university of groningen

University Medical Center Groningen

University of Groningen

The ISEBEL Project Collecting International Narrative Heritage in a Multilingual Search Engine

Meder, Theo; Himstedt-Vaid, Petra; Meyer, Holger

Published in:
Fabula

DOI:
10.1515/fabula-2023-0006

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2023

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Meder, T., Himstedt-Vaid, P., & Meyer, H. (2023). The ISEBEL Project Collecting International Narrative Heritage in a Multilingual Search Engine. *Fabula*, *64*(1-2), 107-127. https://doi.org/10.1515/fabula-2023-0006

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 25-03-2025

Theo Meder, Petra Himstedt-Vaid, and Holger Meyer

The ISEBEL Project

Collecting International Narrative Heritage in a Multilingual Search Engine

https://doi.org/10.1515/fabula-2023-0006

Abstract: ISEBEL is an online database for belief legends. The acronym stands for: Intelligent Search Engine for Belief Legends. The database contains more than 70,000 traditional legends in Dutch, Frisian, Danish and German, while another 6,000 Icelandic legends are currently being added. The initiative for this project was taken several years ago by the Meertens Institute in Amsterdam, the University of Rostock/the Wossidlo Research Center and UCLA/UC Berkeley. The ambition is to create a European database, with an intelligent search function and geographical visualizations. What makes the search engine 'intelligent' is that it can always be searched in English, thanks to high-quality automatic translations in the background. The legend material that is brought together can also form the basis of sophisticated graphs that network themes, motifs, narrators, repertoires, and places. This article includes examples of international legends about mermaids and werewolves.

Zusammenfassung: ISEBEL (Intelligent Search Engine for Belief Legends) ist eine Online-Datenbank für geglaubte Ausdrucksformen, besonders von Sagen und sagenähnlichen Erzählungen. Die Datenbank beinhaltet über 70 000 derartige Erzählungen auf Niederländisch, Friesisch, Dänisch und Deutsch, die zurzeit um weitere 6 000 isländische ergänzt wird. Initiiert wurde das Projekt vor einigen Jahren vom Meertens Instituut in Amsterdam, der Universität Rostock/Wossidlo-Forschungsstelle und der UCLA. Ziel ist es, eine europäische Datenbank mit intelligenter Suchfunktion und geographischen Visualisierungen fertigzustellen. Die „Intelligenz" der Suchmaschine besteht darin, dass sich die Datenbank – dank hochwertiger automatischer Übersetzungen im Hintergrund – immer auf Englisch durchsuchen lässt. Die hier versammelten Erzählüberlieferungen können auch als

Theo Meder, Professor Folktales and Narrative Culture, University of Groningen, Almere, Nederland. E-mail: theo.meder@meertens.knaw.nl
Petra Himstedt-Vaid, Wossidlo Forschungsstelle für Europäische Ethnologie/Volkskunde, Universität Rostock. E-mail: petra.himstedt-vaid@uni-rostock.de
Holger Meyer, Database research group, CSEE Dept., University of Rostock, Germany. E-mail: holger.meyer@uni-rostock.de. https://orcid.org/0000-0002-5803-4911

Grundlage für komplexe Graphen benutzt werden, die Themen, Motive, Erzähler, Konstellationen und Orte miteinander verknüpfen. Der Beitrag stellt internationale Meerjungfrauen- und Werwolferzählungen als Beispiele vor.

1 Introduction

It all started with an idea at a conference of the International Society for Folk Narrative Research in June 2009 in Athens, Greece. Theo Meder presented a paper entitled "From a Dutch Folktale Database towards an International Folktale Database".[1] The idea was to populate an already functioning folktale database with additional folktales[2] from around the world. At the time, the idea was that various folktale databases, which were already up and running in different countries, would be merged into one large, central "super-server," for example in Amsterdam. During the discussion after the presentation, Tim Tangherlini, a folklorist working with computational methods and Danish archival records, wondered whether all of these regional and national institutions would be enthusiastic about simply handing over their stories and metadata to a central database over which they would have little control. Tangherlini suggested it may be technically more convenient to develop a "harvester" that would occasionally retrieve stories and metadata from folktale databases via an online connection so that these stories could subsequently be searched centrally.

This idea ripened for several years. Then, in 2017, the opportunity arose for funding for a collaborative project through the Transatlantic Digging into Data program, jointly funded by the National Endowment for the Humanities (USA) and the European Union.[3] The three collaborating parties would consist of:

1. The Meertens Institute, Amsterdam: Theo Meder and Vic Ding
2. UCLA (now UC Berkeley): Tim Tangherlini and Peter Broadwell
3. University of Rostock/Wossidlo Research Center: Christoph Schmitt, Petra Himstedt-Vaid; Database research group: Holger Meyer, Alf-Christian Schering

The aim of the project would be to develop an international archival data harvester under the name ISEBEL: Intelligent Search Engine for BElief Legends.[4] The appli-

1 Lateron published as Meder 2010.

2 In particular the Nederlandse Volksverhalenbank (Dutch Folktale Database): www.verhalenbank.nl (January 13, 2023).

3 See https://diggingintodata.org/ and https://transatlanticplatform.com/ (January 13, 2023).

4 The search engine can be found here: http://search.isebel.eu (January 13, 2023).

cation was awarded for a period of three years: 2017–2020.[5] During this period, ISEBEL would harvest three databases of folktales to be managed by the participants in the project, namely:

1. The Dutch Folktale Database (Meertens Institute, Amsterdam)[6]
2. The Danish Folktale Database (UC Berkeley, USA)[7]
3. Wossidlo Digital Archive (Rostock, Germany).[8]

2 Digital databases

Although each of the three databases contained folktales (in short, digital archives with narrative heritage), there were clear differences as to scope and underlying structure. The Dutch Folktale Database covers a period from the Middle Ages (c. 1200) to the present day and contains various collections, with a predominant focus on the nineteenth and especially twentieth century collections. The languages range from Middle Dutch and seventeenth century Dutch to modern Dutch, Frisian, and various regional languages. The genres catalogued in the system are also diverse: fairy tales, traditional and contemporary legends, riddles, and jokes. In all, the Dutch Folktale Database contains more than 48,000 narratives.

