Digital Folkloristics: Approaching Variation And Stability In Folklore With Computational Methods

Mari Sarv (mari@haldjas.folklore.ee), Estonian Literary Museum, Estonia and Theo Meder (Theo.Meder@Meertens.knaw.nl), Meertens Institute, Netherlands and Kati Kallio (kati.kallio@finlit.fi), Finnish Literary Society, Finland; University of Helsinki, Finland and Berit Janssen (b.d.janssen@uu.nl), University of Utrecht, Netherlands and Peter van Kranenburg (peter.van.kranenburg@meertens.knaw.nl), Meertens Institute, Netherlands and Risto Järv (risto@folklore.ee), Estonian Literary Museum, Estonia and Eetu Mäkelä (eetu.makela@helsinki.fi), University of Helsinki, Finland

Variation is a complex phenomenon engaging almost all aspects of folklore. Every cultural performance in daily life gets adapted to time and place, circumstances and audience. In this panel we are going to explore complex phenomena of variation and stability in folklore on the basis of textual and musical representations of oral tradition with the help of digital and computational methods.

In many cases, variations can be interpreted as intentional and meaningful. However, folklore seldom changes beyond recognition: there is always a part of narratives and songs that remains stable. Detecting in which points lies the stability in folklore sources, reveals to us what the very essence, the core of tradition is. Approaching the material from the other end, it needs to be analyzed how variation is produced, where the adaptability and creativity of folklore lies, and which are the meaningful possibilities for variation within the limits of tradition. As far as we are dealing with texts or melodies, we can determine in what respect oral performances can be labeled as traditional and to what extent folklore is the product of individual creativity and improvisational skills. After determining what parts of folklore remain the same, what changes and what parts are left out, we need to come up with an explanation: what does this all mean in the light of the culture of daily life?

Our core material consists of narratives and songs: epics, poetry, myths and other folktales, life testimonies, and folk songs (both texts and melodies). Millions of folklore texts and performances, collected in the folklore archives and nowadays available in digital form can, together with the existing metadata, be used as data for finding out the regularities and irregularities in folklore - a universal kind of natural communication with its specific functions in society.

1. Computational Analysis of Life Stories

Theo Meder (Meertens Instituut & University of Groningen)

In 2013, a new project was started by the Humanitas Foundation, department Almere, which was called “Levensboek” (Life Book). Volunteers from Humanitas would conduct interviews with elderly people who would tell their life stories. This life story was then recorded or edited by a volunteer and, with photos, then printed as a booklet in a limited edition. The booklets with life stories were mainly meant as testimonials for the children and grandchildren, other family and friends. One of the initiators of Humanitas, Veronica Stutvoet, contacted Theo Meder of the Meertens Institute with the question whether such Life Books were also interesting for archiving and studying. Since the study of contemporary folk culture is one of the core tasks of the Meertens Institute, Humanitas also decided to offer a booklet for the archive. Due to privacy legislation, a contract was added in which the narrators could indicate when the book could be studied freely. And from the Meertens Institute a list was drawn up with subjects that would interest the researchers, such as folktales, songs, games, festivities and rituals. The first Life Book was received in May 2013 in a festive manner: it concerned the book Met hart en ziel (With heart and soul) by Mrs Elly IJsendijk. After proven success, the Humanitas departments in Apeldoorn and Zaandam also started to produce life books, and after five years 20 booklets were produced. In addition to a paper copy, the Meertens Institute also receives a digital copy on request, so that the stories can also be subjected to a computational analysis. This may include structure analysis, research into motifs or sentiment analysis. Research into gender is also possible; do women talk about other subjects than men? The storytellers were, without exception, born in the 1920s, 30s or 40s - meaning that some experienced the crisis years as a child, while some were born shortly after the Second World War. In any case, the war has left a mark on many children, even if they only heard the stories. The life stories are always linear: they often only heard the stories. The life stories are always linear: they often start with the parents, then the childhood, the (aftermath of the) war, school, friends, education and profession, marriage, children and grandchildren, holidays, illnesses and deaths of loved ones. And yet the stories are always different, through the emphasis on certain themes, and through many unique personal experiences. Perhaps most revealing are the themes that all (or most) narrators ignore or leave out. In my research, I analysed the digital life books on structure, sentiments, themes and the distribution of motifs, using tools such as AntConc and LIWC2015 (Linguistic Inquiry and Word Count 2015).

2. Stability in folk song transmission
Berit Janssen (Digital Humanities Lab, University of Utrecht)

In folk song traditions, melodies are circulated through transmission. In this process, parts of melodies may change, while other parts remain stable, meaning they resist change. Stability has been a long-standing point of interest in folk song research: how can stability be quantified, and can we predict which parts of a melody are stable? In the past, this question has been addressed through experimental research, in which artificial transmission chains were observed. While this direction of research is inspirational, the ecological validity of such approaches may be questioned. With the current computational means and rich digitized corpora of folk songs, we can study the results of real-life transmission of melodies by comparing variants of the same song.

