Scientific Report of Short Term Scientific M

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

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

Duration: 5 working days.

COST Action: IS1305

STSM type: Regular (from FINLAND to ESTONIA)

STSM Title: Development of Sketch Grammar and GDEX (Good Dictionary E

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

Host: Jelena Kallas, Institute of the Estonian Language, jelena.kallas@eki.ee

1. Purpose of the STSM

The aim of the STSM was to learn about the corpus query tool Sketch Engin and to write grammatical rules that operate on a morphologically preanalyz full potential of the Sketch Engine can be put to use. The goal was to produ sketches, which are one-page, corpus-derived summaries of a word's gramm behaviour (ibid.) for a representative selection of Finnish headwords and co the current entries in the Dictionary of Contemporary Finnish. The ultimate new and more efficient ways to update the dictionary. An additional purpos outline preliminary good example sentence specifications for the automatic (Kilgarriff et al., 2008; Kosem et al., 2011, Kosem et al., 2013). The idea beh a lexicographer's work more efficient by suggesting a small number of exam lexicographer to choose from, instead of full concordances. At the end of th opportunity to participate in the EAAL (Estonian Association of Applied Ling

Finnish Word Sketch Grammar is geared towards the specification of the
(available at https://the.sketchengine.co.uk/auth/corpora/).

Day 3: Grammar writing, Word Sketches and other functionalities of the Ske
Finnish and Estonian Word Sketches.

Day 4: Preliminary introduction to GDEX with an quick overview on how the
for the Estonian Collocation Dictionary (Kallas &Tuulik, 2014). We also discu
with Iztok Kosem. On top of that, I learned about a forthcoming dictionary o
Estonian from chief editor Margit Langemets. The Estonian dictionary will b
corpus-based while ours on Finnish is going through only minor updates bas
sources.

Day 5: Defining parameters for Finnish GDEX and writing a preliminary GDE
the basis of the one used for the Estonian Collocation Dictionary. Changes w
conditions on word frequencies and the lists of the words to be excluded. T
script would be the next step. I also continued writing rules on grammatica

During my stay in Tallinn I had a chance to participate in the EAAL conferen
(invited speakers Simon Krek and Iztok Kosem). In this context, on day 4, I g
the conference audience about the updating process of the Dictionary of Co
what kind of methods and sources we use.

As a result of the STSM, Finnish Sketch Grammar with about 70 rules was de
parameters for GDEX were set as well. The results are presented in the app
Finnish Sketch Grammar; Appendix 2: GDEX configuration for Finnish).

3. References

Kallas, J. (2013). Eesti keele sisusõnade süntagmaatilised suhted korpus
[*Syntagmatic relationships of Estonian content words in co
lexicography.] Tallinn: Tallinn University. Dissertations on Humanitie
http://e-ait.tlulib.ee/303/1/kallas_jelena.pdf.

toc/euralex_2008/.

Kilgarriff, A. (2013). Using Corpora and the Web as Data Sources for Dictio
The Bloomsbury Companion to Lexicography. Bloomsbury, London. Ch

Kosem, I.; Husák, M. & McCarthy, D. (2011). GDEX for Slovene. In *Proceedin*
159. Available at: http://elex2011.trojina.si/Vsebine/proceedings/eLe

Kosem, I., Gantar, P. & Krek, S. (2013). Automation of lexicographic work:
lexicographers and crowd-sourcing. In: I. Kosem, J. Kallas, P. Gantar,
M. Tuulik (eds.) *Electronic lexicography in the 21st century: thin*
Proceedings of the eLex 2013 conference, 17–19 October 2013, Ta
Available at: http://eki.ee/elex2013/proceedings/eLex2013_03_Kose