The Danish Folktale Database consists of a single, large collection: the folkloric life work of Evald Tang Kristensen (1843–1929). Around the turn of the century, Tang Kristensen, partly as a paid collector and researcher, amassed an extensive collection of folk songs, fairy tales, jokes, legends, riddles, proverbs, and traditions, largely collected on the Danish peninsula of Jutland. The Danish Folklore Database mainly brings together his legends and, to a lesser extent, his fairy tales, ballads, and descriptions of everyday life: all in all, more than 31,000 stories.

The Northeast German Folklore Database also contains the life's work of one folklore collector: Richard Wossidlo (1859–1939). He collected numerous folktales and traditions in Mecklenburg, and also worked on a regional dictionary. All his material was eventually preserved in the Wossidlo Archive in Rostock, which contains an estimated 2 million documents. The material is organized by an extensive and dynamic note box system that regularly involves hierarchical and linked

5 https://diggingintodata.org/awards/2016/project/intelligent-search-engine-belief-legends-isebel (January 13, 2023).

6 www.verhalenbank.nl (January 13, 2023).

7 See https://scando.ist.berkeley.edu/folklorenexus/ and https://scando.ist.berkeley.edu/dfl2db/story_view.php?story_id=1 (January 13, 2023).

8 https://apps.wossidia.de/webapp/run (January 13, 2023).

folders with field notes and excerpts from ethnographic publications that are related to each other.[9]

3 ISEBEL and comparative research

It seemed sensible that traditional legends were selected as the target corpora for ISEBEL[10], as those collections and their accompanying metadata since each of the databases had such records in abundance and such a selection would make comparative research possible. An additional benefit of selecting these subsets of the collections is that a large number of these belief legends have never been published, essentially lying dormant in the archives, and many are not identified in national or international catalogues, while they are adequately catalogued in the metadata; consequently, ISEBEL would increase accessibility to the underlying collections while also supporting opportunities for comparative research. The resulting target corpora can be characterized as "belief legends", with stories concerning a broad range of supernatural phenomena such as hauntings, ghosts, gnomes, trolls, devils, witches, wizards, magic, werewolves, mermaids, nightmares, and giants. These are supplemented with the other narratives about hidden treasure, famous robbers, underground tunnels, sunken monasteries, and abandoned castles.[11]

In all, the ISEBEL system accesses more than 31,000 Danish legends, 26,000 Dutch legends and 14,000 Mecklenburg (almost Low German) legends, which represents a total of more than 71,000 legends.[12] Although this represents a considerable collection, an important caveat is that not all the stories are contemporaneous with each other. The German and Danish material all date from several decades before and after 1900. During that period, there was collecting in the Netherlands going on as well, but not as intensively, and the harvest was much smaller at the time: at most 700 legends. The great collecting efforts in the Netherlands did not take place until the 1960s and 1970s. For example, the Frisian collector Dam Jaarsma (1914–1991) collected more than 16,000 folktales in that period, and 23 other collectors approximately the same amount.[13] Of course, fairytales and jokes were collected during this period in the Netherlands as well, but legends comprised the majority of the collections. For comparative research, it is worth noting the implications in

9 See Schmitt 2019.

10 For the website see: http://www.isebel.eu/site/ (January 13, 2023).

11 For a first description of the project in Dutch, see Meder 2018.

12 See Schmitt/Tangherlini 2018 for the participation of the Wossidlo Digital Archive and the Danish Folktale Database in ISEBEL.

13 Dekker 1978, 22–26.

the difference in the time of collection: while it seems unlikely that the Dutch storytellers contacted by Jaarsma and others believed in the phenomena reported in their stories, it seems likely that many of Tang Kristensen and Wossidlo's informants believed in the supernatural phenomena and events reported in their stories.[14]

Not all folklore collections look the same, and that had repercussions for developing the ISEBEL system for harvesting, even in the limited context of working with the three target databases. The Danish and Dutch material consists mainly of stories and the necessary metadata related to collection and indexing, while the Wossidlo collection reflects some of the idiosyncrasies of that collection. This spread in the local archiving of the materials and their representation in the local data structures has had important consequences for the structure of the creation of datasets for harvesting as well as the structure of the ISEBEL harvested collection that is used for indexing and search. Consequently, it was necessary to develop a minimal description of each story which could then be consistently harvested by the ISEBEL system. Creating this minimal set was initially conceived of a question-answer template to make the process consistent and easily comprehensible. The template included questions such as: who is the narrator and what is their reported gender? When and where was the story told? With which keywords can the story be characterized? Which places appear in the story? What is the subgenre? Who is the collector? And, finally, which motifs and tale types (if applicable) have been assigned to the story?

Although this template was relatively easy to discover in the Danish and Dutch materials, given their consistent internal structures, discovering this information from the Wossidlo Archive was slightly more difficult, given the physically distributed nature of much of this information in the collection's drawer cabinets, folders and envelopes. Because the leaders of the digitization efforts at the Wossidlo archive, Christoph Schmitt and Holger Meyer, had decided that all notes and cards would form the basis for the online database, a linear approach based on simple SQL queries, such as those that could be deployed against the Danish and Dutch material, was out of the question. Because of the structure of the Wossidlo data, each record had to be constituted through hyperedges linking multiple nodes in a graph structure. Consequently, the digital Wossidlo Archive, WossiDiA, was based on a hypergraph model.[15] Many of the technological challenges for developing the harvester and harvesting nodes was built on the well-established OAI-PMH protocol, with some minor adjustments. Perhaps the biggest departure from standard

14 See Wossidlo 1928/1929.

15 For more information on the hypergraph technology see Meyer/Schering/Heuer 2017; its implementation in WossiDiA see Meyer/Schering/Schmitt 2014, 72–83, and the contribution of Schering and Schmitt in this volume.

OAI-PMH implementations was to treat transcriptions and translations of stories as forms of meta-data.

Beyond the underlying, "back end", tasks of creating a system to harvest consistently structured data, the project was also charged with developing an easy-to-use user interface that would allow for search and basic analytics. Although a web page, including logo, for ISEBEL was created relatively quickly,[16] designing and building a well-functioning and aesthetically satisfactory search engine took a lot more effort.[17] Ultimately, the team settled on implementing CKAN, an open-source data management system, given its configurability and its various search facilities.

