Annex 7: Habilitation thesis reviewer's report

Masaryk University	
Faculty	Faculty of Informatics, MU
Habilitation field	Informatics
Applicant	Pavel Rychlý, Ph.D.
Unit	Faculty of Informatics Masaryk University, Brno
Habilitation thesis	Statistical Processing of Text Corpora
Reviewer	Prof. Dr. Piek Th.J.M. Vossen
Unit	Faculty of Arts VU University Amsterdam, The Netherlands

Reviewer's report (extent of text up to the reviewer)

Sketch Engine developed by Rychlý is a major tool for corpus-based linguistics and lexicography. It brings a whole series of statistical methods with proven usefulness within the reach of scientific and business communities that would not have had any access to these techniques otherwise. The many users from diverse backgrounds and the spread to many languages other than English can be seen as strong indication of the success of the tool and systems. The development of Sketch Engine and the underlying platform are the major contributions of Rychlý. Especially its focus on providing lexicographers a work bench for collecting empirical data on words and word meanings is a unique asset, unique in its kind. It bridges a gap between the purely statistical data from large corpora and the purely abstract generalisations found in dictionaries about the properties of words, collocations and multi word expressions. Sketch Engine, thus has set a high standard for the future of lexicography.

Separate from Sketch Engine I also consider Chapter 11: Word's Burstiness in Language Models very interesting and intriguing. It seems to me that the author is hinting on a very important finding that needs further exploration and discussion. Clearly more experiments and discussion is needed on this topic as this paper is very 'sketchy'.

A critical note on the provided material for the habilitation addresses two aspects of the work:

1. At many places in the articles, proper evaluation is lacking where it could have been provided. I will mention a few examples:
   1.1. Chapter 2, A lexicographer-friendly association-score: the author introduces logDice as a more stable and robust measure of association. When discussing the results, the author uses phrases as "one can see..., Dice score gives very good results", "The logDice score has a reasonable interpretation, scales well ...., is stable and the values are in reasonable range." There are however no formal criteria mentioned on which this evaluation is based. "
   1.2. Chapter 4: Sketch Engine. on pages 32 and 33 the authors describe a pattern-based grammar. Why is the output of these patterns not validated against the output of a full-fletched parser, at least for English for which there are plenty available, or on hand-annotated corpora? It is good that Sketch Engine implements this pattern-based tool to make it easy to apply the analysis to al kinds of texts (e.g. tweets) and other languages. However, the user would like to know if there is a price to be paid and how it is.
   1.3. Chapter 6: An efficient algorithm for building a distributional thesaurus. This paper has been set up in a more scientific way by testing the more efficient algorithm for deriving similarity sets from different sizes of corpora and by running it with different settings. Instead of a discussion on the plural usage of "constraint" at the end of this paper, I would rather have seen a discussion on quality of the output of running the systems which is indeed very efficient. It would be perfect to compare the precision, recall and coverage of the similarity sets against WordNet or some similarity rankings derived independently (there are many of such sets) to see what the impact is of the efficient implementation, the size of data and the different settings. Even if this is not feasible, it would be good to describe the output in quantitative terms: how many sim-sets, how large, how different across the different runs.

2. As a major software tool with many users and great impact, it would be good to get more specific technical details about the implementation of Sketch Engine, especially the underlying machinery. How does the indexing and operations compare to tools such as elastic search.

Reviewer's questions for the habilitation thesis defence (number of questions up to the reviewer)

1. **Sketch engine being a valuable tool has demonstrated its quality through the many users and application to many languages. However, it is important that the value is also established empirically for science and for defining its state-of-the-art to direct future research and developments. One can imagine different ways for doing this:**
   - **1.1. speed and quality (precise data, more complete data) of dictionaries derived**
   - **1.2. testing the output against gold-standards: wordnets or similarity sets for thesaurus building (including selecting appropriate names for sets), similarities and differences between related words in the context of translation equivalences.**
   - **1.3. logDice and Burstiness: capability to generate higher precision and recall lists of collocations and unigram statistics**
   - **1.4. Cross-corpus analysis: comparing observed differences against random divisions of text collections, both for the reference corpus as the domain corpus**

   Can the candidate elaborate on these possibilities, report on any performed evaluations and the directions of future research for this?

2. Fully automated dictionary creation: sketch engine generates one-page data structures for words. Why is a lexicographer needed, why is not it enough to maintain many-page data structures for words as an empirical repository for their behaviour? Is there is added value, what does that mean for the acquired data? Can it be improved and taken to the next level? Any ideas or suggestions?

Conclusion

The habilitation thesis submitted by Pavel Rychlý entitled *"Statistical Processing of Text Corpora"* **meets** the requirements applicable to habilitation theses in the field of Informatics.

In Amsterdam on ..... Piek Th.J.M. Vossen (signature)

19 - Sep - 2015