STSM SHORT SCIENTIFIC REPORT

COST MC Chair: Agata Savary, agata.savary@univ.tours.fr

COST STSM Reference Number: COST-STSM-IC1207-15258

PERIOD: 10.06.2014 – 7.07.2014

COST Action: IC1207 PARSEME

STSM Type: Regular (from Poland to France)

STSM Applicant: Monika Czerepowicka, University of Warmia and Mazury in Olsztyn, czerepowicka@gmail.com

STSM Topic: SEJF development

Host: Agata Savary, Université François Rabelais de Tours, IUT de Blois

1. Purpose of the STSM

The aim of the STSM is development of SEJF, which is a grammatical dictionary of Polish multi-word units (MWE).

SEJF, as an electronic dictionary, uses Toposław, which is a Java application for creating inflectional dictionaries of multi-word units with a user-friendly interface for a lexicographer. The inflection in Toposław is modeled using Multiflex – a graph-based cross-language morphosyntactic generator of multi-word units. All inflected forms and their variants are presented in one graph. Toposław uses a Unitex graph editor. To obtain a solid grammatical description of Polish MWE it is necessary to use a specialized morphological analyzer of Polish. Toposław uses an analyzer called Morfeusz, which seems to be the most effective analyzer for Polish, with its database comprising more than 4 million inflected forms.

SEJF is strongly connected with the National Corpus of Polish (NKJP). Both SEJF and NKJP use the same morphosyntactic tagset of flexemes. Thus SEJF could be used for parsing MWE in corpus.

SEJF contains mainly nouns but there are also adjectives, adverbs and compound conjuctions. The most important purpose of the mission is to create an inflectional description of verbal multi-word units in the dictionary. The latest research shows it is a complex problem which needs a special treatment to find a way to provide an acceptable algorythmization of the description.

2. Description of the work carried out during the STSM

During the STSM we diagnosed the following problems connected with describing verbal MWEs:

- notation of a paradigm's completeness;
- encoding of subordinate phrases (if any and/or set of the phrases);
- discontinuity of verbal MWEs in Polish.

We made a lot of linguistic and technical experiments to code multi-word verbs using, firstly, various graph-editor capabilities and, secondly, different versions of the morphological analyzer Morfeusz. During our research we discovered it is absolutely

necessary to redefine morphological classes and categories in the dictionary. The morphosyntactic tagset from the National Corpus of Polish used in SEJF turned out to be insufficient to generate paradigms of Polish verbs. Therefore it is proposed to use units from a level higher than the morphosyntactic level of linguistic description in the dictionary. Having attempted to prepare a new classification we decided to use the classification of 'syntactic words' used in the National Corpus of Polish. Thus SEJF still fits well with the corpus research.

Creating new sophisticated graphs allows us to describe the problem of discontinuity of phraseological verbs. By many attempts it is possible to code groups of verbal forms and join them as units of a higher level of the linguistic description. In a current version of Toposław separate entries are connected to each other through a system of relations. Relations are also used to show a similarity of meaning in verbal phraseological lexemes.

Details of the SEJF development have been described in a separate paper by Czerepowicka and Savary in Polish.

During the STMS we also discuss a classification of MWEs by Baldwin and Kim (2010) and its implementation with reference to the Polish language. We made an attempt to fill a syntactic classification of MWE in the Wiki profile initiated by Manfred Sailer. Complementing the classification is part of activities of the Working Group no. 1 defined on the 2nd General Meeting of the PARSEME action in Athens, March 2014.

3. Description of the main results obtained

Thanks to the activities taken as part of the mission SEJF has developed considerably. Toposław allows one to code almost all flexemes of verbal MWEs in the dictionary. Only some forms, such as agglutinative of past tense or conditional forms, need special treatment to find a way to provide an acceptable algorythmization of the description. This means that Toposław software needs to be developed further. Experiments undertaken during the mission tend to show that there is also a need to use separate software to generate all inflected forms of MWE verbs with their variants.

Another result of the mission is complementing the syntactic classification of Polish nominal MWEs at the WIKI profile of PARSEME Working Group no.1 (http://wiki.studiumdigitale.uni-frankfurt.de/FB10_Parseme/index.php/MWE_Template:_Polish)

4. Future collaboration with the host institution (if applicable)

In joint collaboration with Agata Savary we plan to work on the classification of Polish MWEs as well as development of Toposław or creating a new tool combining both morphological and syntactic level of description.

5. Foreseen publication/articles resulting from the STSM (if applicable)

We plan to present and publish our proposal of the Polish MWEs classification and current work on SEJF development on an international linguistic conference.

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

Agata Savary: I consider the STSM by Monika Czerepowicka very fruitful for both the invited and the inviting party. We could carry on intensive discussions, on a daily basis, concerning the extension of our previous collaboration on the SEJF dictionary (electronic dictionary of Polish contiguous multi-word units) to verbal MWEs. We thoroughly studied the underlying morphological model of Polish, particularly complex in Polish verbs. We also interacted with the Language Engineering Group from the Polish Academy of Sciences in Warsaw in view of converging the efforts of our three institutions towards a future common project on a large-scale description of Polish MWEs at the level of both valency and inflection paradigms. This would bring substantial progress to the modeling and processing of MWEs in Polish but also potentially in other richly-inflected languages.

Another outcome of this STSM directly connected to the common work in PARSEME WG1 was the first approach to the MWEs classification in Polish. It made us realize that such a classification should be done in several dimensions, including semantic compositionality, syntactic idiosyncrasy, morphological irregularity, grammatical class, and syntactic structure. A classification of this kind has been initiated according to the patterns proposed by WG1. It will be further enhanced within PARSEME.

7. Other comments (if any)

The mission proceeded in a very good atmosphere, both scientific and social. Most time was spent solving problems through a deep discussion. The host institution was prepared very well to the mission, it even ensured the invited party an individual place to work at the University.

Olsztyn, 5.08.2014

Monika Czerepowicka