SVENSKA LANDSMÅL OCH SVENSKT FOLKLIV

Tidskrift för talspråksforskning,
folkloristik och kulturhistoria

Redigerad av Fredrik Skott
och Mathias Strandberg

2017
Årgång 140

Utgiven av Kungl. Gustav Adolfs Akademien för svensk folkkultur
i samarbete med Institutet för språk och folkminnen

“Hvad der byggedes om dagen, blev revet ned om natten…”

Word Sequence Repetition in Danish Legend Tradition

By Peter M. Broadwell, Peter Leonard, and Timothy R. Tangherlini

Abstract

The repetition of phrases or entire sequences of words is a common feature in many folklore genres, including ballads, rhymes, and fairy tales. Often, these types of repetition are either tied to formal features of the genre, such as formula or refrains, or otherwise reinforced by various metrical constraints. Conversational genres, such as the legend, do not include the repetition of word sequences to the same degree, yet the extent to which word sequences appear in legend is an open question. Such repetitions may provide important information about individual storytellers’ repertoires, language use in a specific region, and the crystallization of language in storytelling. Using the Evald Tang Kristensen legend collection as a target corpus, we develop a method and user interface for the discovery of word sequence repetitions across tens of thousands of legends. The matches are deliberately “fuzzy”, thereby taking into account individual storytellers’ linguistic predilections, as well as the impact of editorial interventions. The method allows us to discover overlap between stories that share similarities even though traditional classifiers would not consider the stories to be topically related.

Keywords: legend, computational folkloristics, motifs, sequence alignment, machine learning, digital humanities, language.

The interrelationship between individual story repertoire and overall legend tradition, particularly in the context of performance, can now be more readily addressed as large collections of legends are made accessible in machine-actionable form (Gunnell 2014; Skott 2017; Tangherlini & Broadwell 2014). An important, and open, question in the study of legend is whether tradition participants exhibit expressive consistencies in language within and across their individual story repertoires. These consistencies may be indicative of certain tradition-based phenomena, such as the crystallization of story elements into a specific linguistic form, or the use of formula as an integral part of the story itself. The recurrence of word sequences may reflect stylistic aspects of an individual’s storytelling or, on a broader scale, storytelling in a specific region. The discovery of consistent word sequences across multiple stories also supports investigations into similarities between stories that are not topically related. Such discoveries might facilitate, in turn, comparative studies across repertoires of individual storytellers, classes of storytellers (e.g. gender), and regions, thus augmenting

---

Peter M. Broadwell, Peter Leonard, & Timothy R. Tangherlini, “Hvad der byggedes om dagen, blev revet ned om natten…”: Word sequence repetition in Danish legend tradition. *Swedish Dialects and Folk Traditions* 2017, pp. 9–25.

more standard comparisons of genre and topic. Finally, the investigation of repeated word sequences in a large folklore corpus may provide insight into the practices of folklore collectors, including how they created fair copy out of their at times chaotic field notes, and how they edited their collections for publication (Christiansen 2013; Tangherlini 2008).

In this short investigation, we devise a method for finding word sequences of arbitrary length that appear in more than one story in a large digital legend corpus. The matching is deliberately "fuzzy", increasing the recall of repeated word sequences when compared to standard string-matching approaches, thereby letting a user explore many more examples than would be possible with standard keyword or string-based searches. We employ two user interfaces to facilitate the navigation of the overall corpus space and the discovered matches. The first interface makes use of a heatmap similarity matrix, while the second provides a side-by-side method for examining the word sequence matches in the context in which they appear (Broadwell, Mimno, & Tangherlini 2017; Malm & Leonard 2016). To illustrate the effectiveness of the approach, we use these interfaces to discover several closely matching word sequences, and evaluate these instances of matched word sequences in the context of legend performance, collecting, editing, and publication.

Target corpus

For this study, we use Evald Tang Kristensen's printed collections of Danish legends (*Danske sagn* and *Danske sagn: Ny række*; Kristensen 1892–1901; Kristensen 1928–1939) and descriptions of peasant life (*Gamle folks fortællinger om det Jyske almueliv*; Kristensen 1891–1894; Kristensen 1900–1902). This subset of his larger collection consists of 31,088 stories that were told or submitted by 4,242 identifiable informants or local collectors (Table 1). The stories were collected in the late nineteenth and early twentieth centuries over the course of five decades, during which Tang Kristensen (b. 1843–d. 1929), a schoolteacher, criss-crossed the Jutlandic peninsula, making over 260 field trips and traveling over 70,000 kilometers, largely by foot (Storm et al. 2017). Over the course of his collecting career, he developed his own shorthand for recording stories, filling approximately 24,000 manuscript pages in 335 individual composition books. These composition books were copied fair and comprised the primary materials for his eighty-plus published volumes.[1]

The target collections we work with here (DS, DSnr, JAH, JAT) were ordered into 72 top-level and 774 second-level topic classes. The printed

[1] For a more detailed discussion of Tang Kristensen's collection, see Tangherlini 2013b.

volumes were subsequently organized into sections according to this classification scheme. For example, volume six of *Danske Sagn* concerns the top-level topic "The Devil and Pacts with Him", which includes 33 second-level topics, including "Taken by the Devil". We consider each numbered record in the publications to be a "story", although we recognize that many records may include multiple stories.

Table 1: Corpus statistics summarizing the aggregate of DS, DSnr, JAH, and JAT.

