**Scientific Report of Short Term Scientific Mission**

**COST STSM Reference Number:** COST-STSM-IS1305- 210216-071459

**Period:** 21/02/2016 to 30/04/2016

**Duration**: 50 working days.

**COST Action:** IS1305

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

**STSM Title**: In-depth Introduction to Automatic Knowledge Acquisition for Lexicography and DWS

**Guest/STSM applicant**: Tanara Zingano Kuhn – Capes Scholarship holder nº 0973/13-0 /University of Lisbon/CELGA-ILTEC

**Host**: Iztok Kosem, University of Ljubljana, Ljubljana(SI), iztok.kosem@ff.uni-lj.si

**1. Purpose of the STSM**

To receive in-depth introduction to Automatic Knowledge Acquisition (AKA) for Lexicography and Dictionary Writing Systems so that such methodology can be adapted for the purpose of my PhD thesis project, which proposes a design of an online corpus-driven dictionary of Portuguese for university students.

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

The acknowledgment of the importance of the corpus for the efficiency of language study and for evaluation of the tools used for corpus processing and analysis led me to decide to devote the initial part of the STSM to discuss theoretical matters related to corpus building with dr. Kosem. After important decisions were made, the compilation of my corpus of academic texts written in Brazilian Portuguese and European Portuguese was accomplished. This corpus was then uploaded into the Sketch Engine, where it was tokenised, lemmatised and POS-tagged with Freeling.

With the Word Sketch function, in the Sketch Engine, each lemma generates a word sketch - “ a one-page, corpus-based summary of a word’s grammatical and collocational behaviour” (Kilgarriff et al., 2004, p.105). The data it provides, i.e., grammatical relations, collocates and examples of collocates, are transferred to the Dictionary Writing System (DWS) in the process of automatic extraction.

These word sketches, in turn, are the result of statistical calculations based on sketch grammar. After a detailed evaluation of the sketch grammars available on the Sketch Engine and discussion with dr. Kosem, I decided to specially devise a sketch grammar for this procedure. I then studied Corpus Query Language (CQL), consulted several reference books of both varieties of Portuguese, and learned how to overcome shortcomings related to corpus processing, such as incorrect sentence segmentation, wrong lemmatisation and POS-tagging errors. Even though I would like to improve this sketch grammar and expand its syntactic relations coverage, the current word sketches it yields are richer than the other ones tested. Thus, I have decided to make my sketch grammar available for the Sketch Engine so that other researchers studying the Portuguese language can make use of it.

The remaining of the time was fully dedicated to setting parameters for automatic extraction, which involved word sketches analysis and experimenting with GDEX configurations. GDEX stands for Good Dictionary Examples and is an automated method for finding examples in a corpus; it is available as a function in the Sketch Engine. Readability and informativeness are the measurements for determining the order in which sentences will be presented, based on automatically given scores. These, in turn, are determined by specially designed configurations that give points to certain aspects and penalise others, according to the language and the purpose of the examples.

This last component, however, could not be developed in depth, so dr. Kosem and I will continue to work on it after this STSM.

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

The main results of the STSM are the sketch grammar for Portuguese that I have devised and the preparation for the procedure of automatic extraction.

**3.1 Background**

My PhD project aims at proposing a design of a corpus-driven, online dictionary of Portuguese for university students. As a digitally-born dictionary, to be made from scratch, it is paramount to make use of the advantages that the most advanced resources, tools, and technologies provide. Besides, with the latest developments in methodologies for lexicographic work, it is a serious drawback in a new dictionary project to ignore them.

One of the most innovative methodologies that have been recently developed is the semi-automated approach, in which corpus data is automatically extracted and transferred to the dictionary writing system, where then lexicographic analysis is conducted instead of in the corpus tool. This approach, developed by dr. Kosem and his colleagues in the context of the Communication in Slovene project, originated from their goal to come up with a method to improve the efficiency of lexical analysis by “automatically extracting corpus data in a format that enabled a direct import into the DWS, where it would be validated, edited and cleaned by the lexicographers.” (Gantar et al., 2016). Comparison of the manual approach against the semi-automated approach has shown that this method is effective and streamlines the lexicographical process without reducing the quality of the information provided in the dictionary.

