"En temmelig lang fodtur": hGIS, Text Mining, and Folklore Collection in 19th Century Denmark

Ida Storm, Lund University & Timothy R. Tangherlini, UCLA

In Scandinavia, the folklore collection of the nineteenth and early twentieth centuries coincided with rapid changes in political, economic, and social organization, and resulted in national collections of extraordinary scope. Although some later folklorists have expressed skepticism about the usefulness of these collections as resources for the study of folklore, ethnography and intellectual history, this skepticism is often based on incorrect notions of how these collections came to be, rather than a deep exploration of the actual practices of the collectors themselves. Methods that support detailed analysis of these collecting practices result in a more nuanced view of the role of folklore and fieldwork in the imagining of the nation. These methods can also help delineate the role folklorists played in developing ethnographic perspectives on the impact of social, political, economic and technological change on the lives of normal people. Building on our earlier work, we show how techniques from historical Geographic Information Systems (hGIS) and text mining, wedded to time-tested archival research methods, can be used to reveal the complex dynamics behind a folklore collection. By detailing the routes taken by the Danish folklore collector Evald Tang Kristensen (1843-1929) over the course of his fifty-year active career, we trace how changes to transportation infrastructure impacted his collecting and how his attitudes toward fieldwork developed over time.

Storm, Ida & Tangherlini, Timothy R.. ""En temmelig lang fodtur": hGIS, Text Mining, and Folklore Collection in 19th Century Denmark" *HUMAN IT* 14.2 (2018): 43–81. © The authors
Published by the University of Borås

Keywords: folklore, historical GIS, topic modeling, text mining, intellectual history

In Denmark, and in many parts of Scandinavia, the nineteenth century[1] is often considered to be "the golden age" of folklore collecting (Ellekilde, 1946, 14). This period coincided with rapid changes in political, economic, and social organization across the region. Although the general contours of folklore collecting efforts in Scandinavia have been relatively well described, and the output of these collecting efforts are well documented in folklore archives across the region, the processes by which specific collections came into being can be difficult to trace. Palle Ove Christiansen, in his work on Denmark's most prolific folklore collector, Evald Tang Kristensen, suggests that an analysis of a collector's fieldwork practice, predicated on a close reading of memoirs and letters, along with attention to field work collecting routes and fieldnotes, can lead to insight into the various factors that influence the creation of a folklore collection and help explain the conceptual shift from the Romantic notion of "collecting" to the modern notion of "fieldwork" (Christiansen, 2011; Christiansen, 2013, 134-135). In turn, such an understanding can provide a more nuanced view into the shifting role that folklore and the ethnographic description of agricultural life played first in the imagining of the nation and, later, in conceptualizing regions and the "almue". In addition, the consideration of the process of creating a collection can also be used to understand the impact of economic, political, social and technological currents on the documentation of the everyday lives of normal people.

Christiansen notes that the analysis of a folklore collector's life work can be quite daunting given the scope of some of these collections, and the difficulty of working with largely handwritten documents dispersed across multiple archives and libraries (Christiansen, 2013). In the following pages, we show how techniques from historical Geographic Information Systems (hGIS) and text mining can be leveraged to help scale time tested archival approaches to describing a folklore collector's entire collecting enterprise. This computational approach bolsters researchers' abilities to consider changes in fieldwork practices that may be influenced by any number of confounding factors, such as political changes (*e.g.* shifting borders due to war), technological changes (*e.g.* the rapid development of passenger rail), agricultural changes (*e.g.* the emergence of dairy cooperatives), intellectual changes (*e.g.* the Modern Breakthrough), and social changes (*e.g.* fissures in religious communities) (Gregory & Ell, 2007; Storm *et al.*, 2017). By detailing the routes taken by Tang Kristensen (1843-1929) over the course of his fifty-year active career, we trace not only his selection biases for geographic areas (and by extension, social and economic classes), but also the impact that intellectual currents, political developments and changes to transportation infrastructure had on his collecting. By aligning these collection routes with his field notebooks, we can make estimates related to fieldwork productivity both by region and by time period. By applying simple text mining techniques to his descriptions of these collecting trips, we can further refine our understanding of his field work practice. The changes we discover in Tang Kristensen's field collecting practices can help us understand broader shifts in the conception of the emerging Danish "nation", as well as help us identify the emerging regionalism that bolstered, for example, the development of the "hjemstavnslitteratur" [regional literature] movement in the early 1900s.

Although many Nordic archives relied on networks of collectors who lived dispersed across the country to create their collections, other prominent collections such as those of Peter Christian Asbjørnsen and

Jørgen Moe in Norway, and Per Arvid Säve in Sweden, were the result of experienced folklorists traveling to meet and interact with individual storytellers (Palmenfelt, 1993).[2] In our previous work with the Tang Kristensen collection, we have revealed how the relationship between where people lived and the places they mentioned in their stories bolster our understanding of how people generated complex representations of the environment in which they lived and worked (Broadwell & Tangherlini, 2017; Tangherlini & Broadwell, 2016). In this work, we align Tang Kristensen's trips with descriptions of his collecting practice and generate summary statistics of what he collected. Although Tang Kristensen traveled over 67,000 km, largely on foot, while creating the collection, the trips were not uniformly spread over his career, nor were they of uniform length. Rather, the length, duration, and number of trips varied – at times dramatically – as did the amount of folklore that he collected. These variations were likely influenced by his changing goals as a folklorist, changes in his employment as a school teacher, and changes in his family life, along with profound changes in the organization of the Danish countryside, not least the dramatic changes in transportation infrastructure and the larger agricultural economy.

A great deal of the archival and printed material needed to reconstruct Tang Kristensen's field collecting trips is "noisy": the descriptions are incomplete, the annotations in his field diaries are at times illegible, and the names of the places he visits are ambiguous. We describe methods for working with this noisy data that allow us to provide a "best guess" as to how, when and where Tang Kristensen traveled in Denmark. For each segment of a collecting trip, we assign a primary means of transportation, either directly or through inference. We project these routes onto appropriate historical base maps to visualize his movement through the countryside.[3] We develop aggregate statistics that allow us to understand, at a granular level, his collecting habits, including the number of stops he made on any given route, and the various means of transportation he used. These statistics can be analyzed at a series of scales: (a) by individual trip; (b) trips aggregated by month or months of collection; (c) trips aggregated by year of collection; and (d) trips aggregated by Tang Kristensen's place of residence. We also align the field trips with descriptions of those trips from his memoirs (Tang Kristensen, 1923-1927).[4] These field trip descriptions are modeled in two ways: (i) as word clouds revealing the highest frequency words for each collection period, and (ii) as probabilistic topic models devised for periods of collection and for the collection in its entirety. In all, we map two hundred and sixty-seven trips, starting in 1868 and ending in 1916, spread across five distinct work periods defined by his place of residence: Gjellerup (1866-1876), Faarup (1876-1884), Brandstrup (1884-1888), Hadsten Station (1888-1897), and Mølhom (1897-1929).[5]

Resources

There are two primary resources for determining the routes that Tang Kristensen followed and the places he visited on his field collecting trips: his memoirs, *Minder og Oplevelser*, published in four volumes (1923-1927), and his field diaries (*DFS 1929/16*) in which he wrote down the stories, songs and other aspects of folk expressive culture that he collected from his informants. The memoirs are unusually linear in construction, and the description of field trips, taken largely from letters he sent home and from his own reconstruction of his trips based on his field notebooks and pocket notebooks (*lommebøger*), provide an excellent first approximation of the extent of his travels, the dates of travel, and his modes of transportation. Although we do not explore it in detail in this paper, the memoirs also provide important information about the social network, particularly local school teachers, on which he relied to find local storytellers and singers.

