Short-term Scientific Mission Report

Host: prof. Joakim Nivre, Uppsala University
Visitor: Kaja Dobrovoljc, University of Ljubljana

1 September to 31 October 2015

1 Purpose of the STSM

The STSM visit at Uppsala University has been carried out as part of my PhD research on spoken multi-word discourse markers in user-generated Slovenian. To achieve a better understanding of the formal and distributional properties of discourse-marking multi-word units in speech, the main objective of the STSM was to build the first treebank of Spoken Slovenian, using the recently proposed Universal Dependencies annotation scheme and adapting it to the syntactic particularities of spoken communication, including the highly frequent multi-word discourse markers.

2 Description of the work accomplished during the STSM

2.1 The Slovenian Universal Dependency Treebank

Given that the Universal Dependencies annotation guidelines had not yet been adopted for Slovenian, the first part of the STSM focused on producing the written *Slovenian Universal Dependency Treebank*. This was done by automatic conversion of the recently developed ssj200k dependency treebank, the largest dependency treebank of Slovenian, annotated according to JOS guidelines that has previously already been converted to comply with guidelines for UD morphology.

Due to the specifics of the JOS annotation scheme that excludes dependency annotation of punctuation, particles, interjections, coordinating clauses, apposition and similar non-arguments of the predicate (attaching them as dependent of root node), the resulting conversion script consists of more than 180 rules including different lexical, morphological and dependency features. As not all previously unattached phenomena could be covered with high level of reliability, the size of first Slovenian UD Treebank is slightly smaller than the original ssj200k treebank.

2.2 The Spoken Slovenian Universal Dependency Treebank

In the second phase of the visit, we focused on construction and dependency annotation of the *Spoken Slovenian Universal Dependency Treebank*. First, we gathered the data by sampling the reference Gos corpus of Spoken Slovenian, a collection of recordings and transcriptions of Slovenian speech in different everyday situations. The first version of the spoken treebank includes approx. 30,000 tokens or 3,200 utterances.

After the construction of the corpus automatically assigned JOS morphosyntactic tags have been manually corrected and converted to UD POS categories and morphological feature-value pairs. The morphologically annotated treebank was then parsed with MaltParser by using the written Slovenian Treebank as the training set. The resulting automatically dependency parsed Spoken Slovenian Treebank has been imported to a dependency annotation tool for subsequent manual corrections. So far, 700 utterances have been manually inspected.

In the process of manual annotation of the Spoken Slovenian Treebank, many particularities of spoken language have been identified, both on the structural level (segmentation, spelling, non-lexical tokens, disfluencies) and pragmatic level (frequent use of discourse markers, parentheticals, predicate ellipsis, etc.). Given the fact these phenomena have not yet been thoroughly addressed in the UD scheme, special guidelines have been drafted, in which we explain the treatment of spoken language features by either extending the application of existing UD labels or adding speech-specific extensions.

2.3 Multi-word discourse markers

In drafting the guidelines used for annotation, a special attention was given to multi-word expressions, functioning as discourse markers, such as *a ne* ('right'), *tako da* ('so'), *to se pravi* ('that is'), *a veš* ('you know') etc., trying to define the initial set of guidelines for their identification and syntactic annotation.

Applying the general DM definition of non-truth-conditional prosodicly distinct, semantically and syntactically detachable expressions to actual data revealed several issues in their syntactic delimitation in relation to other multi-word units, such as modal adverbials, connectives and parentheticals, and the underspecification of the adverbial-connective distinction in the UD scheme in general. This problem was solved by three intermediate solutions: the introduction of bias towards the non-DM interpretation in case of doubt; the introduction of a lexicon, in which possible syntactic relations for each of the potential multi-word expressions are explicitly defined; and the introduction of a new annotation layer, in which potential discourse markers are annotated regardless of their syntactic interpretation.

In terms of their syntactic function, both phrasal and clausal non-ambiguous discourse markers have been labeled as *discourse* and attached to the relevant unit. In terms of their syntactic structure, non-clausal multi-word discourse markers have been treated as fixed multi-word units (e.g. *so that*), while clausal multi-word discourse markers retain their compositional anaylsis (e.g. *you* being the subject in the clause *you know*)

The need for actual distinction of the class of discourse markers on syntactic level (instead of adopting a default lexically-motivated interpretation as adverbials, connectives etc., leaving their identification and categorization to semantic- and discourse-level annotation), will be tested as part of future dependency parsing experiments.

3 Description of main results obtained

The work accomplished within the STSM visit resulted in several important contributions, namely:

- the script for conversion from JOS annotation scheme to UD annotation scheme;
- the Slovenian Universal Dependency Treebank, released as part of the UD treebanks release v1.2;
- the construction of the Spoken Slovenian Treebank;
- the Universal Dependencies annotation guidelines for Spoken Slovenian Treebank;
- the full morphological and the initial syntactic annotation of Spoken Slovenian Treebank;
- the initial lexicon of multi-word discourse markers in spoken Slovenian.

4 Future collaboration with host institution

After the completion of the dependency annotation of the Spoken Slovenian Treebank, we plan to collaborate in running initial experiments in dependency parsing of Spoken Slovenian, to gain feedback on the newly proposed speech-specific extensions of the Universal Dependencies annotation scheme, including multi-word discourse markers.

5 Forseen publications resulting from the STSM

We have submitted a conference paper describing the construction of the Spoken Slovenian Treebank, with a special focus on adaptation of the Universal Dependencies annotation scheme to particularities of spoken (Slovenian) language. A conference paper presenting the annotation and distribution of multi-word discourse markers in particular has been accepted to the LPTS 2016 conference in Valencia.

Given that (i) two important new language resources have been created for Slovenian and (ii) this has been the first application of Universal Dependency scheme to spoken data, other publication of research deriving from results of this STSM is possible.

6 Confirmation by the host institution of the successful execution of the STSM

Joakim Nivre, Uppsala University: Kaja Dobrovoljc's visit has been very fruitful and productive, resulting in a number of significant results as well as a promising plan for future collaboration. During her visit to the computational linguistics group in Uppsala, Kaja has not only done excellent work on the written and spoken treebanks of Slovenian, with a special focus on multiword discourse markers, but has also contributed greatly to our activities and been a valued member of the group. In this way, we have built a platform for future research collaboration that will also extend to the treatment of multiword expressions in spoken language parsing.