STSM SHORT SCIENTIFIC REPORT

COST MC Chair: DR AGATA SAVARY, agata.savary@univ-tours.fr

COST STSM Reference Number: ECOST-STSM-IC1207-060114-042514

Period: 2014-01-06 00:00:00 to 2014-03-15 00:00:00

COST Action: IC1207 PARSEME

STSM Type: Regular (from Slovakia to Czech Republic)

STSM Applicant: Daniela Majchráková, Slovak National Corpus (SNK), Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia, danam@korpus.sk

STSM Topic: Building Lexical Resources: Construction of the Czech-Slovak Valency Lexicon based on the PDT-Vallex

Host: prof Jan Hajič, Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University (CUNI), Prague, Czech Republic, hajic@ufal.mff.cuni.cz

1. Purpose of the STSM

The purpose of the STSM was primarily aimed at building a Slovak valency lexicon using syntactically annotated corpus (dependency treebank). The lexicon was inspired by and based on the Czech PDT-Vallex which is a valency lexicon based on the Prague Dependency Treebank (PDT). In the first stage, the Czech PDT-Vallex was automatically translated into Slovak and the valency frames for Slovak were automatically created based on the Czech ones, obtaining Slovak Valency Lexicon (SVL) in the process. Subsequently, the translations of verbs and their valency frames were manually proofread for accuracy, especially those related to multi-word expressions (MWE). There are two types of MWE: light verbs combinations and verbal phraseological units in the valency lexicon whose valency frames are marked with special functors that helped us to identify and extract them.

2. Description of the work carried out during the STSM

The idea has been suggested by prof Jan Hajič. The existing Slovak Dependency Treebank (SDT) (compatible with PDT with minor differences) contains 50 000 sentences manually annotated at the surface layer of syntactic representation and therefore could serve as a basis for an automatic annotation at the tectogrammatical level (the deep syntactic level in PDT). The tool was primarily developed for the Czech PDT, thus this application includes conversion of the Slovak morphological tagset into a positional variant suitable for automatic tectogrammatical annotation based on existing Czech PDT tools.
The automatic annotation did not include the annotation of the MWEs. Therefore, a list of MWEs candidates extracted from the SVL was further used for the automatic process of identification and annotation of MWEs in the Slovak Dependency Treebank. In the next step, we evaluated the results and accuracy rate of the chosen method.

The given tasks were successfully accomplished. Special thanks go to Mária Šimková

(SNK) for the work on the Slovak Treebank, Radovan Garabík (SNK) for the automatic translation of PDT-Vallex, Eduard Bejček (CUNI) for creating the treebank repository, Agáta Karčová (SNK) together with Radoslav Brída (SNK) for preparing syntactic data for automatic deep-layer annotation and Ondřej Dušek (CUNI) for preparing a tectogrammatical level of the syntactic corpus.

3. Description of the main results obtained

The Slovak Valency Lexicon is considered the first successful outcome. The lexicon contains more than 700 valency frames. The manual proofreading of the automatic translation in the lexicon and contrastive analyses of equivalent Czech and Slovak MWEs proved that there was a huge overlap of light verbs constructions in Czech and Slovak. On the other hand, the results demonstrate diversity of the lexical and syntactic representation of identical semantic content from a sizeable number of phraseological units.
The lexicon might be further used for the purpose of contrastive analysis of syntactic and semantic properties of both languages. The list of MWEs can be used for creating various syntactic patterns of MWEs and for automatic verification of the forthcoming Lexicon of Slovak Verbal Collocations.

Another outcome of the mission is an application of the list of Slovak MWEs in automatic identification and annotation of MWEs and subsequent semi-automatic and manual validation of the results. Evaluation of the accuracy of the automatic annotation is still in process.

4. Future collaboration with host institution

In joint collaboration with the host institution we plan to work on extending features for automatic identification of MWEs in the syntactic corpus.

5. Foreseen publications/articles resulting or to result from the STSM

There are plans to present and publish the results of the project at an international scientific conference.

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

The Host institution, Institute of Formal and Applied Linguistics of the Faculty of Mathematics and Physics, Charles University in Prague, confirms that the STSM took place as planned, and that the results have been obtained as described above. We are also looking forward to future cooperation with Daniela Majchráková and the SNK team.