namely, it is indicative that the most frequent categories are mostly in overlap between original Babel and its transcribed version and that conjunctions and residual, i.e., foreign or out-of-vocabulary word forms are more frequent in the non-transcribed version. Adjectives, nouns and prepositions occur more frequently – approximately 6%, 11% and 3% more – in newspaper corpora, while verbs and pronouns are slightly more frequent in Babel (2% to 5%). Punctuation is heavily used in Babel – approximately 10% more than in the two newspaper corpora, mainly due to its usage in accentuation and creating emoticons.
2.2. Syntax

HOBS was annotated according to the PDT model (Hajič et al., 2000), which consists of 70 analytical functions. In the process of manual HOBS annotation, the annotators have observed (Berović et al., 2012) that this tagset size and corresponding annotation guidelines are too complex and not entirely in accordance with syntactic rules of Croatian language. This fact is further supported by observing inter-annotator agreement for HOBS and SETimes.HR treebank (Agić and Merkler, 2013). For similar reasons, Slovene Dependency Treebank project (Džeroski et al., 2006), which used the PDT guidelines, was abandoned and a simpler formalism with only 10 syntactic functions was introduced in the JOS Corpus of Slovene (Erjavec et al., 2010). For these reasons, a new and simpler annotation model is introduced here. The syntactic tagset is reduced to 15 analytical functions and enables easier manual annotation, since the guidelines are more compliant – although still with certain compromises – with grammatical rules of Croatian. Babel corpus served as the first test case for the new annotation model and thus its syntactic formalism is briefly described here. This model is subsequently implemented in SETimes.HR treebank, as Babel treebank predates it.

| Function | Babel original (%) | Babel transcribed (%) | SETimes.HR (%) |
|----------|--------------------|-----------------------|---------------|
| Adv      | 6.68               | 5.22                  | 4.98          |
| Ap       | 0.60               | 1.02                  | 3.03          |
| Atr      | 12.67              | 11.39                 | 26.34         |
| Atv      | 1.41               | 1.38                  | 1.66          |
| Aux      | 6.10               | 4.54                  | 6.49          |
| Co       | 3.81               | 4.90                  | 3.34          |
| Elp      | 7.56               | 7.17                  | 0.57          |
| Obj      | 10.12              | 10.24                 | 7.40          |
| Oth      | 1.83               | 2.70                  | 1.75          |
| Pnom     | 3.13               | 2.88                  | 1.65          |
| Pred     | 11.86              | 13.89                 | 9.29          |
| Prep     | 6.01               | 5.93                  | 9.62          |
| Punc     | 18.50              | 19.67                 | 13.31         |
| Sb       | 6.60               | 5.81                  | 7.07          |
| Sub      | 3.12               | 3.26                  | 3.49          |

Syntactic function Pred is used to annotate every predicate in every sentence and clause. However, predicates in Croatian are often structurally complex and may (and usually do) consist of several words. In such cases, the Pred tag is assigned to the head element of the complex predicate structure. Two classes of complex structures are differentiated by the formalism.

The first class addresses complex tenses, in which the predicate consists of a main verb and a set of auxiliary verbs. The main verb is annotated as Pred, while all auxiliary verbs are assigned with the syntactic tag Aux. Aux is also used to annotate other auxiliary elements such as negation and reflexive pronoun se when they participate in the predicate set in terms of meaning conveyed (Fig. 1).

The second class of complex predicate structures involves semicopulative verbs in relation with main verbs in infinitive. Since semicopulative verbs denote tense, person and number in such structures, they are annotated with Pred, while the verb in infinitive is by convention denoted as Atv. In such structures, an auxiliary verb is often included and annotated as Aux (e.g., On_Sb je Aux mogao_Pred tretati_Atv satima_Adv (en. He could run for hours)). Moreover, syntactic tag Atv is used to annotate conventional verbal complements such as participes, as well as nominal predicate complements that extend the predicate but agree with subject or object (Fig. 1).
Fig. 1. Complex predicate annotation (en. *Mate works sleepy, but he would not rest.*)