The current contribution describes such research on a corpus of 4120 Dutch folk song melodies. Two melodic units were investigated: folk song phrases and motifs. For the phrases, the goal was to predict the occurrence of a phrase in a family of related songs: a phrase occurring in many variants in almost identical form was considered more stable, and was expected to have different melodic properties from less stable phrases. To determine the occurrence of folk song phrases, a pattern matching method was developed in Python, which was optimized on a training set of annotated phrase occurrences. Several similarity measures were compared, and those approaching human judgements on phrase occurrences most closely were combined to detect phrase occurrences in the full set of folk songs. For the motifs, a set of motifs considered characteristic melodic material of 360 melodies was compared against random melodic patterns, with the expectation that the characteristic motifs would have different melodic properties from the random melodic material.

We evaluated prediction success through Generalized Linear Mixed Models. The results show a number of successful predictors for stability of melodic segments in transmission: the length, position and number of repetitions in a melody, conformity to musical expectations, and the presence of repeating motifs can help us to predict whether or not a given melodic segment is stable. Both for folk song phrases and folk song motifs, the melodic predictors explain between 5% and 10% of the variation, constituting a medium-sized effect. Other factors might influence stability in folk song transmission: preference to copy performances of individuals based on their status in society (prestige bias), or preference to copy the most common variants of a melody (conformity bias). Given that such factors cannot be controlled in the current dataset, the extent to which stability can be explained purely on the musical properties of melodic segments is impressive, and shows that stability is certainly not a randomly occurring phenomenon, but arises from the resonance of melodic structures with our cognitive capacities to perceive and memorize music.

3. Rule Mining for Melodic Cadences

Peter van Kranenburg (Meertens Institute, Amsterdam)

The availability of large collections of digitized folk songs enables an empirical approach to the study of various aspects of melodic structure. In this contribution, we focus on melodic patterns that are used to indicate a cadence, or ‘end of phrase’. Most existing approaches for modelling cadential patterns are either based on predefined rules or on statistical learning. Rule based approaches include Narmour’s Implication-Realization model (Narmour, 1992), and Cambouropoulos’ Local Boundary Detection Model (Cambouropoulos, 2001), which both are grounded in principles from Gestalt Theory. Statistical approaches include Rens Bod’s Data Oriented Parsing (Bod, 2001), Huron’s ITPRA-model (Huron, 2006), and the IDyOM model by Pearce et al. (2010). The current study takes a hybrid approach by employing a rule-mining algorithm to infer a model of melodic closure (cadence) from a collection of folk melodies. There are many machine learning methods that could be used to learn models from data. The advantage of a rule-mining algorithm is that the resulting model is highly interpretable, as it consists of a series of rules.

We employ a collection of more than 4,000 melodies in Western tonal idiom from the Meertens Tune Collections (Van Kranenburg, 2014). Since the digitized melodies in these data sets include annotations of phrase boundaries, these are well suited to train cadence-detectors. The data set for rule mining consists of all pitch tri-grams from all melodies. The tri-grams are labelled as either ‘cadential’ or ‘non-cadential’. We represent each tri-gram as a vector of feature values. Features include scale degrees of the three pitches, melodic contour, and metric weights.

We use the RIPPER algorithm (Cohen, 1995) to perform the rule mining. The output of the algorithm consists of a series of rules to separate the cadential tri-grams from the non-cadential tri-grams.

In a first run, we obtain a F-measure of 0.789 on a separate test-set. The three most important rules describe cadences on the first and the fifth degree of the melodic scale. The first rule states that a tri-gram which ends on the tonic, has a high metric weight for the third pitch, and has a descending contour is a cadential tri-gram. This rule reflects common knowledge from music theory.

By closely examining the cases in which the discovered rules fail, we are able to identify possible other features to include. In particular, we find that the position of the tri-gram in a melodic phrase is of importance. Therefore, we include this in our feature set and perform a next run of the algorithm. The newly discovered model achieves a F-measure of 0.839 on a separate test-set. Adding the feature, clearly improved the discovered model.

From this study, we conclude that cadence patterns obey general rules, and that it is possible to derive these rules from melodic data when including the right features. The advantage of a rule-based model is its interpretability in musical terms.
4. Browsing the corpus of Finnic oral poetry

Kati Kallio & Eetu Mäkelä (University of Helsinki)

With a versatile corpus of Finnic oral poems in several related languages and dialects and a wide variety of different orthographies, a central question is how to gather relevant items for each research setting. How to find similar poetic formulas or themes and trace intertextual relationships in a linguistically and poetically heterogenous corpus of oral poetry, and what theoretical possibilities does digital reading offer for the research?

In this paper, we compare searches made with research interface Octavo (https://github.com/jiemakel/octavo) and the present interfaces of two corpora of historical Finnic oral poetry in runo-song meter (www.skvr.fi, www.folklore.ee/regilaul) to the analyses that were made earlier manually with these collections, discussing both the practical and theoretical possibilities and implementations given by digital browsing possibilities.

