STSM Scientific Report

Agata SAVARY

University of Warmia and Mazury in Olsztyn, Poland, 15-24 July 2015

1 Purpose of the STSM

The purpose of this STSM was twofold :

- to reinforce a previous collaboration on the lexical description of Polish Multi-Word Expressions (MWEs), and in particular to take further steps towards the paradigmatic description of verbal MWEs ;
- to carry on a classification of Polish MWEs within the multilingual template set up by Working Group 1 in PARSEME.

2 Work carried out

The first part of the work consisted in completing the design of a framework for a lexical paradigmatic description of Polish verbal MWEs, as initiated in the previous STSM with the hosting person Monika Czerepowicka during her stay in Blois in 2014.

A paradigmatic description of verbal MWEs is the one in which we are primarily interested in the morphological constraints that the head verb is subject to when occurring within a MWE rather than as a stand-alone verb. For instance, *uciąć komuś głowę* 'to cut one's head' is an idiom meaning to punish somebody severely or to strongly express one's disagreement with someone else's behavior, as in example (1). While the head verb *uciąć* 'to cut$_{perfective}$' admits all inflected forms in past and future tense (perfective verbs have no present forms) and an aspectual pair *ucinać* 'to cut$_{imperfective}$', the same verb allows no past tense and no aspectual pair when occurring within the MWE, as shown in examples (2) and (3), respectively.

(1) Jeśli nie wrócę przed jedenastą, utnie mi głowę.
If not come-back.PERS1.FUT before eleven, cut.PERS3.FUT.PERF me head.
(lit.) If I do not come back before eleven, she will cut my head.
'If I do not come back before eleven, she will be extremely angry with me.'

(2) *Gdy nie wróciłam przed jedenastą, ucięła mi głowę.
When not come-back.PERS1.FEM.PAST before eleven, cut.PERS3.FEM.PAST.PERF me head.
(lit.) When I did not come back before eleven, she cut my head.

(3) *Gdy nie wracam przed jedenastą, ucina mi głowę.
When not come-back.PERS1.PRES before eleven, cut.PERS3.PRES.IMPERF me head.
(lit.) When I do not come back before eleven, she cuts my head.

A full-fledged lexical description of verbal MWEs requires a formalism whose expressive power is close to the one of deep grammars (e.g. HPSG, LFG, TAG, etc.) since verbal expressions often admit non-contiguities, open slots with unconstrained syntactic groups, and a large range of syntactic transformations (passivization, nominalization, pronominalization, etc.). If however the description focuses mainly on paradigmatic constraints, less expressive, simpler formalisms can be used. Here, we adapt an existing formalism (Multiflex) and tool (Toposław) used previously for the of creation SEJF, an electronic dictionary of Polish nominal, adjectival and adverbial MWEs. While the core engine (Multiflex graphs, as well as morphological analysis and generation of single words by the state-of-the-art Polish tool Morfeusz) remains the same, new facilities are needed in Toposław due to a complex Polish morphology, manifested especially in verbs. Namely, the tagset for Polish, used in Morfeusz and in Multiflex, admits that elementary lexical entries are 'flexemes', i.e. sets of forms that share the same morphosyntactic characteristics. Thus, e.g., finite present forms of a verb belong to a different flexeme than its past forms since the latter but not the former inflect for gender. Traditional lexemes (such as verbs or nouns) are morphosyntactically heterogeneous, and can only be obtained by grouping together sets of flexemes. Precisely this operation of abstracting over flexemes in order to obtain a (less labor-intensive and more user-friendly) lexeme-oriented representation is needed in Toposław in order to adapt it to the paradigmatic description of verbal MWEs. Additionally, technical efforts are needed in order to integrate the new, more expressive and lexically richer, version of Morfeusz both in Multiflex and in Toposław. A large part of the STSM was dedicated to discussing these issues from both theoretical and technical points of view, as well as to programming efforts. Meetings with the main Toposław developer, Piotr Sikora, as well as with the local linguistic team (users of the future Topoław version) were made possible due to this STSM.

The second part of the STSM program consisted in largely enriching the classification of Polish MWEs within the multilingual template[1] set up by Working Group 1 in PARSEME. A rich classification (with both generic and Polish-specific constructions) was proposed and documented with numerous glossed examples of different syntactic types.

Finally, we have set up a draft of a common publication to be considered as a reference paper for SEJF, the previously created electronic dictionary of Polish nominal, adjectival and adverbial MWEs.

3 Main results obtained

The main outcomes of the STSM include :

- En enriched classification of Polish MWEs within the multilingual template set up by Working Group 1 in PARSEME. It now contains dozens of syntactic types, as well as glossed and translated examples. Some Polish-specific aspects of fixedness/flexibility were also added.
- Specifications for a new version of Toposław adapted to paradigmatic description of Polish MWEs.
- A C++-to-C wrapper for the new Morfeusz generator, allowing it's future integration in Multiflex.
- Specifications for the enhancements needed in the new Multiflex interface with Morfeusz.
- First draft of a reference paper on SEJF, the previously created electronic dictionary of Polish nominal, adjectival and adverbial MWEs.

[1] http://wiki.studiumdigitale.uni-frankfurt.de/FB10_Parseme/

4 Future collaboration with host institution

The collaboration between the University of Tours and the University of Warmia and Mazury will continue quite naturally since Agata Savary is the main developer of Multiflex, and Monika Czerepowicka and her colleagues are now among the most active users of this tool. Moreover, other close synergies around Polish MWEs within PARSEME exist between the two universities.

5 Foreseen publications to result from the STSM

The submission of a common publication on SEJF (the previously created electronic dictionary of Polish nominal, adjectival and adverbial MWEs) is planned either for the LTC conference in Poznań in 2015 or to LREC 2016 in Slovenia.

Other future publications will concern the paradigmatic description of verbal MWEs within Toposław.

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

I consider that the STSM by Agata Savary was very fruitful for both parties - the invited and the inviting one. We could carry on the extensions of Toposław which is an application for creating electronic dictionaries of Polish MWEs. Toposław was successfully used in SEJF - an electronic dictionary of Polish nominal MWEs but to using it to coding verbal units it needs complex programming works. Thanks to meeting with Toposław's author, Piotr Sikora, during the STSM some of them were carried out and it makes an important step to create a new version of the application. Another outcome of this STSM, directly connected to the common work in PARSEME Working Group 1, was filling up the MWE classification in Polish (PARSEME Wiki). We completed the description of MWEs from syntactic point of view and started working on the part of template dedicated to fixedness/flexibility of MWEs. (Monika Czerepowicka)

7 Other comments

The paradigmatic description of Polish verbal MWEs is funded mainly by the Polish ministry of research via VERBEL, a national Polish project, spin-off of PARSEME.