Stories	Informants	Tokens (words)	Characters	Mean words per story	Story bundles	Mean stories per bundle
31,088	4,242	3,943,116	21,128,430	127	1,869	16

In previous work, we have explored how flexible classifiers can allow for the discovery of stories that share commonalities with stories classified separately (Abello, Broadwell, & Tangherlini 2012; Broadwell & Tangherlini 2012, 2016). Tang Kristensen was well aware of this problem, writing in the introduction to one of his first collections of legends, *Sagn fra Jylland*, about the difficulties he encountered with stories that could have been placed in multiple categories:

> I must ask the reader for forbearance on several fronts. First, the ordering of the stories, which has its difficulties; but it should be noted that where a story is opaque or distorted, then the classification is based on a judgment. For example, tale 388 (classified in "On prophecy and portents") could have been put in section ix ("Fairytale-like legends"), 389 ("On prophecy and portents") in section iv ("On revenants and all types of ghosts"), and 438 and 439 ("Religious legends") in section vi ("On witchcraft") (Kristensen 1880: i).

These classification challenges treat stories as topically similar entities, which is understandable as there are clear indications from repertoire studies that storytellers conceived of their stories as such (Pentikäinen 1978; Dégh 1969). Our work proposes to explore stories across these boundaries, a continuation of work we have done on other approaches to classification (Abello, Broadwell, & Tangherlini 2012; Broadwell, Mimno, & Tangherlini 2017).

The interrelationship between stories at the word sequence level is understudied for short conversational genres such as the legend. Word-sequence-level comparison is better known for genres such as the epic (Lord 1960) and the ballad (Gummere 1907), where rhymed expression and metrical constraints can lead to the emergence of formula or other forms of repetition. These features may be closely tied to aspects of memory, facilitating performance and ensuring stability over time (Rubin 1995). For metrical genres, repeated word sequences are often verbatim, or rely on the substitution of words that still conform to the formal features of the genre. In fairy

tales, these constraints are somewhat relaxed, yet repeated word sequences such as "Der var en gang" [Once upon a time] are so well established that one can search on precise string patterns.

Although there are well-theorized explanations of why repeated word sequences are integral features of certain rhymed genres such as counting-out rhymes, epics, ballads, and, as noted, fairy tales (Rubin 1995), there are few explorations of repeated word sequences in legend or personal experience narratives (Tangherlini 2003). For these genres, it is important to recognize that repeated word sequences are usually not precise string matches, yet have a reasonably high degree of word-level similarity. A word sequence here is considered to be a linguistic utterance greater than three words. While the methods we present can readily be applied to corpora where repeated word sequences are expected to appear, the goal of this work is to see how repetition occurs in conversational genres where repetition is not a formal feature of the genre. In this short investigation, we identify a series of repeated word sequences, and then explore what may be driving their repetition.

Methodology

Analogous to our methods for identifying phrasal repetition are methods for discovering textual reuse, such as those deployed in plagiarism detection systems. Traditional methods of textual reuse detection analyze words, or strings of words, to find overlaps (MacCartney 2008). Unfortunately, these methods are vulnerable to errors in the target data, including those caused by optical character recognition (OCR) tasks, an inevitable consequence of digitizing historical collections of folk materials given the large variations in the accuracy and quality of the printing. The methods are also sensitive to variations in orthography as the algorithms attempt to find matches at the level of words – misspelled words or variations in orthography generate words that do not match. Consequently, with these earlier methods, the algorithms potentially miss large numbers of interesting "fuzzy" matches. Indeed, the Tang Kristensen collection is notorious for alternate spellings of words. Similarly, the publication pipeline and subsequent digitization have introduced additional errors in spelling. In an attempt to address this problem, we consider letters, rather than words, as the fundamental unit of analysis.

After preprocessing the corpus through string normalization to facilitate fuzzy matching, we analyze each story with a sliding "window" of a set length, using a default value of fourteen words. The window scans a group of words, as a reader might move a magnifying glass over each line, and then increments to the next window according to a "step" value, which we

have set to four words. As a result, the second window starts four words past the start of the previous window, and continues in this fashion, iterating through the collection until the end of the corpus is reached. Each set of fourteen words is then transformed into a digital index, the constitution of which facilitates matching of word sequences in which a match does not require identical strings of characters. The actual processing involves further decomposition of the word sequence into a set of concatenated, overlapping three-word strings, so that the word sequence "Så kom han til Klemmen og vilde have ham til at vise dem igjen" becomes a list containing the substrings "såkomhan", "komhantil", "hantilklemmen", and so on. This set is then transformed into a "hash" – a numerical representation that is likely to share certain characteristics with other word sequences in the corpus that are broadly similar to it – via the application of the MinHash algorithm provided in the datasketch software library (Zhu 2018). This technique, known as Locality-Sensitive Hashing, allows us to efficiently measure textual similarity between millions of narrative segments (Datar et al. 2004). Locality-Sensitive Hashing has shown itself to be particularly suitable to the detection of "near duplicates", a generalization of the problem we address here (Manku 2007).

To ensure that we can display the user interface pages quickly, we bundle the stories into collections of approximately 2,000 words each, resulting in 1,869 bundles (Table 1). Since our bundles do not cross top-level classifications, some bundles are shorter than the 2,000-word limit, yet we can be certain that each bundle includes stories that Tang Kristensen found topically related in some manner. We use these admittedly inadequate classifications to assist in the visual display of the data as part of our heatmap similarity matrix visual navigation and exploration system.

User interfaces

To assist in the discovery and comparison of repeated word sequences, we employ two user interfaces: a similarity matrix heatmap with drill-down (available at http://bit.ly/etk_heatmap), and a side-by-side comparison interface with summary pages for informants, stories, and most commonly repeated word sequences (available at http://bit.ly/etk_intertext).

