Language Technologies in Humanities:
Computational Semantic Analysis in Folkloristics

Gregor Strle, GNI ZRC SAZU
Matija Marolt, UL FRI

JT DH
29. 9. 2016
Folk Song Lyrics

• Can we analyze **lyrics** and infer
  – song type (e.g. love, moral, legendary, drinking ...)
  – relations between songs

• Melodies in oral traditions are often borrowed, transferred between songs

- love?
- moral?
- legendary?
- death?
- drinking?
- family?
Goal

- Three experiments on a corpus of Slovenian folk song lyrics
  - can we discover topics and conceptual structure of songs?
  - can we classify/group songs according to the topics they describe
Corpus

- Newly created from books *Slovenske ljudske pesmi I-V* ZRC SAZU (1970-2007)
  - scan/OCR

- 4095 Slovenian folk narrative poems
  - from 18th century on
  - 349 variants
  - from 1 to 180 songs per variant
Conversion

- Separate lyrics, metadata
1. **Replacement Rules**  
symbols characteristic of dialect groups (semivowels, diphthongization, pitch accent etc.) are replaced by their grammatical equivalents

2. A **dialect dictionary** is used to translate the words into literary language  
   >18000 words/forms

3. Morphosyntactic tagger for the Slovenian language **Obeliks** was used for lemmatization  
   - tags the words with morphological features  
   - provides **lemmas**

---

A  
Nač predowga, nač prekratka, sej ne bom plesala u nji.

...  
bešta  
bat  
beteg  
...  

tecita  
biti  
bolečin

C  
nič predolg nič prekratek  
saj ne biti plesati v on
Experiments

- Narrow context, just 2 song families:
  - love and fate conflicts
  - family fates and conflicts

- Themes related to death, murder, suicide, infidelity, punishment, e.g.
  - Death of a bride before wedding
  - Nun’s suicide for love
  - Unfaithful student
  - Poisoning of own sister
  - ...

- Strong intertextuality
  - traveling of verses, motifs, and thematic patterns from one song to the other
Experiment one

- **LSA**
  - not **as good in detecting heterogeneity** (three variant types detected)
  - the resulting semantic space generalizes towards the most salient aspects of the corpus

- **LDA**
  - can associate topics with different variant types
  - more even distribution across topics

<table>
<thead>
<tr>
<th>LSA variant types and dimensions</th>
<th>LDA variant types and topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEATH OF A BRIDE BEFORE WEDDING</td>
<td>DEATH AT A REUNION</td>
</tr>
<tr>
<td>d1: mother child young baby shepherd wreath blood</td>
<td>t1: heart boy Breda head sad hunter Danube</td>
</tr>
<tr>
<td>d4: Ljubljana linden lover boy seduce chamber</td>
<td>MURDER OUT OF JEALOUSY</td>
</tr>
<tr>
<td>Tonček</td>
<td>t2: love sword kneel sharp neighbor boyfriend blame</td>
</tr>
<tr>
<td>d5: Breda Ljubljana groom mother-in-law linden baby Turk</td>
<td>BRIDE INFANCIDE</td>
</tr>
<tr>
<td>d6: Breda accident evil house mother-in-law sister groom</td>
<td>t3: home shepherd Mary uncle birth shred rockcradle</td>
</tr>
<tr>
<td>d8: Ljubljana brother linden sea shirt prefer wash lover</td>
<td>UNFAITHFUL STUDENT/NEW PRIEST</td>
</tr>
<tr>
<td>NUN'S SUICIDE FOR LOVE</td>
<td>t4: undertaker love priest parish love promise letter</td>
</tr>
<tr>
<td>d2: convent Ursula nun baptism godmother ring blood</td>
<td>NUN'S SUICIDE FOR LOVE</td>
</tr>
<tr>
<td>d3: convent Ursula nun baptism godmother shepherd wreath</td>
<td>t5: love Uršika convent boy Jesus farewell sword</td>
</tr>
<tr>
<td>HUNTER SHOOTS HIS LOVER AND HIMSELF</td>
<td>REJECTED LOVER</td>
</tr>
<tr>
<td>d7: newpriest grave bury church rifle hunter student</td>
<td>t6: seduce blood house Vida linden Ljubljanians death</td>
</tr>
<tr>
<td>d9: Ljubljana linden rifle grave hunter shaking leaves</td>
<td>ABANDONED ORPHANS</td>
</tr>
<tr>
<td>d10: rifle hunter shaking Tonček leaves face pale</td>
<td>t8: bury window chamber wound grow crying dead</td>
</tr>
<tr>
<td>PUNISHMENT FOR THE WICKED SONS AND DAUGHTERS-IN-LAW</td>
<td>MISTRESS' LOYALTY REPAID</td>
</tr>
<tr>
<td>t9: gold sea mountain rooster fear crying darling son</td>
<td>t10: boy fenced heart nosegay dead grieve loyal</td>
</tr>
</tbody>
</table>

Voronoï diagram represents topological projections of both methods
Experiment two

- Do LDA topics correspond to song families?
  - can we distinguish between love and fate conflicts vs. family fates and conflicts
  - difficulty: intertextuality, themes in both are similar

- Agglomerative **hierarchical clustering** to cluster variant types according to
  - similarity of their average topic distributions

- Result
  - the semantic space does include some notion of song families
  - enables us to place individual (also new or unknown) songs into this space and study their relations to existing materials.

<table>
<thead>
<tr>
<th>family clusters 1 (2:6) and 4 (13:31)</th>
<th>love clusters 2 (17:11) and 3 (6:4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>happy earth unfortunately rifle son mother remember</td>
<td>field three maid sun golden like ark sea lover</td>
</tr>
<tr>
<td>noble castle son stand cry dress letter dress give mother wife children find gold adultery measure colorful stick boy</td>
<td>things husband voice eat say young white know sin school</td>
</tr>
<tr>
<td>mountain will water mother hero angry dam girlfriend mother-in-law</td>
<td>mistress unlock boy saint window pot die lie</td>
</tr>
<tr>
<td>brother father house dear ours sister see</td>
<td>stepmother run home getup graveyard rough get out go home</td>
</tr>
<tr>
<td>tender live leave quickly name call barely crown world beg</td>
<td></td>
</tr>
</tbody>
</table>
Experiment three

• Can LDA detect major themes characteristic for individual variant types

• Supervised learning: **Labeled LDA**
  – predefined labels for topical distributions
  – LLDA learns topic distributions for the labels

• Manually annotated selected variants with labels (18% of the corpus)
  – trained the model

• Inference on the entire corpus
  – yields distributions over labels for each song
Experiment three

- Most variants share multiple topics, with the main topic for each shown as most salient
  - e.g. Mother prevents her son’s marriage
- Disambiguation of similar topics (e.g. unhappy love)
Side project - TextExplore

- Enable non-programmers to experiment with topic models
Side project - TextExplore

- Enable non-programmers to experiment with topic models
  - import corpus
  - create topic models (Mallet)
  - visualize documents, topics, time, location
Conclusion

• LDA can uncover typical characteristics of individual variant types
  – enables classification of unknown materials
  – discover relationships (similarities and differences) in the corpus

• Future work:
  – more song families
  – further develop visualization, exploration
  – relations between lyric and melodic spaces