I Editorial

Gerhard Lauer

Computational folktale studies. A very brief history

https://doi.org/10.1515/fabula-2023-0001

Not many scholarly traditions are so close to computational approaches as the tradition of folklore respectively folktale studies. Since the days of Antti Aarne and Stith Thompson folktales have been studied based on large collections of stories, comparison of genres and motifs, and historical and cultural stemmatologies. Even though no one was talking about data or data modelling in those heydays of folktale research, the formalization of tales was still the heart of folktale studies right from the start when around 1900 academic societies and researchers systematise the field of research. One could even point to Johann Georg von Hahn's early attempts in the nineteenth century to support this claim. 'Formulae' or story radicals, as Hahn's wording was translated, are basically already in 1864 breaking down complex stories into motifs, count and classify the motifs, correlate the sites of finding with stories and compare the motifs over large geographical and cultural distances. It is of no exaggeration to point out that since more than one hundred years, the research agenda of folktale studies has been noticeable similar to today's computational approaches (Voigt 1976).

Several examples bear this out. The *Dansk Folkemindesamling, The National Collection of Folklore in Copenhagen* by Axel Olrik from 1910 and Antti Aarne's *Verzeichnis der Märchentypen*, also from 1910, as well as Aarne's *Leitfaden der vergleichenden Märchenforschung,* using Kaarle Krohn's classification system from 1913, are iconic examples of this formalism in folkloristics. Aarne's advice in his *Leitfaden* how to do research in folktales studies – "Zu diesem Zweck ist die Erzählung in ihre Hauptteile zu zerlegen, die Teile in ihre Hauptzüge" (Aarne 1913, 67) – is analytic on nearly all levels of research, as it is to compare, classify and count, e.g., fruits or animals in folktales. Even statistical arguments were already in place like more frequent variation of a tale might present the older form or a more carefully preserved or more elaborate variation have a better chance to survive.

Prof. Dr. Gerhard Lauer, Gutenberg-Institut für Weltliteratur und schriftorientierte Medien, Abteilung Buchwissenschaft, Johannes Gutenberg-Universität Mainz, Mainz, Germany. E-Mail: gerhard.lauer@uni-mainz.de. https://orcid.org/0000-0003-0230-2574

The list goes on. In 1932 Thompson's *Motif-Index of Folk-Literature* complemented the Aarne-Thompson Index, and both indices are examples for the formalism of counting, comparing, and classifying. The typologies and classifications were of course often debated and alternative classification of folktales were suggested, such as those by Johannes Wilbert and Karin Simoneau, Walter Heissig or Rüdiger Schott. The third revision of the Aarne-Thompson Index by Hans-Jörg Uther in 2004 along with *Enzyklopädie des Märchens* are impressive results of the formalistic research agenda in folktale studies since its beginnings. Also, the limits of the catalogues, indices and encyclopaedias were more than once criticized, and the critique is not outdated yet (e.g., Uther 2001). Even though this was still the analogue world, approaches to formalise and model research problems were already in practice and 'mechanical help' welcomed (Bronson 1949) long before computational approaches were in reach.

With the advent of computer and internet a wide variety of digital collection from Gutenberg Project or Fairy Tale Corpus to single author pages for Hans Christian Andersen or Halil Bajgorić, from Dee L. Ashliman's to David K. Brown collections of folktales, all open new doors for folk-narrative research (Fialkova & Yelenevskaya 2001; Meder 2008). 'Computational folkloristics' (Abello, Broadwell & Tangherlini 2012) became soon an umbrella term for new attempts to use computational methods and take the internet itself as a field of folktale research (Laudun & Goodwin 2013). However, only few of these collections meet research standards like the Dutch and Flemish Folktale Databases do. A lack of precise metadata, missing transcripts of oral transmission, inconsistent or indefinite classification along the ATU standard are some of the often-mentioned limits. The committee for "Folktales and the Internet" addresses since 2005 the issues at stake.

New platforms and search engines enrich folktale research ever since such as MOMFER, the search engine of Thompson's motif index of folk literature (Karsdorp et al. 2015). Attempts for building an international Folktale Database (Meder 2010; Meder 2014) or a platform for comparative narratology (Dadunashvili 2014) are further examples for the improvements in building databases. The Multilingual Folk Tale Database or the ECHO Collection of Folk Tales are further examples as it is the Yashpeh International Folktale Collection and meta collection like Folktale Collections, to mention a few out of many. The series of new attempts in building digital corpora and improving databases for folktale studies has been continued in recent years, e.g., by attempts to convert the Ashliman Folktexts Collection into a public dataset of annotated tale texts meeting the standards of reproducible research (Hagedorn & Darányi 2022).

