Early Stage Researcher's Final statement

My research project was situated in work package 4 "Next generation domain modelling", subproject "Harmonisation of terminological resources". The research was conducted at Tilde SIA, Riga, Latvia, and included a secondment at Saarland University, Saarbrücken, Germany, at the Institute of Applied Linguistics and Translation, chaired by Prof. Dr. Elke Teich, professor of English Linguistics and Translation Studies.

In its outset, my research project aimed at improving the usability of multilingual terminology databases that are available on the web. Web-based termbases, as opposed to paper dictionaries, electronic dictionaries, custom glossaries, translation memories, parallel texts, lexicons and knowledge-bases, text corpora or generic search engines, constitute a maybe under-researched and under-used part of the linguistic resources that are available to language workers (i. e. translators, interpreters and technical writers). However, it is usually assumed that they tend to contain higher-quality data and provide better coverage for less-resourced language pairs or domains.

In order to investigate how the usability of online termbases can be improved, a multilingual termbase such as EuroTermBank (www.eurotermbank.com) was analysed with respect to basic functionalities and compared to various resources including other multilingual and broad-coverage termbases on the web, namely IATE (iate.europa.eu) and Termium Plus (http://www.termiumplus.gc.ca), as well as user-generated and widely used online dictionaries such as LEO (www.leo.org), web concordancers and wikis (termwiki.com). Recommendations for practical terminography from the literature and relevant ISO standards (cf. http://www.isocat.org/) were also taken into account. It was found that many features that are considered necessary and useful in terminological theory are not yet consistently implemented in online termbases. Moreover, inconsistent or incorrect data can further limit the usability of multilingual online termbases. Automated resource merging, update and clean-up remain widely unresolved problems.

My further research concentrated on attempts at making a termbase more user-friendly by providing more information about its entries, especially more semantic information. Relevant ISO standards as well as the specialised literature agree that termbase users need to be presented with both linguistic (i. e. term-related) and semantic (i. e. concept-related) information about terms. However, whereas terminology studies were able to make considerable progress in lexical acquisition tasks by incorporating research results from corpus linguistics and computational linguistics (e. g. terminology extraction), relatively little is known about the automated semantic enrichment of terminological resources. Work in this field has to deal with two basic difficulties, namely the lack of knowledge about the mechanics of knowledge transfer by means of language (How is semantic knowledge expressed in texts?) and the lack of knowledge about how humans structure and categorise knowledge (Which "nuggets" of knowledge are needed for understanding a concept and, consequently, which "nuggets" of knowledge should be included into a terminological database?).

For my research, the concept of "knowledge-rich context" (KRC), that is, a context containing relevant pieces of information about a concept, was chosen as a theoretical starting point and adequately refined to render it operational. Russian and German were chosen as object languages of my study, since richer insights into linguistic means of knowledge transfer are to be expected from a multilingual study and both German and Russian can still boast only relatively sparse (in the case of German) or no (for Russian) research in the field. Moreover, it is expected that these languages can be used as a test bed for the usefulness of existing extraction or processing methods, since they present challenging linguistic features such as free word order. Since language workers use the internet and very often simply utilize search engines for terminological research, internet texts were chosen as the object of my study. Consequently, my research helps to evaluate to which extent information extracted from the internet can be used for terminographic purposes. By now, practical results of my research are:

- a pilot study on KRC extraction from Russian and German web corpora using simple processing methods,
- the implementation of a simple method for the supervised ranking of extraction results and experiments with various data sets,
- the creation of a Russian and German gold standard for KRCs from small web corpora that cover multiple domains,
- a detailed investigation of linguistic properties of KRCs as compared to generic sentences.

Upon its conclusion, the contribution of my research project to scientific work in the field will consist in:

- a theoretically well-grounded, systematic study of KRCs in authentic texts,
- a methodology for creating a gold standard resource for KRCs in web corpora and its discussion,
- a comparative study of KRCs in Russian and German texts,
- the setup and evaluation of KRC extraction experiments for both languages,
- the evaluation of current KRC extraction methods with respect to the usability of extraction results in terminography.

During my CLARA fellowship, I was able to enhance my knowledge about terminology and semantics theory and gain new knowledge about corpus linguistics and computational linguistics. Relevant programming (Perl, Java, Python) and language processing skills (various taggers and parsers, XML, CQP, WEKA) were acquired and I was able to add Latvian to the range of my actively spoken languages. During the secondment in Saarbrücken, I familiarised myself with other types of translation-related, corpus-based studies, e. g. contrastive studies and corpus-based work on the properties of translated text (“translationese”). I was also introduced to new tools and resources, including annotation tools for German and German text corpora. A presentation about my work was given at the institute’s Phd students’ workshop on January 18, 2013 and I participated in supervising a terminology exercise with Master’s level students. Further cooperation in the terminology field is planned. As a Phd student, I am enrolled at the “Zentrum für Translationswissenschaft” of the University of Vienna, Austria, and supervised by Prof. Dr. Gerhard Budin. My Phd dissertation will be finished by the end of this year.

Saarbrücken, February 4th, 2013,
Anne-Kathrin Schumann