By way of illustration, consider a typical yet relatively short field trip. The description of the trip from the memoirs reads as follows:

> *Jeg gik over Hadbjærg og Rud og Voldum, og var saa inde hos lærer Sørensens, inden jeg om eftermiddagen gik til Søby. Sørensen mente, at jeg skulde gaa om i byen og tale med en husmand der, Johannes Pedersen. Han boede i et langt hus i den sydøstlige del af byen, altsaa østen for bækken. Det gjorde jeg da ogsaa, og jeg fik et par smaating skrevet op, men da tiden ikke tillod mere, gjorde jeg aftale med manden om at komme igjen, naar jeg gik tilbage. Jeg maatte nu skynde mig, da mødet skulde være den samme aften. Der blev nu ikke særlig tilstrømning af folk, og jeg havde ikke nogen synderlig glæde af mødet. Jeg bad folkene til slutning give mig et eller andet vink, hvis der var nogen af deres bekjendtskab, der var gode til at fortælle, men det førte nu ikke til noget (Tang Kristensen, 1923-1927, vol 3, 299).*
>
> [I walked on past Hadbjærg, and Rud, and Voldum, and then I stopped in at Teacher Sørensen's place, before I walked over to Søby in the afternoon. Sørensen felt that I should go through town and speak to a cotter, Johannes Pedersen, who lived there. He lived in a long single-length house in the south-eastern part of the town, a little east of the stream. So I did as he suggested, and I managed to get a few small things written down but, since time didn't allow for more, I made an agreement with him that I'd come again when I was on my way back. I now had to hurry, since the meeting was to be held that same evening. There wasn't really an influx of people and I didn't get much out of that meeting. I asked the people to give me a wave if any

> of their acquaintances were good storytellers, but that didn't lead to anything.]

The field trip can be extracted, aligned with the field diary, fit to historically accurate maps, transportation networks, and disambiguated placenames, and represented by the following map (fig. 1):

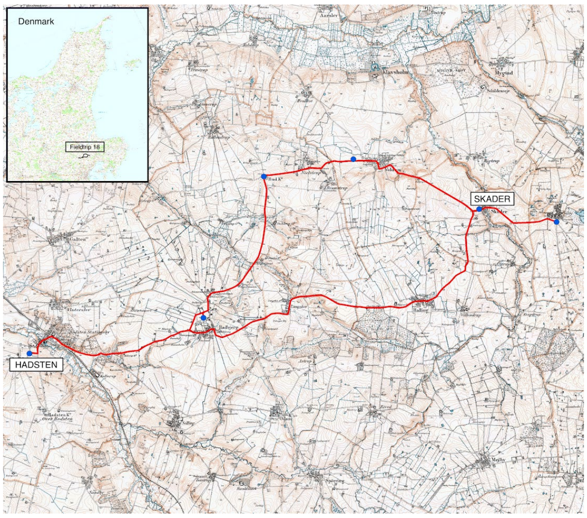
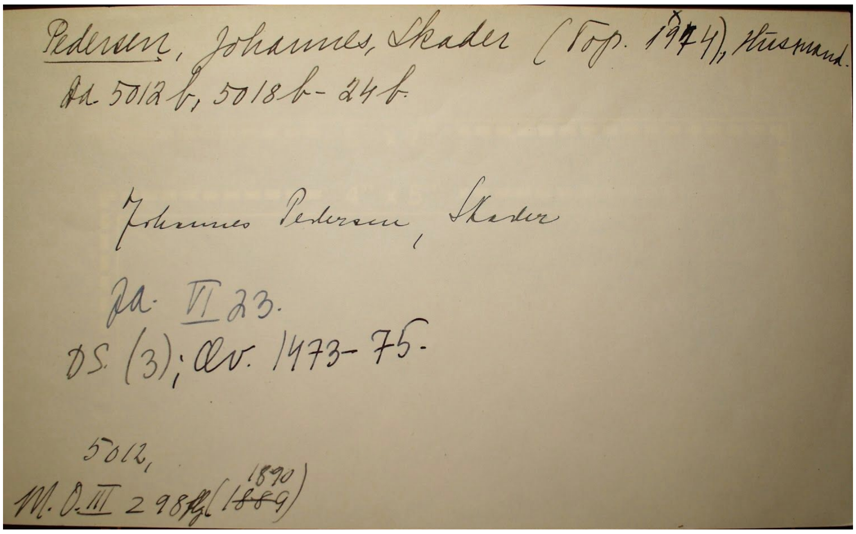


Figure 1. Map of field trip 18. The town of Skader, where Johannes Pedersen lived, is highlighted as is Hadsten, the start and stop point of the trip.

Embodied Action

The informant, Johannes Pedersen, appears in the informant index with a card which details his place of residence, his occupation (husmand [cotter]), the field diary pages that record what he told Tang Kristensen, the various collections in which edited versions of these stories or descriptions are published, and a reference to where he appears in the memoirs (Tang Kristensen, 1923-1929, vol. 3, 298ff), along with the year in which Tang Kristensen visited him (figure 2):



Pedersen, Johannes, Skader (Top. 1974), Husmand.
Da. 5012 b, 5018 b - 24 b.

Johannes Pedersen, Skader

Da. VI 23.
DS. (3); Æv. 1473-75.

5012,
M. O. III 298ff. (~~1889~~ 1890)

Figure 2. The informant card from DFS 1929/129 with information about Johannes Pedersen.

Tang Kristensen's field diary recordings for this brief encounter appear on page 5012b, and include three stories told by Johannes, the first two about hidden folk, and the third a description of a forest that had disappeared (fig. 3):

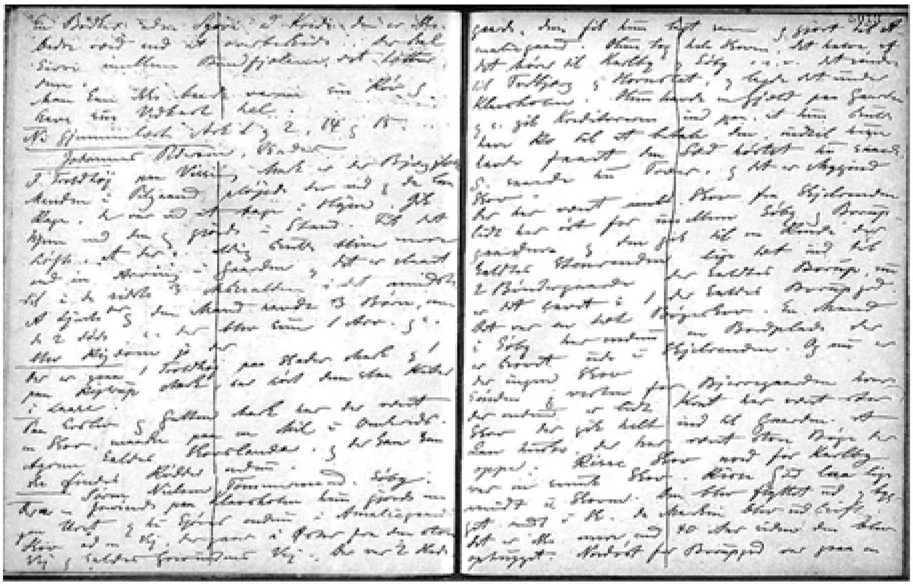


Figure 3. Records for Tang Kristensen's first encounter with Johannes Pedersen, which start one quarter the way down on 5012b (left hand side), and end five lines from the bottom.

On his way home from the meeting to which he alludes in his description of the field trip (right hand side of the route map), Tang Kristensen visited Johannes a second time, and these stories are recorded several pages later in the field diary (5018b-5024b). With the various people he meets on this trip, Tang Kristensen collects folklore that occupies eleven two-sided pages in his field diaries.

