Syntactic Annotation of Slovene CMC: First Steps

Špela Arhar Holdt*♦, Darja Fišer*‡, Tomaž Erjavec‡, Simon Krek‡
* Faculty of Arts, University of Ljubljana
Aškerčeva 2, 1000 Ljubljana
♦Institute for Applied Slovene Studies Trojina
Trg republike 3, 1000 Ljubljana
‡ Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana
E-mail: spela.arharholdt@ff.uni-lj.si, darja.fiser@ff.uni-lj.si, tomaz.erjavec@ijs.si, simon.krek@ijs.si

Abstract

This paper presents the first steps towards the syntactic annotation of Slovene CMC, namely the annotation of 200 Slovene tweets with the JOS dependency model. After a presentation of the dataset we present the selected annotation model, the annotation procedure, and results. The focus of the paper is on the decisions regarding the annotation of CMC-specific elements that required special treatment: Twitter-specific features, foreign language elements, ellipsis and fragments, non-standard use of punctuation, and other non-standard language features. The dataset, together with the CMC-adapted annotation guidelines, can be used for further annotation of language data (from Twitter or other CMC genres), and in the second step to train a parser for the selected CMC domain(s). The large-scale corpus-based research of non-standard Slovene syntax, which will be facilitated by the described activities, will help disprove the myths surrounding CMC that are still present in the field of Slovene studies.

Keywords: computer mediated communication, syntactic annotation, JOS dependency model, Slovene language, tweets

1. Introduction

With the advent of digital media and the Internet, communication practices began to change significantly, challenging the traditionally established dichotomies of public vs. private, formal vs. informal, written vs. spoken, and standard vs. non-standard language use. Initially, the linguistic research community observed the new situation with a somewhat reserved attitude, whereas in the last years, more and more studies aim to disprove the myths surrounding computer mediated communication and its possible negative impact on the evolution of language (Crystal, 2011). Since computer-mediated communication is a global phenomenon, work on languages other than English soon followed (Myslin and Gries, 2010; Storrer, 2013; Chanier, 2015).

While studies have been performed on Slovene as well, they mostly focused on orthographic (Jakop, 2008; Arhar Holdt and Dobrovoljc, 2015), lexical (Michelizza, 2015; Zwitter Vitez and Fišer, 2015) and processing issues (Ljubešić et al., 2016a; Ljubešić et al., 2016b) whereas no larger-scale corpus-based work exists on the syntax of Slovene CMC. The goal of this paper is to present the first steps in bridging this gap, the annotation of 200 Slovene tweets with the JOS dependency model (Erjavec et al., 2010), which will serve as the groundwork for syntactic annotation and analysis of Slovene CMC.

2. Dataset

A dataset of 200 tweets (475 sentences) was extracted from the Janes corpus of Slovene CMC (Fišer et al., 2015), sampled to include an equal amount of linguistically and technically standard and non-standard tweets (Ljubešić et al., 2015). The dataset only includes tweets longer than 120 characters published by private individuals. This material was lemmatized and POS-tagged with the tools described in (Erjavec et al., 2005; Ljubešić et al., 2014). In the next step, the sentence segmentation and tokenization was manually corrected, the tweets were normalised on the lexical and morphological level (Čibej et al., 2016a), and finally, the attributed lemmas and POS-tags were manually corrected (Čibej et al., 2016b).

3. The JOS Dependency Model

For the annotation, the JOS dependency model was used. The system, which was designed in the project "Linguistic Annotation of Slovene" (Erjavec et al., 2010), is based on syntactic dependencies. The categories of the system are presented in Table 1.

Groups of labels	Labels	Description
First level labels link elements in different types of phrases (green and yellow colour in the visualisation).	*dol*	Links heads and modifiers in phrases.
	del	Links parts of verbal phrases.
	prir	Links heads in coordinate structures within clauses.
	vez	Links words or commas in conjuctive function.
	skup	Links (function) words in frozen multi-word structures.
Second level labels link sentence elements (red colour in the visualisation).	*ena*	Clause subject.
	dve	Clause object.
	tri	Adverbial of manner.
	štiri	Other adverbials.
Third level label links all other structures (blue colour in the visualisation).	*modra*	Links to the root, punctuation, fragments, etc.