According to Gantar et al. (2016), the approach is language-independent and all it requires is “a relatively extensive corpus (minimum 100 million words), POS-tagged as accurately as possible, a sketch grammar, GDEX configuration(s), and parameter settings for data extraction”.

Since my PhD research involves corpus data analysis and pilot-entry writing, I decided to follow the latest trend in lexicographical work and test the semi-automated approach in my project. The procedure requires some preparation, which will be described in the next section.

**3.2 Preparation of the procedure of automatic extraction of corpus data**

The corpus I have compiled for my research is composed of texts published in online academic journals from Scielo (Scientific Electronic Library Online), which is an open access platform with collections from Brazilian journals (Scielo.br) and Portuguese journals (Scielo.pt). Besides the advantage of gathering journals from different areas all in one website, Scielo’s strict criteria for admission and permanency of journals in their collection guarantee the quality of the texts that comprise my corpus.

Due to the need of balance between the two Portuguese varieties, the determination of the size of the corpus came from the smaller text collection, the Scielo.pt. The corpus now contains 10,883 texts, distributed among six areas of knowledge, totalling 53,516,445 tokens. This size makes my corpus ideal for evaluation and tool development, which are usually done on sample corpora with sizes varying from 50 million to 100 million tokens.

The corpus was then uploaded into the Sketch Engine and automatically tokenised, lemmatised and POS-tagged with the preloaded tool Freeling. Based on metadata from the corpus, two subcorpora comprising the varieties of Portuguese were made.

Dr. Iztok and I adjusted the procedure of automatic extraction of data from my corpus in order to cater for the specificities of the Portuguese language varieties and incorporate recent developments in the Sketch Engine, like the longest commonest match and cluster functions. The procedure consisted of the following steps:

1. Selecting a sample of lemmas;
2. Writing Sketch Grammar for Portuguese;
3. Designing GDEX configurations especially for this corpus of Portuguese;
4. Setting parameters values for the extraction script for the corpus as a whole;
5. Extracting data and metadata from the corpus;
6. Writing script for extraction of data from each subcorpus;
7. Extracting information from each variety subcorpus;
8. Writing script for merging datasets;
9. Merging datasets;
10. Setting clustering parameters values for the clustering script;
11. Writing script for clustering and for extracting and adding longest-commonest match information;
12. Adding clustering and longest-commonest match information to the merged data;
13. Setting values for labels assignment;
14. Excluding information below cut-off points and adding labels;
15. Exporting extracted data into iLex (Dictionary Writing System);
16. Evaluating results.

**3.3 Phases development and outcomes**

For lemma selection, a list of lemma was yielded by using the Word List function in the Sketch Engine. Next, manual analysis was performed in order to select 85 lemmas for each of the following categories: nouns, verbs and adjectives. For adverbs, a smaller set of 65 lemmas was chosen. The main criterion was to provide a sample containing lemmas with diversified grammatical behaviour and medium frequency.

As earlier mentioned, I devised a sketch grammar specially for this procedure. There were in total 48 grammatical relations (*gramrel*, in the Sketch Engine): 8 for adjectives, 10 for nouns, 25 for verbs, and 5 for adverbs. I then listed all the gramrel names, separated according to each category.

Dr. Kosem and I discussed language characteristics and the user profile of the dictionary in order to design specific GDEX configurations for my corpus. Although we could not devote much time for this activity, after some testing, we reached satisfactory results. Nevertheless, we intend to continue collaboration and keep working on the GDEX for Portuguese.

The API script for automatic extraction of data from the corpus in the Sketch Engine was written by a programmer. He used the lemma list and the list of gramrels to extract lemmas, collocates, information on grammatical relation frequency and salience, and information on frequency and salience of each collocate.

A script was written for extracting collocates and 3 examples per collocate from each subcorpora. Another script had to be elaborated to merge these two datasets.

The cluster function, which gathers collocates according to their similarity, was manually evaluated, and it was decided that minimum similarity between cluster items would be set at 0.30. The longest commonest match, another recent function of the Sketch Engine, provides an example of the most salient combination between lemma and collocate (which might include words to the right or to the left) right below the collocate in the word sketch result panel. The programmer then wrote a script to extract cluster information and longest commonest match and add that information to the merged datasets.

