**Host Report Short Term Scientific Mission COST-STSM-IS1305-34353**

**COST STSM Reference Number**: COST-STSM-IS1305-34353

**Period**: 22-09-2016 to 28-09-2016

**COST Action**: IS1305

**STSM type**: Regular (from Italy to Slovenia)

**STSM Title**: NLP4CMC & Lexicography

**Guest/STSM applicant**: Egon Stemle, EURAC, Bolzano/Bozen, Italy

**Host**: Simon Krek, Jožef Stefan Institure, Ljubljana, Slovenia, simon.krek@ijs.si

**Duration**: 5 working days.

Report:

The main goal of the STSM was to adapt and apply tools and methods for creating innovative e-dictionaries to CMC (Computer-Mediated Communication) data. Originally, the workflow used in this STSM was defined for the purposes of the ENeL Training School 2016, organised by the STSM host. It consisted of three tasks: collecting and cleaning textual data, automatically extracting data, and editing and publishing extracted data in online dictionary writing systems or other publishing platforms. In the context of STSM these tasks were adapted: collecting data from CMC and social media, automatically extracting data, manually checking the resulting data as an editing step.

For the first step existing tools were used (STyrLogism2 tool, Heritrix4, COW etc.). Automatic data extraction was performed by comparing lists of words. Finally, the candidates were manually checked. The experiment showed that the overall number of novel words exceeds expectaions expecially if we take into account that only five web sites were used to collect data. This shows the need for an efficient tool for neologism detection adapted for lexicography which can be performed on a large-scale using web crawling and data processing. There is still a lot of room for optimising these tools for precision, usability and speed. The experiment also showed the need for the tools to be adapted for and tested on other languages.