Host Report Short Term Scientific Mission IS1305-26615

COST STSM Reference Number: COST-STSM-IS1305- 26615

Period: 20-04-2015 to 24-04-2015

COST Action: IS1305

STSM type: Regular (from FINLAND to ESTONIA)

STSM Title: Development of Sketch Grammar and GDEX (Good Dictionary Example) for Finnish

Guest/STSM applicant: Tarja Heinonen, Institute for the Languages of Finland

Duration: 5 working days.

Report:

The Institute of the Estonian Language (Tallinn, Estonia) hosted Tarja Heinonen from the Institute for the Languages of Finland (Helsinki, Finland) for the period from the 20th of April to the 24th of April 2015. During the Short Term Scientific Mission it was planned that Tarja Heinonen will learn more about the methods and tools that lexicographers in Tallinn have been using in producing corpus-based dictionaries. The aim of the STSM was to learn about the corpus query tool Sketch Engine (Kilgarriff et al., 2004), to write Sketch Grammar for Finnish language and to develop GDEX classifiers for Finnish Language. Finnish Word Sketch Grammar was planned to be geared towards the specification of the fiTenTen [2014] corpus (available at https://the.sketchengine.co.uk/auth/corpora/).

The work plan for the visit was as follows:

Day 1: Sketch Engine functions Word Sketch and GDEX. Working with reference list on the theme of Sketch Grammar writing and GDEX development.

Day 2: Finnish corpora in Sketch Engine. Tagset analysis. Specifying grammatical relations for Finnish Word Sketches.

Day 3: Writing Sketch Grammar. Basic rules. Preliminary evaluption.

Day 4: Analysis of DGEX configuration for Estonian. What should be investigated in order to write GDEX scriptfor Finnish.

Day 5: Defining parameters for Finnish GDEX and writing a preliminary GDEX script for Finnish.

As a result of the STSM Sketch grammar for Finnish was developed (it contains 70 rules) and basic parameters for GDEX were developed as well. The results are presented in the appendices of scientific report (Appendix 1: Finnish Sketch Grammar; Appendix 2: GDEX configuration for Finnish). Sketch Grammar was tested on the 41 million word grammar development corpus (access was provided by Jan Michelfeit and Miloš Jakubíček from Lexical Computing Ltd).