Calculating Topics in Estonian Folksongs:
Problems with Texts in Nonstandard Language

One of the most specific genres of Estonian folklore is runosongs, an archaic indigenous tradition Estonians shared by several Finnic peoples. The knowledge of this genre, collected in various ages and modes (texts, melodies, sound and video recordings) has been gathered together into the Estonian Folklore Archives, where the database of runosong texts has been work in progress since 2003 (Oras, Saarlo, Sarv 2003-2019). By the current moment the database contains approximately 2/3 of runosongs ever collected in Estonia, i.e. ca 100,000 texts (or 6 million words) together with basic metadata on the time and place of collection, collector, and performer, if available.

Availability of so many texts in digital format has opened up new insights into the research of runosong tradition. In pre-digital times no one was able to analyse the whole body of texts to find out how the folkloric variation has been functioning: where lie the main regional divisions in this tradition, and on what are they based. My studies have shown that layers of metrical, typological and word use variation behave in notably different ways and it is not easy (or even possible) to draw the general regions of runosong.

My research question in the current study is if it is possible to detect, using topic modelling, what is the topical structure of runosong corpus. Here, we immediately face the problem of linguistic variation. When applying the algorithm on the whole corpus, the variants of different dialects come up as topics. The language of runosong is based on colloquial dialectal language, it varies similarly to dialects, but uses a specific, archaic poetic register. Therefore, the songs cannot be automatically lemmatized, it is not easy to eliminate stopwords etc. Using the method of stylometry (with R application Stylo, Eder et al. 2013) for the study of linguistic variation, and network
community detection algorithm (Blondel et al. 2008) used in network analysis application Gephi (Bastian et al. 2009), I have detected the runosong regions with similar language use (which do not exactly overlap with dialectal areas) – of course the statistics based on word form frequencies includes both, the linguistic as well as content aspects of songs. Within such a region, however, the use of topic modelling (wherefore I have been using MALLET application, McCallum 2002) gives meaningful results.

References