As described in earlier work, we manually tag all of the field trips for stops, means of transportation, and dates based on the descriptions in the memoirs (Storm *et al.*, 2017). Importantly, these brief descriptions also include Tang Kristensen's evaluation of the areas through which he traveled, and his impressions of the people he met. By themselves, these evaluative comments are minimal but, when considered in the aggregate, they provide interesting insight into his feelings about classes of storytellers, regions, and groups of people (such as the Inner Mission groups whom he disliked intensely). These brief fieldwork descriptions also provide Tang Kristensen's own estimation of the value of the material he collected on these trips. Finally, the memoirs provide us with important biographical information, thereby allowing us to break Tang Kristensen's productive field collecting work into various periods.

The field diaries can be aligned with field trips based on dates, places, and people. With this alignment, we are also able to provide a coarse estimate of the amount of folklore collected on any single trip. These aggregate page counts per field trip are helpful for providing a first pass at the productivity of a particular field trip, although they should only be seen as loose proxies for the collecting itself. Currently, we have not aligned all of the entries in the informant index (DFS 1929/129) with the field diary pages, and therefore the pages assigned to any particular field trip are approximate. Furthermore, this method of estimating fieldwork productivity does not include Tang Kristensen's loose-leaf collections (DFS 1929/12), and therefore likely underestimates his productivity both in the aggregate and for individual trips.[6]

One of the central challenges in hGIS work is place identification. Our work has been helped considerably by the multi-year project at the University of Copenhagen, Digdag (Digitalt atlas over Danmarks historisk-administrative geografi), that developed robust resources for Danish hGIS work, including historically accurate and unusually thorough geolocations for all Danish place names (Digdag, 2009). Using Danmarks stednavne (Denmark's placenames) as a starting point, we developed an address locator for use in the GIS software we used for this project with over one hundred and fifty thousand geolocated place names in Denmark.[7] Importantly, the *Danmarks stednavne* dataset includes the periods in which particular names were used for particular places, as well as variant spellings for those names, allowing us to disambiguate placenames using time data (*i.e.* we could ignore matches that were not in use during the time of the collecting trip). The Danish Cadastral Survey's release of a series of very high resolution maps from the late nineteenth century further allowed us to align a modern road network derived from Open Street Maps with the extant road network from the period during which Tang Kristensen was collecting.

Although the development of the railway network in Denmark started in the 1840s, it gained significant momentum in the 1860s. Up through the end of the nineteenth century, thousands of kilometers of railway were laid, and private railways were coordinated into a much larger national network. In prior work, we developed a series of hGIS shapefiles describing the routes and stations of the Danish railway network. The data includes the period during which particular sections of the railway were in operation, as well as when stations opened along these railway sections. It is worth noting that the railway in Jutland began with the Aarhus-Randers line in 1862, and continued to develop rapidly precisely during the period when Tang Kristensen was engaged in his folklore collecting. We augment the railway network with ferry and steamboat routes derived from historic ferry schedules, and maps tracing the routes of these ferry and steamboat services. Given the number of islands in Denmark, there is a close connection between the railways and the ferry lines, particularly with the development of railway ferries, starting in 1883.

Methodology

We devised two three step pipelines of interrelated work: (a) route segmentation and field diary alignment; (b) transportation assignment with inference; (c) productivity calculation; (d) textual preprocessing of field trip descriptions; (e) simple word cloud modeling of field trip descriptions based on relative word frequency counts; and (f) LDA probabilistic topic modeling of field trip descriptions. We describe each step below.[8]

Route segmentation and field diary alignment

After we had developed the routes for the two hundred and sixty-seven trips that we identified in *Minder og Oplevelser*, we split each route into inter-stop segments. By doing so, we could develop more detailed statistics regarding segment length and travel mode. To divide the field trips into inter-stop segments, we loaded a previously calculated field trip shapefile into ArcMap. The routes were then extracted into inter-stop segments by starting an edit session on the individual routes, with segments defined as the shortest path between any two stops. This processing step resulted in a series of route segments with identifiable start and stop points for each field trip. Tang Kristensen reported sporadically and inconsistently on the dates of his travels, but often made some mention of one or more specific dates for route segments. This sporadic mention of dates allowed us to assign specific dates, often at the level of day, month and year, to route segments, thereby providing a fairly precise time window for each trip. Individual informants were also mentioned in the memoirs in the context of particular stops; this allowed us to align individual trip segments with specific field diary pages through the intermediary step of consulting the informant cards (see fig. 2 above for an example). Working from the "inside out" in this manner we aligned field trips with the field diaries, and subsequently generated estimates of the first page and last page in the field diaries associated with each collecting trip.

Transportation inference

Although Tang Kristensen preferred to walk on his trips, and was quite proud of his ability to walk long distances, he also wanted to cover as much ground as possible. Consequently, there was an ongoing tension between his love of walking and his desire for coverage, which required him to make use of other means of transportation. We have been able to identify five means of transportation that he used: (1) walking; (2) horse wagon, such as the postal wagon or day wagon; (3) train; (4) ferry or steamboat; (5) bicycle. There is only one segment assigned to bicycle, and Tang Kristensen notes that it was the only time he ever rode a bicycle (Tang Kristensen, 1923-1927, vol. 4, 202). Tang Kristensen makes inconsistent mention of the means of transportation he used for specific route segments, and therefore we needed to infer the means of transportation for many of these segments. Certain inferences were quite simple. To cross water, for example, he had to travel by boat. Either he traveled on steamboats, such as the one that plied the route from Aarhus to Copenhagen, or ferries, such as the train ferry that linked Nyborg on Fyn to Korsør on Sjælland which started service in 1883. In a few cases, he used sailboats or enlisted the help of a local fisherman. Although in later years automobiles began appearing in Denmark, he makes no mention of them in his memoirs.

We attempt to assign all inter-stop segments to a primary mode of transportation. First, we assign the primary mode of transportation between stops to whatever Tang Kristensen identifies as that mode. So, for instance, if he says, “Fra Rødding gik jeg saa til Krejbjærg” [From

Rødding I then walked to Krejbjærg] (Tang Kristensen, 1923-1927, vol. 3, 70), we assign that segment to "walking." We perform this same task for all transportation related descriptions in all of the field trip descriptions.

We recognize that Tang Kristensen may have made use of multiple modes of transportation to get from a point A to a point B. For example, in one instance, he describes a segment that starts with him walking to the train station in Odense to catch the train to Ringe: "Næste Morgen gik jeg ned til Toget, men var nær kommen for sildig, da Togplanen var bleven ændret, og Toget gik nu 10 Minutter før. Saa kom jeg til Ringe og søgte ind til Lærer Munch" [The next morning, I walked down to the train, but almost came too late, as the train schedule had been changed, and the train now left ten minutes earlier. Then I got to Ringe, and sought out Teacher Munch] (Tang Kristensen, 1923-1927, vol. 4, 363). Since we do not break inter-stop segments into individual transportation mode segments, we assign the mode of transportation based on the mode that carried him the farthest distance on that segment, here the segment from Odense to Ringe. As a result, this segment is assigned to "train", even though he starts by walking to the station in Odense, and presumably walks from the train station in Ringe to Teacher Munch's house.

