**Scientific Report of Short Term Scientific Mission**

**COST STSM Reference Number:** COST-STSM-IS1305-33190

**Period:** 12-06-2016 to 12-07-2016

**Duration**: 30 days (21 working days)

**COST Action:** IS1305

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

**STSM Title**: Automatic extraction of Good Dictionary Examples for Estonian learners dictionaries

**Guest/STSM applicant**: Kristina Koppel, Institute of the Estonian Language / University of Tartu

**Host**: Dr. Iztok Kosem, University of Ljubljana, iztok.kosem@trojina.si

**1. Purpose of the STSM**

The initial Estonian GDEX configuration was developed in 2014 in the Institute of the Estonian Language and in collaboration with Lexical Computing Ltd. We used GDEX in the automatic extraction of the Estonian Collocations Dictionary (ECD) (2014–2018), aimed at language learners at the upper intermediate and advanced levels. We extracted 5 example sentences for each collocation and our goal was to choose the best one out of the 5 and use it as an example sentence. Once the database was compiled, we started editing the entries and it turned out that Estonian GDEX output was poor – often there were no good sentences to choose from.

When developing the first Estonian GDEX configuration, we missed the evaluation part. I have written an article (Koppel & Kallas 2016) where I have pointed out some of the classifiers that gave poor results and needed to be further tested.

The main purpose of my STSM was to improve the existing Estonian GDEX configuration and identify which classifiers are the most important in order to detect good examples in a mass of corpus data and to evaluate different configurations, always keeping in mind language learners as the target group. I also wanted to develop different configurations for each word class, as Iztok Kosem et. al (2011, 2013) research showed – it gives better results.

**2. Description of the work carried out during the STSM**

Before arriving to Ljubljana, we prepared two different databases in the Institute of the Estonian Language out of the example sentences from the compiled entries of the ECD database:

1. General database containing **all selected** (so-called *good*) examples (44038 sentences);
2. General database containing **all rejected** (so-called *bad*) examples (128239 sentences);

In Ljubljana I divided these databases into 8 different datasets:

1. database containing **selected** sentences for **adjectives** as keyword (6043 sentences),
2. database containing **selected** sentences for **adverbs** as keyword (6933 sentences),
3. database containing **selected** sentences for **substantives** as keyword (28506 sentences),
4. database containing **selected** sentences for **verbs** as keyword (2556 sentences),
5. database containing **rejected** sentences for **adjectives** as keyword (24029 sentences),
6. database containing **rejected** sentences for **adverbs** as keyword (9831 sentences),
7. database containing **rejected** sentences for **substantives** as keyword (86130 sentences),
8. database containing **rejected** sentences for **verbs** as keyword (8249 sentences).

These datasets were then tagged and statistically analysed. As a result I had detailed statistical information about different characteristics of sentences per each dataset. This information enabled me to modify the existing classifiers for general GDEX configuration, but also add new ones. The analysis was also very useful in developing different configurations for different word classes.

The analysis gave me information about:

1. average sentence length and sentence minimum/maximum in each dataset
2. average word length and word length minimum/maximum in each dataset
3. word frequency median in each dataset
4. headword position in each dataset
5. keyword repetition in each dataset
6. sentence initial tags in each dataset
7. minimum frequency threshold for in each dataset
8. etc.

I then analysed different configurations developed for different languages and marked out classifiers that I hadn’t used in developing the initial Estonian GDEX configuration. I tested each of these classifiers to see whether they had any impact on the output. The classifiers I tested were:

1. sentence initial word(s)
2. keyword repetition
3. second collocate
4. penalty for lemmas with a frequency of less than X

I implemented each classifier one after another. In case of soft classifiers I played around with the penalty to see if and how the output changes. For evaluation process I used GDEX Editor[[1]](#footnote-1) developed by Jan Michelfeit within another EneL STSM. If the results were satisfying, I implemented the classifier into my configuration.

When I established all the classifiers that are relevant in detecting good example sentences for language learners, I had to set weights to each of them, i.e. determine, which classifiers are more important than the others.

