# Uncovering Semantic Structures within Folk Song Lyrics

**Gregor Strle**
Institute of Ethnomusicology
Research Centre of the Slovene Academy of Sciences and Arts
gregor.strle@zrc-sazu.si

**Matija Marolt**
University of Ljubljana
Faculty of Computer and Information Science
matija.marolt@fri.uni-lj.si

## ABSTRACT

In this paper, we focus on computational methods for natural language processing (NLP) and evaluate some possibilities that NLP methods offer to folkloristics. Due to inherent dialectical diversity and strong intertextuality, folkloristic materials have generally proven to be very challenging for NLP. Our goal was therefore to evaluate different NLP methods and study the semantics generated by respective approaches and their practical implica-tions. Three experiments analyzing a collection of Slovenian folk narrative poems are presented and results discussed.

## 1. INTRODUCTION

Growth of digital collections in recent years has motivated interdisciplinary research that connects various fields of computer science and humanities. For example, machine learning techniques can today complement researcher's analysis to uncover latent semantic structures in music, visual materials and text. They go beyond limitations of human (manual) analysis and are especially useful in processing and classification of materials, as they can inspect large amounts of data in relatively short time.

In this article, we focus on computational methods for natural language processing (NLP). NLP is used to solve a wide variety of tasks: machine translation, optical character recognition (OCR), parsing (grammatical analysis), speech recognition and semantic analysis (word-sense disambiguation, topic recognition). We are interested in the latter.

We present three experiments of using NLP to analyze a collection of Slovenian folk poems. In the first, the goal was to get an insight into the conceptual structure of the materials and at the same time discover advantages and disadvantages of two main NLP approaches. In the second experiment, we wanted to assess whether the automatically obtained topics correspond in any way to annotated song families. In the third experiment, we wanted to assess whether topic distributions in any way correspond to the major themes of individual variant types.

## 2. CORPUS

For our experiments, we selected 1,965 variants of Slovenian folk narrative poems (Golež Kaučič, Kumer, Terseglav, & Vrčon, 1998; Golež Kaučič, Kumer, Šivic, Terseglav, & Vrčon, 2007), part of our multimedia digital library EthnoMuse (Strle & Marolt, 2012). The selection includes narrative poems about love and fate conflicts and about family fates and conflicts. The songs date back to 18th and 19th century, with some variant types represented by only one variant, whereas others have up to 180 versions and are still sung today. Thematically, the variants are closely related, as they share similar stories about death, murder, suicide, infidelity, punishment, etc. Moreover, strong intertextuality is present through the whole corpus, which reflects a characteristic folk song phenomenon: traveling of verses, motifs, and thematic patterns from one song to the other. This has strongly affected the results, as most occurring themes and motifs dominated over rarer variant types.

## 3. EXPERIMENT 1

Our first experiment has focused on the general characteristics of Slovenian folk song lyrics at the level of poetic (variant) types and topics that intertwine within them. We wanted to get an insight into the conceptual structure of the materials and at the same time discover advantages and disadvantages of two main NLP approaches: a statistical associative approach using Latent Semantic Analysis (LSA (Landauer & Dumais, 1997)) and a probabilistic topic-modeling approach using Latent Dirichlet Allocation (LDA (Blei, Ng, & Jordan, 2003)). There are significant differences between the two in terms of semantics and context they generate.

The synthetic nature of Slovenian language with many morphological rules, as well as strong dialects in singing, which are reflected in transcriptions, made it necessary to lemmatize song lyrics before analysis. We first replaced special characters used for encoding characteristics of dialect groups (such as semivowels, diphthongization, pitch accents etc.) by their grammatical equivalents. A dialect dictionary was then used to translate the words into literary language and finally a statistical morphosyntactic tagger for the Slovenian language (Grčar, Krek, & Dobrovoljc, 2012) to lemmatize the text.

### 3.1 Results

LSA and LDA were both performed on lemmatized documents that were converted into word-document matrices with word counts weighted according to the tf-idf statistics. For both types of analyses, we limited projections of the word-document space to 10 dimensions/topics, as the size of the corpus is relatively small. Table 1 shows the most salient words and documents for each method according to variant types they mostly represent. While

| *LSA variant types and dimensions* | *LDA variant types and topics* |
|---|---|
| DEATH OF A BRIDE BEFORE WEDDING | DEATH AT A REUNION |
| d1: mother child young baby shepherd wreath blood | t1: heart boy Breda head sad hunter Danube |
| d4: Ljubljana linden lover boy seduce chamber Tonček | MURDER OUT OF JEALOUSY |
| d5: Breda Ljubljana groom mother-in-law linden baby Turk | t2: love sword kneel sharp neighbor boyfriend blame |
| d6: Breda accident evil house mother-in-law sister groom | BRIDE INFANTICIDE |
| d8: Ljubljana brother linden sea shirt prefer wash lover | t3: home shepherd Mary uncle birth shred rockcradle |
| NUN'S SUICIDE FOR LOVE | UNFAITHFUL STUDENT/NEW PRIEST |
| d2: convent Ursula nun baptism godmother ring blood | t4: undertaker love priest parish love promise letter |
| d3: convent Ursula nun baptism godmother shepherd wreath | NUN'S SUICIDE FOR LOVE |
| HUNTER SHOOTS HIS LOVER AND HIMSELF | t5: love Uršika convent boy Jesus farewell sword |
| d7: newpriest grave bury church rifle hunter student | REJECTED LOVER |
| d9: Ljubljana linden rifle grave hunter shaking leaves | t6: seduce blood house Vida linden Ljubljanians death |
| d10: rifle hunter shaking Tonček leaves face pale | 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 |

**Table 1.** Top words and the corresponding variant types for LSA dimensions and LDA topics.

similarities between words in relation to variant types are relatively similar for both analyses - compare for example LSA and LDA analysis of variant type 'Nun's suicide for love' in dimensions 2-3 and topic 5 - there is a significant difference in the detection of variant types across the corpus. As shown in the table, LSA could detect only three variant types that dominate across semantic space, whereas LDA successfully detected heterogeneity of the corpus and associated each topic with a different variant type. As LSA cannot account for topical distribution, and hence lacks additional hierarchy, it has difficulty detecting heterogeneity and the resulting semantic space repeatedly generalizes towards the most salient aspects of the corpora. LDA's projections, on the other hand, are more balanced: less overlapping and better at detecting heterogeneity.