For unspecified segments, we check if Tang Kristensen could have possibly taken the train. If not, we assign "walk" to the segment if the inter-stop segment distance is less than 20 km and "wagon" if the segment is more than 20 km. For any journey over water, we assign "ferry/steamboat." We infer that Tang Kristensen preferred to take the train whenever available, given the train's comparative comfort and the relatively predictable departure and arrival times (apart from the trip segment described above!). The train could only be used if there were open stations at either end of the segment, or along the segment. To ascertain whether Tang Kristensen could have taken the train for the inter-stop segment or some portion of the segment, we perform a series of tests. First, we use our train data to determine whether the track segment that lay along the route was in operation at the time of the field trip. Then we check which stations were in operation along the segment during the field trip time. We draw a two kilometer buffer around the open stations, as two kilometers seems to be a reasonable estimate of how far Tang Kristensen was likely to walk without making mention of it – we assume a walk of greater than two kilometers would result in the addition of a route segment (*i.e.* he would have mentioned a start and end place) or a specific mention of walking to the train station as in the example above (in which case we would not have to infer the mode of transportation) (fig. 4).

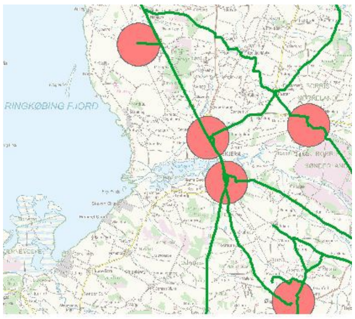


Figure 4. Two kilometer buffer drawn around open stations on the train network at the time of a specific fieldtrip.

Finally, we intersect the inter-stop route segment with the buffered stations and ascertain that the number of intersections between the route and the buffered zones is greater than or equal to two kilometers. To facilitate this work, a specific tool, TrainGremlin, was developed for use in the ArcMap toolbox.[9] If any one of these conditions is not met, then the segment is assigned to wagon.

Collecting productivity

To measure productivity, we find the first and last person mentioned associated with a field trip in the field diaries using the index of informants to find their associated pages in the field diaries as noted above. We label these the "interim first" and "interim last" informants for any given field route. We then check entries in the diaries before and after these interim first/last informants, and adjust the identity of the first/last informants if we can ascertain that prior or subsequent informants were also part of the target field trip. This provides us with a start page and end page for each field trip. We then sum the number of pages in that range as a fieldwork productivity measure. It is worth noting that Tang Kristensen's handwriting density is quite consistent over his years of collecting, and so this coarse metric does not vary much by time.

This measure is necessarily coarse for four reasons. First, Tang Kristensen did not use his field diaries exclusively to record what his informants told him, using loose sheets of paper and a series of small pocket books that he also carried with him. He began using the field diaries in 1868 and used them somewhat sporadically until 1871, at which point the diaries became his main method for recording what his informants told him. He stopped using the diaries as his primary collecting record in 1916, about the time that he stopped doing fieldwork. Second, Tang Kristensen did not collect exclusively on field trips, but also from local informants who were not visited on identifiable trips. We do not count these local collecting efforts separately, but rather assign all field diary pages to a field trip, leading to an overcount of pages for some collecting trips. Third, the 21,046 field diary pages are numbered consecutively, from 1 to 10523, with the recto receiving the designation "a" and the verso, "b"; consequently, for each page number, there are two diary pages. Because separating the a and b pages by informant is time consuming and would require manual confirmation of each informant to page assignment, we assign any trip that ends or starts on a verso page to the next page. Assuming that the number of field trips that start or end on a verso page is more or less equal to those that start or end on a recto page, the discrepancy is minimal. Fourth, we do not take into account the fact that the records from a field trip may start or end at some place other than the top or bottom of a page. Again, the distribution of these phenomena throughout the field diaries is such that the effect of this miscount is minimal. Since we sum productivity not only by trip, but also by year and by place of residence, at these aggregate scales, errors in counting are smoothed out.

Field trip description preprocessing and word frequencies

The second set of procedures focus on text mining. For each field trip, we identified the volume and page numbers from *Minder og Oplevelser* describing the trip. To generate a corpus for textual analysis, we find the sentence that contains the phrase indicating Tang Kristensen began a field trip and excerpt the text from that point to the point in the text where he indicated that he had returned home or the trip had otherwise ended. These "field trip chunks" are used for the word frequency word clouds and for topic modeling.

We preprocess the text using Lexos by removing punctuation, converting all letters to lowercase, and removing any other unusual formatting (Kleinman *et al.*, 2016). We also perform simple orthographic normalization. We remove a series of stop words, including prepositions, personal names, pronouns, articles, modal auxiliary verbs, and conjunctions, and we perform lemmatization and concatenation of certain terms. We then group the texts into five bins that align with Tang Kristensen's places of residence during his collecting career. These texts are subsequently loaded into Voyant tools, a suite of online text analysis tools to generate word clouds as a visually helpful method for assessing word frequencies (Sinclair *et al.,* 2012). Although each of the residence-based corpora varies in size, these frequency-based methods normalize them by calculating a simple rank list of the most frequent words for each of these groupings. Given the small size of the corpora, these relative ranks are sufficient for this analysis.

Field trip description topic models

Topic modeling has gained wide acceptance in the Humanities. In our modeling of the Tang Kristensen fieldtrips, we consider each field trip description in *Minder og Oplevelser* a "document" and use this collection of documents to constitute the corpus of field trip descriptions. Using a probabilistic topic modeling algorithm (LDA), we model these descriptions at a level of k= 10, 20, 30, and 50 topics, to uncover latent topics in his descriptions, for the entire corpus, and for fieldtrips associated with individual places of residence (Blei *et al.*, 2003).[10] To limit the amount of noise in the topics, we use the same stopword filter for these descriptions as was used in the word frequency calculations, although we do not apply lemmatization or concept concatenation.

Topic modeling presents another method for aggregating field trips, since field trips are grouped by topic, as opposed to some other classification such as date or region of trip. We can then explore the characteristics of field trips associated with a particular topic. We use a simple topic modeling interface, Subcorpus Topic Modeling, devised for literary analysis (Tangherlini & Leonard, 2013), which makes use of Mallet's implementation of LDA (McCallum, 2002). Here, the texts describing field trips for each place of residence are constituted as sub-corpora, while the overall corpus constitutes all of the descriptions of the field trips in the aggregate. An advantage of using this particular approach to topic modeling is that it allows us to curate our models, and to explore the relationship of one subcorpus of field trip descriptions to the field trip descriptions as a whole.

Results and Discussion

Fieldtrips and modes of transportation by collecting period

With the segmentation method and transportation inference methodology presented above, we are able to provide a refined assignment of field trip segments to transportation modes. This approach results in the assignment of an additional seven hundred and sixty-nine segments to the train (fig. 5). These refinements also provide new counts for the number of route segments, as well as number of kilometers, traveled by the various modes of transportation.

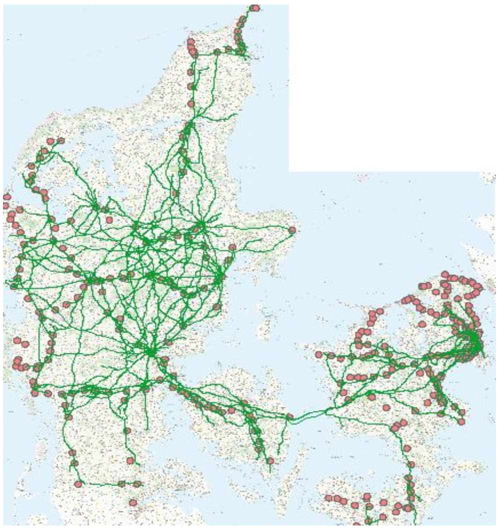


Figure 5. 769 route segments assigned to the train via our modeling. The pink dots represent 2 km buffers around stations, while the green lines are segments assigned by inference to the train.

Given the dates associated with each field trip, we can align the aggregate collecting activity, including mode of transportation and total distance traveled, by place of residence (fig. 6).

