ARCHER TAYLOR MEMORIAL LECTURE 2012

# The Folklore Macroscope

## *Challenges for a Computational Folkloristics*

Timothy R. Tangherlini

ABSTRACT

*Folklorists are poised on the cusp of an exciting new era. The digital revolution has swept over the field of folklore, vastly increasing the amount of accessible research material. To take advantage of these changes, folklorists must develop consistent methods for digitizing, storing, retrieving, displaying and interpreting these materials. Computational methods for the study of traditional culture can help us address these issues, and are essential for the future success of our field. In this essay, I present some of the main challenges for a computational folkloristics, and propose some preliminary approaches to addressing these challenges. KEYWORDS: computational folkloristics, archives, research methods, future of the discipline, digital humanities*

> Essentially, all models are wrong, but some are useful.
>
> George E. P. Box (Box and Draper 1987)

## INTRODUCTION: TOWARD A COMPUTATIONAL FOLKLORISTICS

In the final paragraph of his mammoth four-volume memoir, *Minder og Oplevelser* [Memories and Experiences], the nineteenth-century Danish folklore collector Evald Tang Kristensen writes, "Fra flere Steder er

Timothy R. Tangherlini is a professor in the Scandinavian Section and in the Department of Asian Languages and Cultures at UCLA.

*Western Folklore* 72:1 (Winter 2013):7–27. 

der ytret Ønske om, at jeg skulde have ledsaget Værket med et Person-Register, og det vil virkelig være højst ønskeligt… Men hvem vil udgive og trykke det? Jeg kan ikke" [There have been requests from many directions that I should have accompanied this work with an index of names, and that would certainly have been desirable… But who would publish and print it? I can't] (Kristensen 1923 IV: 442). This concluding lament, clearly written in a fit of intellectual exhaustion, is a bit more profound than one might imagine at first glance.[1] With this one short remark, Tang Kristensen inadvertently calls into question the processes of collection, archiving, classification, and indexing, as well as the shifting ground—even at the beginning of the twentieth century—of the publishing market for folklore, let alone folklore indices. If he had included considerations of the analysis of folklore and how best to present both the collection and the results of analytical work, he would have constructed the perfect accidental sentence interrogating the foundations of the very field to which he had dedicated his life and the 1680 pages of his memoirs.

What he did accomplish with this brief complaint—although this was not remotely on his mind—was to help delineate four main challenges to folkloristics as a modern discipline which grow more pressing as we move further and further into an "algorithmic" age, challenges that underpin the development of a computational folkloristics. Broadly conceived, these four challenges are (1) collection and archiving; (2) indexing and classification; (3) visualization and navigation; and (4) analysis. Tang Kristensen would have been the first to recognize that these categories are not mutually exclusive, but rather mutually constitutive and, furthermore, that a holistic rather than atomistic approach to folklore is a necessary foundation for the field. Of course, as Tang Kristensen's gripe makes clear, these challenges have been around since the inception of the discipline; they have simply become more acute in an age of "Big Data" as we confront a proliferation not only of "born digital" resources for the study of folklore but also recognize the possibilities that arise from liberating older resources from the static realm of handwritten archives, printed collections, and other "off-line" repositories.

As the television show "Hoarders" teaches us, finding things becomes increasingly difficult the more things one has. Had Tang Kristensen been content with collecting a few dozen or a few hundred stories from a handful of people in his immediate neighborhood, perhaps constraining himself to ballads and fairy tales as his mentor Svend Grundtvig had suggested, finding things in his collection would not have been terribly

difficult, and his memoirs would have been considerably shorter. But his collecting endeavor, spanning over half a century and encompassing tens of thousands of stories collected from thousands of people from across most of Denmark prefigures in many respects the Internet, a complex and dynamic information resource where the ones who hold the keys to the realm are those who can find things accurately and quickly; in short, the ones with the best indices.

When I first read Tang Kristensen's complaint on a typically rainy afternoon in September 1999, I thought that producing an electronic index for *Minder og Oplevelser* would be a reasonable challenge for a group of graduate students in my folklore methodologies seminar at the University of Copenhagen. But as this small group began work on the index, it unleashed a series of desiderata that made "If you give a mouse a cookie" look like child's play (Numeroff 1985). While at the end of this version of the story—which one can perhaps label "if you give a mouse an unindexed memoir"—the mouse predictably winds up with yet another unindexed memoir, along the way he has worked through a super computer, several terrabytes of storage, a high-speed digital scanner and a crew of engineer refugees from Yahoo!

Although Tang Kristensen had only raised the challenge of indexing the names of people mentioned in the memoir, my students recognized that even this simple task was intractable. After the first chapter or two, the proliferation of people was paralyzing not only because many people shared the same or similar names (an artifact of Danish naming conventions) but also because they played numerous roles at different times in the collection. Similarly, places and dates were as frequently mentioned as people, and it became increasingly clear that these needed to be indexed as well. To make the task even more complex, Tang Kristensen kept pointing outside of the memoirs to other aspects of his collection which constituted indexable works in their own right—his field-collecting trips, his field collections, his editorial work, his publication endeavors, and his correspondence. Consequently an index for one resource necessarily required indices for these other resources.