4 Searching in English: dirty translations and domain specific keywords

A researcher who masters the languages represented in ISEBEL could search separately in the German, Danish and Dutch databases. If searches had to be in each of the target languages, ISEBEL would provide little added value, apart from the fact that a search could be executed in a single place. The reason why ISEBEL is called intelligent, however, is searchability of the data in English along with the target languages. Additional analytics, such as the visualization of the geographic distribution extend the output of such a broad search.

Multi-lingual search is a well-known challenge in search and retrieval. For ISEBEL, the initial goal was to search this heterogeneous and often noisy data in English, returning results from all of the languages. Although a small number of German stories had already been manually translated into English, and approximately three percent of the Danish stories had English keywords, none of these provided support for the search functionality envisioned by the project. Manually translating everything was unthinkable given the scale of the various corpora. Moreover, if more databases were to join ISEBEL at a later stage, automatic translation would be the only workable solution. The problem was therefore mainly with the automatic translation of all texts into English with some acceptable degree of accuracy: with standard Danish, German, and Dutch this went reasonably well, but with Low German and Frisian, for example, it became more difficult, given the low-resource nature of these languages, and the lack of accurate language models for these languages. Orthographic variation, particularly in these low resource

16 http://www.isebel.eu/site/ (January 13, 2023).
17 http://search.isebel.eu/ (January 13, 2023).

languages, significantly increased the challenge of devising automatic translations that were sufficient for search.

Despite the increasing accuracy in commercially available neural machine translation methods such as Google Translate turned out not to be a viable option. That was not due to the quality of the Danish, German, and Dutch translations, because they were usually quite good, especially if Google Translate was fed with complete stories and not just separate words, such as the keywords. Complete stories provided sufficient context for an adequate translation. But the bulk of texts quickly became too large: for such quantities one had to pay for automatic translation, and that would only get worse in the future, if more databases were to join and budgets were low. So ISEBEL had to be trained to translate independently.

Peter Broadwell and Tim Tangherlini took responsibility for the machine translation process, with the team managing to achieve highly acceptable translations through machine learning techniques. It was assumed from the outset that these would be so-called "dirty translations", which would remain invisible to the user. For the human reader, the "dirty translations" can be somewhat disorienting to read, as these translations reflect well-known errors in syntax, translation, and other semantic gaps. Despite these faulties, the "dirty translations" are sufficient for search.

To increase the accuracy of the search over the dirty translations, we developed an augmentation step to account for domain specific terms (here, from the domain of folk belief) that are difficult for language models trained on corpora such as parliamentary records, subtitles and news. These terms, such as "duivelsdrek" (devil's dung, used as defense against witchcraft), "stalkaars" (jack-o'-lantern or will-o'-the-wisp), "tsjoenster" (witch), "karismatiker" (faith healer), "gammen" (gryffin), "Wesselbalg" (changeling), "Freikugel" (magic bullet) or "Hexenproben" (witch test), often confound the model and do not get translated properly.

In Danish, the word "fanden" means the devil, but it can also be used as a swear. As an example: in a Danish legend[18] a man is very ill. His wife warns that the door can only be opened after three knocks. If the door is opened prematurely, there will be a problem …

> "…for det er Fanden, der vil ind og gjøre min mand fortræd."

In the earliest experiments with neural machine translation, using "off the shelf" models, this passage was rendered, rather humorously yet incomprehensibly, as:

> "…for it is the fuck that will come in and make my husband upset."

18 http://search.isebel.eu/dataset/da-etk-ds_06_0_00378 (January 13, 2023).

Yet the mistranslation of the specific character, here the Devil, rendered the translation useless for search – in other words, a search on the Devil would not have returned this story. Importantly, later models render the phrase far more accurately:

> "…for it is the Devil who wants to come in and harm my husband."

To address these problems of low frequency words with high value in the domain of folk belief, a "domain specific word list" was created for words where automatic neural machine translation tends to fail.

This two-stage rocket approach to "dirty translation" augmented with a "domain specific word list" yielded better search results, but several confounders still existed. First, spelling variants threw a spanner in the works: "duvelsdrek" was not recognized as "duivelsdrek", for example. Second, stories often referred to phenomena that were not explicitly named: a woman with witchy behavior who is never called "witch", or someone taking his own life, while the word "death" is not mentioned. A monk may appear in a story, while one finds nothing on "clergy", or the protagonist is a soldier, while searching for "military" yields nothing.

An interim solution is co-translating keywords already attached to stories in a target corpus: since existing keywords are always in the standard language, making translation more accurate, they also contain more abstract concepts such as "death" and "military". As a simple test, a search for the term "witch" was performed: without translation of the keywords, the search returned ~700 stories, while with the translated keywords, the search returned ~3,000 stories, the number expected based on a manual search of the databases. In some cases, more stories than expected are found, reflecting the considerable increase in *recall* afforded by these methods; the searched stories are now being found, and a large part is therefore no longer "lost in translation."

5 Dispersion of mermaids and werewolves

As an illustration of the capacities of ISEBEL as a platform to support research, we present several examples of research driven queries and their results.[19] To begin with, we look at a simple example focused on the distribution of stories referencing mermaids (*havfrue, meermin, zeemeermin, Seejungfer, Wassernixe* etc.) by searching

19 Also see Jansen 2021.

for the English term "mermaid", with a result of 232 legends.[20] The distribution map mainly shows the coastal regions as places where stories have been told about mermaids. This remains the case when we zoom in on the map. For the Netherlands, the further inland we look, the fewer stories about mermaids have been recorded. Details from the north of Denmark show a similar phenomenon along the coastline.