Table 1: The labels in the JOS dependency model. (http://eng.slovenscina.eu/tehnologije/razclenjevalnik)

As the JOS dependency model is based on the principle that the relations inducible from the tags on lower levels (lemmas and POS) are not annotated again on the syntactic level, it is significantly simpler and more robust than similar models, e.g. the Prague Dependency Treebank (Böhmová et al., 2003). The main features of the model are described in Erjavec et al. (2010), and in more detail in the annotation guidelines (Holozan et al., 2008). The model was applied in the "Communication in Slovene (SSJ)" project to annotate the ssj500k training corpus (Krek et al., 2015), on the basis of which a parser for Slovene was trained (Dobrovoljc et al., 2012). Additionally, a specialised program was developed for the visualisation, manual annotation and search of the data (the screenshots on Figures 1 to 4 are from this program, the author of the program is Janez Brank).

4. Annotation and Results

The dataset, described in Section 2, was automatically parsed and imported into the SSJ annotation program. Syntactic annotations were then manually corrected, following the guidelines for the annotation of the jos500k corpus. During annotation, the majority of the problems could be adequately addressed by the existing guidelines, while for some specific questions, the guidelines had to be complemented by additional rules. In the remainder of this paper, we present the decisions regarding the annotation of: Twitter elements; foreign language; syntactical fragments and ellipsis; non-standard use of punctuation; and other non-standard language features. The implement solutions are exemplified in Figures 1 to 4. The examples are in Slovene, with English translation provided in the corresponding figure title.[1]

4.1 Twitter Elements

We considered hashtags, usernames, URLs, and emoticons of two kinds. The elements that were syntactically part of a sentence were annotated in accordance with their function, while function-free elements (typically appearing at the beginning or the end of the tweet) were connected to the node. This decision is in accordance with similar projects (Kong et al., 2012), and the annotation of the dataset indicates that the separation of the two groups is sufficiently straightforward. Figure 1 presents an example, where the first hashtag (*#zooljubljana*) connects to the node, while the second one is annotated as a part of a noun phrase (*a plane to #sochi*).

4.2 Foreign Language

Foreign language elements (primarily from English and related South Slavic languages) appear in Slovene tweets as single words, word phrases, or longer segments/clauses. Different levels of adaptation to Slovene can be observed regarding the spelling and morphology of these elements. The questions about how to lemmatise and POS-tag them (including the question how to separate the ones to be tagged as *foreign* from the ones to be treated as *Slovene*) were addressed at the earlier stages of the project (Čibej et al., 2016b). On the syntactic level, we followed a principle that single words and two-part phrases with a clear dependency relation are attached into the syntactic tree, whereas in longer phrases and segments, all of the foreign elements get attached to the node instead. Figure 2 presents an example of the first type, where the English phrase *personal message* is connected to the tree.

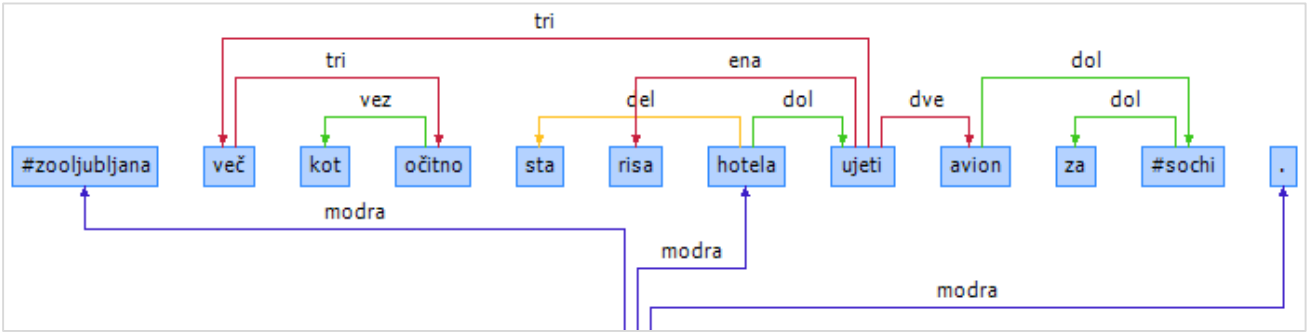

Figure 1: ***#zooljubljana*** *more than obviously the lynx wanted to catch* ***a plane to #sochi****.*