Similarity matrices compare individual objects against all the other objects in a corpus (Figure 1). The matrix allows one to compare the entire corpus to itself in a visually compact form; in these visualizations, darker colors off the diagonal indicate a correspondence that is not due to self-comparison. In our system, the two axes of the similarity matrix arrange the "bundles" of stories in the order in which they were published (the sequences run from top to bottom and left to right), an ordering that

is roughly chronological and also corresponds to the top-level classifications for each of the collections. "Boxes" around the diagonal reveal similarities within topic regions since the axes are based on the ordering of the volumes.[2] This visualization also provides a helpful method for discovering features that diverge from the diagonal – these divergences represent objects of particular interest, since they may indicate a failure of classification, where items classified under one topical index are found to be similar to items assigned to a different index. The heatmap, with its use of contrasting colors, therefore makes it easy to identify regions of potential interest.

We take advantage of the fact that in standard similarity matrices the two "triangles" of the visualization on either side of the diagonal are identical, and combine two visualizations into a single heatmap. For the lower right of this heatmap, we calculate an n-gram text similarity score across the story bundles. This easily computed measure, known as cosine similarity, provides a coarse overview of potential overlaps between text bundles (Singhal 2001). In this section of the heatmap, pixels are darker if the cosine similarity of the 1/2/3-grams of each pair of bundles is higher than others. In the upper left of the heatmap, we visualize the matched word sequences that we discovered with the locality-sensitive hashing algorithm discussed above. In this half of the visualization, pixels are darker if the number of text-reuse matches between all texts contained in the pair of bundles being compared is higher than others. It is worth noting that most bundle-to-bundle comparisons have zero text-reuse matches, while the matches in n-gram similarities are significantly greater.

The heatmap is interactive, allowing one to select any combination of the four target collections. It also allows one to pan and zoom in to different parts of the matrix. Selecting a spot on the heatmap identifies the story bundles at the intersection of the two axes, and provides access to the texts in those bundles; the "fuzzily" matched word sequences are displayed in bold to the right (Figure 2).[3]

Results of the dynamic identification process from the heatmap similarity matrix interface can then be used to "seed" searches in the second user interface (Figure 3). This component is actually the user interface portion of *Intertext*, an open-source processing tool developed for the study of literary texts that was used to perform the "fuzzy" text reuse analysis discussed in the Methodology section above. Its visualization features allow for side-by-side close reading of identified "fuzzy" text matches (Duhaime

---

[2] In future work, we plan to allow the user to choose different axes or axis orderings based on other classifications of the material.

[3] The software is open source and available for download on GitHub (https://github.com/broadwell/ReuseMapper).

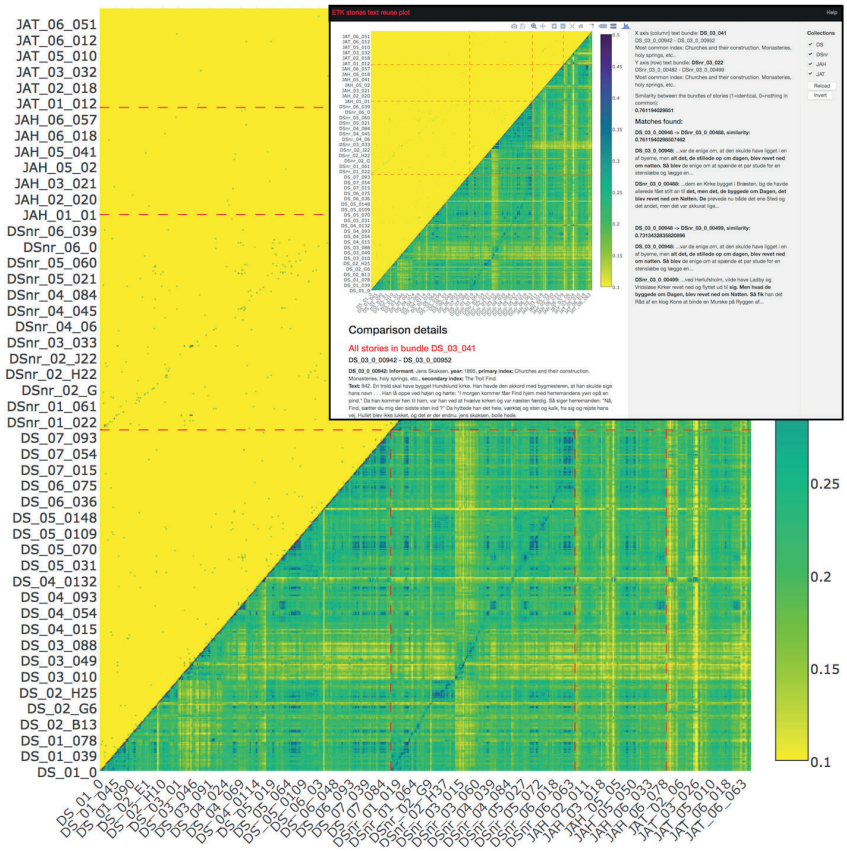

Figure 1: The Heatmap Similarity Matrix user interface. In the upper left of the interactive heatmap graphic are matches based on the locality-sensitive hashing algorithm, visible as dark specks against the yellow background. In the lower right are matches based on n-gram similarities. The surrounding interface (inset) provides drill-down into the selected text bundle below the similarity matrix, while the panel to the right is reserved for information about the pairwise comparison at the location of the pointer on the heatmap.

2018; Malm & Leonard 2016).[4] The two passages, and their immediate preceding and following contexts, are presented side by side on the screen, with the story publication information, date of publication, and, when known, the name of the storyteller. A small circle between the two texts provides a measurement, the Jaccard similarity index, for the match between the highlighted word sequences (Jaccard 1901).

Moving between the two interfaces allows for the examination of similarities at various scales, from corpus-wide investigation to story bundles to individual word sequences, thereby instantiating a key feature of macroscopic research in folklore (Tangherlini 2013a).