Fig. 1: ISEBEL: the spread of legends about mermaids in the Netherlands, Denmark and Northeast Germany

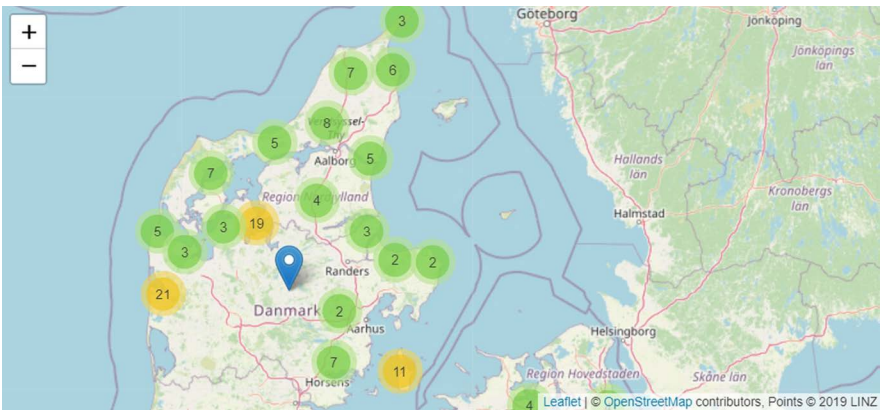

Fig. 2: ISEBEL: detailed map of legends about mermaids in the north of Denmark

20 http://search.isebel.eu/dataset?q=mermaid&button= These searches have been performed in the Spring of 2022; in the meantime, the search results may differ slightly due to extra legends added to ISEBEL.

A second, equally straightforward example centers on the werewolf (*varulv, weerwolf, Werwolf*), with an English query returning 570 results.[21] These legends were predominantly told by men (314), with far fewer told by women (89).[22] Geographic distribution is uneven as well, with 56 Danish stories, 85 German stories and no fewer than 431 Dutch stories, the majority of which can be situated in the south of the Netherlands. Based on the Dutch distribution, one might suspect, as further elaborated below, that werewolf legends were especially popular in the Catholic areas.

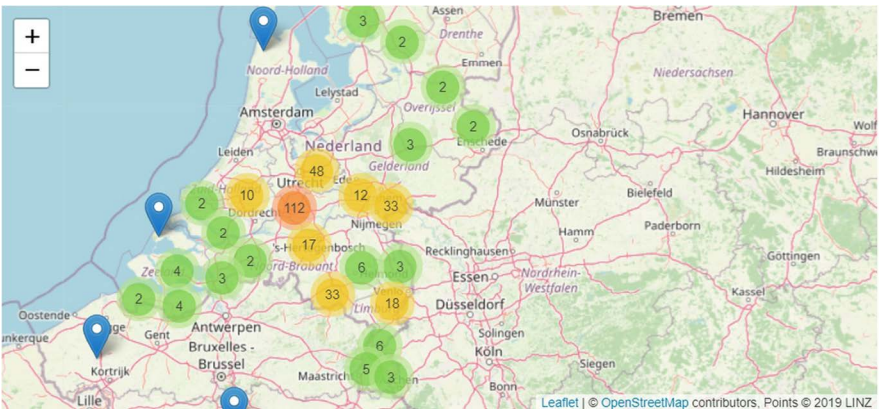

Fig. 3: ISEBEL: detailed map of werewolf legends with a center of gravity in the south

In Dutch werewolf legends, it often happens that a suitor turns out to be a werewolf. For example, a young couple is walking through the forest, when the boy separates from the girl for a while. He advises her that, if a werewolf comes, she should throw her handkerchief at him. Indeed, a werewolf arrives a little later, the girl flees and throws her handkerchief, which is then torn by the werewolf. When the boy later rejoins the girl, she sees the threads of the handkerchief between his teeth. It can also happen that the werewolf assaults the girl's clothing; he then rips part of her skirt or apron, and the threads are visible again later. ISEBEL contains 27 Dutch versions of the handkerchief variant alone. Among the Danish werewolf legends, we find 16 variants with threads between the teeth: for example, because an apron or

21 http://search.isebel.eu/dataset?q=werewolf&button=.

22 The gender of the storyteller is not always known (147 legends). The gender of the collector may be an important factor in this skew in the data, as male collectors were inclined to interview more men than women.

Fig. 4: ISEBEL: detailed map of North German werewolf legends with a cluster in the forest areas

a shirt has been bitten. Four versions of this particular werewolf legend are found in the German material.[23]

Another werewolf story tells of a man who becomes a werewolf when he puts on a werewolf belt or strap. Disenchantment takes place after the belt is burned in the fireplace. If one searches for "werewolf AND strap" in ISEBEL, one finds 28 German versions, 13 Dutch (the oldest from the early seventeenth century), and none from the Danish (in Danish legends the werewolf is more often disenchanted by identifying it with a simple statement – sometimes trebled: "You are a werewolf").[24]

One of the problems for researchers is, of course, explaining the geographical dispersion of a phenomenon. In the Netherlands werewolf legends are mainly found in the southern, traditionally Catholic, part of the country. One possible explanation for this concentration is that, in the past, prosecution of werewolves went on much longer among the Catholics than among the Protestants in the North. The practice of prolonged prosecution may explain why the legends survived over a longer period of time. However, this conclusion cannot explain the distribution of werewolf legends in Northeast Germany because a division between Catholics and Protestants can hardly be found in predominantly Protestant Mecklenburg. It looks like, in the German case, there is a concentration of werewolf tales in forest areas: belief in werewolves survived longer in the woodlands (see the cluster of 17 werewolf stories in the woodlands near Lübtheen/Ludwigslust).[25] Forests again could not serve as the explanation for the Netherlands, since the country hardly

23 http://search.isebel.eu/dataset?q=werewolf+AND+teeth&button=.
24 http://search.isebel.eu/dataset?q=werewolf+AND+strap&button=.
25 Himstedt-Vaid 2022 and 2023.

had any forests; in those areas that used to be forested, the occurrence of werewolf tales is sparse.

The 85 German werewolf legends in ISEBEL, with their metadata including as motif, gender and profession of the narrator and narrator's place, offer a good basis for graph visualization. The graphs are built upon the metadata about the werewolf legends harvested using OAI-PMH[26]. Some details of the transformation are given in the next section.