In nominal predicates, the nominal part of the predicate is annotated as Pnom, while the (semi-)copulative verb is annotated as Pred. If the nominal predicate involves a complex tense in which the (semi-)copulative part consists of several elements, the more meaningful verb is denoted as Pred, while the other elements are denoted as Aux (e.g. *On_Sb je_Aux postao_Pred sportaš_Pnom* (en. *He became an athlete*)). Furthermore, a nominal predicate structure may be assembled by using a semicopulative verb, auxiliary verb, (semi-)copulative verb in infinitive and a nominal word (e.g. *Mogao je_postati sportaš*) (Fig. 2). In such cases, nominal words are annotated as Pnom and introduced by (semi-)copulative infinitive verbs, while the remaining part of the structure is annotated as described for complex verbal predicates.

Fig. 2. Annotation of nominal predicates, complex conjunctions and subordination
(en. *Being that Ante runs whenever he can, he might have become an athlete.*)

Syntactic functions Sb, Obj, Adv, Atr, Ap and Prep are in turn used to annotate subjects, objects, adverbs, attributes, appositions and prepositions in full compliance with contemporary grammars of Croatian (Silić and Pranjković, 2005).

Syntactic function Co is used to link coordinated clauses and coordinated sentence elements (e.g. *Lana i Žana pjevaju*) (Fig. 3). This implies that Co is usually assigned to a specific subset of Croatian conjunctions. However, it may be assigned to punctuation in coordinating role (e.g., *Jedi, moli, voli*) (Fig. 4a). Punctuation not involved in coordination is annotated as Punc. In contrast to Co, syntactic tag Sub is assigned to sentence elements that introduce subordinate clauses. Other than the respective subset of conjunctions, certain pronouns and adverbs may be denoted as Sub (e.g., *treći kad može*) (Fig. 2).

Syntactic function Elp denotes all those sentence structure elements that are directly dependent to words left out from a sentence, e.g., in sentences without predicates (*Dva kotača, jedna ljubav*) (Fig. 4b). When compared with newspaper text, such sentences are more frequent in the public messages corpus. Elp is also used for inserted
sentences or sentence parts that form digressions in the main sentence structure (e.g., Oni_Sb su_Aux _,Punc kažu_Elp _,Punc došli_Pred u_Prep Zagreb_Adv (en. They have, so they say, come to Zagreb)). Oth is basically used to annotate all sentence elements that are left unannotated by other syntactic tags. More specifically, it is assigned to particles, intensifiers and modal particles that often take particle roles in specific sentences, while belonging to a different part of speech, e.g., pronouns and conjunctions (pleše sve bolje) (Fig. 3). Oth is also used to annotate single-function multiword structures such as complex conjunctions (e.g. budući da) (Fig. 2).

![Fig. 3. Annotation of coordination and intensifier particle (en. Lana and Žana are singing and Jelena dances better and better.)](image)

![Fig. 4. (a) Coordination without conjunctions (en. Eat, pray, love.) (b) Ellipsis (en. Two wheels, one love.)](image)

The annotation of Babel corpus, both original and transcribed, was done by using the DgAnnotator tool (http://medialab.di.unipi.it/Project/QA/Parser/DgAnnotator/). Basic annotation stats are given in Table 3, which displays the distribution of syntactic functions in both versions of the Babel treebank compared to SETimes.HR. Direct comparison to HOBS is excluded from Table 3 as the formalisms do not entirely match and would possibly introduce misleading differences. Table 3 shows that attributes are much more frequent in SETimes.HR newspaper text than in Babel, while – as was the case with morphosnytactic annotation – Babel has much more punctuation and a slightly higher ratio of nominal and verbal predicates. These differences are most likely reflecting the differences in text domains, as public messages are substantially shorter than newspaper sentences and thus leave no room for optional arguments such as attributes and apositions. Moreover, ellipses are also much more frequent in Babel, as public messages sometimes contain verbless statements such as the ones often seen in graffiti, e.g., Ljubav, ne rat! (en. Love, not war!). Such sentences are typically not frequent in newspaper text.

