**BROADWELL — ElfYelp**

<3 figures>

**ElfYelp: Geolocated Topic Models for Pattern Discovery in a Large Folklore Corpus—Motivation: A New Historic-Geographic Method for Folklore Studies**

Broadwell, P. M. and Tangherlini, T. R.

Folklore, by its very nature, is deeply linked to place. This link is particularly strong in legends (Gunnell, 2008; Tangherlini, 1994; Lefebvre, 1974). Consequently, being able to discover the correlation between places mentioned in legends and the topics those legends address offers folklorists a powerful method for exploring the complex relationship between storytelling and place. By projecting these correlations into a historically accurate cartographic viewing environment, ‘ElfYelp’ affords researchers an opportunity to explore large, complex cultural archives through a geographically tuned lens.

Early folklorists were intrigued by the possibility of discovering geographic correlations between traditional stories and the places where they were found (Krohn, 1926). These early studies imagined that the mapping of cultural artifacts could help locate the ‘origin’ of those forms, an approach critiqued by C. W. von Sydow (von Sydow, 1948). One of its main failings was the mobility of peoples: where something was collected had little to do with where it originated. Von Sydow’s rebuff had the unfortunate collateral effect of leading many folklorists to abandon cartographic endeavors and to adopt the equally flawed assumption that folkloristic phenomena are evenly distributed across the landscape. With the advent of relatively easy-to-use GIS software and accurate, machine-actionable historical gazetteers, there is now a renewed interest among folklorists in the power of maps.

In this new cartographic method, maps are used to discover latent geographic patterns in the overall corpus, allowing researchers to explore how storytellers have conceptualized the landscape through narrative. For example, are there places in the landscape that are home to more supernatural creatures than others? Discovering patterns leads to important second-order questions: Do certain areas, rich in a set of topics, have topographic or structural features that lead to these associations? Are there political, social, or economic events in these areas that might help explain these patterns?

**Project Overview and Related Approaches**

In the current work, we extend our preliminary research into geo-semantic browsing of Danish folklore (Broadwell and Tangherlini, 2014) by developing a tool to discover and compare different regions’ topic ‘fingerprints’. Our results are further enhanced via a more sophisticated labeling of the Kristensen collection (Broadwell et al., 2014; Kristensen, 1980) and a refined method for the interpolation of geographic regions.

This work builds upon emergent methods in computational data mining and spatial search to support micro-targeted marketing and location-aware search and recommendation services. Sizov’s proposal to generate ‘folksonomies’ of geolocated photograph captions in social media (Sizov, 2010) led to Yin et al.’s technique of ‘Latent Geographical Topic Analysis’ (LGTA), which establishes a ‘location-text joint model’ for geo-LDA (Yin et al., 2011). This model assigns a probability for each topic (based on Gaussian spatial distributions) that any point on the map might ‘generate’ a text containing that topic in some proportion. A refinement to this approach uses polygonal regions that avoid topic overlap as much as possible (Kling et al., 2014). Considerable modifications of these techniques are necessary to address folklore questions. The insistence of the latter approach on avoiding topic overlap, for example, works against the goals of folklorists, since many interesting research questions are based on overlap, such as: Do areas that have a high concentration of ‘ghost’ stories also have high concentrations of ‘wise minister’ stories?

**Data Processing and Analytical Methods**

In our data, each of the 20,431 Danish legends in the target corpus is treated as a locus of place/content co-occurrence. We resolved place references in the stories to approximately 1,750 distinct latitude/longitude pairs via historical gazetteers (DigDag, 2008). Several content-based attributes are associated with each story: a vector of prominent keyword frequencies as well as vectors of 36 high-level and 772 fine-grained categories assigned by the collector and his assistants. We regularized the inconsistencies of the high-level index by using it to train a Naive Bayes classifier that we then applied to the full corpus (Broadwell et al., 2014). To reduce sampling bias from the collector’s habits, we aligned historical census data with administrative records to calculate the population densities for the approximately 2,000 parishes in Denmark, which we used to normalize the co-occurrence matrices by population.

