The author's stylistic fingerprint translated? Using keyword analysis to find individual feqtures in lexicon

Keyword analysis is used for obtaining lists of words specific for a given group of texts. The idea of the method is to compare frequency lists compiled from a research corpus and a control corpus using Dunning's log likelihood test and thus to find the words with significantly different frequencies. The availability of the test as a utility in the Word Smith Tools software package has made the method popular among researchers. Keyword analysis is used for many different purposes from marketing to terminology and political studies (Sardinha 2000, Kemppanen 2008, Probirskaja 2009). It may also help to find individual features in lexicon of an author or group of authors. With a parallel corpus, it is also possible to check, whether the observed features remain intact in translations.

I will show the application of the method on the example of a small case study on Russian-Finnish material. Two pairs of text corpora were compiled for the purpose: 1) Russian data: original works by Mikhail Bulgakov (B-ru) vs. a collection of other works of fiction by other Russian authors of the same generation (RC-ru); 2) Finnish data: translations of Bulgakov's works into Finnish (B-fi) vs. translations of RC-ru into Finnish (RC-fi).

The unlemmatized word lists were processed with the Keywords utility of Word Smith Tools. After comparing the frequency lists of B-ru with RC-ru and B-fi with RC-fi the utility produced the list of 362 textforms for the first and 369 for the last.

The method does not work well with low frequency words, so they have to be removed from the lists. Proper names, titles, etc. are not relevant for the study, because they are not connected with style. Dispersion of the candidate words across texts of the corpora was another important criteria to check. After excluding non-relevant items, the Russian list of keywords shrank to 131 and the Finnish list to 134.

Although the number of keywords in the two lists is close, the Finnish keyword list does not look at all like a translation of the Russian list. E.g. Bulgakov's love for verbs of speech observed in the list of Russian keywords cannot be seen as clearly in the keyword list generated from the Finnish translations of the novels. The frequencies of translation equivalents do not coincide and even can differ substantially due to difference in synonymic and thematic relations in the source and the target languages. On the other hand, the similarities in frequencies can be accidental. The influence of the style of the originals is smoothed up by differences between lexicons, grammar, and stylistic norms of the two languages. Thus, the facts obtained from the keywords list need further investigation with the

help of other tools. Usage examples and collocation tables can confirm researchers' conjectures or reject them and promptly provide new data.

The strength of the method is detection of concrete lexical features of the research data. The weakness is in problems with defining control data and in obtaining sufficient amount of texts.

References

Kemppanen, Hannu (2008). Avainsanoja ja ideologiaa: käännettyjen ja ei-käännettyjen historiatekstien korpuslingvistinen analyysi. University of Joensuu, Joensuu.

Oakes, Michael (1998). Statistics in corpus linguistics. Edinburgh University Press., Edinburgh.

Probirskaja, Svetlana (2009). Rajankäyntiä: Suomen ja Venäjän kahdenväliset valtiosopimukset käännöstieteellisen avainsana-analyysin valossa. Tampere University Press, Tampere.

Sardinha, Tony Berber (2000). Comparing corpora with WordSmith Tools: How large must the reference corpus be? *The Workshop on Comparing Corpora*. Hong Kong, Association for Computational Linguistics, pp. 7-13 <http://www.aclweb.org/anthology/W00-0902>

Word Smith Tools online manual (2013). http://www.lexically.net/downloads/version5/HTML/?keywords_calculate_info.htm