

Univerza v Ljubljani  
Fakulteta za računalništvo  
in informatiko



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?



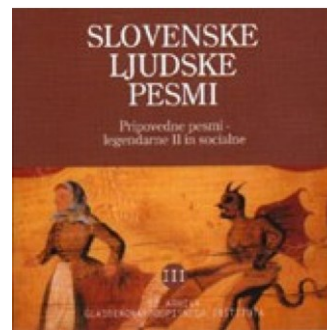


# Corpus

- 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



V





# Conversion

- Separate lyrics, metadata

mikrofon vključen. Sprva raztrgana, skromna obnova je mimogrede prerasla v samostojno verzijo naše pesmi — B. V nji najdemo nekaj novih prvin, ki jih tu natisnjena verzija (A) nima:

Matjaža so dali v ječo. Linčica se je vanj zaljubila, menda je bil lep človek. Linčica je dala »dór-mido« (uspavalni napoj ali prašek) očetu in materi, kasneje pa je uspavala tudi stražo. O mostu, ki je slonel na lažnih opornikih in ki sta ga ubežnika

zadnji hip pred prihodom zasledovavcev zažgala, slišimo, da je bil na meji (turške dežele).

Za pesem, ki je na sarnem začetku okrnjena, pomenijo ti dodatki dragoceno dopolnilo. Ob nadržnosti, da je Linčica dala očetu in materi napoj za spanje (dopolnilo pevkine sestre Marije), se je Ana celo glasno začudila, kot bi hotela reči: »Glej glej, tega pa nisem vedela!« ali »Na to pa sem pozabila!«

6.

Kraj: Bila v Reziji

Pela in pripovedovala: Ana Buttolo, vd. Zanetti, Pécawa (1894)

B

Z p.s.: Matičetov-Vodušek, 20. junija 1963

$\text{♩} = 122 - 136$  Poco rubato

1. Lin-či-ca tur-kin-če-ca, na wze-la no-ga Won-gar-ja.  
 3. — Kar je bil pro-žo-ni?, (i.d.)

<sup>1</sup> Od 4. kitice dalje na tem mestu »«.

- |  |  |
|--|--|
| 1. Linčica Turkinčica<br>na wzela noga Wóngarja.                     | 9. Dwa dni na jě mu púlela<br>nú na jě mu fys raklá:                             |
| 2. To bila noga kraja šči<br>nu won to bil de Wóngar krei,           | 10. »Da ti me vliku fes plažáš<br>nu ti maš víde' me womožé»                     |
| 3. ke an je bil prožonír<br>nu za tri dni je bil zažán.              | 11. ke ja éon te wiğá' wozdē,<br>ja bon te fis wiğál' wozdē.«                    |
| 4. Nu Linčici je zaplažél,<br>na je prosila grač(jico). <sup>1</sup> | 12. »Káko béj maš mé wóženē,<br>ke já si wžž wóženjen,                           |
| 5. Nu won je jo koncédinal,<br>na drě delá pokléknula                | 13. a) ma já man žče dwa bratra tapar<br>iši,<br>bōj lipča nekuj mle.«           |
| 6. nu nōge na mu búšnula<br>nu drě ga búšnula pa njağá.              | 13. b) man dwa lipča bratra někoj mle, <sup>2</sup><br>ke tédwa tó bo te wzéto.« |
| 7. »Da Linčeca Turkinčeca,<br>wse tō ke boš me bārala,               | Nu dwa dni na je mu púliła jěst<br>anu kjūče na je ískala                        |
| 8. já ja bon te kontantəl,<br>ke ti si moja sama ščíl«               | wod te vílíkih dur tow Fránčiji,<br>ke grejo skuz,<br>na vílika galeríja skuz,   |

6. B <sup>1</sup> Linčica Turkinčica / je vzela Ogra, / <sup>2</sup> je bila kraljeva hči / in on je bil kralj Oger, / (5) <sup>3</sup> ki je bil jetnik, / in za tri dni je bil zaprt / <sup>4</sup> in Linčica se je zaljubila vanj, / Ona je prosila [očeta] dobroto / [da bi nosila 3 dni jetniku jest v ječo]. / <sup>5</sup> In on je privolil. / (10) Ona je dol pokleknila / <sup>6</sup> in noge mu poljubila / in brě poljubila še njega. / <sup>7</sup> »Oj Linčica Turkinčica, / vse, kar me boš prosila, / (15) <sup>8</sup> jaz te bom zadovoljil, / ker ti si moja edina hči!« / <sup>9</sup> Dwa dni mu [jetniku] je nosila [jest] / in ona mu je rekla: / <sup>10</sup> »Ti si mi prav močno všeč / (20) in ti glej, da se oženiš z mano, / <sup>11</sup> ker jaz te bom rešila od tod, / jaz te bom resnično rešila od tod!« / <sup>12</sup> »Káko se boš omožila z mano, / saj jaz sem že oženjen! / (25) <sup>13</sup> a) Ampak jaz imam še dva brata doma, / lepša od mene.« / b) Imam dva lepša brata kot sem jaz, / ta dva te bosta vzela!« / In dva dni mu je nosila jest, / in ključe je iskala / (30) od velikih duri v Franciji, / ki gredo skozi, / skozi velik



# Conversion

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 Nəč predowga, nəč prekratka,  
sej ne bom plesala\_u nji.