To map the aggregated co-occurrence frequencies of a given story attribute to geographic locations, we calculated the spatial ‘decay’ of the observed z-values based on the standard deviational ellipses for five sample storytellers (Tangherlini, 2010). This model of story ‘decay’ is predicated on an idea of how far a storyteller’s stories might ‘reach’ into the surrounding community. We favor a simple Kriging approach to interpolating between observed points given the decay radius, rather than merely layering Gaussian curves to create ‘hot spots’; importantly, Kriging has the potential to take into account landscape features such as hills and marshes that may impede the ‘reach’ of stories (Cressie, 1993).

Comparing the maps of the original indices to those of the alternative NB classifier reveals hidden geographical affiliations in the collection, much as the differing story labelings between these two classifiers highlighted interesting ‘borderline’ stories (Broadwell et al., 2014). For example, the alternative index assigns more stories mentioning the area around Breum Kilde to the ‘Witches’ category. Previous studies have linked this area to several stories containing oblique references to witchcraft that were not reflected in the collector’s high-level index; these references likely derive from the location’s historical associations with a Catholic monastery that was also the site of the last witch burning in Denmark (Tangherlini, 2000) (see Figure 1).

Yet story labels like those described above, as well as standard topic modeling, do not use as a formative basis the idea that folklore topics may exhibit affinities for particular geographical regions. To explore this aspect of folklore, we employed a version of LGTA that finds a specified number of geographical regions and assigns a set of ‘geo-topics’ to each region in varying proportions; the geo-topics are themselves built from term vectors associated with each story. Story labels based on the 772-label secondary indexing scheme proved particularly effective in constructing geo-topics.

Building co-occurrence matrices of places to topics, labels, and keywords also enables the comparison of arbitrary locations. Ranking a location’s associated attributes allows the identification of regions that share similar ranked lists of topics (e.g. witches, cunning folk, animal disease) via a standard vector comparison algorithm (cosine similarity). LGTA provides such a ranking for geo-topics; for other region features such as keyword co-occurrence counts, we use a weighting system similar to the term frequency-inverse document frequency (TF-IDF) scores commonly used for web searches (Broadwell and Tangherlini, 2012).

**Experiments**

To test the usefulness of our models, we devised two experiments. The first asks, ‘Where are the elves?’ A population density–corrected Kriging map reveals various regions that have a high concentration of stories labeled as ‘elf stories’ by the original collector or the Naive Bayes classifier, or regions that have ‘elf’ geo-topics high in their topic ranking (Figure 2). Consulting a base layer of historically accurate maps allows the user to interrogate the landscape surrounding these areas and ask, ‘What else is here?’ (streams, woods, etc.). Similarly, the population density maps provide a second set of considerations: Do people situate elves in places that are less populated? Finally, the experiment allows a user to consider other areas that have similarly high concentrations of elf stories. Are there certain landscape features that may help explain why certain storytellers associated a particular place and similar places with a specific type of threat?

A second experiment inverts the problem, asking, ‘What do I find here?’ A user draws an arbitrarily large bounding box on the map, and the system returns a ranking of keywords and geo-topics for that geographic area (Figure 3). It also identifies similar regions, based on the ranked list of topics. The underlying premise is that areas for which certain topics appear highly ranked will share features, be they environmental, institutional, or demographic (or a combination of all three). As such, this approach can help stimulate research questions: Are there types of places that were related, in the minds of storytellers, with certain types of stories?

**Conclusion**

ElfYelp provides a productive research environment for interrogating the relationship between storytelling and the concept of place. The tool, which we believe can be a key component of a ‘folklore macroscope’ (Tangherlini, 2013; Börner, 2011), raises more questions than it answers, but in so doing allows for the sustained exploration of how people use stories to comment, construct, and negotiate concepts of space and place.