Ultimately, we realized that proposing an index to the memoir was simply another way of proposing an index to the entire collection. But even that seemingly straightforward task raised questions: What constituted the collection? How was the collection currently organized? Where were these resources located? What benefit would developing an electronic index have for research? It also raised the vexing question of what is meant by an index. How, for example, does one index

a field-collecting trip? As a result, the initially modest goal of indexing *Minder og Oplevelser*, an exercise that I imagined would take a few weeks at most for my talented students, led to a series of theoretically rich discussions but no index to the memoirs *per se*.

## THE FOLKLORE MACROSCOPE

An important outcome of these discussions was a new wish list that conceptualized the collection as a broadly accessible digital collection coupled to a series of study tools that could assist in the classification, presentation, and analysis of the collection assets. Foremost among these desiderata was an integrated computer-based study environment for the Tang Kristensen collection and, by extension, any folklore collection. As this theoretical environment began to take shape—sketched out on whiteboards, scribbled on scraps of paper, traced on napkins, etched on beer coasters, modeled out of origami and string—it converged on the contours of what Katy Börner has enticingly labeled the "macroscope" (Börner 2011). For Börner, "Macroscopes provide a 'vision of the whole,' helping us 'synthesize' the related elements and detect patterns, trends, and outliers while granting access to myriad details. Rather than make things larger or smaller, macroscopes let us observe what is at once too great, slow, or complex for the human eye and mind to notice and comprehend" (Börner 2011:60). The macroscope holds the promise of wedding "close reading" approaches, which have been a fundamental analytical approach in folkloristics since the beginning of the field, to what Franco Moretti has called "Distant Reading" where "Distance . . . *is a condition of knowledge*: it allows you to focus on units that are much smaller or much larger than the text: devices, themes, tropes—or genres and systems" (Moretti 2000:57).

Importantly, the macroscope as we conceived it would move considerably beyond the simple ability to switch views from the close to the distant, and incorporate methods to analyze the social and information networks through and across which folklore is transmitted and collected. By bringing in outside resources such as census lists, voting and enlistment rolls, church books, insurance inventories, and probate records, it would let researchers consider the historical currents that necessarily impact the contours of any tradition. By bringing in detailed geo-referenced historical maps, it would let researchers view the collection in relationship to the changing physical environment in which the tradition participants lived and worked and through which the collector moved. The folklore macroscope, in short, would model the complex dynamics

of a folklore collection taking into account not only the texts but also, as Alan Dundes proposed, the context and texture (Dundes 1964). Clearly, achieving these goals would require computational tools to assist in pattern discovery from a distant, comparative approach, to help interrogate those patterns at changing scales of engagement and from various vantage points, and to visualize the results of these engagements in meaningful ways. In so doing, the folklore macroscope would keep the researcher aware of the fundamental premise that folklore is created by people who live their lives in complex societies, embedded in both time and place.

First, I will briefly explore how we eventually developed the Danish Folklore Nexus, a proof-of-concept research and presentation environment that represents a first attempt at developing the folklore macroscope. I will also highlight additional work that focuses on some of the computational tools for pattern discovery that we have applied to this research corpus as a means for charting a course toward DFL 2.0, an integrated research environment for the entire Tang Kristensen collection. DFL 2.0 will incorporate aspects of the "plug-and-play" environment that Börner proposed as a means for keeping the macroscope up to date as new analytical approaches become available. Along the way, I will explore each of the four challenge areas for computational folkloristics that Tang Kristensen inadvertently enumerated in his final complaint—collecting and archiving; classification; presentation and navigation; and analysis—highlighting some of the main computational challenges in each area and proposing some of the fundamental theoretical premises from folklore that can guide this work. Although the examples that I will discuss are drawn largely from work on the Tang Kristensen collection, it is not hard to draw comparisons to analogous problems from other collections. Indeed, some of our exploratory work has been done with corpora as diverse as the Shoah Foundation Institute's Visual History Archive, the Nordic language books in the Google Books collection, the close-captioned feeds of 100,000 hours of evening newscasts, multi-lingual blogs and Tweets related to the Iranian uprisings, and rumors about health threats culled from a large number of health-care related websites. Nevertheless, the following work is necessarily limited in scope and hardly a comprehensive charter for a broad computational folkloristics. Even though the Tang Kristensen collection also includes photographs, sketches, wax recordings, material culture such as house beams, and rare books such as *Cyprianus* that hover somewhere between the written and the printed (Ohrvik 2011), in our current work we deal primarily with textual resources in a single language.

## CHALLENGE #1: DIGITIZATION AND ARCHIVING

One of the most pressing challenges for a computational folkloristics is that many of our research collections do not exist in machine-readable form; at the start of our work, the Tang Kristensen collection was no exception. A quick assessment of the status of both Tang Kristensen's memoirs and the overall collection revealed that few if any of the resources were available in electronic form, and that the vast majority of the collection was only accessible as hand-written manuscripts at the Danish Folklore Archives. Even Tang Kristensen's published oeuvre was not available as a complete set outside of the collections of the Royal Library. The collection survey brought into stark relief the first challenge for developing an electronic version of the collection, namely that many of the target resources not only were not "born digital" but also were either in such a fragile condition that it was unclear whether they could survive even the most gentle of scanning protocols or stored in hard to access archives. Methods for scanning written and photographic resources and the digital conversion of moving image and audio resources are well established and these issues are generally considered to be "solved problems," but that does not mean that these are not difficult, expensive, and complicated undertakings.