...  
bešta                      tecita  
bət                         biti  
beteg                      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

**LSA variant types and dimensions**

**DEATH OF A BRIDE BEFORE WEDDING**  
 d1: mother child young baby shepherd wreath blood  
 d4: Ljubljana linden lover boy seduce chamber Tonček  
 d5: Breda Ljubljana groom mother-in-law linden baby Turk  
 d6: Breda accident evil house mother-in-law sister groom  
 d8: Ljubljana brother linden sea shirt prefer wash lover

**NUN'S SUICIDE FOR LOVE**  
 d2: convent Ursula nun baptism godmother ring blood  
 d3: convent Ursula nun baptism godmother shepherd wreath

**HUNTER SHOOTS HIS LOVER AND HIMSELF**  
 d7: newpriest grave bury church rifle hunter student  
 d9: Ljubljana linden rifle grave hunter shaking leaves  
 d10: rifle hunter shaking Tonček leaves face pale

**LDA variant types and topics**

**DEATH AT A REUNION**  
 t1: heart boy Breda head sad hunter Danube

**MURDER OUT OF JEALOUSY**  
 t2: love sword kneel sharp neighbor boyfriend blame

**BRIDE INFANTICIDE**  
 t3: home shepherd Mary uncle birth shred rockcradle

**UNFAITHFUL STUDENT/NEW PRIEST**  
 t4: undertaker love priest parish love promise letter

**NUN'S SUICIDE FOR LOVE**  
 t5: love Urška convent boy Jesus farewell sword

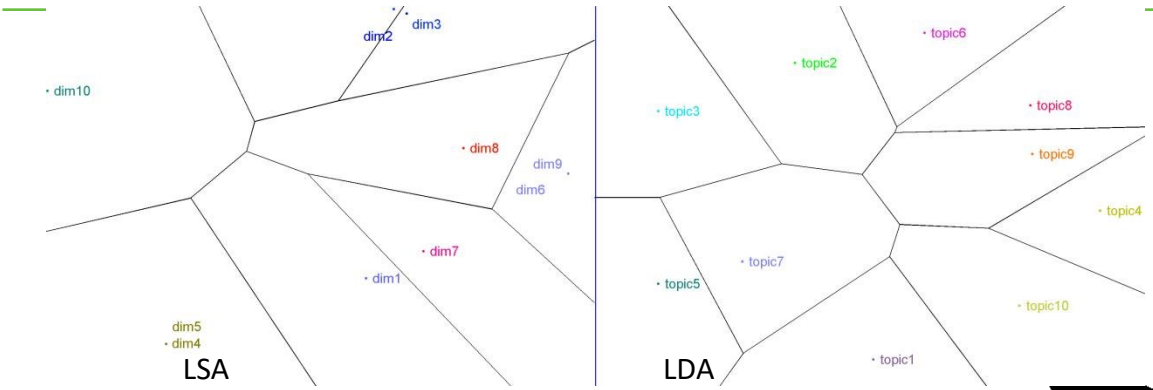
**REJECTED LOVER**  
 t6: seduce blood house Vida linden Ljubljanians death