**Figures**

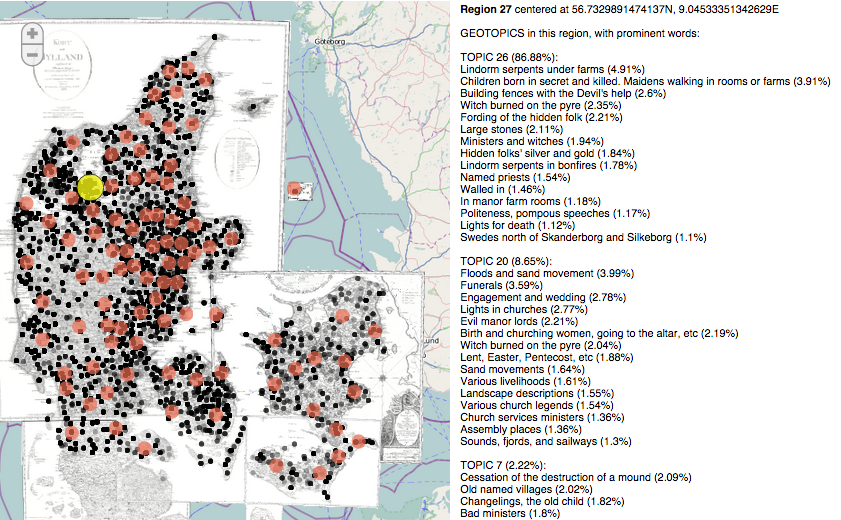


Figure 1. The ElfYelp geo-LDA interface, showing the geo-topics and most prominent labels of a region (highlighted) historically associated with witches.

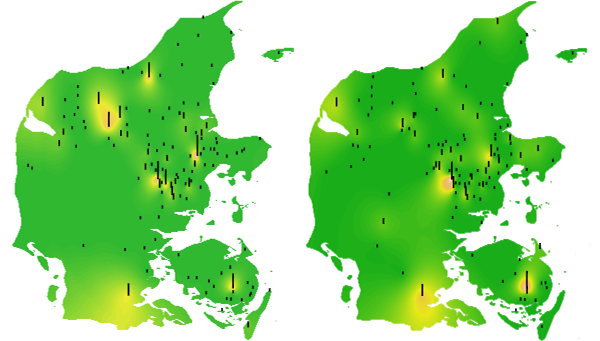


Figure 2: A comparison of the place co-occurrence maps for stories assigned to the ‘Ellefolk’ (Elves) category by Tang Kristensen and his assistants (left) and by a Naive Bayes classifier (right). Simple Kriging was used to interpolate the z-values between the observed locations.