[4] The software is open source and available for download on GitHub (https://github.com/ YaleDHLab/intertext).

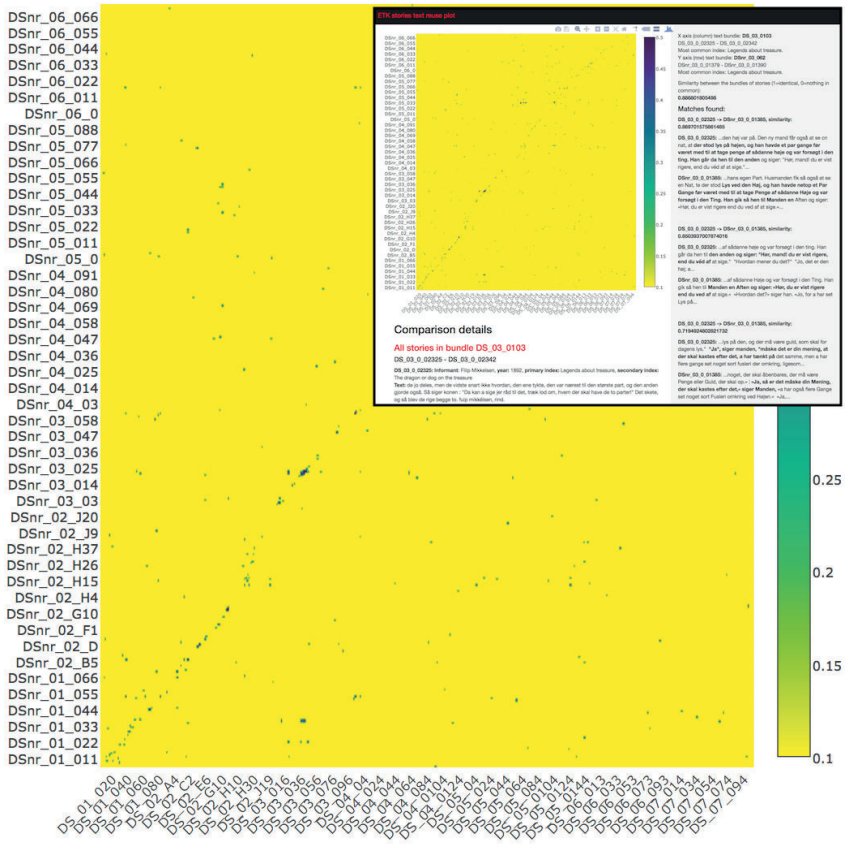

Figure 2: Zooming in on parts of the heatmap allows one to rapidly discover reused word sequences across topic classifications. The reused word sequences are highlighted in the text that appears on the right-hand panel (inset).

Figure 3: The *Intertext* interface displaying side-by-side comparison of two legends, with a "fuzzily" similar word sequence shown in bold.

Results

We discovered four main types of word sequence repetition in the corpus: (1) repetition of formulaic expressions, which is often related to the reporting of direct speech, irrespective of whether that speech is rhymed or not; (2) repetition related to specific types of actions or phenomena which often

cross topic domains; (3) the reprinting of stories in different collections; and (4) repetition based on the consistent expansion of abbreviations in the field diaries. We believe that there may be repetitions that are based on specific storytellers' individual styles, or common use in specific regions, but have not been able to test these hypotheses. One of the more interesting patterns that are particularly clear in the comparison of *Danske sagn* to *Danske sagn, Ny række* is the appearance of thread-like structures off the diagonal of the heatmap that show the appearance of matched word sequences between corresponding topic domains across the two collections, which share the same topical ordering. There may be other broad classes of repetition in the corpus, and we invite others to explore these interfaces and to develop additional testable hypotheses.

The repetition of formulaic expressions constitutes one group of noticeable repetitions in the legend corpus. Here, the term *formulaic* needs to be slightly relaxed from Parry's classic definition of the formula: "a group of words which is regularly employed under the same metrical conditions to express a given essential idea", given that the legend has no metrical conditions (Parry 1930, p. 80). Nevertheless, the matches of these word sequences are very precise, and can include end rhyme between phrases in the word sequence. An example of this is found in various treasure-hunting legends, such as DS 1 654. Here, the word sequence "taget mig så sødt og lagt mig så blødt" [picked me up so carefully and placed me down so softly] is identified in numerous other legends, where it takes similar, albeit not identical, forms (Table 2).

Table 2: Word sequences similar to the target word sequence from a treasure-hunting legend in the first row.

Legend	Informant	Speaker	Word sequence	Classification
DS 1 654	Rebekka Stovstrup	hidden folk (rød dreng)	"Havde du ikke **taget mig så sødt og lagt mig så blødt**, så skulde du have faaet andet at tænke på."	Hidden folk: hidden folk's gold and silver
DSnr 3 1385	Filip Mikkelsen	dog	Så sagde Hunden: "Havde du ikke **taget mig så lettelig og lagt mig så blødt**, så havde du ikke kommet så godt fra det."	Treasure: Dragon or dog on the treasure
DSnr 6 217	Ane Marie Nielsen	dog	Så siger den: "Havde du ikke **tagen mig så sødt og lagt mig så blødt**, så skulde du have fået med mig at bestille."	On the Devil and Pacts with Him: Dog games
DSnr 1 842	Anders Pedersen	calf	Så siger den: "Ja, havde du ikke **tagen mig så sødt og lagt mig så blødt,** skulde du ikke så nemt have rendt med mine Penge."	Hidden folk: Dogs etc. on the treasure
DS 3 2332	LN Bertelsen	dog	"Havde du ikke **taget mig så sødt og lagt mig så blødt**, skulde du ikke have kommet sådan fra det"	Treasure: Dragon or dog on the treasure