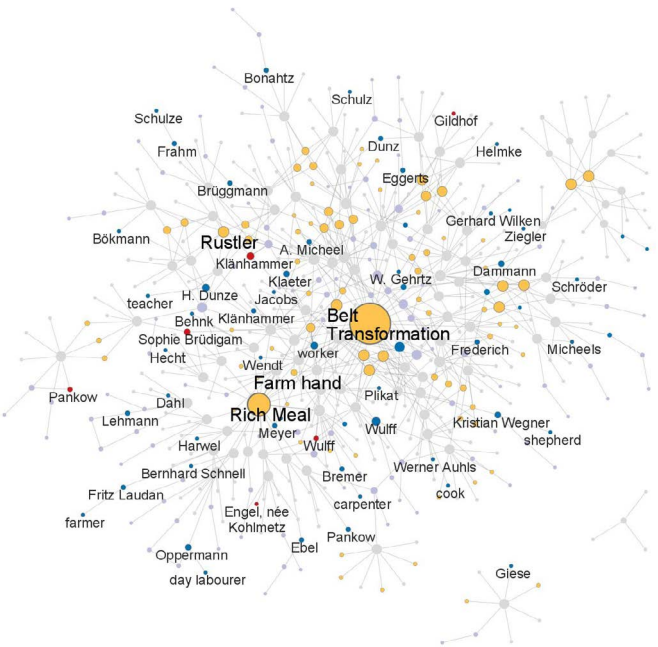

Fig. 5: Motifs of German werewolf stories: Belt, Transformation, Farm Hand, Rich Meal, and Rustler are frequently present

In the graph "Motifs of German werewolf stories", which has been laid out using a simple force-directed algorithm, all information of the 85 legends of werewolves is represented as typed nodes. These are story, person, (especially narrator by gender), place, keyword, and date nodes. They are connected by typed edges representing the role a certain node plays in a story, e.g., place-of-narration, narrator, etc.

26 https://en.wikipedia.org/wiki/Open_Archives_Initiative (January 13, 2023).

The female narrators (red), the male narrators (blue) and motifs (yellow) are labeled with their names. The node size is scaled according to in-degree (the number of connections – edges – received) representing somehow their importance (or frequency). The keyword "werewolf" has been removed from the graph as it is present in all legends and tends to occlude the main themes of these stories: belt, transformation, farm hand, strap, and meal. The keyword "transformation, werewolf" forms the centre of the graph, as this is where most of the overlaps in the legends can be found. The "rich meal" is a highly represented node among the subjects, and these are connected to the "farm hand" and the "belt". These motifs are also shared by another set of stories. The graph allows the most important subjects of the werewolf legend to be visualized so that the main themes can be recognized at a glance.

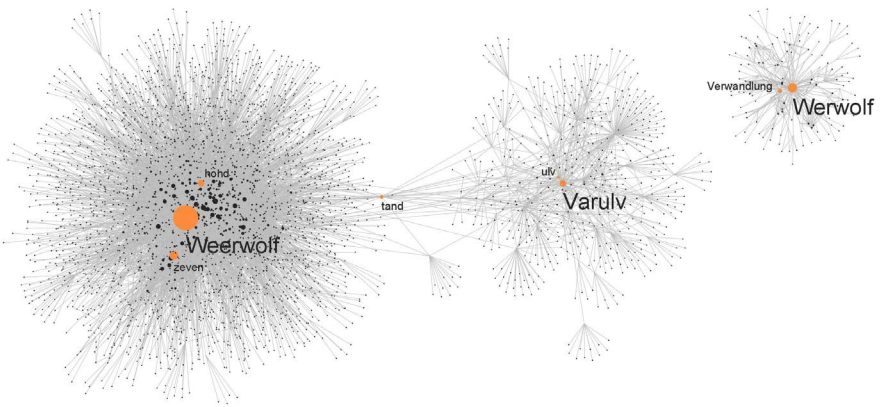

Fig. 6: ISEBEL Werewolf Graph: Visualization of the size of the Dutch, Danish and German werewolf subgraphs based on the X2G graph extraction method

6 Mining ISEBEL data on werewolves with graphs

Although there are several approaches for mining the ISEBEL story data including text search and more complex database/faceted query-based approaches, one of the great affordances of ISEBEL is the potential for graph mining on the data. Below we outline some of the challenges and the benefits in exploiting graph mining techniques on the underlying data. Although ISEBEL harvests all of the data from contributing data bases as valid XML data as described above, the first challenge for making this data available for graph mining is to render these XML data into graph

data, e.g., represented by the property graph model.[27] We have designed a tool, X2G[28], for this task. X2G uses a set of rules for extracting and filtering the XML data and generating graph data in various formats, e.g., comma separated values for the nodes, edges and labels of the graph.[29] As a first step, the rule language uses XPath[30] expressions for selecting and extracting parts of interest from story documents. These parts usually occur in XML elements, sub-elements, attributes, and element content. In the second step, graph data is generated, i.e., nodes, edges, labels and properties are constructed from the extracted data (see Figure 6). Since the keywords assigned to each story are language specific, each of the subgraphs are largely separated. While the German werewolf subgraph is isolated, the Dutch and Danish share the tenuous connection through the node "tand", the word for "tooth" in both languages. The werewolf subgraph also makes apparent that the Dutch collection has many more details, including stories, than either the Danish or German collections. Not terribly surprisingly, the most frequent keyword in each language is werewolf: Dutch *weerwolf*, Danish *varulv*, or German *Werwolf*. The second most frequent keywords are also labeled in the graph, here *zeven* (seven)[31] and *hond* (dog)[32] in the stories from the Netherlands, *ulv* (wolf) in the Danish and *Verwandlung* (metamorphosis) in the German stories.

There are several ways of exploring the resulting graph data. Gephi or Cytoscape and similar tools for visualizing graphs can be used for visualization and simple analysis. Both platforms support graph loading, filtering, colouring, and provide several state-of-the-art graph layout algorithms. Using these platforms, a researcher can easily explore medium sized graphs, i.e., graphs consisting of several thousand nodes and edges. For example, the graph in figure 6 consists of 5,430 nodes and 9,043 edges and was produced by Cytoscape. Simple analytics including calculations of degree, betweenness and centrality can also be used to adjust the visualizations for legibility.

More complex analyses make use of more advanced graph techniques including community detection and other types of clustering algorithms to discover topological features of the graph.[33] In community detection, a community is based on the intuition that "communities" consist of sets of nodes which are much more con-

27 Cf. Bonifati et al. 2017.

28 Meyer 2022.

29 Cf. Fig X2G tool.

30 https://en.wikipedia.org/wiki/XPath (January 13, 2023).