Although our initial target had been *Minder og Oplevelser*, which we had realized could provide an excellent basis for charting the development of Tang Kristensen's collection as it grew through time, we had also realized that any such digital edition would be nothing more than a baby step on the way toward developing a digital representation of the entire collection. Scanning the memoirs was a relatively trivial, if boring, task, but the scanning immediately brought into relief problems associated with Optical Character Recognition (OCR). Again, while OCR is largely a solved problem, even a cursory glance at a non-English book in Google Books' enormous collection of OCRed digital texts reveals significant accuracy problems for non-English books. For Tang Kristensen's published oeuvre, these problems were compounded by the inexpensive paper and ink that his various printers had used, the use of *fraktur* as the main font for some of the published volumes, as well as Tang Kristensen's decision to record and transcribe some stories in his own idiosyncratic dialect. Although we were able to produce a working version of the memoirs relatively quickly, it was nearly four years before we had a corrected and tagged version of *Minder og Oplevelser* as part of our digital resources.

Despite these challenges, the experience with the memoir encouraged us to scan and OCR all of Tang Kristensen's publications. While this current digital collection is not as "clean" as the memoirs, it does represent an important facet of the collection, comprising approximately 40,000 printed pages and, if our initial "chunking" of this material is accurate, 68,000 individual story records. Common wisdom proposes that Tang Kristensen published approximately two thirds of his hand written collection—our work should be able to eventually confirm this while also making the entire collection available to a much broader audience.

Once we had produced a relatively clean version of the memoir, we experimented with various approaches to indexing the work, from simply taking advantage of built-in indexing in Adobe Acrobat to developing a specific TEI tag-set for the collection. Given our main interests, namely a model of the people, places, and field diary records that Tang Kristensen produced over the course of his numerous field trips, we needed at the very least to be able to consistently identify people and place names, and resolve them to disambiguated lists, presumably the hand-written people and place name indices at the Danish Folklore Archives (these have now been scanned, but since they are hand-written, transcription will necessarily be manual). We also wanted to align place names with a geo-referenced historical gazetteer so that we could accurately project those references onto maps. Finally, we wanted to be able to recognize dates, and when the dates did not include years, propose solutions for those unresolved dates based on surrounding information. Dates would prove particularly useful as we worked on describing the detailed routes of Tang Kristensen's more than 200 field-collecting trips.

Many of these problems fall under the broad rubric of "Named Entity Recognition" and "Entity Resolution" or disambiguation and are generally considered to be "open problems" in Natural Language Processing. We only partially succeeded in implementing automatic detection and resolution for the memoirs; a lot of our processes were "supervised" where automatically assigned solutions had to be hand-corrected to get a reasonable level of accuracy. Developments in this area are encouraging, and the unique and complex challenges posed by folklore corpora might entice computer scientists to help us address this surprisingly complex task. These challenges in NLP extend to our work with place name referents in stories themselves and also reverberate through our work on other corpora.

Our second target for digitization was the field diary collection since, ultimately, the goal was to develop a method for easily navigating the

collection based on a series of researcher-driven criteria. Fortunately, at the Center for Folklore at the University of Copenhagen, we had a microfilm copy of the field diaries that Bengt Holbek had used while writing *Interpretation of Fairy Tales* (Holbek 1987). Unfortunately, the microfilm was deteriorating rapidly, evidenced by a faint vinegar smell permeating the folklore suite.[2] Since the center's microfilm reader had been stolen and the university was unwilling to purchase a new one for the center, the associate dean for the humanities proposed digitally scanning the microfilm. Unfortunately, the resulting 24,000 tiff images were returned to us with no other index than a sequential numbering of the image files. For a brief moment we panicked, but then we solved the problem by using an inexpensive off-the-shelf image database that attached meta-data created during the scanning process to each of the image files, and stored pointers to those files in a standard database format (MySQL). This simple solution also allowed us to automatically assign page numbers corresponding to Tang Kristensen's idiosyncratic recto/verso page numbering system, which we subsequently manually corrected to account for duplicate numberings or numbering gaps in the underlying resource, and to add missing pages that had been overlooked when the microfilm was initially made.

Over the course of several short months, we had gone from theoretical discussions about making a simple index to what now appeared to be a laughably small four-volume memoir, to confronting a proliferation of digital resources that were not only in need of an index but also an appropriate data structure and a means for secure storage. This realization once again changed the direction of what was quickly becoming a fairly schizophrenic seminar and precipitated extensive discussions of the relationships between the various resources with which we were now working. As a result of these discussions, we were able to summarize the overall structure of the collection: the field diaries and loose-leaf files were the unedited collections of songs, ballads, proverbs, fairy tales, legends, recipes, prayers, and descriptions of everyday life that Tang Kristensen used as the basis for his myriad published collections. He had collected this material during his numerous field trips as he traveled from place to place meeting and talking to various people. These trips, as well as his life as a folklorist, were described in his memoirs that relied heavily on his correspondence, a contemporaneous record of his dealings with academics, teachers, administrators, politicians, friends, family, and, oddly enough, himself. This simple conceptualization of what had initially appeared to be an inconceivably complex collection

allowed us to move rapidly forward with designing a transparent system for organizing our data, an organization that is predicated on the theory that folklore always consists of traditional expressive forms circulating through and across social networks. This fundamental notion of social interaction in informal settings as the underpinning of all folklore collections often gets lost in the study of archival resources, even though it should not. Indeed, if one conceives of Facebook as a dynamic self-archiving folklore collection, one can in one fell swoop recognize the importance of this conceptualization of the folklore archive and explain the popularity of Facebook.

