COST Office Science Officer: Dr Luule Mizera, Luule.Mizera@cost.eu
COST MC Chair: Prof Martin Everaert, Martin.Everaert@inl.nl

COST STSM Reference Number: COST-STSM-IS1305-34353
Period: 2016-09-22 to 2016-09-28
COST Action: IS1305
STSM type: Regular (from Italy to Slovenia)

STSM Applicant: Mr egon stemle, The European Academy of Bozen/Bolzano (EURAC), Bolzano (IT),
egon.stemle@eurac.edu
STSM Topic: NLP4CMC & Lexicography
Host: Simon Krek, Centre for Language Resources and Technologies, Ljubljana (SI), simon.krek@guest.arnes.si

Budget Request: Year-2016

Travel	190 Euro
Subsistence (hotel/meals)	350 Euro
Total	540 Euro

Short CV:

I am a Researcher at EURAC's Institute for Specialised Communication and Multilingualism in Bolzano, Italy. As a cognitive scientist I am comfortable in environments with partners from the humanities and formal sciences where I contribute with my research focus in the area where computational linguistics (CL) and artificial intelligence (AI) converge. I work on computer aided fabrication of ontologies from large document repositories, the technological feasibility thereof and the utilization of cross-linked structured data in applications, as well as on tools for editing, processing, and annotating linguistic data.

I have worked on collecting and cleaning web scale corpora of different language varieties for general and specialised use, as well as on collecting CMC corpora. For the bottom-up analyses of these corpora I employ methods from CL and AI in cooperation with colleagues from linguistics, lexicography and terminology. For the processing of these corpora I actively participate in adjusting and designing CL tools to process non-standard (far beyond newspaper texts) data.

I was a trainer at the past WG3 training school for "Tools and methods for creating innovative e-dictionaries", I am re-elected secretary of the ACL's Special Interest Group on Web as Corpus, invited speaker, organising and programme committee member of conferences and workshops for web and CMC corpus topics.

Work Plan Summary:

ENeL's WG3 concerns innovative e-dictionaries with a focus on the development of digitally born dictionaries, and the past training school introduced participants, among others, to collecting, analysing, and automatically extracting data from web corpora. We know that "new vocabulary is characteristic for CMC discourse, e.g. ‘funzen’ (an abbreviated variant of the German verb ‘funktionieren’, en.: ‘to function’) or ‘gruscheln’ (verb denoting a function of a German social network platform, most likely a blending of ‘grüßen’, en.: ‘to greet’ and ‘kuscheln’, en.: ‘to cuddle’)" [1] and so the goal of this STSM is to apply these methods and tools to CMC data - a related but separate task which has been deliberately excluded from the training school's programme. Also, at the time of this STSM we will have looked into the interoperability of lexicography with language technology (mid September's WG3 meeting in Brno is dedicated to this topic) but from the submissions for presentations at the meeting we see that no one will address the interoperability of lexicography with language technology for CMC data; this STSM will partly remedy the situation and we will be able to supplement the meeting notes for the following dissemination.

[1] Michael Beißwenger, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer, and Angelika Storrer; DeRiK: A German reference corpus of computer-mediated communication; Lit Linguist Computing 2013 28: 531-537.

I request the approval of a COST Short Term Scientific Mission as described above

Applicant:
Mr egon stemle 31 May 2016

COST is supported by
the EU Framework Programme
Horizon 2020