The first clause generally includes the rhyme between the two adverbs *sødt* and *bløDt* (with some variation in the first term), while the second part of the word sequence is less consistent. The word sequence is always uttered by a supernatural creature, but the topic classifications for the stories can vary fairly widely. The rhymed quality of the word sequence provides it with a degree of stability, and likely accounts for its broad attestation across the corpus. There are a series of other word sequences of this nature. Frequently reported as direct speech, they are at times tightly connected to a particular place or event, as is the case with a dialect word sequence exclaimed by a dragon during an unsuccessful treasure hunt; as the treasure sinks into a lake, the Dragon taunts, "må a int i Stowhøw blyww, skal I åller mæ å Sørup sø dryvvw" ['if I cannot remain in Stovhøj, then you will never get me out of Sørup lake'], found in DS 3 2316 (Gerhard P Andersen) and DS 3 2317 (Jokum Kristensen).

Word sequences related to the destruction of a church or some other building such as a barn, and the eventual discovery of a location where it can be built, are frequently repeated in the corpus. In these stories, the structure is built by day, and destroyed by night, either by an identifiable actant or, more often than not, by some unseen force. Here, the first word sequence ends with the adverbial "om dagen" and the second with "om natten", offering a modicum of rhyme on the definite common gender enclitic article.

Table 3: Word sequences related to the destruction of churches

Legend	Informant	Word sequence	Classification
DS 3 908	J Larsen	hvad der byggedes om dagen, blev revet ned om natten	Churches and their Construction: Trolls disturb the work
DS 3 970	Anna Rasmussen	og begyndte at bygge; men hvad de byggede op om dagen, blev revet ned om natten,	Churches and their Construction: Finn the Troll
DS 3 802	P Jensen	hvad man byggede op om dagen, blev revet ned om natten, så	Churches and their Construction: The Sign
DSnr 3 497	Mogens Kr. Ottosen	men alt hvad de byggede om Dagen, det blev revet ned om Natten. Så	Churches and their Construction: The two steer, lambs, horses, etc.
DSnr 1 411	Louise Hansen	Men hvad de byggede op om Dagen, blev revet ned om Natten,	Hidden Folk: Cessation of the destruction of mounds

While the top-level classification is generally stable, the motif of the destruction of a church or a building has a degree of flexibility across numerous stories. Consequently, the word sequence similarity approach may be particularly suited to the alignment of motifs across various topical domains.

There are other word sequences that also exhibit crystallization, but do not include rhymed expression. One example of this type of fuzzy-

match word sequence is found in stories where the Devil is forced to act as the fourth wheel of a wagon. In these stories, a wagon is unable to move, or a priest must return rapidly from the site of a conjuring. The Devil is enlisted to serve as a fourth wheel. The act of taking the back wheel off and placing it into the wagon shows a high degree of stability.

Table 4: Word sequences related to the Devil as fourth wheel.

Legend	Informant	Word sequence	Classification
DS 5 796	J. Jensen	Hr. Michael fik ham da til at stå af vognen og **tage det ene baghjul af og lægge det op i vognen**.	Different types of haunting and the conjuring of revenants: Places where revenants have been conjured down
DS 4 1159	Søren Hansen	så må du godt gå ned og **tage det ene hjul af og lægge det op i vognen**	Priests: Driving with three wheels
DS 5 740	C Weiss	at stå af og **tage det nærmere baghjul af og lægge op i vognen**	Different types of haunting and the conjuring of revenants: Driving with three wheels
DS 4 1034	A Olesen	kusken stige af og **tage det ene baghjul ud og lægge dette i vognen**	Priests: Named priests
DSnr 5 413	JP Wammen	Så siger Præsten: "Du skal stå af og **tage det ene Baghjul af og lægge op bag i Vognen**."	Ghosts and Revenants: Driving with three wheels

These word sequences provide additional confirmation of the inadequacy of the one story–one classification indices that are a feature of Tang Kristensen's print collections. Interestingly, the appearance of the word sequence across stories related to both the Devil and ghosts further supports the equivalence that was made in Lutheran Denmark between the Devil and the theologically unsound ghosts of folk tradition (Tangherlini 1999). The divergence of topics given the stability of the word sequences reinforces the suitability of this method for tracing motifs across a range of legend topics.

We can attribute a third group of repeated word sequences to either accidental or deliberate republication of the same story. A secondary set of screens on the comparison browser allows one to see ranked lists of stories or informants whose stories include high word sequence reuse values (Figure 4), with those to the right of the graph (in red) likely being cases of accidental republication of the same story.

An excellent example of this is provided by two stories printed nearly four decades apart, although both were told by Jacob Jensen in 1894 (DS 3 1817 and DSnr 3 1136). In these cases, the level of alignment is very high, owing to the essentially identical nature of the stories; identifying stories of this kind may provide some further insight into the various work practices related to Tang Kristensen's editing and publishing endeavors.

