Desperately Seeking House Elves

Problems from Computational Folkloristics

Timothy R. Tangherlini
U.C.L.A.
A story…

It was the old counselor from Skaarupgaard who came riding with four headless horses to Todbjærg church. He always drove out of the northern gate, and there by the gate was a stall, they could never keep that stall door closed.

They had a farmhand who closed it once after it had sprung open. But one night, after he'd gone to bed, something came after the farmhand and it lifted his bed straight up to the rafters and crushed him quite hard. Then the farmhand shouted and asked them to stop lifting him up there. "No, you've tormented us, but now you'll die..."

I heard that's how two farmhands were crushed to death. He wanted to close the door and then they never tried to close it again.
Some meta-data

- Told by Margrete Jensdatter
  - Weaver woman
  - 1813-1902
- Told
  - in 1889
  - in Mejlby
- Mentions
  - Todbjærg (Todbjerg)
  - Skaarupgaard (Skårupgård)
  - The old counselor
- Classified as
  - DS IV: Manor lords, ladies and mistresses
What can I do with this story?

- Some standard questions:
  - Role of ghosts in late 19th century Denmark?
  - Origins of the story?
  - Structure of the story?
  - Who, what, where of this story?

- With all these great questions, is there really a need for computational folkloristics?
How do I find a story?
How do I find similar stories?

- Collections and archives
  - Largely printed or hand written
  - Problems of access
- Collection indices
  - Largely idiosyncratic
  - Domain specific
- Type indices
  - Not uniformly applied / applicable
    - Emic – Etic categories
  - One text – one classifier
Evald Tang Kristensen Collection

- Evald Tang Kristensen (1843-1929)
  - Actively collected from 1865-1923
- Stories collected
  - from ~4500 informants
  - in ~2200 places (mostly in Jutland)
  - mentioning ~12,000 places (also mostly in Jutland)
- 24,000 manuscript pages
  - Covering ~210 fieldtrips
  - 6,000 additional loose leaf pages
  - 2,000 additional pocket diary pages
- 80 published volumes
  - ~100,000 entries
  - 12 separate indices
Connections…

- In 2000, none of this data was digitized
  - Now ~80% of the data is digitized
- Most of this data is “unconnected”
  - Indices to individual collections:
    - Informants and stories
    - Stories and places mentioned
    - Stories by topics
      - Topics are idiosyncratic and not consistently applied
  - Each published collection has its own index
    - Inconsistent orthography
    - Incomplete
    - Published collections are organized largely by genre (Danske sagn, Folkeæventyr fra Jylland, Gamle folks fortællinger om det jyske almueliv, etc)
Archival finding aids

- Archival finding aids are hand-written
  - The master index at the Danish folklore archive is entirely on hand-written cards
  - Cannot be easily OCR’ed
- There is no clear or consistent geographic coding of these materials
Several index cards...
Memoirs as resource

- Tang Kristensen’s memoir, *Minder og Oplevelser* acts as a 4 volume “Rosetta Stone” to the archive
  - Has no index
  - Was not machine readable
- Solution:
  - We OCR’ed and hand corrected the 1200 page memoir
  - We added tags for personal names and place names
  - We added tags for end and start points of field trips
  - Currently:
    - aligning personal names with informant index
    - aligning place names with gazetteer
    - extracting field trip itineraries and aligning with field diary pages
Goal: Use all the information

- Develop biographical information about storytellers
  - Gender
  - Education and Occupation (from MO)
  - Places of residence (from MO)
  - Family information (from MO, Church records and Census records)
  - Dates of birth, confirmation, death (from MO, Church records and Census records)
  - Economic information (from MO, fire insurance protocols and probate records)
- Connect storytellers to places
- Connect storytellers to stories
- Connect stories to places
- Connect storytellers to fieldtrips
  - Gives a date for collection
  - Gives a range of field diary pages for each storyteller
    - Provides an informant index for the field diaries
- Connect stories to field diary pages
- These relations allow us to generate adjacency matrices very quickly based on all or subsets of the data
So that…

- given a particular research question, one can
  - rapidly find and retrieve resources that can help answer that question
  - visually browse the archive on multiple criteria
  - drill-down to underlying resources when one finds something of interest

- existing barriers to the type of research questions that can be posed the archive are removed
Problems, problems, problems
Non-standard resources

- Scanning, OCR, chunking
  - These are not terribly interesting
- Disambiguation / deduplication
- Geocoding challenges
- Connecting everything together
  - Editing / adding material
Disambiguation / Deduplication

The problem of disambiguation is considerable

- Duplication of informants and place names across collection indices
- Informant and Place names with variant spellings
- Multiple places with same name
- Multiple people with same name
Geocoding challenges

- Place names in the collection:
  - when indexed use an idiosyncratic topographical index used only by the Danish Folklore Archive (Skjelborg 1967)
- The topographic index does not easily align with existing gazetteers
  - Orthographic differences
  - Different names for same place
- Historical organization no longer in use / cannot rely on contemporary resources
  - Current Kommune organization different from Amt/Herred/Sogn divisions of the 19th century
  - Boundaries have shifted dramatically
- Existing Danish gazetteers use Easting/Northing
  - Requires conversion for use with most web-based mapping
Some fixes

- Using NLP (Mallet and Lingpipe) to recognize names in unindexed resources

- Using ddupe to disambiguate names
  - See Getoor et. al. 2006
  - Working to use Geo-ddupe to speed process (Getoor et. al. 2007)
  - Find it [here](#)

- Figuring out processes to combine indices by using information about:
  - Places
  - People

- Relatively fast workflow / High accuracy combination of the indices
Adding more information

- **Keyword indexing**
  - Extract keywords from texts
    - Requires language expertise
  - Named Entity Detection
    - Indices are incomplete / additional place names and personal names

- **Shallow ontology**
  - Devise a two-level ontology for collection (Paneva, Rangochev, Luchev 2007; AFS Ethnographic Thesaurus 2007)

- **Genre classification**
  - Rapid classification based on Tang Kristensen’s collection ordering

- **Topic modeling (experimental)**
  - Supervised learning – statistical classifiers
  - LSI and LDA
The Pilot Project: Can one do this in a useful way?