**WIDOWER ON BRIDE'S GRAVE**  
 t7: tender abandon blood bread jesus rockcradle married

**ABANDONED ORPHANS**  
 t8: bury window chamber wound grow crying dead

**PUNISHMENT FOR THE WICKED SONS AND DAUGHTERS-IN-LAW**  
 t9: gold sea mountain rooster fear crying darling son

**MISTRESS' LOYALTY REPAID**  
 t10: boy fenced heart nosegay dead grieve loyal



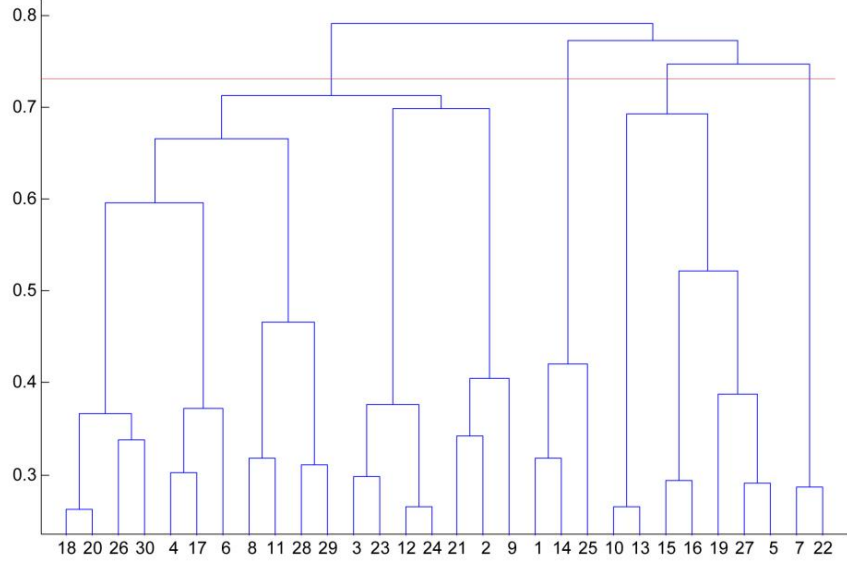
Voronoi diagram represents topological projections of both methods





# 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.



**family clusters 1 (2:6) and 4 (13:31)**

- hunter earth unfortunately rifle **son mother** remember
- noble castle **son** stand cry dress letter dress give
- **mother wife children** find gold adultery measure colorful stick boy
- mountain will water **mother** hero angry dam girlfriend mother-in-law
- **brother father** house dear ours sister see
- tender live leave quickly name call barely crown world beg

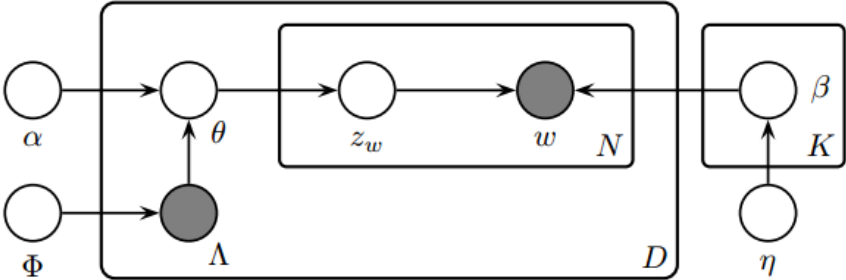
**love clusters 2 (17:11) and 3 (6:4)**

- field three maid sun golden like ark sea **lover**
- things **husband** voice eat say young white know sin school
- **mistress** unlock boy saint window pot die lie
- stepmother run home getup graveyard rough get out go home



# Experiment three

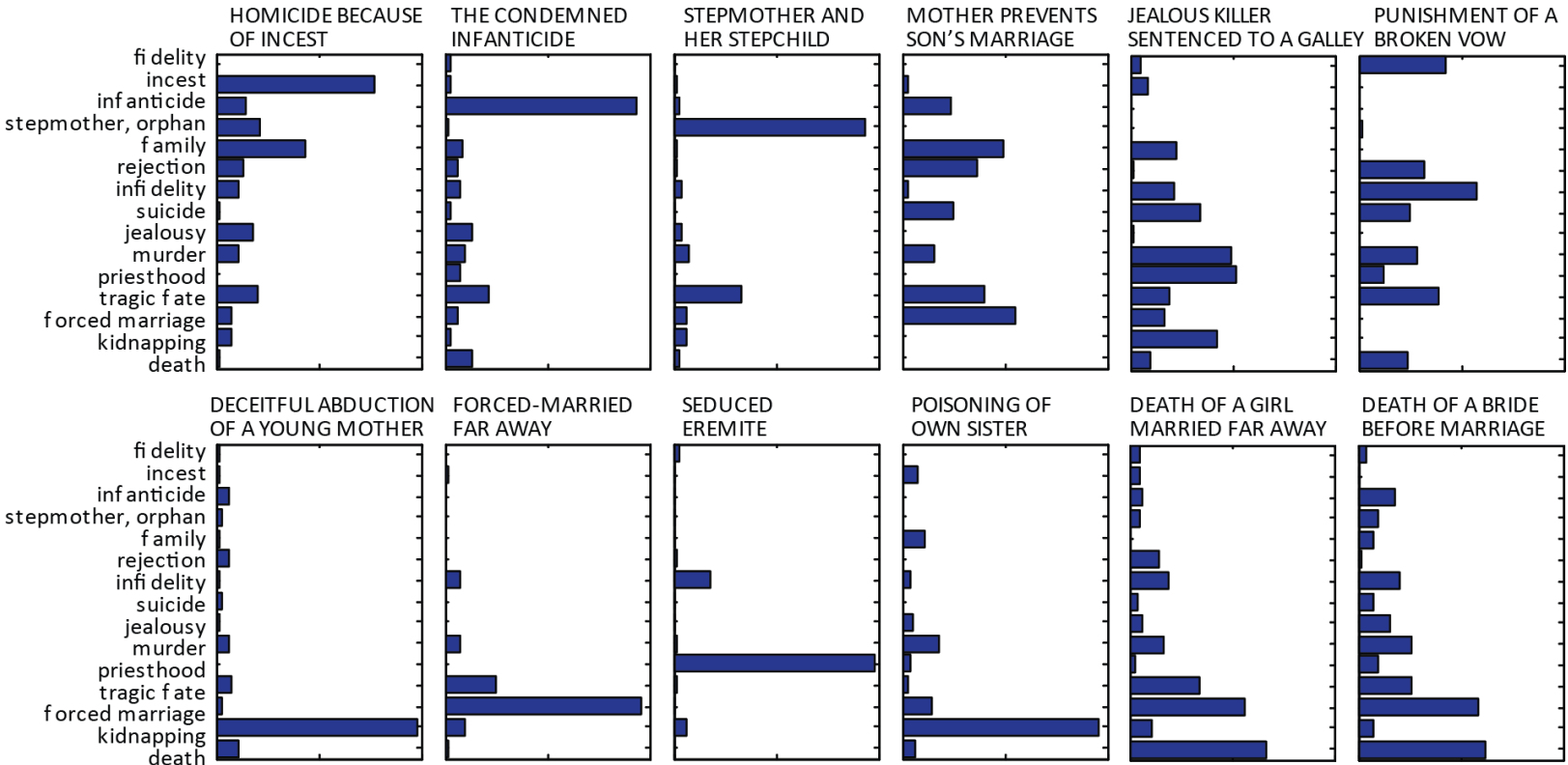
- Can LDA detect major themes characteristic for individual variant types
- Supervised learning: **Labeled LDA**
  - predefined labels for topical distributions
  - LLDA learns topic distributions for the labels
- Manually annotated selected variants with labels (18% of the corpus)
  - trained the model
- 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