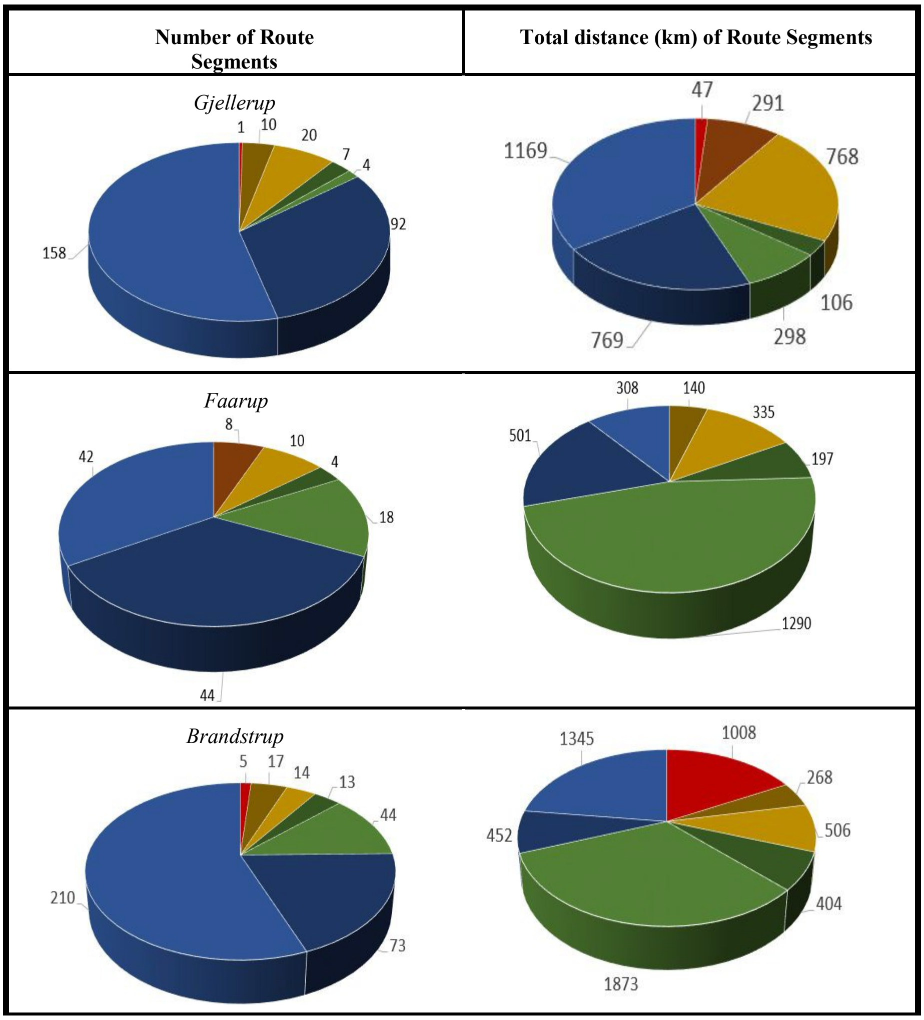


Figure 6. Pie charts indicating the distribution and absolute number of segments per transportation mode. Ferry Wagon Wagon (model) Train Train (model) Walk Walk (model)

These representations help illustrate Tang Kristensen's reach, as well as the frequency with which he made excursions to collect folklore, gives talks at local *højskoler*, as well as trips to Copenhagen to meet with colleagues and to study. With the refined segmentation, we are also able to provide an aggregate count of stops he made on his trips by year, further classified by his places of residence (fig. 7).

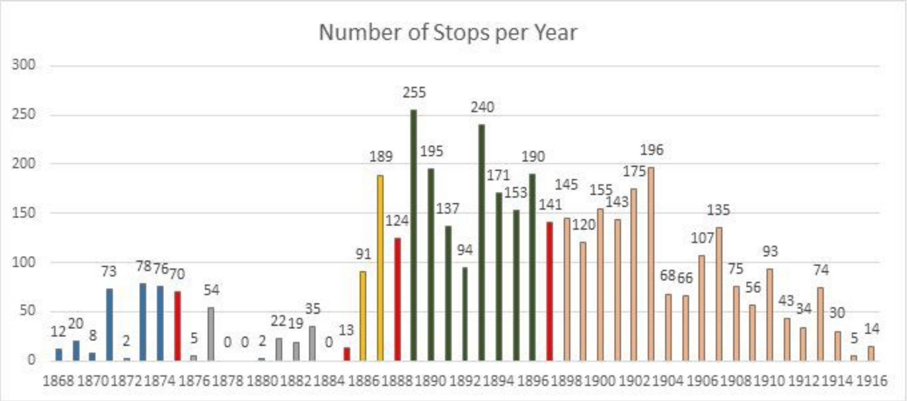


Figure 7. Number of stops per year (field trips aggregated by year), with *place of residence indicated by color. Years in which Tang Kristensen moved are presented in red. Gjellerup Faarup Brandstrup Hadsten Mølholm*

These calculations of distance traveled by year and by place of residence confirm Tang Kristensen's increasing reliance on the train, as well as his continuing dedication to walking. It is worth noting the progression of distances and modes of travel. In his first collecting period, he travels mainly by foot, and makes no trips over water, remaining close to home and the local area. When he lives in Faarup, he travels a bit further, but the number of trips he makes is very low. His reliance on the day wagon and post wagon is still quite high. When he moves to Brandstrup, he begins to shift increasingly to the train and, by the time he has moved to the station town, Hadsten (moving in across the street from the station), the train has become fully integrated into his fieldwork practice.

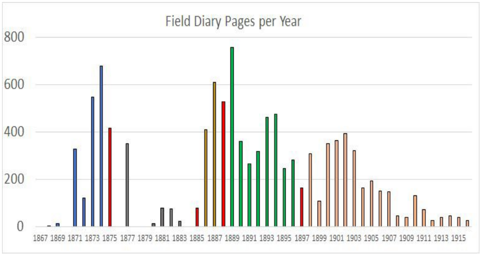
Interestingly, the charts reveal how a technology, such as the train, can gain rapid acceptance and have significant impact on cultural documentation. As the train became fully integrated into everyday life, it facilitated not only the transportation of goods and people, but also allowed for Tang Kristensen to more thoroughly cover his field area (here predominantly Jutland), and create a richer, deeper collection. Paradoxically, the same technology that allowed him to reach more informants and record in many more places also facilitated the movement of peoples across Denmark, and contributed, along with technologies such as the radio, to the homogenization of Danish regional cultures and, quite possibly, the decline in rural populations who actively participated in these traditional forms of expression.

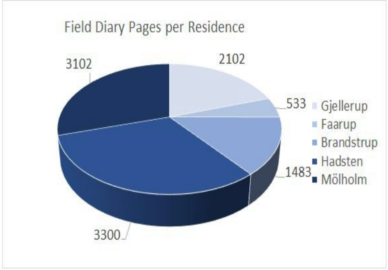
Although the field trip segments and stop counts provide a measure of Tang Kristensen's travels, they provide little information on his actual collecting productivity. It is already well-established that Tang Kristensen's years in Faarup (1876-1884) were among his least productive, as his attentions were consumed by family, school, and local administrative matters. Nevertheless, this period includes his singly most productive trip in terms of field diary pages, with trip 44 in January and February 1877, resulting in the collection of ~349 field diary pages (Tang Kristensen, 1923-1927, vol. 2, 266-276). While the distances he traveled during this period reflect this, the travel distance and stops per year also suggests that his time in Gjellerup was equally unproductive, which turns out to be incorrect.

By aligning the field diary pages that were recorded with field trips, we devise a rough measure of productivity by period and by field trip. Although the Gjellerup years were marked by very short field trips, and ones with very few stops, and therefore very low per trip productivity, the aggregate number of diary pages he recorded during this period, particularly for the years 1873-1874, rival his productivity at other times in his career. At an average rate of approximately ~210 pages per year, Gjellerup ranks only behind Brandstrup and Hadsten for sheer volume of folklore collected per year (table 1). If one corrects for his earliest collection, and limit the Gjellerup years to 1870-1876, the yearly productivity jumps to second place.