31 It is often mentioned in Dutch legends that the seventh son is a werewolf.

32 In Dutch legends it is often remarked that the werewolf resembles a black dog.

33 Since many of these clustering algorithms were primarily developed for social network analysis, they are called community detection algorithms.

nected to each other than to those outside of the community. Mark Newman's work (2006) on modularity presaged a revolution in these types of community detection methods. Modularity describes what a "good" cluster in a graph is and is used in most graph clustering algorithms. A good cluster has a large number of node pairs linked within the cluster and only a few links from nodes inside going to nodes outside the cluster.

For our work, we use the Louvain algorithm which is an efficient and fast implementation of a modularity-based method for finding communities. Applied to our ISEBEL werewolf stories, the Louvain algorithm[34] finds clusters which share a set of specific motifs (keywords from the controlled vocabulary).[35] Figure 8 *Graph clustering* depicts the result of the community detection algorithm by inking clusters differently: each colour represents a motif group here. Figure 7 *Graph cluster* shows details of one of these clusters, here the motif group of "ample meal". An interesting confirmation of this technique comes from a comparison with a close reading analysis of the stories.[36] Here, we confirm that most of the clusters found by the algorithm are identical to the groupings found in that analysis. This is a promising result and argument for using graph mining techniques for supporting distant reading.[37]

The steps in producing the werewolf graphs are in general as follows: (1) We extracted all werewolf stories from the collected ISEBEL XML data. (2) The resulting set of XML files are input to the X2G tool. X2G renders the XML data into nodes and edges with labels and property sets based on transformation rules supplied by the user. Output are node and edge lists as comma separated values. (3) The graph is loaded into Cytoscape (Gephi would be an alternative tool). Some nodes are eliminated from the graph at this stage, especially the werewolf keyword as the central motif to all stories. (4) The resulting graph is visualized using certain algorithms, e.g., different layout and clustering algorithms such as Louvain. Node sizing and coloring take place after graph analysis. By this, one can distinguish different node types, places, narrators, keywords/motifs namely. Clusters are made visible by spatial proximity in the graph. In our example, one can identify clusters around the "ample meal", the "transformation" or the "rustler" motifs.

The visualized representation of the werewolf legends offers further possibilities, e.g., to present the different narrative repertoires of men and women using the links between narrator and story. Similarities show up in clustered nodes. In

34 Blondel et al. 2008.

35 Khorsand, 2021 found that Louvain superseded Girvan-Newman with respect to larger graphs by keeping the same precision and recall.

36 Himstedt-Vaid 2022.

37 Jänicke et al. 2015.

the graph, only a few aspects were highlighted for illustration, but the possibilities of focusing on different aspects (narrative repertoire by motif, by gender, by place, etc.) are almost limitless. Here, we used a very straightforward mapping of the ISEBEL story data to a graph with person, places, and keyword nodes which were linked to the story nodes. Changing this mapping can offer completely new views on the story data.

In a small corpus manual story analysis is feasible but in a large corpus, e.g., for about thousands of stories about witches, a challenge. By complementing traditional analysis with graph analysis techniques and clustering algorithms can provide an alternative starting point for the researcher. The graphs also offer new views on relationships, offering insight into the corpus (or subcorpus) as a whole, and providing new avenues for comprehension.

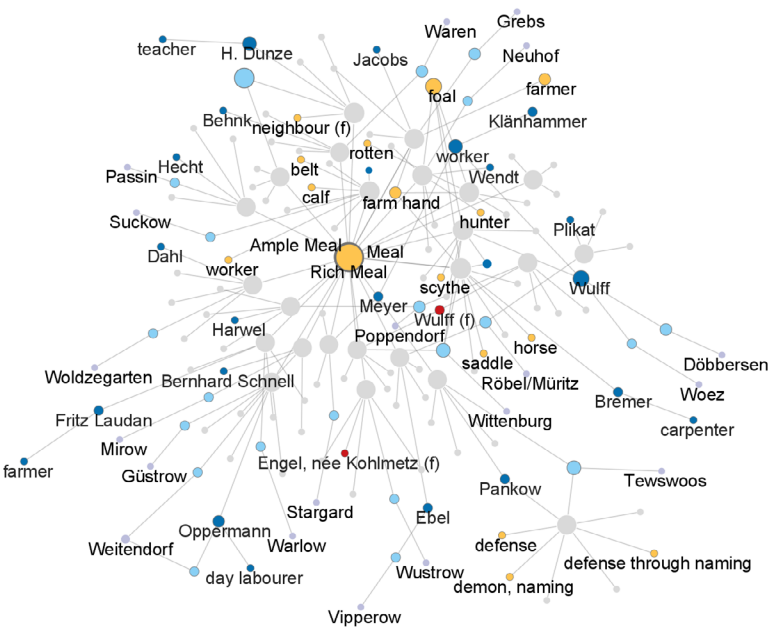

Fig. 7: Graph cluster: Details of one cluster of German stories with the rich meal motif

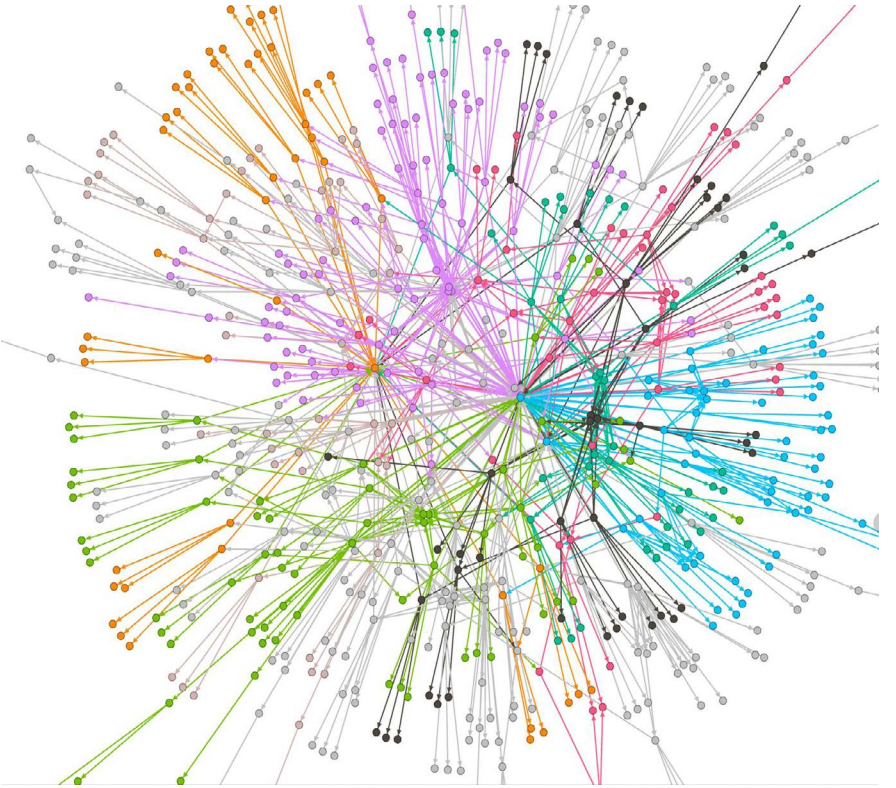

Fig. 8: Graph clustering: Result of an overall cluster analysis of the German Werewolf stories Different color means different story cluster, e.g., pink for rustler, blue metamorphosis and green for the rich meal motif