## CHALLENGE #2: INDEXING AND CLASSIFICATION

It is, of course, not enough to digitize and organize a folklore collection in a database structure, since moving around the collection in a meaningful way requires more than simple query tools. Modern databases thankfully include sophisticated relational structures and query mechanisms that have developed well beyond the important early conceptualizations of the important early folklore database proposed by James LaVita and John Lindow (LaVita and Lindow 1986). But even queries in these recent database systems rely to some degree or another on features already attached to the stored assets. These features are all related to questions of classification.

The classification of the various expressive forms that usually constitute the main focus of a collection is one of the longest standing problems in folklore. From the early classificatory work on the ballad by Svend Grundtvig (1966), to the type and motif indices of the folktale (Uther 2004; Thompson 1955), to the early work on genre by Jolles (1968), to the structural work of Propp (1968), Dundes (1980) and Holbek (1987), the history of folklore studies has been inextricably linked with questions of classification. This connection makes sense, given the comparative nature of folklore study and the concomitant necessity of methods for decomposition, aggregation, and classification. It also means that it is a fundamental challenge in our discipline and an important question to be addressed by a computational folkloristics.

From an information retrieval perspective, there are some fundamental problems with most folklore classification schemes, not least of which is their relative inflexibility. Generally, folklore's "one story–one classification" systems do an excellent job if the research question aligns with the classifier. Consequently, if one is looking for Cinderella tales across many traditions, then the ATU index works well. In the

Cinderella information retrieval problem, the ATU 510a classification points to a series of tales that all share certain features of character and action, allowing the researcher to rapidly discover variants of the tale in collections that subscribe to the ATU classification system. But successful use of schemes such as the ATU index presuppose two things: first, that research questions align with the classifier and second, that everything of interest has already been classified consistently and accurately according to the schema. In the very few instances where this is the case, deploying the classifier returns highly accurate results with few extraneous materials. In the parlance of information retrieval, these classifiers have very high precision rates with very low recall rates. In other words, a search on a corpus properly indexed according to one of these schemes—be it a type index, a motif index, a genre classifier, or an ontological classifier (such as the AFS Ethnographic Thesaurus) and so on—will return all of the items that fit into that classification and only those items. Of course, many of these classifiers are based on a tree structure, and so with a little work one can access items classified in the immediate neighborhood by going up a level and then down the next set of branches, an excellent strategy if the returned results are too sparse. This same observation applies on a micro-scale to the topic indices that are a feature of most published folklore collections; there the problem is slightly different, because collection indices tend to be highly idiosyncratic and do not propose to index materials outside of the very limited scope of the publication itself.

All of these classificatory systems fail if the research question does not align with the classifier, at which point the trade-off between precision and recall is moot. Similarly, these classificatory systems do not scale well. Since most of the folklore classifiers are manually applied, they are not only costly to implement but also have difficulty keeping pace with the increasing size of digital collections. It is one thing to classify a dozen or even one hundred tales according to the ATU index. It is quite another to classify thousands of tales; a case in point is the Tang Kristensen collection, where only a fraction of the tales that could be classified using the ATU system have been assigned numbers—and even many of these assignments are suspect.

Computational approaches to, in this case, text classification may be able to address these problems in a comprehensive manner if we agree that the goal of folklore classifiers is to identify items in a target corpus for intensive analysis while also capturing the relationships between those identified items and the corpus as a whole. For example, one

might have several reasons for wanting to study ATU 510a tales. First, one might want to understand the variant forms the tale takes as different people perform it at different times in a particular tradition group. Second, one might want to contrast aspects of those performances with the performance of fairy tales in the tradition group in general and in the repertoires of a class of storytellers in particular. A sophisticated system that would be part of the folklore macroscope—and that we might want to label the "story space navigator"—would identify stories that share certain features—be they structural or semantic, related to performance or the biography of the informant—with a seed story, and propose other stories that might share features with the seed story and other closely related stories.

Classifying text is not a problem unique to folklore and has been an area of intensive research over the past decade. Many of the existing computational classifiers for unlabeled text data rely on weighted text vectors, usually making use of the TF*IDF (text frequency, inverse document frequency) algorithm for describing the individual texts in a corpus. These vector representations can then be used for various statistical classification engines, such as support vector machines (SVM). Although these approaches provide good first-level approximations of text affinities, they tend to break down at the fine-grained level necessary for folklore analysis. At the same time, supervised machine learning approaches such as SVM and Naïve Bayes classifiers tend to work best on collections much larger than the average folklore collection. Unsupervised methods—such as the much-touted "Principle Components Analysis" (PCA)—are not well suited to the classification of text, since the approach is not easily grasped and the results are very hard to duplicate (Hoover 2008).[3]