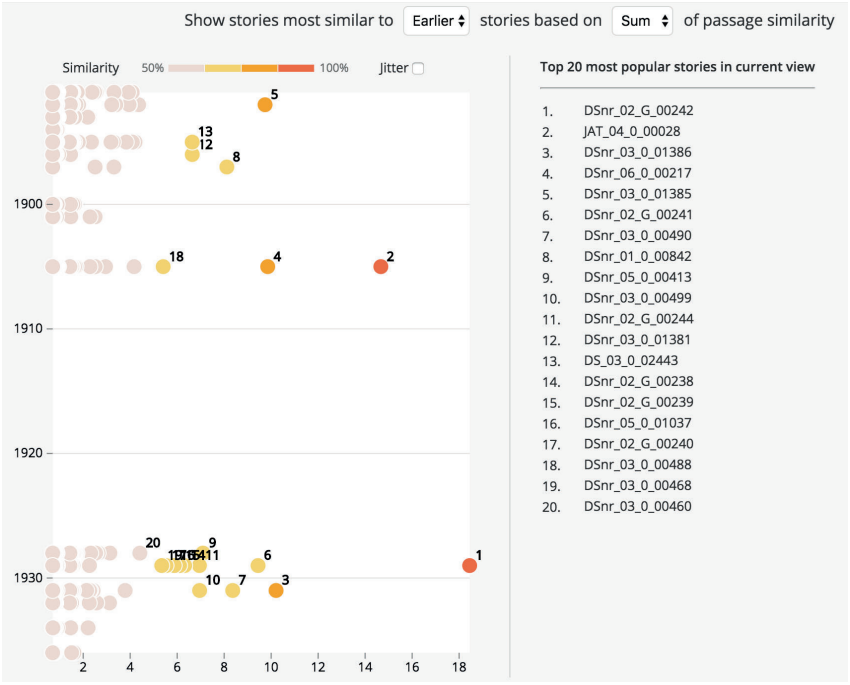

Figure 4: A view of the stories with high levels of word sequence reuse. Stories toward the right of the graph are likely reprinted versions of the same story.

There are also some unexpected discoveries that the method helps surface. DSnr 2 G 242, the story ranked as "most similar" to earlier stories in the collection, is an intriguing example of a "himmelbrev", which had been copied by Nis Jensen Christiansen, and its most frequent match is with DS 2 G 314. This latter record provides additional attestations to word sequences in Nis Jensen Christiansen's letter and other similar letters. Tang Kristensen, in an unusual footnote to the record printed in *Danske sagn*, notes "Denne afskrifts original... er i mit eje… Ved ælde og megen brug er papiret blevet gult og forslidt, og skriften afbleget, så den mange steder næppe kan læses… En næsten ligelydende optegnelse foreligger ved P. Jensen… En tredje opskrift foreligger ved Anton Andersen… En fjerde opskrift ved Maren Bonde haves i flere huse i Vedersø og gjemmes sædvanligvis i Bibelen…" [The original of this copy is in my possession. With age and a great deal of use, the paper has become yellowed and worn, and the writing faded, so in many places it can barely be read… A nearly identical copy is with P. Jensen… A third copy with Anton Andersen… A fourth copy is with Maren Bonde and can be found in many houses in Vedersø, and it is usually kept in the Bible] (DS 2, pp. 331–332). The Nis Jensen Christiansen version thus appears to be a later addition to this emerging collection of folk letters,

a genre often overlooked in folkloristics, but one that had caught Tang Kristensen's attention.

A final group of repeated word sequences appears to be related to the expansion of field diary abbreviations. Tang Kristensen used abbreviations both in his field notes, which were then expanded in fair copy, and in his published collections, where he truncated certain records through the use of ellipses. While these latter examples are not discoverable in the word sequence repetition system here (since they depend on the elision of repeated word sequences or segments), the former may be discoverable. Unfortunately, given limitations related to the interface, there is no immediately accessible manner in which one can compare repeated word sequences to their original manuscript form.

Conclusion

The current study has shown how the detection of repeated word sequences in folklore genres not generally considered to be likely to include such linguistic consistencies can lead to productive avenues of inquiry. Some of the repeated word sequences are linguistically similar to the formula identified by Parry and explored later by Lord and other researchers interested in rhymed genres (Parry 1930; Lord 1960). Rubin suggests that this type of rhyming is closely related to the physiological features underlying human memory (1995). Other repeated word sequences appear to be closely linked to specific motifs, and may be reflective of the stability of motifs across the tradition group. While there are indications that some repeated word sequences are closely correlated to individual or regional storytelling style, we reserve those investigations for future interface developments that support geographically based or informant-based queries into the underlying similarity matrices. A final type of text reuse provides insight into collecting and editorial practices, offering an opportunity to discover stories where the collector's hand has brought the published versions closer to each other in terms of phrasing.

There are numerous avenues for future development of this approach. Given the linguistic similarity across Scandinavia, it might be possible to discover matching word sequences across these closely related traditions. Such investigations could help illuminate aspects of the circulation of stories throughout the Nordic region. Other refinements related to the user interfaces might lead to a more fluid experience. An interface for selecting a word sequence from any legend, and then revealing close matches to that word sequence, would help researchers who have a particular target legend or group of legends in mind. Integrating the two interfaces we present here would allow for a seamless movement across scales of investigation. Similarly, the incorporation of geographic mapping of the repeated word se-

quences would assist in the study of geographically correlated phenomena. Finally, the ability to select an informant or a class of informants would allow researchers to explore questions related to individual repertoire, or comparisons across different groups of informants. Incorporating more sophisticated dating of the legends (currently a story is tied to the date of publication of the volume in which it appeared, and not the date of collection) might offer an opportunity to explore the dynamics of word sequence similarities in a well-documented tradition. Although word sequence similarity is a time-tested approach in the study of certain genres in folklore, from the epic and ballad to the fairy tale, our brief investigations here reveal the intriguing questions that emerge when the approach is applied to the seemingly infinitely flexible genres of legend and personal experience narrative.

References

Abello, James, Broadwell, Peter M., & Tangherlini, Timothy R., 2012: Computational Folkloristics. In: *Communications of the Association for Computing Machinery,* 55(7), pp. 60–70.

Broadwell, Peter M., & Tangherlini, Timothy R., 2012: TrollFinder: Geo-Semantic Exploration of a Very Large Corpus of Danish Folklore. In: *Proceedings of the Workshop on Computational Models of Narrative*. Istanbul.

— 2016: WitchHunter: Tools for the Geo-Semantic Exploration of a Danish Folklore Corpus. In: *Journal of American Folklore,* 129(511), pp. 14–42.

