Short Term Scientific Mission COST-STSM-IS1305-37407

Host Report

**COST STSM Reference Number:** COST-STSM-IS1305-37407

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

**Duration**: 5 days.

**COST Action:** IS1305

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

**STSM Title:** Exploring methodology and editing tools for Corpus Pattern Analysis (CPA) in Dutch **Guest/STSM applicant:** Ms Lut Colman, Instituut voor de Nederlandse Taal, Leiden (INT)
**Host:** Sara Moze, Research Institute of Information and Language Processing (RIILP), University of Wolverhampton, Wolverhampton (UK)

The host institution (RIILP, University of Wolverhampton) has an international reputation for research in computational linguistics and lexicography. On the lexicographical side, it has developed the methodology of ‘Corpus Pattern Analysis’ (CPA), which has yielded several academic papers, a monograph, conference presentations, and an online Pattern Dictionary of English Verbs (PDEV).

Ms Colman worked with Prof. Hanks and Dr. Moze on several English verbs and was given ‘on-the-job’ training. She had already got a good understanding of the principles of CPA, so there was a harmonious meeting of the minds. However, much of Ms Colman’s time was taken up with testing the lexicographic software (the PDEV Editor, the corpus tagging system, and the links between them). This testing was very beneficial to our team, as it demonstrated that although the software works well under good conditions, it is not robust. Therefore, we shall instruct our subcontractor to make necessary adjustments when funding allows.

Ms Colman explored the CPA shallow ontology of semantic types used in PDEV and established that it is suitable for the Dutch project, with certain adjustments. We tentatively agreed that on a future visit Ms Colman will compare the ANW ontology (noun categories) with the CPA ontology, which we believe would lead to mutually beneficial results. Prof. Hanks emphasized that it would be a mistake to overinterpret the names of the semantic types, as they are no more than addresses for storing lexical sets. He explained that a future project with the University of Lancaster, if funded, will use machine learning techniques interactively with lexicographical expertise in order to build such lexical sets. Ms Colman agreed that this would be beneficial for her Dutch pilot project.

Ms Colman gave a presentation on behalf of INT, explaining its role in researching the Dutch language, both historically (the WNT dictionary) and for future research, especially with an emphasis on phraseology and collocations. Her talk was very well received by the audience.

From the point of view of RIILP, future collaboration with Ms Colman and INT would be very welcome if funding permits.

**Dr. Sara Moze**

Research Associate in Lexicography

Editorial Assistant for the Cambridge Journal *Natural Language Engineering*

Research Group in Computational Linguistics

Research Institute of Information and Language Processing

University of Wolverhampton

MC133

Stafford Street

WOLVERHAMPTON

WV1 1LY

*Tel: +441902 322 409*