During the last decades, the folklore archives in Estonia and Finland have digitised two large sources of historical Finnic oral poetry, consisting of c. 181,000 poems in various dialects of small related languages around the Baltic sea: Karelian, Izhorian, Votic, Estonian and Finnish. The poems were recorded in 1564–1939 with various orthographical systems. Some words may appear in hundreds of different written forms. The stories and main characters may exist in various ways, with individual, local and regional peculiarities. The language may contain archaisms or special word forms, syllables and words used only in songs. The poetic system is complex and versatile, and there are no comprehensive dictionaries or ready-made parsers for the data.

Yet, the research history provides a point of comparison. During the first half of the 20th century, a great amount of detailed studies on geographical variation of individual song types was made, with the aim of taking all the collected examples into account. Although the theoretical understanding of oral poetry has since changed, making these studies partly invalid, these studies are still relevant depictions of variation within the data. When compared with searches made with digital tools, they give a baseline for evaluating the possibilities and limitations of the present tools. The paper focuses on three examples:

1) Analysis by Väinö Kaukonen (1956) of the manuscript sources of oral poetry used by Elias Lönnrot when composing the Finnish national epic Kalevala.

2) Historical-geographical analysis by Martti Haavio (1948) of the vernacular Death Song of Saint Henrik, the medieval patron saint of Finland.

3) Typological-stylistic analysis by Matti Kuusi (1949) of the Karelian mythological Sampo-epics.

Is it possible to find digitally all those variations and intertextual links — or more — of a particular theme or poetic formula that were gathered manually by the past researchers? This is approached 1) with word and collocation searches, 2) by checking the results by using a thematic index of the SKVR-corpus (using also visualisation with Palladio), and 3) finally comparing both strategies with the findings of earlier manual research.

5. Potential of Stylometry in Studying Folkloric Variation: Content, Style, Language

Mari Sarv (Estonian Literary Museum, Tartu)

Stylometry - a statistical method comparing sets and share of most frequent words (or other units) in different texts - has been most notably used in the field of authorship attribution, but also in genre studies, in translation studies etc. The main idea lies in the assumption that individual style of an author is represented in the way he/she (unconsciously) uses the most frequent words (usually grammatical function words) or other units. In applying stylometry for the large historical corpora of literary writings one can detect development of style, which is not clearly distinguishable of the changes in natural (and thus also literary) language use (see e.g. Eder and Görski, 2016; Eder, 2018).

The current paper addresses the potential of stylometry in studying variation in folklore texts. Stylometric analysis could possibly help us to find answers to many questions concerning the nature of folklore and variation inherent to it, say the individuality versus traditionality of performers/creators, similarities/differences of different folklore genres. In addition, stylometry could be used as clustering tool for detecting tradition areas within a bigger area, and even folkloristic text-types within the text corpora when focusing on content words in the analysis.

At first glance stylometry seems to be an extremely useful and feasible method for getting better knowledge on variation in folklore; there are additional difficulties to solve though. First, the linguistic (dialectal) variation and folkloric variation are inseparable and overlapping. Different words and word forms used in different (micro) dialects do not have to mean differences in content aspects, like modes, genres, types. Non-standard language and non-standard orthographies present in folklore texts do not make the task of comparison easier either. Moreover, the folklore texts are usually not written down by performers themselves, but collectors who have left prints of their personal style into recordings. The complexity of variation in folklore makes it a challenge to tackle, and evokes questions if the variation we are able to detect using stylometry (comparing the presence and share of most frequent words in different text groups), form part of dialectal, stylistic or folkloric variation; is it individual, or reflects the peculiarities of genre, thematic or functional groups of texts.
My experiments with multilingual corpora of folksongs (www.skvr.fi, www.folklore.ee/regilaul) in several Finnic languages (dialects) on the basis of word forms have revealed that both, linguistic as well as content aspects play a role in clustering, reflecting main dialect boundaries in first instance, but revealing for example also regional predominance of lyric and epic mode in songs, and different thematic accents in regional groups.

6. The network of characters in Estonian animal tales

Risto Järv (Estonian Literary Museum, Tartu)

If folklore is characterised by variation and milieu-morphological adaptation of characters, the adaptation of internationally spread animal tales will retain some established dominant characters – it is certain fixed characters that appear as certain types. The presentation observes the variability of animal characters, using network analysis. The study is based on the Estonian folk tale text corpus created by the Estonian Folklore Archives of the Estonian Literary Museum and the Department of Estonian and Comparative Folklore at the University of Tartu. The corpus contains 13,000 Estonian texts, of which approximately a fifth is made up by animal tales. While the Estonian folklore scholar Pille Kippar has noted in an earlier discussion of characters of animals tales (Kippar, 1989) that the characters can be easily interchangeable within the limits of their stereotypes, I am analysing a sample of selected tale types to check how predominant such variability is, which characters in particular appear as interchangeable, and which regularities emerge in the variability as concerns versions of specific tale types as well as versions by particular storytellers.

I analyse which sets (pairs) of characters are most likely to vary within the tradition and whether there are causal relationships between this feature and the animal characters being active or passive. As several animal tales appear as cycles in the folklore tradition, which combine different tale types within one tale, also this characteristic is taken into account to detect whether any distinct features of character variability emerge in these cases.

Appendix A

Bibliography