Broadwell, Peter M., Mimno, David, & Tangherlini, Timothy R., 2017: The Tell-Tale Hat: Surfacing the Uncertainty in Folklore Classification. In: *Journal of Cultural Analytics*, 1(2). DOI: 10.22148/16.012

Christiansen, Palle O., 2013: *Tang Kristensen og tidlig feltforskning i Danmark: National etnografi og folklore 1850–1920.* Copenhagen.

Datar, M., Immorlica, N., Indyk, P., and Mirrokni, V. S., 2004: Locality-sensitive hashing scheme based on p-stable distributions. In: *Proceedings of the Twentieth Annual Symposium on Computational Geometry*, pp. 253–262. ACM.

Dégh, Linda, 1969: *Folktales and Society: Story-Telling in a Hungarian Peasant Community.* Bloomington.

Duhaime, Douglas, 2018: *Intertext: Detect and visualize text reuse in large document collections*. https://github.com/YaleDHLab/intertext

Gummere, Francis Barton, 1907: *The Popular Ballad*. New York.

Gunnell, Terry, 2014: *Sagnagrunnur.* http://sagnagrunnur.com/en/

Jaccard, Paul, 1901: Étude comparative de la distribution florale dans une portion des Alpes et des Jura. In: *Bulletin de la Société Vaudoise des Sciences Naturelles,* 37, pp. 547–579.

Kristensen, Evald Tang, 1880: *Sagn fra Jylland.* Copenhagen.

— 1891–1894: *Gamle folks fortællinger om det jyske almueliv, som det er blevet ført i mands minde, samt enkelte oplysende sidestykker fra øerne*. 6 volumes. Kolding (JAH).

— 1892–1901: *Danske sagn, som de har lydt i folkemunde, udelukkende efter utrykte kilder.* 7 volumes. Århus (DS).

— 1900–1902: *Gamle folks fortællinger om det jyske almueliv, som det er blevet ført i mands minde, samt enkelte oplysende sidestykker fra øerne. Tillægsbind.* 6 volumes. Århus (JAT).

— 1928–1939: *Danske sagn, som de har lydt i folkemunde. Ny Række.* 7 volumes. Copenhagen (DSnr).
Lord, Albert B., 1960: *The Singer of Tales*. Cambridge.
Malm, Mats, & Leonard, Peter, 2016: Marknadens intertextualitet: Kulturarv och återbruk 1840–1900. In: Gunnel Furuland, Andreas Hedberg, Jerry Määttä, Petra Söderlund, & Åsa Warnqvist (eds.), *Spänning och nyfikenhet: Festskrift till Johan Svedjedal*. Möklinta. Pp. 28–36.
MacCartney, Bill, Galley, Michel, & Manning, Christopher D., 2008: A phrase-based alignment model for natural language inference. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
Manku, Gurmeet Singh, Jain, Arvind, & Das Sarma, Anish, 2007: Detecting near-duplicates for web crawling. In: *Proceedings of the 16th International Conference on World Wide Web*. ACM.
Parry, Milman, 1930: Studies in the Epic Technique of Oral Verse-Making. In: Homer and Homeric Style. In: *Harvard Studies in Classical Philology,* 41, pp. 73–148.
Pentikäinen, Juha, 1978: *Oral Repertoire and World View: An Anthropological Study of Marina Takalo's Life History*. Helsinki.
Rubin, David C., 1995: *Memory in Oral Traditions: The Cognitive Psychology of Epic, Ballads, and Counting-out Rhymes*. Oxford.
Singhal, Amit, 2001: Modern Information Retrieval: A Brief Overview. In: *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4), pp. 35–43.
Skott, Fredrik, 2017: *Sägenkartan*. http://www.sprakochfolkminnen.se/om-oss/kartor/sagenkartan.html#/places
Tangherlini, Timothy R., 1999: "Who ya gonna call?" Ministers and the Mediation of Ghostly Threat in Danish Legend Tradition. In: *Western Folklore*, 57, pp. 153–178.
— 2003: And all anyone heard: Crystallization in paramedic storytelling. In: Lotta Tarkka (ed.), *Dynamics of Tradition: Perspectives on Oral Poetry and Folk Belief*. Helsinki. Pp. 343–358.
— 2008: "And the wagon came rolling in…": Legends and the Politics of (Self-)Censorship in Nineteenth Century Denmark. In: *The Journal of Folklore Research,* 45, pp. 241–261.
— 2013a: The Folklore Macroscope: Challenges for a Computational Folkloristics. The 34th Archer Taylor Memorial Lecture. In: *Western Folklore*, 72(1), pp. 7–27.
— 2013b: *Danish Legends, Folktales and Other Stories.* Seattle.
Tangherlini, Timothy R., & Broadwell, Peter M., 2014: Sites of (re) Collection: Creating the Danish Folklore Nexus. In: *Journal of Folklore Research,* 51(2), pp. 223–247.
Zhu, Erik, 2018: *datasketch: Probabilistic data tools for Big Data.* https://ekzhu.github.io/datasketch/minhash.html

Sammenfatning

Repetition af ordsekvenser er et kendetegn for mange folkloristiske genrer, især genrer som indeholder rim (f.eks. viser, gåder, børnerim, sanglege, osv.). Formularer og faste udtryk er også kendt fra eventyr og andre nært

beslægtede genrer. Samtalebaserede genrer, som sagn, opfattes som mere frie i sine udtryk, således at repetition af ordsekvenser kun forekommer sjældent. Det er et åbent spørgsmål, hvorvidt ordsekvenser repeteres i sagn som et kendetegn af den enkelte fortællers fortællestil, som kendetegn for fortællinger i en region eller som del af en større tradition. Hvis repetition af ordsekvenser i sagn kan identificeres, kan de fundne sekvenser danne grundlag for nye sammenligninger af sagn på tværs af de nuværende emnebaserede klassifikationer. Samtidig kan de fundne sekvenser give indblik i sprogbrug blandt enkelte fortællere eller større grupper. Man kan evt. bygge videre på disse resultater til at sammenligne sprogbrug blandt forskellige fortællergrupper (f.eks. mandlige og kvindelige fortællere) og stille helt nye spørgsmål: Forekommer repetition hyppigere blandt mænd end kvinder? Findes repetition oftere i sagn om overnaturlige væsener? Samtidigt kan disse undersøgelser hjælpe med at uddybe vores forståelse af de indsamlings- og redigeringsprocesser, der knyttede sig til samlingernes udgivelse.