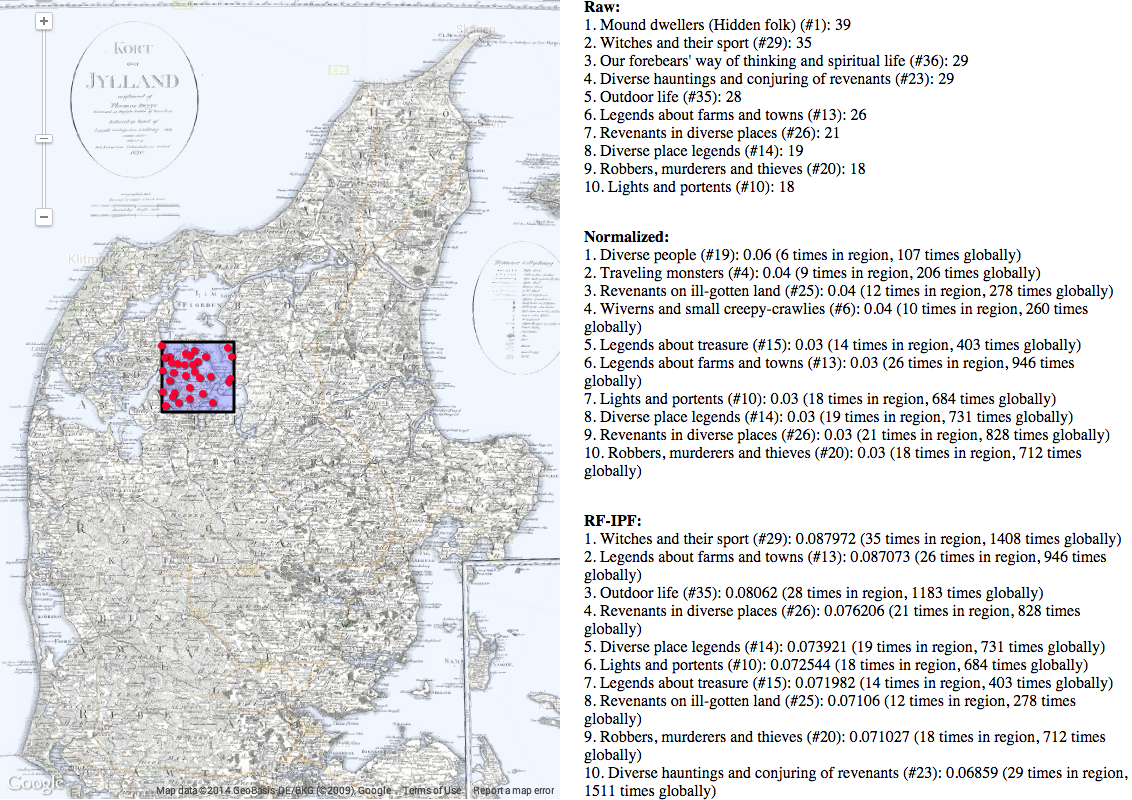


Figure 3. Rankings of Naive Bayes story classifications for the specified region according to the ElfYelp *Spøgelsesskop* (‘spooks-scope’) interface.

**References**

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

**Broadwell, P. M., Mimno, D. and Tangherlini, T. R.** (2014). The Tell-Tale Hat: Reverse Engineering a Folklore Expert. *Proceedings of Digital Humanities 2014*, 7–12 July 2014, Lausanne.

**Broadwell, P. M. and Tangherlini, T. R.** (2012). TrollFinder: Geo-Semantic Exploration of a Very Large Corpus of Danish Folklore. *Proceedings of the Workshop on Computational Models of Narrative*, 26–27 May 2012. Istanbul.

**Broadwell, P. M. and Tangherlini, T. R.** (2014). Forthcoming. WitchHunter: Tools for the Geosemantic Exploration of a Danish Folklore Corpus, *Journal of American Folklore*, special issue on computational folkloristics.

**Cressie, N.** (1993). Statistics for Spatial Data. John Wiley and Sons, New York.

**DigDag: The Digital Atlas of Denmark’s Historical-Administrative Geography.** (2008). http://digdag.dk.

**Gunnell, T. (ed.)** (2008). *Legends and Landscape.* University of Iceland Press, Reykjavik.

**Kling, C. C. et al.** (2014). Detecting Non-Gaussian Geographical Topics in Tagged Photo Collections. *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, 24–28 February 2014, New York.

**Kristensen, E. T.** (1980). *Danske sagn som de har lydt i folkemunde.* Nyt Nordisk Forlag, Copenhagen. Originally published 1892–1908.

**Krohn, K.** (1926). *Die folkloristische arbeitsmethode.* Aschehoug, Oslo.

**Lefebvre, H.** (1974). *La production de l’espace.* Éditions Anthropos, Paris.

**Sizov, S.** (2010). GeoFolk: Latent Spatial Semantics in Web 2.0 Social Media. *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, 3–6 February 2010, New York.

**Tangherlini, T. R.** (1994). *Interpreting Legend: Danish Storytellers and Their Repertoires.* Garland Publishing, New York.

**Tangherlini, T. R.** (2000). ‘How Do You Know She’s a Witch?’: Witches, Cunning Folk, and Competition in Denmark. *Western Folklore,* **59:** 279–303.

**Tangherlini, T. R.** (2010). Legendary Performances: Folklore, Repertoire, and Mapping. *Ethnographia Europaea,* **40:** 103–15.

**Tangherlini, T. R.** (2013). The Folklore Macroscope: Challenges for a Computational Folkloristics. The 34th Archer Taylor Memorial Lecture. *Western Folklore,* **72**(1): 7–27.

**von Sydow, C. W.** (1948). *Selected Papers on Folklore.* Rosenkilde and Bagger, Copenhagen.

**Yin, Z., et al.** (2011). Geographical Topic Discovery and Comparison. *Proceedings of the 20th International Conference on the World Wide Web*, 28 March–1 April 2011,

Hyderabad.