To address some of these problems with existing text classification methods, we constructed a multi-modal network classifier for a small subset of the Tang Kristensen collection to explore the potential for such an approach. For each story, we developed a vector representation of the story based on a series of "features." Since folklorists over the years have spent considerable time and effort classifying collections, our approach included these pre-existing classifiers as one of the feature sets, an important consideration in the context of developing computational tools for folkloristics. There is no need to take an either/or approach in the digital realm, where the cost of adding something such as a pre-existing classification scheme is quite small, and the gains can be quite large. In addition, we added simple keywords as features,

derived from a lemmatized list of the overall corpus vocabulary, along with a shallow ontology for the collection that was based on my earlier statistical analysis of "trends" in the tradition (Tangherlini 1994). Finally, we included place names mentioned in the stories as well as the identity of the storyteller as nodes in the network. We then made connections between the nodes representing stories, places and informants based on these features. The advantage of this multi-modal network approach is that the researcher can then decide which features (and modes of the network) to include in any exploratory work. The disadvantage, of course, is that for even a relatively small collection of 942 stories, one can easily generate, as in our case, a network of 2973 nodes with 52663 edges. By the standards of the Internet, this network is miniscule and, since the mathematics of graphs is very well understood, one can discover meaningful structures in what would otherwise be nothing more than a hairball of connections. In other work, we have shown how this method is useful for reclassifying texts in the collection based on researcher driven criteria, and how such a reclassification can highlight unexpected and otherwise undiscoverable similarities between stories (Abello, Broadwell, and Tangherlini 2012). Another advantage of this method is that it scales extremely well, with most feature discovery performed automatically.

One can imagine adding several other features for each item to take advantage of advances in NLP and text classification from other disciplines. A great deal of attention has been focused recently on latent topic discovery for unlabeled text data. In our work, we have explored the application of Latent Dirichlet Allocation (LDA), a probabilistic topic-modeling algorithm that breaks a corpus into a series of topics and assigns documents to multiple topics based on the underlying semantic connections between words in the document, essentially representing each document as a mixture of topics (Blei, Ng, and Jordan 2003; Chang et al. 2009). There are several interesting benefits to this approach: topics at varying levels of granularity (an algorithmic instantiation of Moretti's distant reading) can be captured as features of each item in the collection, possibly weighted inversely to the level of the topic modeling. Edges could be drawn, again weighted according to various thresholds, between items that share topics. Topic modeling also allows one to rapidly assess areas of a collection that one might be interested in and, if implemented properly, drill-down to the underlying resources that contribute to a specific topic; these collections of documents related by latent semantic criteria frequently cut across pre-existing classifiers.

So, for instance, applying LDA with fifty topics to *Danske Sagn*, reveals a forest topic that pulls items from numerous volumes of the collection, and cuts across many of Tang Kristensen's pre-existing categories (Kristensen 1892). One can then interrogate why these stories include the forest topic. Additional unsupervised learning methods could be applied with potentially equal information gains and incorporated into the network model of the corpus.

Another approach that holds significant promise falls under the rubric of automatic story decomposition. While most of the other automatic classifiers work on the semantic level of individual words or the probability of the co-occurrence of words, these algorithms aim at discovering underlying structural patterns in the documents and aggregating these at a higher semantic level. Reminiscent of the structural theories that have played such a formative role in our discipline, stories are broken into constitutive elements, and these elements are then aggregated at a structural level. For instance, preliminary work on an algorithm labeled "Analogic Story Merging" by Mark Finlayson has been able to automatically extract Propp's morphological structures from a corpus of folk tales (Finlayson 2009). Such an approach could be used to identify not only the vocabulary of motifemes in a narrative tradition but also the range of allomotifs for each of those motifemic slots acceptable to a specific tradition group at a specific time, thus instantiating both Dundes' algebraic morphology (Dundes 1980) and providing a test bed for Albert Eskeröd's notion of "tradition dominants" (Eskeröd 1947).

While Finlayson's approach is tuned specifically to narrative, one can easily imagine other algorithms that are tuned to non-narrative space that allow for the identification of structurally equivalent elements of any expressive form such as baskets, dances, or latkes. These structural elements could then be used as the basis for yet another mode in a multi-modal network model of the folklore corpus. Similarly, the various allomotifs discovered during the automatic story decomposition could be included as another class of nodes in a "structuralist" mode of the network. The additive nature of the network classifier approach is appealing as it allows not only for a representation of the multiple layers of complexity that are a characteristic of any folklore collection but also for the discovery and interrogation of patterns based on that complexity. The imagined "story space navigator" also aligns with the "plug-and-play" architecture that Börner envisions, ensuring that the research environment can keep pace with the rapid development of algorithms related to

textual and network analysis by simply plugging new algorithm modules into the network classifier section of the macroscope.

## Challenge #3: The Visual Navigation of a Folklore Corpus

Once we have untangled the relationships between people, places, and things and discovered the underlying timeline for a folklore collection, we need to develop systems that allow us to navigate this space. The navigation must add value, so that one is not simply sailing blind through a digital version of a formerly analog archive—in short, we need to have tools for finding patterns and alerting us to those patterns. While traditional folklore scholarship has relied on our ability to decipher handwriting, to remember where we saw something, and to have access to really big oak tables with Lazy-Susans, computers allow us to rapidly visualize a great deal of our data in complex ways, allowing us to explore parts of the collection from the very close-up views necessary to decipher handwriting to the very distant views necessary to see broad geographic patterns. The folklore macroscope allows us to take advantage of the fact that humans are excellent at detecting visual patterns while well-designed applications can prevent us from over interpreting these visualizations. A great deal of this work accrues to computational areas such as data and network visualization, and geographic information systems. The basic question of this work is how do we leverage the huge amount of work that we have done in acquiring, organizing, and providing dynamic labels for our research materials?

