Computational source language detection: Can it help identify indirect translations?

Laura Ivaska
University of Turku

Because information regarding translations' source languages is sometimes incomplete or absent, identifying indirect translations (ITr) may be difficult. Methods to identify ITrs include the analysis of paratextual material and the comparison of different language-versions. These methods, however, depend on the availability of material, are time-consuming and/or require proficiency in several languages. A computational analysis to detect the source languages of translations could provide a more efficient method for identifying ITrs.

Previous studies suggest that the language of translations into language X is different from original texts in language X, and that translated language contains traces of the source language (Toury 1995; Mauranen 2004). The features that are carried over from the source to the target language can be used to computationally detect the source language of a translation (e.g., Islam and Hoenen 2013), but what happens when ITrs are put under computational source language detection – will the analysis detect traces of the ultimate source language, the mediating language, or neither?

To explore this question, I use the package Stylo (Eder et al. 2016) in R and a corpus that contains non-translated Finnish prose, Finnish prose translations from English, German, French, Modern Greek, and Swedish, as well as indirect Finnish translations of Modern Greek literature via English, German, French, and Swedish. Results show that there is coherence within a group of texts translated directly from the same source language and variation between the groups of texts with different source languages. Testing the method with ITrs yields mixed results: six of the thirteen ITrs are similar to direct translations from the ultimate source language, two to translations from their mediating languages, and five to neither. However, these preliminary results are encouraging; with a more robust corpus, they are likely to become more accurate.

REFERENCES

Eder, Maciej, Jan Rybicki and Mike Kestemont. 2016. "Stylometry with R: a package for computational text analysis." *R Journal* 8 (1): 107–121.
Islam, Zahurul and Armin Hoenen. 2013. "Source and Translation Classification using Most Frequent Words." *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 1299–1305.
Mauranen, Anna. 2004. "Corpora, universals and interference." In *Translation Universals: Do they exist?*, edited by Anna Mauranen and Pekka Kujamäki, 65–83. Amsterdam: John Benjamins.
Toury, G. 1995/2012. *Descriptive Translation Studies and Beyond*. Amsterdam: John Benjamins.