7 Turning the wind and the witch wreath

In another experiment, we focused on magic and the concept of "turning the wind." In many rural communities up through the nineteenth century, certain people were believed to be capable of more than other people, especially at the level of the supernatural and magical. One can think of wizards, witches, fortune tellers or Freemasons, but certain religious persons were also assumed to have magical powers. For example, there are many legends in which it is said that someone could turn the wind. If there was a fire somewhere and the fire was threatening to reach a farm, for example, such a person was brought in to turn the wind so that the fire would go another direction. In the Netherlands, we see a special distribution of this type of legend: they occur almost exclusively in the south, and hardly or not at all

in the north. If we look at the content of the legends, it appears in almost all cases to be a pastor or priest who can turn the wind. In folk belief, Catholic clerics were therefore regularly thought to be a kind of magician: in addition to being able to exorcize the devil, banish evil spirits, or magically heal people, they were also able to turn the wind. On the map of ISEBEL we see all those wind-turning clerics concentrated in the Southern Netherlands. For Denmark, the ISEBEL map shows much less of a clear geographical distinction, and here we are therefore dependent on the content of the legends.[38] On closer inspection, the Catholic-Protestant distinction also appears to be evident in Denmark. Those who can turn the wind in a fire are always called *præst* (priest) and *pastor*.[39] Just like in the Netherlands, Protestant ministers were of no help in turning the wind.

One form of evil magic for which a Protestant cleric is never called in to provide a solution involves legends about children or adults being sickened by witchcraft. This always concerns patients in whom a so-called 'witch wreath' is found in the pillow. Remarkably enough, we hardly find such legends in ISEBEL in the Danish and German[40] material, but in abundance in the Dutch material.[41] As far as distribution in the Netherlands is concerned, it is striking that the stories are mainly found in the northern part of the country. This seems to indicate that these mainly concern legends from the Protestant areas, and that (again) no magical assistance from ministers was to be expected. For the defence against witchcraft, the Protestant population was left to its own devices. Considering the large number of legends, it seems for the time being that the Netherlands is a hotspot for such tales and folk belief.

Many of these stories come down to this: a child or adult is sick and does not get better. A suspicion of witchcraft arises, and the pillow is cut open to check. A witch wreath is found between the feathers in the pillow (in the form of a rose, a star or – to a lesser extent – a chicken). The wreath is not yet fully completed, because if it were, the victim would have died. In some stories, finding the wreath is enough to override the harmful witchcraft: the sick person gets better. But in many Northern

38 https://search.isebel.eu/dataset?q=fire+AND+wind+AND+turn&button=.
This query yields some false hits for Danish. Searching for the specific legend type "Turning the wind" yields a better selection:
https://search.isebel.eu/dataset?q=ETK+AND+%2228.577%22&sort=score+desc%2C+metadata_modified+desc. The map now mainly produces coastal areas, but also legends about sailing ships and wind.

39 Incidentally, some Danish legends testify that certain wizards could put out a fire by walking around the seat of the fire three times. They are not mentioned as being clergy.

40 For the "witch wreath" in North German legends see Himstedt-Vaid 2021, 326–328.

41 http://search.isebel.eu/dataset?q=sick+AND+wreath&sort=score+desc%2C+metadata_modified+desc.

Netherlands variants, more action is required. For healing or disenchantment, the feather wreath must be boiled in a pan (sometimes thrown into a fire), and in some cases it was only effective if a live (black) chicken was placed in the pan. This ritual was intended to break the magic, but in many cases also to hurt and entice the culprit, the witch. Burning the feathers or the chicken was a form of sympathetic magic, with the burning and pain passing to the witch, who felt compelled to seek out the source of the pain. Sometimes it is mentioned that the witch now also had burns, and sometimes the witch came in to beg to stop the torture. On this the inhabitants could demand that the witch would end her sorcery. If this happened, then the sick person was healed.[42]

Something similar is found in German material, but usually the discovery is an unfinished wreath of straw found in a straw sack on which the sick person slept. Laying the wreath outside the door of the alleged perpetrator could be enough to cancel the harmful magic. In some North German legends, it is said that the wreath of feathers must be burned. Then the witch comes running. But then you do not let the witch into the house. So, the sick person must be taken out of the witch's sphere of influence.

For the time being, the legend material in ISEBEL shows the Protestant Northern Netherlands as a hotspot for the many legends about witch wreaths, as well as the belief in them, but this picture may change if more legend collections are added to the database.

8 Future developments

It took considerably more time than expected to build ISEBEL and create an agreeable design for the user interface. Although each of these databases was well-described, it took considerable programming to get all of the (meta)data from the three databases, and to correctly represent it in ISEBEL. The development of a search engine making it possible to search in all languages from English also took a considerable amount of time, largely because of two factors: first, there was no clear blueprint for such a model and, second, the development of useful NMT models[43] for low resource languages with unusual vocabularies required considerable development of a hybrid method for supporting multilingual search. The final,

42 Meder 2001, 119, 527–528; Dekker 1987, 246–247. In the Catholic parts of the Netherlands a priest could sometimes heal a patient through prayer and blessings.

43 Models for Neural Machine Translation: https://en.wikipedia.org/wiki/Neural_machine_translation (January 13, 2023).

three-stage rocket of "dirty translations"', "translating keywords" and the multilingual "domain specific word list" ultimately produced satisfactory results.