This new dependency-syntactic formalism for Croatian was inspected for manual annotation quality by assessing inter-annotator agreement and comparing it with the PDT-style formalism of HOBS in (Agić and Merkler, 2013). Two expert annotators were given 100 overlapping sentences from HOBS and the other two expert annotators were assigned with 100 sentences from SETimes.HR. Labelled and unlabeled attachment score (LAS, UAS), linear label attachment (LA) and Cohen’s kappa coefficient κ(LA) was calculated. The new formalism was shown to
consistently outperform the PDT-based one regardless of the metric: 86.11 LAS (7.22 points over HOBS), 91.29 UAS (+2.13), 92.51 LA (+8.44). Cohen's kappa for label attachment was calculated at 0.920 (+0.081 over HOBS). It should still be noted that these scores were observed on newspaper text from HOBS and SETimes.HR. However, due to temporal constraints and limited human annotator availability for an under-resourced language such as Croatian, it is argued here that the properties obtained on newspaper text are not genre and domain dependent, i.e., that the differences in annotation quality between the two formalisms would be maintained on data from the public messages corpus.

3. Tagging and parsing

To the best of the authors' knowledge, no experiments with morphosyntactic tagging and dependency parsing of non-standard Croatian text were conducted at the time of developing Babel treebank. Here, HunPos tagger (Halacsy et al., 2007) and MSTParser (McDonald et al., 2006) were trained on SETimes.HR corpus (Agić et al., 2013) and SETimes.HR treebank (Agić and Merkler, 2013) training sets and tested on the entire Babel treebank, its transcribed version and on SETimes.HR corpus and treebank test sets. Accuracy on part of speech only (POS) and full morphosyntactic tagset (MSD) was observed for tagging, while parsing accuracy is expressed in labeled (LAS) and unlabeled attachment score (UAS) and linear label attachment (LA). The results are given in Table 4.

Tagging and parsing scores on newspaper text from SETimes.HR present the state of the art in Croatian processing. The decrease in POS and MSD tagging accuracy is substantial when moving from newspaper text to transcribed, i.e., standardized public message text: POS tagging accuracy decreases by 13 points and MSD tagging accuracy by almost 17 points. The decrease for non-standard public message text is even more substantial: 21 points for POS and 26 points for MSD tagging. The first set of differences indicates a strong bias of tagging model trained on newspaper text towards texts from the same domain, while the second set of differences emphasizes the inadequacy of standard text language models in natural language processing of non-standard text. These differences are further accentuated by dependency parsing scores – 17 and 38 points lower LAS moving from SETimes.HR to standardized and non-standard Babel supports arguments for domain adaptation and elaborate pre- and post-processing strategies for natural language processing of non-standard text.

| Corpus            | Tagging | Parsing |
|-------------------|---------|---------|
|                   | POS     | MSD     | LAS     | UAS     | LA      |
| Babel             | 75.60   | 60.17   | 39.54   | 47.54   | 54.95   |
| Babel transcribed | 83.79   | 69.53   | 60.22   | 71.32   | 75.56   |
| SETimes.HR        | 96.78   | 86.24   | 77.13   | 83.08   | 88.82   |

4. Conclusions and future work

The process of manual construction of a dependency treebank of public messages written in Croatian – the Babel treebank – is presented in the paper. The treebank is developed on basis of the Babel project of Francesco Jodice and the Museum of Contemporary Art in Zagreb. The work on treebank construction presents a new direction in treebank development for Croatian as a set of syntactic functions is suggested that deviates from the formalism used in Croatian Dependency Treebank in terms of overall simplicity and closer correspondence with contemporary grammars of Croatian. This course of action is also implemented in a new treebank of Croatian newspaper text, the SETimes.HR treebank. A simple experiment in morphosyntactic tagging and dependency parsing of Babel treebank using SETimes.HR language models was conducted, indicating the need of more elaborate language modeling for processing non-standard Croatian text.