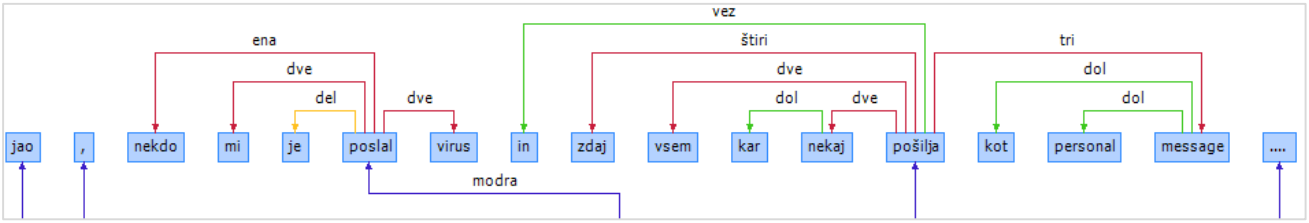

Figure 2: *Jeez, somebody sent me a virus and now it's sending random stuff to everyone as a {****personal message****}.*

[1] The translation is somewhat word-by-word to facilitate comprehension of annotated syntactic relations to non-Slovene speakers, however certain adaptations were obviously required due to language differences.

4.3 Ellipses and Fragments

Tweets are especially challenging to annotate syntactically due to their fragmented nature and a large number of ellipses. One possible solution to this problem is to use a system that allows for the orphan node to be promoted to the place of the missing parent (Dobrovoljc and Nivre, 2016). Another alternative is to use a system that attaches such elements directly to the root (Kong et al., 2012). The JOS dependency model was designed as the latter: while in regular clauses only the head of the predicate is attached to the root (or the relevant ordinate clause), in fragments or clauses without the predicate, each separate phrase head attaches to the node as well (Figure 3).

4.4 Non-standard Use of Punctuation

The annotation of the dataset revealed that lexical and morphological normalisation of tweets and subsequent manual correction of lemmas and POS tags successfully eliminated many of the potential problems for syntactic annotation. However, the non-standard use of punctuation in tweets remains an important factor of negative influence. The existing parser is trained on standard Slovene language, where punctuation – especially the use of the comma – plays an important role in determining the borders between clauses and other types of sentence segments. Omitted, redundant, and misplaced commas thus as a rule lead to parsing mistakes, and with the comma being notoriously difficult to master for Slovene speakers, such instances are frequent. For the annotation of the dataset, the parsing errors were manually corrected, however the findings indicate that it might be beneficial to include a step of punctuation normalisation before the attempts on the syntactical level (some work for Slovene has been presented by Kranjc and Robnik Šikonja, 2015).

4.5 Other Non-standard Language Features

Last but not least, the annotated dataset exhibits a number of other syntactic features that have been previously attributed to non-standard written Slovene (Michelizza, 2015), e.g. atypical word order, non-standard use of conjunctions, cases, grammatical number, high number of demonstrative pronouns and certain particles. A preliminary analysis of the annotated data reveals that 49 % of the (linguistically and technically non-standard) tweets exhibit at least one of the listed features. While it is clear that these phenomena need to be linguistically addressed in the future, they did not pose a problem for the annotation. Figure 4 presents an example, where the predicate *to be similar* is accompanied by two objects in dative (*he is similar to the members of the parliament* and *similar to me* = *to me he seems similar*). While valence in this example is atypical, the annotations are relatively straightforward.

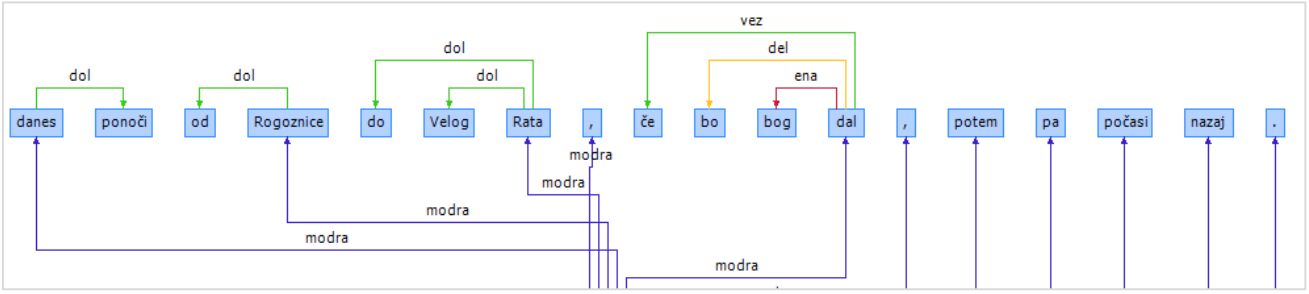

Figure 3: *During the night from Rogoznica to Veli Rat, if god allows it, and then slowly back.*