Table 1: Average number of field diary pages collected per year by residence.

Residence	Total Diary Pages	Avg pages / year
Gjellerup (1866-1876)	2102	210 (1870-76: 350)
Faarup (1876-1884)	533	67
Brandstrup (1884-1888)	1483	321
Hadsten (1888-1897)	3300	367
Mølholm (1897-1916)	3102	163 (1887-96: 310)



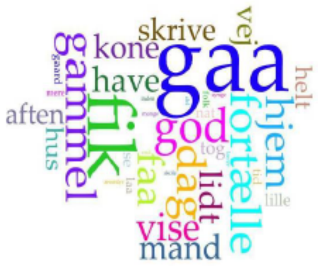


We also generate average numbers of pages per field trip per residence period, and calculate 95% rough confidence intervals over this productivity. Although the confidence intervals for very small samples, such as the five collecting trips during the Faarup period are so large as to be of little use, they do reveal the substantial swings in productivity per field trip. While the earlier periods are marked by a fairly stable productivity in the ~70 pages / field trip range, Hadsten is characterized by both many more fieldtrips of greater length, with an average of ~50 pages / field trip. This metric holds true for the collecting he undertook while he lived in Mølholm until 1903, after which Tang Kristensen's field trips taper off. His overall productivity captures an interesting feature of his later collecting – as he visits many of the same informants as earlier in his career, his productivity becomes less variable per trip, but also results in fewer pages of recordings (Storm *et al.*, 2017).

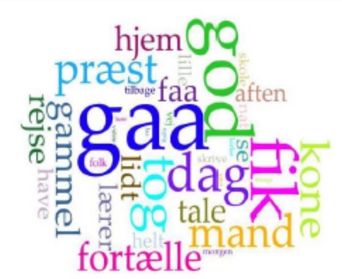
These calculations support in part the two divergent views on folklore collecting that were prevalent at the time. Tang Kristensen's close-to-home work while he lived in Gjellerup confirms that local expertise can be a boon to fieldwork, since the opportunity for establishing rapport with a cultural group that the collector knows well, and the possibility of developing a rich network of storytellers and perhaps other local collectors can lead to productive collecting with deep coverage of a local area. In general, the regionalism of these collections can be offset by a large number of local collectors working productively in their own regions contributing their work to a central archive. This networked approach to local fieldwork was a key element in the development of much larger national collections throughout Scandinavia. Yet local collecting is not the only solution. Tang Kristensen's later collecting confirms the benefits of the other approach, namely that of the experienced field worker going much further afield and activating a network of local contacts to create a consistent collection for a very large region or country.

Fieldtrip Descriptions and Topics

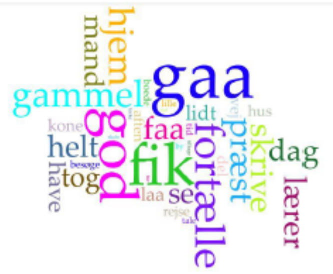
A macro-scale analysis of Tang Kristensen's field trip descriptions provides insight into how he conceptualized these trips. The word clouds provide a first-level approximation of the highest frequency words in these descriptions (fig. 8). Although there is clear stability in the highest ranked word(s) which, for all periods were forms of the verb *at gaa* (norm. *gå*) meaning to go or to walk, there are subtle changes in the other highly ranked words (top-35 frequency).[11] Not surprisingly, one can trace a shift in the vocabulary of travel over the years (see fig. 8, cell 6), where the train rises in prominence (as measured by the word "toget") over the course of the five periods, while the word "vogn" (and its related words such as "dagvogn" and "postvogn") drop in rank, most notably correlated with his move from Gjellerup, a transition that marked his increasing reliance on the train. One of the biases that the word cloud makes clear is Tang Kristensen's emphasis on the "gammel" [old], with that adjective, along with the adjectives "god" [good] and "lidt" [little quantity], ranking in the top three adjectives across all five collecting periods. Clearly his focus on finding "good" material, either old or from old sources, was a constant concern. The word "lidt" is interesting – he often uses it in the phrase "ikke saa lidt" [not so little], a modest way of addressing what he considered to be significant productivity.



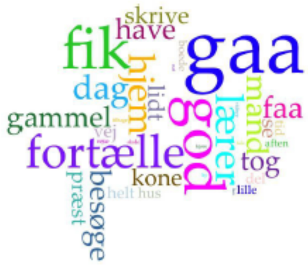
Gjelleru



Faaru



Brandstrup

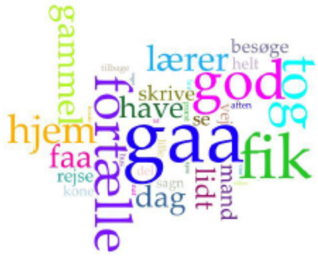


Hadsten Station

Top 5 ranked transportation words by residence with rank in brackets

Faarup	Gjellerup	Brandstrup	Hadsten	Mølholm
rejse [14]	vej [12]	vej [19]	vej [19]	rejse [15]
vej [27]	rejse [46]	rejse [24]	rejse [27]	vej [18]
kjøre [43]	toget [89]	kjøre [70]	kjøre [35]	toget [34]
vogn [72]	kjøre [116]	toget [76]	toget [36]	kjøre [35]
toget [82]	vogn [317]	vogn [140]	vogn [154]	vogn [147]

*minus the words "gaa" and "tog" (see discussion)



Mølholm

Figure 8: Word clouds and rank lists for field trip descriptions, aggregated by place of residence.

Using text analysis on the field descriptions is most informative when representing the stable aspects of his practice such as, for example, whom he considered to be good informants, as well as the shifting goals informing that practice, most notably the pivot toward legend and away from ballads and fairy tales. A simple ranked list of (a) genres and performance styles and (b) classes of contacts and informants, reveals these subtle changes in practice (table 2). Here, words for males (men, boys) are concatenated under the word "mand", which can also mean husband, and words for females (women, girls) are concatenated under the word "kone", which can also mean wife.

Table 2: For each place of residence, column "a" presents a ranked list of genre and performance words, and column "b" presents a ranked list of informant and contact class words. Actual rank in brackets.

Faarup		**Gjellerup**		**Hadsten**		**Brandstrup**		**Mølholm**	
Genre	**Inform**	**Genre**	**Inform**	**Genre**	**Inform**	**Genre**	**Inform**	**Genre**	**Inform**
fortælle [7]	præst [6]	fortælle [3]	kone [14]	fortælle [4]	lærer [5]	fortælle [4]	præst [5]	fortælle [3]	lærer [6]
vise [57]	kone [8]	vise [10]	mand [15]	sagn [53]	mand [11]	sagn [67]	lærer [8]	sagn [21]	mand [14]
synge [70]	mand [9]	æven-tyr [31]	folk [26]	vise [78]	præst [16]	æven-tyr [87]	mand [15]	vise [30]	kone [20]
sagn [103]	lærer [16]	synge [33]	lærer [28]	æven-tyr [100]	kone [17]	synge [152]	kone [20]	synge [37]	præst [26]
æven-tyr [110]	folk [24]	sagn [62]	præst [36]	synge [128]	folk [42]	vise [196]	folk [40]	æven-tyr [41]	folk [58]

Even at this very coarse level of analysis, it becomes clear that Tang Kristensen's collecting shifted from an emphasis on ballads to an emphasis on legends, with fairy tales rounding out the targeted genres. Intriguingly, this pattern is surfaced as a latent feature of his descriptions of field trips – nowhere does he state this shift in focus; rather his descriptions speak for themselves.