Before exporting the data to iLex, labels were assigned and collocate cut-off limit was implemented. Parameters for determining labels assignment were the following: a collocate would receive a “typical\_(Br or Pt) variety” label if the ratio between their frequencies was over 15. In addition, if a collocate had 0 frequency, the other variety would immediately receive a BR or PT label. These labels are alerts for the lexicographer and should not be taken as definitive assertions to be directly incorporated in the entry. Finally, collocates with a frequency below 6 were eliminated.

These data were then exported to iLex. However, I only initially began to evaluate the results. Further work on this sample of automatic extracted data will be done after the STSM.

**4. Foreseen publications resulting from the STSM (if applicable)**

The work on the preparation for automatic extraction of data for my research project will result in a publication co-authored with Dr Iztok Kosem. It was also suggested that work developed in different phases of the procedure be presented in conferences, symposia or ENeL meetings.

**5. Concluding remarks**

This STSM has equipped me with knowledge not only to apply the semi-automated approach to my research, but also to evaluate the outcomes of its application and proceed with eventual adjustments.

I can now point out its strengths and limitations, and I can propose improvements for the detected shortcomings.

**6. References**

Gantar, P., Kosem, I., & Krek, S. (2016). Discovering automated lexicography: the case of the Slovene lexical database. *International Journal of Lexicography*. Advanced access. doi:10.1093/ijl/ecw014

Kallas, J., Kilgarriff, A., Koppel, K., Kudritski, E., Langemets, M., Michelfeit, J., Tuulik, M.,& Viks, Ü. (2015). Automatic generation of the Estonian Collocations Dictionary database. In Kosem, I.; Jakubíček, M.; Kallas, J. Krek, S. (eds.). *Electronic lexicography in the 21st century: linking lexical data in the digital age.Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom*. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd., pp. 1–20.

Krek, S. and A. Kilgarriff. 2006. ‘Slovene Word Sketches’ In Erjavec, T. and J. Zˇ ganec Gros (eds), *Jezikovne tehnologije: zbornik 9. mednarodne multikonference Informacijska druz› ba IS 2006, 9. do 10. oktober 2006 = Language technologies: proceedings of the 9th International Multiconference Information Society IS 2006, 9th-10th October 2006, Ljubljana, Slovenia*. Ljubljana: Institut "Joz›ef Stefan", 62-67.

Kilgarriff, A., & Kosem, I. (2012). Corpus tools for lexicographers. In: Granger, S.; Paquot, M. *Electronic Lexicography*. Oxford: Oxford University Press, p. 31–55.

Kilgarriff, A., Pomikalek, J., Jakubíček, M., & Whitelock, P. (2012). Setting up for Corpus Lexicography. In *Proceedings of the 15th Euralex* (pp. 778–785)

Kilgarriff, A.; Jakubícek, M.; Pomikalek, J.; Sardinha, T. B.; Whitelock, P. (2014). PtTenTen: A corpus for Portuguese lexicography. In Sardinha, T.B.; Ferreira, T. L. S. B. (eds.). *Working with Portuguese Corpora*. London, New York: Bloomsbury, p. 111- 127.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography: Journal of ASIALEX*, *1*, p. 7–36.

Klosa, A. (2013). The lexicographical process (with special focus on online dictionaries). In R. H. Gouws, U. Heid, W. Schweickard, & H. E. Wiegand (Eds.), *Dictionaries. An international Encyclopedia of Lexicography. Supplement Volume: Recent Developments with Focus on Electronic and Computational Lexicography.* (pp. 517–524). Berlin, Boston: de Gruyter.

Kosem, I.; Gantar, P.; Krek, S. (2013). Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing. *Electronic Lexicography in the 21st Century: Thinking Outside the Paper : Proceedings of the eLex 2013 Conference, 17-19 October 2013, Tallinn, Estonia*, p. 32–48.

Kosem, I.; Gantar, P.; Logar, N.; Krek, S. (2014). Automation of Lexicographic Work Using General and Specialized Corpora: Two Case Studies. In *Proceedings of the XVI EURALEX International Congress: The User in Focus*, p. 355–364.