In the last two decades a lot has been done not only in developing better databases, but also in computational folktale analysis. For example, automated motif identification following Propp's seminal work and the Russian Formalists'

proto-narratologist approach is one of the main strands in computational folktale studies (Schmid 2009). The identification systems are developed by annotating semantic roles, co-reference, temporal structure, event sentiment, and dramatis personae for the identification of function groups in fairy tales (e.g. Finlayson 2016). For typifying motif sequences linked with tale types researchers make more and more use of machine learning techniques (e.g., Ofek, Darányi & Rockach 2013). Sentiment analysis helped to identify affective sequencing patterns in larger folkloristic corpora (e.g., Alm & Sproat 2005; Aman & Szpakowicz 2007) and quantitative approaches reveal the folk-zoological knowledge embedded in folktales (Nakawake & Sato 2019) or demonstrate that they can enhance folkloristic understanding of culture at scale (Kenna, MacCarron & MacCarron 2017). Finally, the cultural evolution approach offers new routes for the historical-geographical approach in folktale studies (e.g., Tehrani 2013).

It is now obvious, how much computational approaches are already a part of folklore studies, although computational folktale studies is still a small field. Three major research attempts are of particular interest within the next years. First, it is of great importance to build reliable databases for any further research, specifically for building training sets for machine learning or deep learning applications. As the saying in computer science emphasize: 'garbage in – garbage out'. And researchers who pick a too small sample leave themselves at the mercy of sampling luck. Secondly, a standardized annotation of the data and metadata along the established routines of research is mandatory to make databases reusable and the results comparable. By doing so, databases will and have to follow the FAIR and CARE principles of today's research standards. And third, in the long run analytical methods and data visualization are the real game changer for the research. The new methods expand the research possibilities and allow to study large and complex traditions of story transmission over long historical and cultural distances not possible to do by hand and eye alone. Already now, contributions to the field of computational folktale studies are coming not only from folklore studies but also from narratology, computer sciences, and often from archaeology and even biology.

The articles collected in this *Fabula* special issue addresses the three major issues of collecting, annotating, and analysing. Yoel Perez presents new databases for the Sephardic tradition of folktales and discusses the expanded functionalities of today's databases for folklorists from around the world. The long history of Polish folktales and recent developments of online folktales collections are the subject of Violetta Krawczyk-Wasilewska's contribution to this volume and to the general topic of the global intangible heritage in the digital age. Christoph Schmitt and Alf-Christian Schering pictured their work to transform Richard Wossidlo ethno-linguistic collection into an advanced hypergraph database, the Digital Wossidlo Archive. Theo Meder, Petra Himstedt-

Vaid, and Holger Meyer worked on an online database for belief legends and developed an intelligent search engine for multilingual narrative heritage. Finally, James Abello, Peter Broadwell, Timothy Tangherlini, and Haoyang Zhang introduce two novel network decomposition methods for the study of folktale collections at corpus scale, which could be adapted to similarly indexed collections. Their test case is the Danish nineteenth-century corpus of Evald Tang Kristensen.

The topic of Jeana Jorgensen contribution to this volume is a quantitative approach to gender by analysing the biased link between depicting beauty and youth, gender and success in folktale plots. The central problem of folkloristics, namely the clear identification of motifs, is addressed in Johan Eklund's, Josh Hagedorn's, and Sándor Darányi's research. They use statistics and Support Vector Machine algorithm on annotated folktales test collection, to predict text membership in their internationally accepted categories. With the example of thousands of Cinderella-like stories Gessica Sakamoto Martini, Jeremy Kendal, and Jamshid Tehrani demonstrate the research potential of phylomemetic methods. They take Anna Birgitta Rooth's Cinderella typologies and make use of Bayesian phylogenetic inference, phylogenetic networks and a model-based clustering method derived from computational biology. The group of Julien d'Huy, Jean-Loïc Le Quellec, Marc Thuillard, Yuri Berezkin, Patrice Lajoye, and Jun'ichi Oda make also extensively use of statistical tools to reconstruct the past of myths and folktales. A maybe new comparative mythology is within the reach of an approach, which model the evolution of myths and mythological traditions as phylogenetic trees. The results of these advanced methodological approaches are encouraging and with all due caution folktale studies has many reasons to embrace the new research possibilities.