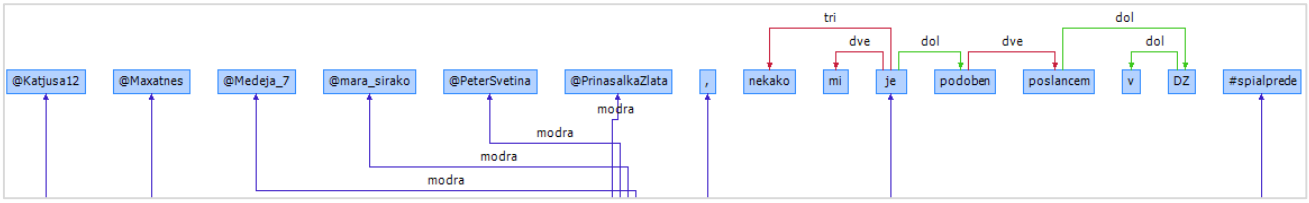

Figure 4: *[…] @PrinasalkaZlata, somehow* ***to me*** *he is similar* ***to the members of the parliament*** *#spialprede.*

5. Conclusion and Further Work

The paper presented the first steps towards a syntactic annotation of Slovene CMC. In this first stage, 200 tweets were annotated with the JOS dependency model and annotator guidelines were supplemented with examples for the annotation of Twitter-specific and non-standard language features. The dataset, together with the guidelines, can be used for further annotation of language data (from Twitter or other CMC genres), and in the second step to train a parser for the selected CMC domain(s).

The JOS dependency model in combination with the SSJ annotation program proved to be adequate for the described task, the main advantages of the system being its robustness and the ability to allow multiple attachments to the root element. A major drawback is that the system is language-specific and as such offers little possibility for cross-lingual comparison. Recent attempts to translate the annotations of the ssj500k corpus to the Universal Dependencies system (Dobrovoljc et al., 2016) suggest a possible solution to this problem in the future.

6. Acknowledgements

The work described in this paper was funded by the Slovenian Research Agency within the national basic research project "Resources, Tools and Methods for the Research of Non-Standard Internet Slovene" (J6-6842, 2014–2017).

7. References

Arhar Holdt, Š. and Dobrovoljc, K. (2015). Zveze samostalnika z nesklonljivim levim prilastkom v korpusih Janes in Kres. In D. Fišer (Ed.), *Zbornik konference Slovenščina na spletu in v novih medijih*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 4–9.

Böhmová, A., Hajič, J., Hajičová, E. and Hladká, B. (2003). The Prague dependency treebank. In *Treebank: Building and Using Parsed Corpora*. Netherlands: Springer, pp. 103–127.

Chanier, T., Poudat, C., Sagot, B., Antoniadis, G., Wigham, C. R., Hriba, L., Longhi, J. and Seddah, D. (2014). The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. *Journal for Language Technology and Computational Linguistics*, 29(2), pp.1-30.

Čibej, J., Fišer, D. and Erjavec, T. (2016a). Normalisation, Tokenisation and Sentence Segmentation of Slovene Tweets. In *Proceedings of the Workshop on Normalisation and Analysis of Social Media Texts (NormSoMe)*. Portorož: ELRA, pp. 5–10.

Čibej, J., Arhar Holdt, Š., Erjavec, T. and Fišer, D. (2016b). Razvoj učne množice za izboljšano označevanje spletnih besedil. In *Proceedings of the Conference on Language Technologies and Digital Humanities.* Ljubljana (in print).

Crystal, D. (2011). *Internet Linguistics: A Student Guide.* London, New York: Routledge.

Dobrovoljc, K. and Nivre, J. (2016). The Universal Dependencies Treebank of Spoken Slovenian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '16).* Portorož, pp. 1566–73.