- Five main informants
  - Told approximately 400 stories
    - Added an additional 500 “variant” stories
  - Project aims to
    - Present stories in rich manner
      - Allow for drill-down
    - Connect collection to maps
      - Contemporary
      - Satellite
      - Historical
  - Offer multiple paths for navigation through the collection
    - By person
    - By place
    - By story (keyword, topic, ontology, genre)
    - By time (fieldtrip)
Distant Reading to Micro Reading

- Danish folklore data nexus
- Allow for multiple views on the same archival resources
An (embedded) network model of Folklore

- Folklore, in this model,
  - emerges from the dialectic tension between the individual and tradition
  - and is expressed within and across social networks
- As such, it is also embedded in time and space
A very brief digression

- Rumor as a “hyperactive” transmissionary state of legend
  - Often does not include a strategy for dealing with threat
  - Pushes call for action out of narrative world into real world

- Rumors in social media
  - Hurricane Katrina
  - Political unrest and upheavals (with Roja Bandari):
    - Iran
    - Egypt
    - Tunisia

- “Rumor monger”
  - Rumor discovery on Twitter, Facebook, Blogs
  - Rumor / disinformation
    - Signal profiles in the network?
    - Burstiness
Using a network model

- For understanding connections
  - Between people, places and texts
- For discovering patterns that might be of interest given a particular research question
  - Homophily and similarity – text discovery
- For generating new research questions
How classifiers have failed us…

- Expensive
- Predicated on a particular research question
  - E.g. The Types of International Folk Tales
- One text ✩ One classifier
  - Frequent misclassification
  - Catch-all categories
Some experiments with supervised learning

- NB and Max Entropy classifiers
  - Used 12,000 texts from the *Danske Sagn* collection (approximately 1/9 of the entire collection)
  - Dirty OCR
  - Chunked into individual stories
  - Only looking at text
    - Can we get the machine to classify texts in the same manner as Tang Kristensen did?
First looks

Naïve Bayes Classifier
Some comparisons of classifiers and splits (training target)
So?

- Certain categories are easily recognized by the text classifier
- Some categories are “catch-all”
  - Hekse og deres idrætter
    - Witches and their sport
  - Forsjellige spøgeri og gjenfærds nedmaning
    - Different types of ghosts and their conjuring
  - Gjenfærd på forskellige steder
    - Revenants in different places
  - Bjærgmænd
    - Mound dwellers
- Can I use the network model
  - to refine these classifications?
  - discover new areas of interest?
Using a network map to help with text classification

- Limit the edgelist to story data
  - Keywords
  - Topic Index
  - Shallow Ontology
  - Named entities
- Sort the network
- Find my target story
- Look for stories in the immediate neighborhood
  - layout is no longer random
Using a network map to help with text classification (unsorted)
Using a network map to help with text classification (sorted)
Networks as text classifier

- Places the text in a neighborhood of like minded stories (stories about ghosts)
- Oddly places the text in a neighborhood of stories that mention butter
- Worth a look...
Remember our ghost story?

- DS IV 650
- Classified as a story about manor lords, not ghosts!
- Impossible to find in the archive if one is looking for a ghost story
- Found it in the network neighborhood along with DS II B 147
A story about a house elf at a farm in Egå...

Ends as follows:

- When they got home, the farmhand was happy because now he’d gotten something to use for feed, and afterward nis could go and feed just as much as he wanted to. Then they got another farmhand, and he didn’t want to let him go on like that. But he got lifted up in his bed and all the way up to the rafters, so he lay there dead when people got up the next morning.
More questions

- Is this just an idiosyncratic illustration?
- How can we make use of all of the other network data in a meaningful way?
- Can we use the folklore data for more than just the study of folklore?
Folklore and Google Books

- Can we discover folklore topics in the Nordic literary corpus, 1800-1923, based on our knowledge of folklore topics in Tang Kristensen?
- Can an awareness of the connections between authors (mined from literary histories) help us trace influence?
A little experiment

- Create a series of topics modeled on the Tang Kristensen collection, using Latent Dirichlet Allocation (Blei et al. 2003)
  - 50 topics
- Chunk the short stories of St. Steensen Blicher
  - author known for using folkloric topics and topics from everyday life
- Map the alignment between folklore topics and Blicher
LDA: 50 Folklore topics in Blicher
Possible directions

- Use knowledge of Blicher from literary histories to increase the target group (iterative)
- Map drop-off in the folkloric across the literary landscape
Thanks / Funding

Thanks to:
- Peter Leonard (UCLA)
- Dr. Peter Broadwell (UCLA)

Funding
- Google Books Humanities Grants
- NSF #IIS-0970179 “Network Pattern Recognition for the Humanities”
- NEH #HT5001609 “Network Analysis for the Humanities”