To take stock, the value and versality of the computational approaches expand the possibilities of folktale research tremendously. Now, tradition and networks of folktales could be studied at a scale in depth and wide unthinkable with established methods. However, the enthusiasm is quickly clouded if one is aware of the enormous amount of work necessary to collect and classify the data and metadata precisely, to annotate and train large models, and to make use of complex analytical tools and visualisations. This could no longer be done by a single researcher but in research groups with members from more than one discipline. What we know as folktale studies became something new, and we can say we were there.

References

Aarne, Antti: Leitfaden der vergleichenden Märchenforschung (FF Communications 13). Hamina 1913.

Abello, James/Broadwell, Peter/Tangherlini, Timothy: Computational Folkloristics. In: Communication of the ACM 55 (2012) 60–70.

Alm, Cecilia Ovesdotter/Sproat, Richard: Emotional Sequencing and Development in Fairy Tales. In: International Conference on Affective Computing and Intelligent Interaction. Heidelberg 2005, 668–674.

Aman, Saima/Szpakowicz, Stan: Identifying expressions of emotion in text. In: International Conference on Text, Speech and Dialogue. Heidelberg 2007, 196–205.

Bronson, Bertrand: Mechanical Help in the Study of Folk Song. In: Journal of American Folklore 62 (1949) 81–86.

Dadunashvili, Elguja: Webplattform der vergleichenden Erzählforschung. In: Corpora ethnographica online. eds. Holger Meyer/Christoph Schmitt/Alf-Christian Schering/Stefanie Janssen. Münster 2014, 129–134.

Darányi, Sándor/Voigt, Vilmos: Automated Motif Identification in Folklore Text Corpora. In: Folklore 110 (1999) 126–141.

Fialkova, Larissa/Yelenevskaya, Maria: Ghosts in the Cyber World. An Analysis of Folklore Sites on the Internet. In: Fabula 42 (2001) 64–89.

Finlayson, Mark: Inferring Propp's Functions from Semantically Annotated Text. In: Journal of American Folklore 129,511 (2016) 55–77.

Hagedorn, Joshua/Darányi, Sándor: Bearing a Bag-of-Tales. An Open Corpus of Annotated Folktales for Reproducible Research. In: Journal of Open Humanities Data 8,16 (2022), DOI: http://doi.org/10.5334/johd.78.

Karsdorp, Folgert/van der Meulen, Martin/Meder, Theo/van den Bosch, Antal: MOMFER: A Search Engine of Thompson's Motif Index of Folk Literature. In: Folklore 126 (2015) 37–52.

Kenna, Ralph/MacCarron, Máirín/MacCarron, Pádraig (eds.): Maths Meets Myths. Quantitative Approaches to Ancient Narratives. Bern 2017.

Laudun, John/Goodwin, Jonathan (2013). Computing Folklore Studies. Mapping over a Century of Scholarly Production through Topics. In: The Journal of American Folklore 126,502 (2013) 455–475.

Meder, Theo: Art. Internet. In: The Greenwood Encyclopedia of Folktales and Fairy Tales. ed. Donald Haase. Westport 2008, here vol. 2, 489–492.

Meder, Theo: From a Dutch Folktale Database towards an International Folktale Database. in: Fabula 51 (2010) 6–22.

Meder, Theo: The Folktale Database as a Digital Heritage Archive and as a Research Instrument. In: Corpora ethnographica online. eds. Holger Meyer/Christoph Schmitt/Alf-Christian Schering/Stefanie Janssen. Münster 2014, 119–128.

Nakawake, Yo/Sato, Kosuke: Systematic Quantitative Analyses Reveal the Folk-Zoological Knowledge Embedded in Folktales. In: Palgrave Communication 5,161 (2019), DOI: https://doi.org/10.1057/s41599-019-0375-x.

Ofek, Nir/Darányi, Sándor/Rokach, Lior: Linking Motif Sequences with Tale Types by Machine Learning. In: 6th Workshop on Computational Models of Narrative. eds. Mark Finlayson/Bernhard Fisseni/Benedikt Löwe/Jan Christoph Meister. Dagstuhl 2013, 166–182.

Schmid, Wolf (Ed.): Russische Proto-Narratologie. Texte in kommentierten Übersetzungen. Berlin, New York 2009.

Tehrani, Jamshid: The Phylogeny of Little Red Riding Hood. In: PLoS ONE 8,11 (2013) e78871. DOI: https://doi.org/10.1371/journal.pone.0078871.
Uther, Hans-Jörg: Klassifikation von Volkserzählungen nach Aarne und Thomson. Zur erneuten Revision von Types of the Folktale. In: Schweizerisches Archiv für Volkskunde 97 (2001) 109–115.
Voigt, Vilmos: Means and Aims of Computer Folklore Research. In: Papers in Computational Linguistics. Berlin 1976, 549–554.