Dobrovoljc, K., Erjavec, T. and Krek, S. (2016). Pretvorba korpusa ssj500k v Univerzalno odvisnostno drevesnico za slovenščino. In *Proceedings of the Conference on Language Technologies and Digital Humanities.* Ljubljana (in print).

Dobrovoljc, K., Krek, S. and Rupnik, J. (2012). Skladenjski razčlenjevalnik za slovenščino. In T. Erjavec, J. Žganec Gros (Eds.), *Zbornik Osme konference Jezikovne tehnologije.* Ljubljana: Institut Jožef Stefan, pp. 42–47.

Erjavec, T., Fišer, D., Krek, S. and Ledinek, N. (2010). The JOS linguistically tagged corpus of Slovene. In: *LREC 2010, 7th International Conference on Language Resources and Evaluations*. Valletta, pp. 1806–1809.

Erjavec, T., Ignat, C., Pouliquen, B. and Steinberger, R. (2005). Massive multi-lingual corpus compilation: Acquis Communautaire and totale. In *Proceedings of the 2nd Language & Technology Conference*. Poznan, pp. 32–36.

Fišer, D., Ljubešić, N. and Erjavec, T. (2015). The JANES corpus of Slovene user generated content: construction and annotation. In *International Research Days: Social Media and CMC Corpora for the eHumanities: Book of Abstracts*. Rennes,.p. 11.

Holozan, P., Krek, S., Pivec, M., Rigač, S., Rozman, S. and Velušček, A. (2008). *Specifikacije za učni korpus.* Kamnik: Projekt »Sporazumevanje v slovenskem jeziku« ESS in MŠŠ.

Jakop, N. (2008). Pravopis in spletni forumi – kva dogaja? In *Slovenščina med kulturami, Zbornik Slavističnega društva Slovenije 19*, pp. 315–327.

Kranjc, A. and Robnik Šikonja, M. (2015). Postavljanje vejic v slovenščini s pomočjo strojnega učenja in izboljšanega korpusa Šolar. In D. Fišer (Ed.), *Zbornik konference Slovenščina na spletu in v novih medijih,* Ljubljana: Znanstvena založba Filozofske fakultete, pp. 38–43.

Kong, L., Schneider, N., Swayamdipta, S., Bhatia, A., Dyer, C. and Smith, N. A. (2014). A dependency parser for tweets. In *Proc. of EMNLP*. Doha, Qatar, pp. 1001–1012.

Krek, S., Dobrovoljc, K., Erjavec, T., Može, S., Ledinek, N. and Holz, N. (2015). *Training corpus ssj500k 1.4, Slovenian language resource repository CLARIN.SI*, http://hdl.handle.net/11356/1052.

Ljubešić, N., Fišer, D., Erjavec, T., Čibej, J., Marko, D., Pollak, S. and Škrjanec, I. (2015). Predicting the level of text standardness in user-generated content. In *10th International Conference on Recent Advances in Natural Language Processing: Proceedings of RANLP 2015 Conference.* Hissar, pp. 371–378.

Ljubešić, N., Erjavec, T. and Fišer, D. (2014). Standardizing tweets with character-level machine translation. In *Computational linguistics and intelligent text processing: 15th International Conference, CICLing 2014, Kathmandu, Nepal: Proceedings: part II.* Springer, Heidelberg, pp. 164–175.

Michelizza, M. (2015). *Spletna besedila in jezik na spletu.* Založba ZRC, ZRC SAZU, Ljubljana.

Myslin, M. and Gries, S. T. (2010). k dixez? A corpus study of Spanish Internet orthography. *Literacy and Linguistic Computing*, 25(1), pp. 85–104.

Storrer, A. (2013). Sprachverfall durch internetbasierte Kommunikation? Linguistische Erklärungsansätze – empirische Befunde. In *Sprachverfall? Dynamik – Wandel – Variation. Jahrbuch des Instituts für Deutsche Sprache 2013*. De Gruyter Mouton, pp. 171–196.

Zwitter Vitez, A. and Fišer, D. (2015). From mouth to keyboard: the place of non-canonical written and spoken structures in lexicography. *Electronic lexicography in the 21st century: linking lexical data in the digital age: proceedings of eLex 2015 Conference*, *Herstmonceux Castle, UK*. Ljubljana: Trojina, Institute for Applied Slovene Studies; Birmingham: Lexical Computing, pp. 250–267.