Also evident in these frequency graphs is the increasing importance of teachers in his network of contacts, providing him access to the local storytellers and ballad singers. While there appears to be no overt bias in his reference to men or women, the high frequency of the mention of women is skewed by his frequent reference to his wife, as well as the wives of informants and colleagues (1518 mentions of "kone" words in *Minder og Oplevelser*). There is little doubt that the overall collection has a strong bias toward male informants; what is not clear is whether this bias is even across the collections sent to Tang Kristensen and his own collection. The descriptions of field trips here suggests that a more substantive exploration of the gender composition of Tang Kristensen's informants from whom he collected directly is worth exploring.

Topic modeling provides a different view on the field trip descriptions. At a level of *k=50*, we can identify a series of meaningful topics related to informants, genre and performance. Interestingly, at the four levels modeled (*k*= 10, 20, 30 and 50), there were no topics that were clearly or predominantly associated with transportation, apart from walking. In the documents associated with the "walking" topic (table 3), Tang Kristensen emphasizes the amount of walking he has done, with highly ranked n-grams underscoring this.

Table 3: Topic modeling over the stop-worded field trip description corpus, showing a topic related to walking at k=10, 20, 30 & 50 topics.

k	**10**	**20**	**30**	**50**
top 3 words	gik, aftenen, fortælle	gik, faa, lidt	gik, vej, aftenen	gik, gaa, mil
top 3 n-grams	lange vandring kop kaffe støvler og strømper	halv mil skrev hjem fik brev	lange vej mulm og mørke benene rørt	halv mil halvanden mil mulm og mørke

Topics related to specific genres provide equally interesting perspectives on how Tang Kristensen viewed his collecting efforts. Here there is slightly more specificity than that provided by word clouds and frequency counts, and the discovered topics point to a potentially differing type of engagement with different genres of folklore. While ballads are identifiable as a separate topic at all levels of *k* of twenty or greater, the discrete topics for legends and fairy tales effectively disappear already at the level *k<50*, with broader storytelling topics concatenating these genre differences (as measured by highly ranked genre words). Indeed, at *k=20*, the fairy tale topic is completely subsumed in a secondary legend topic that is predominantly focused on "telling" (table 4).

Table 4: Topic modeling over the stop-worded field trip description corpus, showing ranked words, n-grams and fieldtrips in topics related to legends, fairy tales, and ballads, at k = 20, 30 & 50 topics. Field trip identifiers are listed in parentheses after the place of residence.

Topic	***k*=**	**20**	**30**	**50**
Legend	top 3 words	fortalte, fortælle, gik	fortalte, fik, tog	fortalte, sagn, fortælle
	top 3 n-grams	fortalte sagn lille hus langt hus	gode ting gode bidrag gode sagn	fortalte sagn fortælle sagn fortalte æventyr
	top 3 field trips	Mølholm (69), Gjellerup (30_10), Gjellerup (30_11)	Mølholm (167) Mølholm (206) Mølholm (195)	Mølholm (204) Gjellerup (30_5) Mølholm (171)
Fairy tale	top 3 words	fortalte, fik, lærer	fik, fortælle, holdt	lange, æventyr, kone
	top 3 n-grams	gode bidrag fortalte sagn par sagn [fortalte æventyr - 6]	holdt foredrag holde foredrag fortælle æventyr	lange veje ømme fødder smaa vaade pletter
	top 3 field trips	Mølholm (171) Mølholm (192) Mølholm (191)	Brandstrup (15) Gjellerup (7) Mølholm (192)	Gjellerup (42) Gjellerup (36_5) Gjellerup (36_4)
Ballads	top 3 words	viser, sang, gamle	viser, gamle, sang	gamle, sang, viser
	top n-gram	gamle viser sang viser synge viser	gamle viser synge gamle viser synge viser	gamle viser synge viser gamle melodier
	top 3 field trips	Mølholm (188) Gjellerup (2) Mølholm (193)	Mølholm (188) Gjellerup (1) Gjellerup (2)	Mølholm (188) Gjellerup (1) Mølholm (187)

Unlike the word frequency counts described above (which aggregate over a corpus or sub-corpus), topic saturation for each description facilitates drill-down into the underlying descriptions, which can then be aligned with the field trip level information that we have derived using hGIS, and through alignment with the field diaries, informant index, and published collections. In this manner, the topics can be viewed as an additional form of classification, creating new views on the collection. For instance, at *k=20*, Tang Kristensen's positive experience with teachers as good local resources becomes apparent, with a topic one could label "the good teacher", with highly ranked n-grams including the phrase "venlig imod", referring to the teachers' friendly reception of the wandering folklorist. One could then use this topic to identify those fieldtrips that Tang Kristensen described as including friendly teachers. These fieldtrips could then be used to explore the areas in which the "friendly teacher" could be found, particularly in the context of rural education and political and religious movements in those areas. Additional topic modeling methods, such as Structured Topic Modeling (Roberts *et al.*, 2013) may help refine this approach, as those methods allow for the consideration of confounding factors such as time.

Conclusion

Our work reveals the shifting parameters of Tang Kristensen's field collecting, from his intensely local focus early on to his more expansive and confident travels at the end of his career, when his collecting was no longer aligned with Romantic nationalist goals but were more in tune with a thick descriptive approach to Jutlandic rural life. By wedding hGIS techniques to text mining methods, we provide a degree of detail about his travels missing in earlier studies. We can discern the changing productivity of his fieldtrips at the same time as we can visualize his increasing range, influenced in large part by the expanding railway network in Denmark. We also show that, while Tang Kristensen's discussions of various modes of transportation are relatively stable, his self-evaluation of his ability to walk long distances is easily captured. Importantly, topic modeling allows us to provide an unsupervised method for classifying field trips based on his discussions of the trips; aggregating fieldtrips by topic may in future work provide some new avenues for accessing the collection (*e.g.* fieldtrips that are highly correlated to the topic of legends). Our interlocking methods enable a macroscopic approach to understanding how a folklore collection came to be, affording us an opportunity to interrogate Tang Kristensen's work at varying scales of resolution (Tangherlini, 2013). With these methods, we can move from the micro-consideration of a single field trip, to a meso-consideration of all trips he took while living in a particular place, to a macro-consideration of all of his trips taken as a whole. Importantly, our approach provides clear insight into the gradual change from folklore collecting that dominated his early, local work, to a more ethnographic engagement with the broader region of Jutland. The approach we outline here is not limited to Tang Kristensen and his collection, but rather can be extended to ethnographic fieldwork collections as a whole, allowing researchers to develop rich, multi-level macroscopic perspectives on complex collections that encompass far more information than critics have suggested.

Ida Storm finished her undergraduate degree at UCLA where she also worked as a research associate with Tangherlini. Their group focused on applying GIS methods to the study of Nordic folklore, the development of Danish transportation infrastructure, and population change in the late nineteenth and early twentieth centuries. She is currently pursuing a master's degree in GIS at Lund University and is involved in various humanities projects where GIS is relevant.

Contact: idastorm1991@gmail.com

Timothy R. Tangherlini is Professor, Scandinavian Section, UCLA. His work focuses on folklore, Nordic literature and literary history, computational approaches to problems in the Humanities, and the study of culture. He has received support from the National Endowment for the Humanities, the National Science Foundation, the National Institutes of Health, the Fulbright Foundation, the Nordic Council of Ministers, the Mellon Foundation, the John Simon Guggenheim Memorial Foundation, the American Council of Learned Societies, Apple, and Google.