As a treebank of non-standard Croatian text, the Babel treebank introduces various courses of future work. The 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 (Žic Fuchs and Tuđman Vuković, 2008) or hrWaC – the Croatian web corpus (Ljubešić and Erjavec, 2011). State of
the art natural language processing tools for non-standard Croatian could be implemented on top of such merged resources. Further insight into processing of non-standard Croatian text is required. Just recently, a large dataset of non-standard Croatian text from blogs, forum posts and newspaper text comments was manually assembled and manually annotated for standardness. Combined with Babel treebank, this dataset could serve as a starting point for implementing more sophisticated methods of non-standard language processing for Croatian.

Acknowledgements

The authors would like to thank Francesco Jodice of the Babel project and Radmila Iva Janković from the Zagreb Museum of Contemporary Art for making a sample of messages from the Babel project available for research purposes.

References

Agić, Ž. and Merkler, D. (2013). Three Syntactic Formalisms for Data-Driven Dependency Parsing of Croatian. Lecture Notes in Artificial Intelligence (Proceedings of TSD 2013), in press.
Agić, Ž., Ljubešić, N. and Merkler, D. (2013). Lemmatization and Morphosyntactic Tagging of Croatian and Serbian. In Proceedings of ACL 2013, to appear.
Agić, Ž. (2012). Pristupi ovisnosnom parsanju hrvatskih tekstova. PhD thesis, University of Zagreb.
Ambati, V. and Chen, W. (2007). Cross-Lingual Syntax Projection for Resource-Poor Languages. CMU.
Berović, D., Agić, Ž. and Tadić, M. (2012). Croatian Dependency Treebank: Recent Development and Initial Experiments. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012) (pp. 1902-1906). ELRA.
Džeroski, S., Erjavec, T., Ledinek, N., Pajas, P., Žabokrtský, Z. and Žele, A. (2006). Towards a Slovene Dependency Treebank. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006). ELRA.
Erjavec, T. (2010). MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010).
Erjavec, T., Fišer, D., Krek, S. and Ledinek, N. (2010). The JOS Linguistically Tagged Corpus of Slovene. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010).
Hajči, J., Böhmová, A., Hajčová, E. and Vidová Hladká, B. (2000). The Prague Dependency Treebank: A Three-Level Annotation Scenario. In Treebanks: Building and Using Parsed Corpora. Kluwer.
Halacsy, P., Kornai, A. and Oravecz, C. (2007). HunPos: An Open Source Trigram Tagger. In Proceedings of ACL 2007 (pp. 209-212).
Ljubešić, N. and Erjavec, T. (2011). hrWaC and slWaC: Compiling Web Corpora for Croatian and Slovene. Text, Speech and Dialogue (pp. 395-402). Springer.
McDonald, R., Lerman, K. and Pereira, F. (2006). Multilingual Dependency Analysis With a Two-Stage Discriminative Parser. In Proceedings of CoNLL 2006.
Silić, J. and Pranjčiković, I. (2005). Gramatika hrvatskoga jezika: za gimnazije i visoka učilišta. Zagreb: Školska knjiga.
Tadić, M. (2005). The Croatian Lemmatization Server. Southern Journal of Linguistics, 29, 206-217. URL http://hml.ffzg.hr/.
Tadić, M. (2007). Building the Croatian Dependency Treebank: the initial stages. Suvremena lingvistika, 63, 85-92.
Tadić, M., Brozović-Rončević, D. and Kapetanović, A. (2011). Languages in the European Information Society: Croatian. META-NET White Paper Series, META-FORUM.
Tadić, M. and Varadi, T. (2012). Central and South-East European Resources in META-SHARE. In Proceedings of COLING 2012: Demonstration Papers (pp. 431-438).
Žic Fuchs, M. and Tuđman Vuković, N. (2008). Communication Technologies and Their Influence on Language: Reshuffling Tenses in Croatian SMS Text Messaging. Jezikoslovije, vol 9, 1-2, 109-122.