The intention in the initial project was to add a series of analytic plugins, such as WitchHunter and GhostScope to support other types of geographically informed search, but adding these plugins will take some more time.[44] Similarly, graph visualization and graph mining techniques using data discovered through ISEBEL search engine are not integrated into the interface at the moment. Indeed, due to the complexity of the mining algorithms, it is necessary to make use of cloud-based platforms to power these approaches.

ISEBEL will be maintained going forward by the Meertens Institute in Amsterdam, where the search engine has been developed and designed in collaboration with the Rostock and UCLA (now UC Berkeley) groups. Several Northern European databases have expressed an interest in joining ISEBEL, in particular Iceland, Norway and Sweden, with Icelandic legends from the Sagnagrunnur collection recently being added.[45]

9 Literature

Blondel, Vincent D./Guillaume, Jean-Loup/Lambiotte, Renaud/Lefebvren Etienne: Fast unfolding of communities in large networks. In: Journal of statistical mechanics: theory and experiment 10 (2008) 1–12.

Bonifati, Angela/Fletcher, George/Voigt, Hannes/Yakovets, Nikolay: Querying graphs. In: Synthesis Lectures on Data Management 10, 3 (2008) 1–184.

Broadwell, Peter M./Tangherlini, Timothy R.: GhostScope: conceptual mapping of supernatural phenomena in a large folklore corpus. In: Maths Meets Myths. Quantitative Approaches to Ancient Narratives. eds. Ralph Kenna/Máirín McCarron/Pádraig McCarron. Heidelberg 2017, 131–157.

Dekker, Adrianus J.: 150 jaar Nederlands volksverhaalonderzoek. In: Volkskundig Bulletin 4,1 (1978) 1–28.

Dekker, Thomas: Heksen en tovenaars in twintigste-eeuwse sagen. In: Nederland betoverd. Toverij en hekserij van de veertiende tot in de twintigste eeuw. eds. Marijke Gijswijt-Hofstra/Willem Frijhoff. Amsterdam 1987, 242–255.

Girvan, Michelle/Newman, Mark EJ: Community structure in social and biological networks. In: Proceedings of the national academy of sciences 99,12 (2002) 7821–7826.

Himstedt-Vaid, Petra: Verrufen, Verhexen und böser Blick: Schadenzauber in norddeutschen Erzählungen. In: Von Mund zu Ohr via Archiv in die Welt. Beiträge zum mündlichen, literarischen

44 For an experiment with Danish material see for instance Broadwell/Tangherlini 2017.

45 See https://sagnagrunnur.com/en/ (January 13, 2023). Due to the recent addition of 2,000 German and 6,000 Icelandic legends, some search results mentioned above, may now differ a bit.

und medialen Erzählen. Festschrift für Christoph Schmitt. eds. Petra Himstedt-Vaid/Susanne Hose/Holger Meyer/Siegfried Neumann. Münster/New York 2021, 311–330.
Himstedt-Vaid, Petra: Der Werwolf in norddeutschen Sagen. In: Fabula 63,1–2 (2022) 143–162.
Himstedt-Vaid, Petra: Of Wolf-Belts, Hungry Servants and Tattered Skirts. The Werewolf in North German Legends. In: Werewolf Legends. eds. Willem de Blécourt/Mirjam Mencej. In print (London 2023).
Jänicke, Stefan/Franzini, Greta/Cheema, Muhammad Faisal/Scheuermann, Gerik: On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. In: EuroVis (STARs). eds. Rita Borgo/Fabio Ganovelli/Ivan Viola. May 2015, 83–103.
Jänicke, Stefan/Geßner, Annette/Scheuermann, Gerik: A distant reading visualization for variant graphs. In: Proceedings of the Digital Humanities (2015) 1–21.
Jansen, Mathilde: Op zoek naar weerwolven en zeemeerminnen. In: E-Data and Research, June (2021). https://edata.nl/2021/05/25/op-zoek-naar-weerwolven-en-zeemeerminnen/ (January 13, 2023).
Khorsand, Zahra: Mining Graph Data in the ISEBEL Project. Master's thesis, University of Rostock, Computer Science Department 2021.
Meder, Theo: Vertelcultuur in Waterland. De volksverhalen uit de Collectie Bakker. Amsterdam 2001.
Meder, Theo: From a Dutch Folktale Database towards an International Folktale Database. In: Fabula 51,1–2 (2010) 6–22.
Meder, Theo: ISEBEL: Intelligent Search Engine for Belief Legends. In: Volkskunde 119,1 (2018) 87–89.
Meder, Theo: Sagen op de internationale kaart in ISEBEL. In: Neerlandia, Nederlands-Vlaams tijdschrift voor taal, cultuur en maatschappij 127,1 (2023), 8–11.
Meyer, Holger: X2g – A tool for mapping NoSQL data into property graphs. Technical report CS-03-22. University of Rostock, Computer Science+Department 2022.
Meyer, Holger/Schering, Alf-Christian/Heuer, Andreas: The Hydra. PowerGraph System – Building Digital Archives with Directed and Typed Hypergraphs. In: Datenbank-Spektrum 17,2 (2017) 113–129.
Meyer, Holger/Schmitt, Christoph/Schering, Alf-Christian: WossiDiA – The Digital Wossidlo Archive. In: Corpora ethnographica online. Strategien der Digitalisierung kultureller Archive und ihrer Präsentation im Internet. Münster 2014, 61–84. https://doi.org/10.18453/rosdok_id00002265.
Newman, Mark EJ: Modularity and community structure in networks. In: Proceedings of the national academy of sciences 103,23 (2006) 8577–8582.
Schmitt, Christoph: From Idiosyncratic Index-Card Machines to Digital Folklore Archives. In: Folkloristics in the Digital Age. eds. Pekka Hakamies/Anne Heimo. Helsinki 2019, 132–157.
Schmitt, Christoph/Tangherlini, Timothy R.: Folklore Archives Online. Zur Sichtbarmachung, Auswertbarkeit und Interoperabilität einer dänischen und einer nordostdeutschen Sammlung. In: Jahrbuch für Europäische Ethnologie [Dritte Folge] 13 (2018) 181–204.
Wossidlo, Richard: Glaubt das Volk noch an seine Sagen? In: Quickborn 22,4 (1928/1929) 115–122.