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	Tomaž Erjavec, Ph.D.
Unit	Department of Knowledge Technologies Jožef Stefan Institute, Ljubljana, Slovenia

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

Dr. Rychlý works in the area of computational linguistics, in particular on the development of software to enable the exploration of (very) large corpora and on statistical properties of language. His greatest achievement are the Manatee / Bonito corpus analysis tools, used worldwide under the name of Sketch Engine (which also adds further functionality to simple corpus searching) by thousands of users.

He is the author of mainly conference publications, and his works have been cited a number of times. His publication track record would be even better if it included more high-impact journal publications but the number of citations on his conference papers proves that his papers do reach a wide and appreciative audience. Dr. Rychlý has participated in a large number of mainly national projects and has significant teaching experience. He has been active as a member of the organising or programme committees of several workshops.

The Habilitation thesis provides a brief and to the point introduction to his field of expertise, showing that corpus processing is a well-established yet very active field of investigation. The included papers mostly focus on the design and features of Sketch Engine, with Dr. Rychlý being its main architect. Some papers also present support tools for language processing (such as a program for rediacritisation of Czech) and of statistical investigations on the properties of language (such as the nature of the burstiness of words in language models). The papers prove that Dr. Rychlý is a mature scientist with original yet often simple ideas which he is also able to put into practice. It should be noted though that the introduction to the thesis and, more surprisingly, the published papers do contain a fair amount of typos and mistakes in English.

In conclusion, Dr. Rychlý has good scientific output which should be measured not only in citations but also in the very significant practical use that is being made of his work. While not very active in other academic pursuits, he nevertheless does participate in teaching, organising events and peer-reviewing publications.

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

1. While Sketch Engine does offer concordances of parallel corpora, their functionality is hampered by the fact that the queried-for word or phrase in the source corpus does not have its translation equivalent highlighted in the target corpus. At the same time SMT is now powerful enough that parallel corpora can be word or phrase aligned with decent accuracy. How could Sketch Engine be improved to take advantage of word / phrase alignment in parallel corpora?

2. There are now more and more corpora that have phrase-based or dependency analyses as part of their annotation, or such annotation could be easily performed. Could Word Sketches work better with a proper syntactic analysis instead of its current simulation with PoS tag expressions? What would be the pros and cons of such an approach?

3. All linguistic annotation contains errors. How do such errors impact on the quality of the results obtained, say on Word Sketches? Esp. in the context of Slavic languages, should more research be devoted to make tokenisation, tagging and lemmatisations better or should we concentrate on new annotation levels?

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 Ljubljana on 28. 8. 2015 Tomaž Erjavec (signature)