The environment for the visual presentation and navigation of a folklore corpus is quite possibly the key element of the folklore macroscope. As such, it must do something that cannot be done in the analog world. It must provide high quality access to one-of-a-kind archival assets, offer a spatial representation of the various assets that comprise the collection, and incorporate intuitive tools for selecting and moving rapidly between different facets of the collection. Ideally, the navigation of the corpus and the display of the changing collection vistas would be conditioned by the vantage point adopted by the researcher as he or she navigates the collection. The goal of our work has been to allow people to have different entry points into the collection, so that they can explore people, places, stories, and Tang Kristensen's role in creating the collection. This ability to change vantage points as one navigates the corpus holds considerable promise, reducing the importance of the collector-centric viewpoints that have tended to implicitly guide investigations of most folklore corpora.

Early folklorists such as Kaarle and Julius Krohn, as well as their student Antti Aarne, recognized the power of the visual display of complex data as they developed the historic-geographic method in folklore (Krohn 1926). Several years ago, in a deliberately provocative talk at the University of North Carolina, I proposed that we develop a "new historic-geographic method." This new historic-geographic method would focus on exploring latent geo-semantic patterns in folklore collections. Over the years, maps have gotten a bad name in folklore partly because they were enlisted in the search for *urformen* and partly because they were reductionist, offering simplistic representations of folklore data that had not only been reduced to simple point data but also aggregated in ways that erased the role of individuals in the creation and perpetuation of tradition. Maps, however, can be pressed into service to display various aspects of a corpus, including the relationship between individuals and their repertoires, the relationship between a collector, storytellers and the local environment, and the concentration of topics in a specific location. In so doing, they can help guide our research questions (Knowles and Hillier 2008; Gregory and Ell 2007).

In the Danish Folklore Nexus, maps play an important navigational role, allowing one to rapidly visualize the distribution of topics across a region, to get an overview of the distribution of place names in an individual's repertoire, to develop a better understanding of individual mobility, as well as to see the routes of collection that Tang Kristensen traveled as he shaped his collection. In the nexus, a transparency slider allows one to overlay historical maps with contemporary satellite maps that make evident the shifting land use patterns in Denmark during the past century. At the level of the individual story, maps from the time of collection help pinpoint places mentioned in the story, and can be used as part of an analysis of the relationship between the narrated local topography and the mapped environment (Tangherlini 2010).

In recent work, we have been expanding these maps as a simple background for display of point data to an analytical environment taking advantage of the more sophisticated statistical algorithms available in geographic information systems (GIS). For example, in a series of experiments, we were able to show that men tended to tell stories that mentioned places with a greater geographic reach than women, and that the main axis through an ellipse defined by their place name referents aligned with the major transportation routes between the nearest market towns (Tangherlini 2010). Women, in contrast, tended to refer to places more evenly distributed around their place of residence. We

were also able to show that, despite his public presentation of his folklore collecting as a largely west Jutlandic enterprise, Tang Kristensen collected frequently and intensively in northern and eastern Jutland as well (Tangherlini 2010) .

In other work, we have been using "heat maps" as a means for visualizing the concentration of topics or other thematic groupings with intriguing results. For example, a heat map representation of stories about witches, identified by a semantic network of the 30,000 legends in the published Tang Kristensen corpus, revealed a substantial hotspot for the witch topic centered on Grinderslev (Broadwell and Tangherlini 2012). Even after correcting for population density in the surrounding areas, the hotspot was pronounced. Although none of the stories mentioned Grinderslev by name, the area was significantly overrepresented in the corpus by the witch topic. Drilling down into the stories and the historical record led to an interesting discovery. Grinderslev kloster was the site of an important holy spring, *Breum Kilde*, but the monastery was abandoned in the aftermath of the Reformation. The spring, however, was still considered to be spiritually powerful and soon became related—at least in storytelling—to witchcraft. Indeed, the last witch burning in Denmark, the burning of Anne Madsdatter and her sister, took place at *Breum Kilde* in 1686 (Tangherlini 2000). The sisters had caught the attention of the local authorities because of their frequent use of the spring and, during questioning, admitted enthusiastically to their use of the spring for witchcraft. Although the witch burning happened two hundred years before Tang Kristensen's collecting, the heat map reveals that in the popular imagination the area surrounding Grinderslev and the holy spring at Breum were still firmly connected to witchcraft.

Of course, maps should not be the only visualization available in the folklore macroscope. I have shown how different interactions with the collection space allow one to get at the underlying documents for various aspects of the collection, from the images of the manuscripts and their transcriptions to the photographic images of the storytellers. We have experimented with other representations of the corpus space as well, such as the visualization of a navigable multi-modal network representation and multi-level topic models, each of which can comprise an important component of a plug-and-play folklore macroscope. But there is an important caution related to all of this visual presentation of the folklore corpus space: Although humans are adept at discovering patterns in visual data, we are perhaps too adept at it and often get seduced by pretty pictures, proposing patterns where none may exist.