EXPLORE

🔍 ubiti , ljubiti ⚙️ REFINE YOUR SEARCH

Location ▾ Year ▾ Corpus ▾

## TOPIC

**topic**

- iti
- žlahten
- priiti
- stati
- trije
- govoriti
- reči
- grad
- toko
- hitro
- dati
- marjetica
- hud
- jest
- storiti
- prinesti
- vzeti
- ven
- kri
- dom

**TOP DOCUMENTS IN** clear selected year

266.

**NEZVESTA GOSPA S TREMI STRAŽARJI - A / 5.**

Ljudska pesem  
1839, okolica Celja

1 *Gospa postavlja vahte tri,  
kedaj bojo šli gospod damo.*

2 *U prvi uri te noči* [...] ▬

22.

**PRED ZMAJEM REŠENO DEKLE / 1.**

Ljudska pesem  
1839, Kranjsko



# Side project - TextExplore

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

266. NEZVESTA GOSPA S TREMI STRAŽARJI - A / 5.

Ljudska pesem  
1839, okolica Celja  
SLP5: DRUŽINSKE PRIPOVEDNE PESMI

PREVIEW	LEMMATISED	TOPIC -	
<b>topic</b>	ljubica iti fant stati reči lipa odpreti micka priti ležati svoj ljub roka kam sinoči hitro vzeti dekcle ljuba mlad		%
<b>topic</b>	srce oče ljub videti mati svoj kri sin gora hoteti vzeti moči storiti kak priti hiša prositi sestra dati smrt		%
<b>topic</b>	svet duša iti anton pekel vrata trije bog menih ven nebo slišati vprašati hitro star priti peter nebesa jest močno		%
<b>topic</b>	iti govoriti sestrica študent jest toko mlad imeti vol dati priti glava trije zlat seja brat bel črn neja bratec		%
<b>topic</b>	trije priti mrtev hanžek iti ležati delati fantič hitro dati zvoniti nesti fant štirje bel bolan živ umreti grob vstati		%
<b>topic</b>	ime roka hoditi imeti drag živeti smeti znati priti črn mlad svoj grob puška gora lovec povedati dekcle vzeti soldat		%
<b>topic</b>	marija jezus svet božji iti nesti pot gora devica nebeški nebesa priti prositi marij pomagati trije ljub vprašati usmiljen jožef		%
<b>topic</b>	mlad dati iti bog ančka bel imeti vida hala sinek olga star kača hlapec stati prelep adam eva peljati gledati		%
<b>topic</b>	voda morje marija grešnik jezus greh stati teči človek gora vstati bister jen barka svet barčica plavati drevo goreti moči		%
<b>topic</b>	bog ležati priti poslati dati vzeti umreti janez bolan iti svet duša živeti deklica mašnik imeti moliti dekcle spovedati milo		%



# 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

