**Scientific Report of Short Term Scientific Mission**

**COST STSM Reference Number:** COST-STSM-IS1305- 240417-085352

**Period:** 24-04-2017 to 28-04-2017

**Duration**: 5 working days.

**COST Action:** IS1305

**STSM type:** Regular (from The Netherlands to United Kingdom)

**STSM Title**: Exploring methodology and editing tools for Corpus Pattern Analysis (CPA) in Dutch.

**Guest/STSM applicant**: Lut Colman, Instituut voor de Nederlandse Taal (INT)

**Host**: Dr Sara Moze and Prof. Dr. Patrick Hanks, Research Institute of Information and Language Processing (RIILP), University of Wolverhampton. Moze, Sara S.Moze@wlv.ac.uk, Hanks, Patrick patrick.w.hanks@gmail.com

**1. Purpose of the STSM**

The purpose of the STSM was to get acquainted with the lexicographic method of corpus pattern analysis (CPA) as it is used in A Pattern Dictionary of English Verbs (PDEV). We also wanted to know if the CPA pattern and entry manager and editor are suitable dictionary wrinting systems for the INT-project 'Duch Verb Patterns'.

Another purpose was to get acquainted with the semantic ontology used in PDEV. The PDEV uses an ontology with Semantic Types (categories of nouns that share particular semantic features), like [[Human]], [[Action]], [[Emotion]], [[Location]]. Each argument slot in a pattern is populated by a Semantic Type (ST) instead of formal phrase types like Noun Phrase (NP) or thematic roles as they are used in generative parse trees with semantic role labelling, like Agent, Experiencer, etc. Thematic roles lack semantic accuracy. For example, an experiencer may be a human or an animal, which is not accounted for in the parse trees with thematic roles.

**2. Description of the work carried out during the STSM**

Day 1:

Sara Moze introduced me to the CPA editor, the pattern annotation procedure and the ontology. Then I tried to edit a verb myself: *misinterpret*. In the beginning I had some technical problems as the editing and annotating procedures are to be followed very strictly. Patterns have to be opened in the editor before one starts annotating the pattern in the concordances. If one begins with adding a pattern number in the concordance page instead, this will result in a bug. Tracing patterns requires a lot of skill and experience. As a beginner in CPA I will probably have overlooked some important issues. Assigning Semantic Types to the argument slots is a demanding task. The ontology is fine-grained, so sometimes it is difficult to assign the appropriate ST to a set of lexical collocates. More practice in using the ontology will certainly help to assign STs that occur frequently in patterns.

Day 2:

Professor Patrick Hanks discussed some interesting issues with me about corpus pattern analysis and meaning. We talked about the lightness of 'meaning' and prototypes in phraseology. In the PDEV, prototypical phraseology is mapped onto prototypical meanings. Verbs only have meaning potential; meaning is assigned through context. Patrick and Sara have written some proposals for a semantic parsing project in which they want to develop technology for assigning semantics to words in context automatically. Large sets of collocates are needed for a machine to learn to assign the most likely ST of a word in an analogical context.

For Dutch Verb Patterns I have a similar interest in sets of collocates. Each slot in a pattern that is assigned a ST should also reveal the prototypical collocates that represent that ST. Having the same interests, we exchanged ideas on a possible joint venture for a Dutch – English phraseological subproject. Funding will need to be found to develop the technology from which both The PDEV and the Dutch project (and possibly other pattern dictionaries) could benefit.

 For the rest of the day I worked on the verb *overhear*.

Day 3:

I spent the day being briefed on procedures in corpus pattern analysis by Prof. Hanks and examining some problematic issues, for example regular alternations of Semantic Types. For example, we talked about the regular alternation between [[human]] and [[institution]] for speech act verbs and the regular alternation between [[ Human]] and [[ Animal]] for certain kinds of action.

I continued the work on *overhear*.

Day 4:

Today I introduced my own project in a presentation: Dutch Verb Patterns Online. After the presentation, there was interesting feedback from the public. Professor Hanks and I decided that it would be very useful to compare the Dutch verb entry list with English equivalent verbs. We would very much like to compare the patterns of these verbs between languages, because the choice of a lexical item in, for example, a translation strongly depends on the pattern in which the translated lexical item will be used. The outcome of this research could certainly give us more insight in the problems of and possible solutions for better machine translation, better bilingual dictionaries and learners' dictionaries.

We negotiated an Erasmus + mobility scheme agreement with Prof. Mitkov, which will allow us to exchange visits between our institutions in the future.

Day 5:

The last day I spent editing some more verbs.

**Results**

It was very inspiring to work with Dr. Sara Moze and Prof. Dr. Patrick Hanks. I had some technical problems with the editing system at first, but this also gave Patrick and Sara more insight in the sensitiveness of the editor. If the system would be more robust, it would less often suffer from bugs. All fields and functionalities for use in a Dutch project are there. Some adjustments will have to be made for Dutch use, but the editor is certainly an interesting option to consider. Developing a similar editor for Dutch from scratch would be an expensive and time-consuming operation. Since both projects have similar interests, it is necessary that we find funding to optimize the tools for CPA and to do more research on semi-automated semantic parsing.

I would very much like to continue the collaboration with Prof. Hanks and Dr. Moze, because I think we would make a good team to do great and important things for the future of lexicography.

Verbs edited:

Misinterpret

Overhear

**3. References**

Hanks, P. (2013). Lexical Analysis: Norms and Exploitations. MIT Press.

Hanks, P. (2012). ‘How people use words to make meanings: Semantic types meet valencies’. In A. Boulton and J. Thomas (eds.) Input, Process and Product: Developments in Teaching and Language Corpora. Masaryk University Press.