Contact: tango@humnet.ucla.edu

Notes

1. A great deal of collecting occurred well into the twentieth century in Denmark and across Scandinavia.
2. Boberg (1953) provides an excellent historical overview of the development of folklore research and collecting in central and northern Europe. See also Tangherlini (2013a), particularly chapter two, "The Rise of Folklore Scholarship." Strömbäck *et al*. (1971) include biographies of most of the important Nordic folklorists, while Hult (2003) provides an accessible consideration of the Norwegian Peter Christen Asbjørnsen's folklore collecting efforts. See also Antonnen (2005), Bringéus (1966), Hodne (1979), Klintberg (1979), Lilja (1996), and Piø (1971) for additional studies of folklore collecting and the intellectual basis for such efforts in the Nordic region.
3. In future work, we intend to estimate segment duration based on average travel times by transportation mode. These segment durations will allow us to estimate overall trip duration as well as stop dates, reconciling that with information from the memoirs. This refinement may allow us to further align trip descriptions with local news reports and weather conditions, approaching the "thick description" of fieldwork that Clifford Geertz proposes as the gold standard for ethnographic work (Geertz, 1973).
4. In future work, we intend to include descriptions of his field collecting from his letters and from his chorographic works, such as *Vindt Mølle* (1887) or *Heden* (1930), to build a richer corpus of fieldwork descriptions. The former are not currently machine actionable, while the latter are difficult to align with individual field trips.
5. Tang Kristensen's active collecting career ended in 1916, and so we use 1916 for most calculations, as opposed to his year of death.
6. These two resources – the informant index and the loose leaf collections – will be used to refine this model in future work.

7. For this project, we use ArcGIS software from ESRI. Other GIS software packages, such as QGIS, could also be used.
8. A more detailed explanation of the procedures, as well as the complete data set of field trip shapefiles, and field trip descriptions are available at <https://github.com/ScandinavianSection-UCLA/hGIS_ETK>
9. TrainGremlin is available on the project's github repository.
10. *k* is used to indicate the number of topics in a model.
11. The Danish verb "at gå" is ambiguous, and can mean "to walk" or "to go" (as in go to a destination). Consequently, without a laborious process of manual tagging, it is impossible to use this verb solely to model walking. A similar confusion concerns the verb "at tage" which, in the past tense, becomes "tog". This verb form causes ambiguity with the Danish noun, "tog", which means train.

References

1929/12 Dansk folkemindesamling – Loose leaf collections.

1929/16 Dansk folkemindesamling – Field diaries

1929/129 Dansk folkemindesamling – Informant index

ANTONNEN, PERTI J. (2005). *Tradition through Modernity: Postmodernism and the Nation-State in Folklore Scholarship.* Helsinki: Finnish Literature Society.

BLEI, DAVID M; NG, ANDREW Y. & JORDAN, MICHAEL I. (2003). “Latent Dirichlet Allocation.” Journal of Machine Learning Research 3: 993-1022.

BOBERG, INGER M. (1953). *Folkemindeforskningens historie i mellem- og nordeuropa.* Copenhagen: Einar Munksgaard.

BRINGÉUS, NILS ARVID (1966). *Gunnar Olof Hyltén-Cavallius som etnolog: En studie kring Wärend och Wirdarne*. Nordiska museets handlingar 63. Stockholm: Ohlsson.

BROADWELL, PETER M. & TANGHERLINI, TIMOTHY R. (2017). “GhostScope: Conceptual Mapping of Supernatural Phenomena in a Large Folklore Corpus.” In, *Maths meets myths: Quantitative approaches to ancient texts.* Eds. Ralph Kenna, Máirín MacCarron, Padraíg MacCarron. Cham, Switzerland: Springer. 131-158.

CHRISTIANSEN, PALLE OVE (2013). *Tang Kristensen og tidlig feltforskning i Danmark. National etnografi og folklore 1850-1920.* Copenhagen: The Royal Danish Academy of Sciences and Letters.

CHRISTIANSEN, PALLE OVE (2011). *De forsvundne: Hedens sidste fortællere.* Copenhagen: Gads forlag.

DIGDAG (2009). *Digitalt atlas over Danmarks historisk-administrative geografi.* Copenhagen.

ELLEKILDE, HANS (1946). “Indledning” in *Svend Grundtvigs Danske Folkesagn, 1839-1883.* Copenhagen: Ejnar Munksgaard. 7-65.

GREGORY, IAN N. & ELL, PAUL S. (2007). *Historical GIS: technologies, methodologies, and scholarship*. Cambridge Studies in Historical Geography 39. Cambridge: Cambridge University Press.

HODNE, BJARNE (1979). *Eventyret og tradisjonsbærerne. Eventyrfortællere i en Telemarksbygd.* Oslo: Universitetsforlag.

HULT, MARTE HVAM (2003). *Framing a National Narrative: The Legend Collections of Peter Christen Asbjørnsen.* Detroit: Wayne State Univ. Press.

KLEINMAN, SCOTT ET AL. (2016). *Lexos. v3.0.* doi:10.5281/zenodo.56751. <https://github.com/WheatonCS/Lexos/> [2018-12-21]

LILJA, AGNETA (1996). *Föreställningen om den ideala uppteckningen: En studie av idé och praktik vid traditionssamlande arkiv – ett exempel från Uppsala 1914-1945.* Uppsala: Dialekt- och folkminnesarkivet.

MCCALLUM, ANDREW KACHITES (2002). “MALLET: A Machine Learning for Language Toolkit.” <http://mallet.cs.umass.edu> [2018-12-21]

PALMENFELT, ULF (1993). *Per Arvid Säves möten med människor och sägner.* Stockholm: Carlssons.

PIØ, IØRN (1971). “Svend Grundtvig og hans folkloristiske arbejdsmetode.” *Danske Studier*: 91-120.

ROBERTS, MARGARET E. ET AL. (2013). “The structural topic model and applied social science.” In *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*.

SINCLAIR, STÉFAN; ROCKWELL, GEOFFREY & THE VOYANT TOOLS TEAM (2012). *Voyant Tools* (web application).

STORM, IDA ET AL. (2017). "Folklore Tracks: Historical GIS and Folklore Collection in 19th Century Denmark." In DH 2016. Proceedings of the International Symposium on Digital Humanities (Växjö, Sweden). Edited by Korajlka Golub and Marcelo Milrad. CEUR Workshop Proceedings, vol 20-21: 75-98.

STRÖMBÄCK, DAG ET AL., EDS. (1971). *Biographica. Nordic Folklorists of the Past.* Copenhagen: Nordisk institut for folkedigtning.

TANG KRISTENSEN, EVALD (1887). *Vindt Mølle*. Viborg: F.V. Backhausen.

TANG KRISTENSEN, EVALD (1923-1927). *Minder og Oplevelser.* Volumes 1-4. Viborg: Forfatterens forlag.

TANG KRISTENSEN, EVALD (1930). *Heden.* København: Woels forlag.

TANGHERLINI, TIMOTHY R (1994). *Interpreting Legend.* New York: Garland Publishing. Reprinted 2017, Routledge.

TANGHERLINI, TIMOTHY R. (2013a). *Danish Folktales, Legends and Other Stories: The Danish Folklore Nexus.* Digital Materials. Seattle: Univ. of Washington Press.

TANGHERLINI, TIMOTHY R. (2013b). "The Folklore Macroscope." *Western Folklore* 72. 1: 7-27.

TANGHERLINI, TIMOTHY R. & BROADWELL, PETER M. (2016). "WitchHunter: GeoSemantic Browsing in a Large Folklore Corpus." *Journal of American Folklore* 129. 511: 14-42.

TANGHERLINI, TIMOTHY R., & LEONARD, PETER (2013). "Trawling in the Sea of the Great Unread: Sub-corpus topic modeling and Humanities research." *Poetics* 41. 6: 725-749.