Without accompanying analytical methods, the best visualizations have very little value.

### CHALLENGE #4: ANALYSIS

Computational approaches to folklore corpora can help guide analyses as well as provide evidence to support hypotheses. Perhaps one of the most exciting venues for computational work on folklore collections relates to the conception of the folklore process as the circulation of expressive forms in and across social networks. Social Network Analysis (SNA) is a rapidly expanding field, in part buoyed by the extraordinary commercial success of social media. Recent applications of social network analysis include work that showed an unusual concentration of Tweets related to the "viral" Kony 2012 video emanating from small towns and cities in Alabama, Oklahoma, and Indiana, with a conspicuous cluster emerging in Birmingham *prior* to the video's release. This analysis of the Twitter data helped reveal some of the methods used by a well-funded Christian organization, the Barnabas Group, to promote the Invisible Children project, an effort characterized by its founder as a "Trojan Horse evangelical project," as if it were a grassroots movement. Other interesting work has focused on the role social media have played in circulating and refuting rumors during the Iranian uprisings, where stories such as "Tanks have rolled into Tehran" spiked on Twitter only to be refuted very quickly by a counter spike. In folklore studies, Rob Howard has begun to identify the social networks and the semantic profiles of people based on their posts engaged in a discussion of conspiracies related to childhood vaccinations on *mothering.com*. This work, and work like it, will doubtlessly help us deepen our understanding of the life of stories in complex communities, virtual and real.

The proposed folklore macroscope will allow one to construct a multi-modal network representation of a collection. Since tradition participants are included in the network, one can develop individual profiles for each network member. For a story-based collection, the profile would be calculated based on the stories that an individual tells as well as the story elements that he or she shares with others across the tradition area. Places could also be "profiled" in this same manner. By incorporating time as a feature of such a network representation, one could explore the system as a dynamic model, important if one wants to study problems such as the lifespan of a rumor on a social media site or across the blogosphere. With this type of model, one could trace not only when a rumor starts to emerge, but also among which types of storytellers and

in which types of places. Furthermore, one could identify story cliques where stories are shared frequently between a group of people but who have little impact on the overall network and, conversely, one could identify "super spreaders," people who have the potential to reach and influence large parts of the network. These analytic approaches are still in their infancy and one can only speculate about future developments. To take advantage of these advances, folklorists will need to be ready, both with increased training and awareness of computational methods and with well-structured collections.

## CONCLUSION: COMPUTATIONAL FOLKLORISTICS AND THE FUTURE OF THE FIELD

The idea of the folklore macroscope is, at the very least, a powerful heuristic that allows us to interrogate our research methods and our analytical goals. Devising such a tool—or suite of tools—holds remarkable promise for working with very large corpora. In the folklore macroscope, the underlying social nature of folklore becomes a basic feature of the study environment. Linking people to places and times, and allowing the collection to rest on this foundation provides a far more useful model of the complex, dynamic system that is folklore than earlier models that privileged one dimension over the other. Indeed, the performance anxiety that characterized the field of folklore might be allayed by a system that allows for the holistic engagement with collections irrespective of when they were created. Developing a folklore macroscope is not easy, but it brings to the fore a series of important challenges for the field, computational or not. The judicious exploration and development of computational methods has the potential to advance our understanding of how people, engaged in the never-ending dialectic dance with tradition, create meaning for themselves and each other through folklore.

## ACKNOWLEDGMENTS

I would like to thank the members of the Western States Folklore Society for honoring me with the invitation to present the 2012 Archer Taylor Memorial Lecture. I would like to thank Peter Broadwell for his energy, his insight and his collaboration over the past five years as we have worked on the Tang Kristensen material together. Collaboration with James Abello (Rutgers/DIMACS) helped turn basic ideas into sophisticated realizations. Fascinating collaborative work with the Peter Leonard (University of Chicago) has focused on developing topic

models and visualizations for the corpus. I would like to thank Barbara Hui for her work on data structures, as well as the staff at UCLA's Center for Digital Humanities for early discussions about functionality of the Danish Folklore Nexus. Conversations with Katy Börner (Indiana University), Fil Menczer (Indiana University), Peter Jones (Yale University), Fernando Diaz (Yahoo! Research), Yannet Interian (Google), David Smith (University of Massachusetts), David Mimno (Princeton University) and David Blei (Princeton University) have been invaluable as we developed this work. Inspiration has also been found in conversations with Roja Bandari (UCLA), Scott Weingart (Indiana University) and Rob Howard (University of Wisconsin). A semester spent at NSF's Institute for Pure and Applied Mathematics at UCLA was transformative; at IPAM, I thank Mark Green, Russel Caflisch and Christian Ratsch for their encouragement and support. This work has been funded by grants from the American Council of Learned Societies, the John Simon Guggenheim Memorial Foundation, the National Endowment for the Humanities (NEH Grant HT5001609), the National Science Foundation (NSF Grant IIS-0970179), and a Google Books Humanities grant.