I vores artikel beskriver vi en metode til at finde repetition af ordsekvenser i et stort korpus af sagn. Vi bruger Evald Tang Kristensens udgivne samling af danske sagn (*Danske sagn* and *Danske sagn. Ny række*) (Kristensen 1892–1901; Kristensen 1928–1939) samt hans samling af beskrivelser af det jyske almueliv (*Gamle folks fortællinger om det Jyske almueliv*) (Kristensen 1891–1894; Kristensen 1900–1902). Tilsammen giver det en samling på 31.088 sagn fortalt eller indsendt fra 4.242 navngivne fortællere eller lokale indsamlere. Alle historierne blev samlet i de sidste årtier af det nittende århundrede og de første to årtier af det tyvende århundrede. Historierne blev ordnet af Tang Kristensen efter hans eget klassifikationssystem, som indeholder 72 overordnede kategorier og 774 underkategorier. I vores arbejde er vi mest interesserede i at finde frem til historier, der kan placeres i flere kategorier eller som går på tværs af mange kategorier (Abello, Broadwell, Tangherlini 2012; Broadwell, Mimno, Tangherlini 2017). Arbejdet her udvider vores muligheder for at identificere beslægtede historier og med at finde frem til ordsekvenser, der forekommer i historier på tværs af de forskellige kategorier.

Vores metode er baseret på tidligere studier, som fokuserer på genbrug af tekst. Dette er hyppigst anvendt i forbindelse med systemer, der bruges til at afdække plagiat. Mange af de eksisterende systemer sammenligner ord eller ordstrenge for at finde steder, hvor to tekster er overlappende. Denne metode er meget følsom over for de OCR-fejl og ortografiske særheder, der ofte forekommer i de historiske data. Resultaterne af denne metode er således ikke tilstrækkelig til at danne et præcist billede af sprogbrug og repetition i sagn. Ordbaserede metoder er mere præcise og egner sig bedre til at identificere »ord for ord«-repetition; tit springer de over upræcise, men interessante sekvenser. Vi anvender en alternativ metode, som søger efter en-

heder. Metoden, »Locality-Sensitive Hashing«, bruger et vindue af fjorten ord, som flyttes gennem en historie med et fire-ordstrin. Da vi sammenligner enheder og ikke ord, er det nemmere for os at inkludere nærtræffere, hvor *ø* og *ö* eksemplevis kan matches. Vi kan effektivt undersøge mange millioner af ordsekvenser og finde de sekvenser, som danner et match over en vis grænse (en grænse som brugeren selv kan bestemme).

Forskere kan dermed finde og sammenligne fundne ordsekvenser ved hjælp af to brugergrænseflader. I den ene viser vi en »similarity matrix«, hvor brugeren hurtigt kan overskue hele samlingen – de mønstre som ligger over eller under diagonalen er historier, hvori ordsekvenser gentages. Grænsefladen er designet, så brugeren kan zoome ind på de steder i matrixen, som har interesse, og brugeren har desuden mulighed for at finde frem til både de underliggende historier og ordsekvensrepetitionerne. I den anden grænseflade får brugeren en grafisk repræsentation over de historier, som udviser hyppigt genbrug af ordsekvenser. Brugeren kan vælge en historie og klikke gennem til en browser, hvor den valgte historie vises side om side med andre historier, som deler en eller flere sekvenser med den valgte historie. Jaccard similarity score viser, i hvor høj grad sekvenserne matcher.

Da vi anvendte vores metode over det samlede korpus, viste der sig at være fire slags ordsekvensrepetition i de undersøgte sagn: (1) repetition af formular, ofte som direkte tale; (2) repetition af beskrivelser af forskellige handlinger, som forekommer i sagn klassificerede underflere forskellige rubrikker; (3) gengivelsen af samme historie i forskellige bind af samme eller andre samlinger; (4) repetition af udtryk, som blev forkortet i dagbøgerne, men udvidet i den trykte samling.

En del af de ordsekvenser, som systemet fandt, var faste udtryk som de, der findes i historier om skattejagt. Den overnaturlige hund, som sidder ovenpå skatten, siger til en af skattejægerne: »Havde du ikke taget mig så sødt, og lagt mig så blødt, så skulde du have fået andet at tænke på.« Her er de rimende fraser nok årsagen til, at repetitionen er temmelig enslydende i de mange historier, hvori sekvensen forekommer. Andre sekvenser er mindre nøjagtige, men alligevel iøjnefaldende, især da de findes i historier, som er blevet klassificeret under andre rubrikker og derfor trykt i forskellige bind af samme samling. Vi finder også en del historier, som er blevet genoptrykt, og nogle få tilfælde, hvor vi kan konstatere, at repetitionen skyldes renskrivningsprocessen, f.eks. i form af udvidede forkortelser eller indsætning af ordsekvenser, hvor dagbøgerne kun anfører tre prikker, mens den fulde sekvens står i den trykte udgave.