**4. Description of the main results obtained**

For general GDEX configuration:

1. I reduced the **minimum sentence length** from **5–20** to **4–20**. It was seen in the GDEX Editor that short sentences which contain three lemmas and which would normally be selected as an example sentence in ECS, get a score lower than 0.5.
2. **Optimal interval** appeared to be a very important classifier. In the initial Estonian GDEX configuration it was set from 10 to 12. According to statistical analysis average sentence length in ECD is 9 tokens. I tested different optimal intervals and finally increased it from **6 to 12**.
3. In the initial Estonian GDEX configuration I had implemented a classifier bad\_adverbs\_first, which blacklisted only sentences starting with five certain adverbs with an anaphoric meaning. I complemented the list with **60** more **words** based on the rejected sentences database and named it **bad\_first\_word**.
4. Based on rejected sentences database I compiled a list of words (**bad\_first\_two**) that also have an anaphoric meaning. That list contains **58 word pairs**.
5. Based on the statistical analysis of the rejected/selected sentences, I implemented a new classifier **penalizing** certain **sentence initial tags**.
6. Based on the statistical analysis of the selected sentences I implemented a set of new soft classifiers:
	1. each sentence containing more than X pronouns get penalized
	2. each sentence containing more than X verbs get penalized
	3. each sentence containing more than X conjunction get penalized
	4. each sentence containing more than X proper name get penalized
	5. each sentence containing more than X number get penalized
	6. each sentence containing more than X adverbs get penalized
7. Based on the statistical analysis I implemented a classifier, which penalizes sentences containing lemmas with a frequency of less than 1000.
8. Based on the analysis of the rejected sentences database, I added some **more** **illegal** and **rare** **characters** to the list.
9. Based on the analysis of rejected sentences and the evaluating of different configurations in GDEX Editor, I added **70 more** **words** to the list of **bad words**.
10. I removed some of the non-relevant classifiers and modified the existing ones.

For GDEX configurations for different word classes:

1. I modified soft classifiers based on the statistical analysis of each dataset for each word class.
2. I implemented a classifier which takes into consideration the position of the keyword.

GDEX Editor proved to be a very useful tool when developing and evaluating different configurations. It helped me to better understand the calculations behind each classifier. The Editor was great help in deteriming weights to each classifier, which helped to imprive the results remarkibly. Here is screenshot of the evaluation process:



**Figure 1.** Configuration evaluation in GDEX Editor

In Figure 1, sentences are ordered according to the new configuration. In figure 1 one can also see, whether the classifier meets the criteria fully (score is set to 1) and if it doesn't, how much it lowers the score.

**5. Future plans**

The results of my STSM will be applied in ECD and possibly in creating SkELL[[2]](#footnote-2) (Sketch Engine for Language Learning) for Estonian.

Dr. Kosem and I have discussed writing a paper on good example extraction for the special issue of the International Journal of Lexicography. I would contribute a section about developing and evaluating GDEX configuration for Estonian (aimed at language learners). I also plan to write a paper for Estonian Papers in Applied Linguistics focusing on the characteristics of sentences in selected/rejected sentences database.

**5. References**

Koppel, Kristina, Jelena Kallas 2016. Õppijasõbralik korpuslause – automaatse valiku võimalusi. [User-friendly corpus sentence: parameters for automatic selection]. Lähivõrdlusi. Lähivertailuja. [*to be published*]

Kosem, Iztok, Milos Husák, Diana McCarthy 2011. GDEX for Slovene. – I. Kosem, K. Kosem (Eds.). Electronic Lexicography in the 21st Century: New Applications for New Users. Proceedings of the eLex 2011 conference, Bled, 10–12 November 2011, 151–159.

Kosem, Iztok, Polona Gantar, Simon Krek 2013. Automation of lexicographic work: An opportunity for both lexicographers and crowd-sourcing. – I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, M. Tuulik (Eds.). Electronic Lexicography in the 21st Century: Thinking Outside the Paper. Proceedings of the eLex 2013, 17–19 October 2013, Tallinn, Estonia, 17–19.

1. <https://beta.sketchengine.co.uk/gdex_editor> (11.7.2016) [↑](#footnote-ref-1)
2. <https://skell.sketchengine.co.uk/run.cgi/skell> (11.7.2016) [↑](#footnote-ref-2)