## NOTES

1. It is unlikely, of course, that many readers would make it to the end of the memoirs to give the sentence any glance whatsoever, first or not, given Tang Kristensen's Buckminster-Fulleresque obsession with detail.
2. While this aroma was intriguing on a culinary level, and led students and staff to experiment with Mediterranean salads and some excellent Spanish wine pairings at lunch, it also spelled imminent doom for the microfilm and our easy access to the field diary collection.
3. The results of PCA are generally clustered using one of the many possible clustering algorithms, further muddying the waters. The end result is not transparent and the results are essentially those of a black-box system. One need only ask someone who uses PCA methods on text how they choose their principle components and which features have generated those components to make this problem clear. Similarly, one could interrogate them over the choice of clustering algorithms.

## WORKS CITED

Abello, James, Peter Broadwell, and Timothy Tangherlini. 2012. Computational Folkloristics. *Communications of the Association for Computing Machinery* 55(7):60-70.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3:993-1022.

Börner, Katy. 2011. Plug-and-Play Macroscopes. *Communications of the ACM* 54(3):60-69.

Box, George E. P. and Norman R. Draper. 1987. *Empirical Model-building and Response Surfaces*. Wiley Series in Probability and Mathematical Statistics. Oxford, England: John Wiley & Sons.

Broadwell, Peter and Timothy Tangherlini. 2012. TrollFinder: Geo-Semantic Exploration of a Very Large Corpus of Danish Folklore. In *Proceedings of LREC*. Istanbul, Turkey.

Chang, Jonathan, Sean Gerrish, Chong Wang, and David M Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. *New York* (31):1-9.

Dundes, Alan. 1964. Texture, Text and Context. *Southern Folklore Quarterly* 28:251-265.

———. 1980. *The Morphology of North American Indian Folktales*. FF Communications. Helsinki: Suomalainen Tiedeakatemia.

Eskeröd, Albert. 1947. Årets Äring. *Etnologiska Studier i Skördens Och Julens Tro Och Sed*. Nordiska Museets Handlingar. Lund: Håkan Ohlsson.

Finlayson, M.A. 2009. Deriving Narrative Morphologies via Analogical Story Merging. *New Frontiers in Analogy Research: Proceedings of the 2nd International Conference on Analogy*.

Gregory, Ian and Paul S. Ell. 2007. *Historical GIS: Technologies, Methodologies, and Scholarship*. Cambridge and New York: Cambridge University Press.

Grundtvig, Svend. 1966. *Danmarks Gamle Folkeviser*. Copenhagen: Universitets-Jubilæets Danske Samfund (Akademisk forlag).

Holbek, Bengt. 1987. *Interpretation of Fairy Tales*. FF Communications. Helsinki: Suomalainen Tiedeakatemia.

Hoover, David L. 2008. Quantitative Analysis and Literary Studies. In *A Companion to Digital Literary Studies*. Oxford: Blackwell Publishing Professional. http://www.digitalhumanities.org/companion/view?docId=blackwell/9781405148641/9781405148641.xml&doc.view=content&chunk.id=ss1-6-9&toc.depth=1&brand=9781405148641_brand&anchor.id=0#ss1-6-9_b2 (accessed 20 August 2012).

Jolles, André. 1968. *Einfache Formen*. 4., unveränderte Aufl. Tübingen: Niemeyer.

Knowles, Anne Kelly and Amy Hillier. 2008. *Placing History: How Maps, Spatial Data, and GIS Are Changing Historical Scholarship*. Redlands, CA: ESRI Press.

Kristensen, Evald Tang. 1892. *Danske Sagn, Som De Har Lydt i Folkemunde*. Århus and Silkeborg: Århus Folkeblads Bogtrykkeri.

———. 1923. *Minder Og Oplevelser*. Viborg: Forfatterens Forlag.

Krohn, Kaarle. 1926. *Die Folkloristische Arbeitsmethode*. Oslo: H. Aschehoug & Co.

LaVita, James, and John Lindow. 1986. Software Tools and the Folklore Archive: A Different Perspective. *Computers and the Humanities* 20(2):97-106.

Moretti, Franco. 2000. Conjectures on World Literature. *New Left Review* 1:54-66.

Numeroff, Laura Joffe. 1985. *If You Give a Mouse a Cookie*. New York: HarperCollins.

Ohrvik, Ane. 2011. Conceptualizing Knowledge in Early Modern Norway: A Study of Paratexts in Norwegian Black Books. Ph.D. diss., University of Oslo.

Propp, Vladimir I. A. 1968. *Morphology of the Folktale*. Publications of the American

Folklore Society. Bibliographical and Special Series. Austin: University of Texas Press.

Tangherlini, Timothy R. 1994. *Interpreting Legend. Danish Storytellers and Their Repertoires*. Milman Parry Studies in Oral Tradition. New York: Garland Publishing.

———. 2000. "How Do You Know She's a Witch?" Witches, Cunning Folk and Competition in Denmark. *Western Folklore* 59(3/4):279-303.

———. 2010. Legendary Performances: Folklore, Repertoire and Mapping. *Ethnographia Europaea* 40:103-115.

Thompson, Stith. 1955. *Motif-index of Folk-literature. A Classification of Narrative Elements in Folktales, Ballads, Myths, Fables, Mediaeval Romances, Exempla, Fabliaux, Jest-books, and Local Legends*. Bloomington: Indiana University Press.

Uther, Hans-Jörg. 2004. *The Types of International Folktales : a Classification and Bibliography, Based on the System of Antti Aarne and Stith Thompson*. FF Communications. Helsinki: Suomalainen Tiedeakatemia.