## 4. EXPERIMENT 2

The goal of our next experiment was to assess whether LDA topics correspond in any way to song families. As described previously, our corpus consists of two main song families: poems about love and fate conflicts and poems about family fate conflicts. Although the themes in both are similar, we wanted to assess whether the topics discovered by LDA have any correspondence to the two families. We first calculated distributions over LDA topics for individual variant types within both families by averaging topic distributions of all variants of each type. The purpose of averaging was to reduce the imbalance of the number of variants of each poem type in our corpus, as some types have many (over 50) variants, while others have just a few. This resulted in a set of 90 topic distributions for all 90 variant types in our corpus.

We then used the cosine similarity measure, often used in text retrieval, to cluster the variant types based on similarities of their topic distributions by using the agglomerative hierarchical cluster tree method. We examined the obtained clusters to find out whether they relate to the division of poems into families. The ratio between love and family ballads taken from 4 major clusters is shown in Table 2. As shown in the table, topics about family relations (e.g. 'son', 'mother', 'brother', 'father', 'wife', 'mother-in-law' etc.) correspond more to clusters 1 and 4, which have a higher ratio of family ballads, whereas clusters 2 and 3 include more love oriented topics. Thus even though the entire corpus contains strong intertextuality and themes in both families are very related, the obtained semantic space does include some notion of song families and 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)* | *love clusters 2 (17:11) and 3 (6:4)* |
|---|---|
| hunter earth unfortunately rifle son mother remember | field three maid sun golden mara ark sea lover |
| noble castle son stand cry dress letter dress give | things husband voice eat say young white know sin school |
| mom wife children find gold adultery measure colorful stick boy | mistress unlock boy saint window pot die lie |
| mountain will water mom hero angry dam girlfriend mother-in-law | stepmother run home getup graveyard rough get out go home |
| brother father house dear ours sister see | |
| tender live leave quickly name call barely crown world beg | |

**Table 2.** Distribution of songs about love and family - clusters (love:family) and topic descriptions (top words) are shown.

## 5. EXPERIMENT 3

The third experiment focused on topic distributions among variants in order to assess whether LDA can detect major themes characteristic for individual variant types. We used a supervised learning method, Labeled LDA (LLDA). The procedure is similar to basic LDA analysis, with few minor differences. LLDA is a supervised topic model that uses predefined labels for calculating the topical distributions of the corpus. Thus, we first manually annotated selected variants with labels corresponding to the major theme (or themes) of particular variant (many of the labels represent the most occurring themes in the corpus). Next, we used this annotated dataset (around 18% of the whole corpus) to train our model. In the final phase, the whole corpus was used for inference to find variants best associated with each label globally. Results are shown in Figure 1, which shows thematic structure for selected variant types.

Most variant types share multiple topics, with the main topic for each type shown as most salient. For example, variants of type '*Mother prevents her son's marriage*' cover a wide range of topics, predominantly 'forced marriage', 'family', 'rejection' and 'tragic fate' of course, which in some variants result in 'murder', 'suicide' or even 'infanticide' due to illegitimately born child. In '*Punishment of a broken vow*', a young bride-to-be is torn between her (or her mother's) vow to Jesus (for her to live in a monastery as a nun) and a vow given to her lover, consequently being punished for her weakness; hence topics 'fidelity'/'infidelity', 'rejection', 'murder', 'death' and 'suicide', with a pinch of 'priesthood' in-between. LLDA can disambiguate different senses of unhappy love. For example, thematically similar to the previously mentioned '*Punishment of a broken vow*' are two variant types '*Death of a girl married far away*' and '*Death of a bride before marriage*', but here the emphasis is on 'forced marriage' and consequently 'death'.

Some variant types have a single dominant topic. For example, type '*Homicide because of incest*' has appropriate topic distribution with 'incest' prevailing over other context-related topics, such as 'family' and 'tragic fate', which are activated to a lesser extent. The extreme cases of single topic dominance are for example the three variant types '*The condemned infanticide*', '*Stepmother and her stepchild*' and '*Deceitful abduction of a young mother*', with the former two having at least 'tragic fate' in common.

An interesting example is variant type '*Poisoning of own sister*'. Here, the predominant topic given by LLDA is 'kidnapping', even though in the second part of the ballad the story slowly gravitates towards enmity between the two sisters, Zarika (Dawn) and Sončica (Sun), and ends with the former murdering the latter. But it is important to point out that 'murder' is only vaguely expressed in the ballad, with no hint before the metaphorical last verse "*The sister does not recognize her / And gives snake's poison for her to drink*", whereas 'kidnapping', and later ransom and rescue from "strong and evil" Turks, is made explicit.

As a rule, topic distributions for variant types with strongly expressed or dominant topics (such as 'infanticide' and 'kidnapping') will often show single topic domination, whereas types with less dominant or more obscure themes (e.g. 'tragic fate' or 'rejection') will commonly share multiple topics. Moreover, due to strong in-

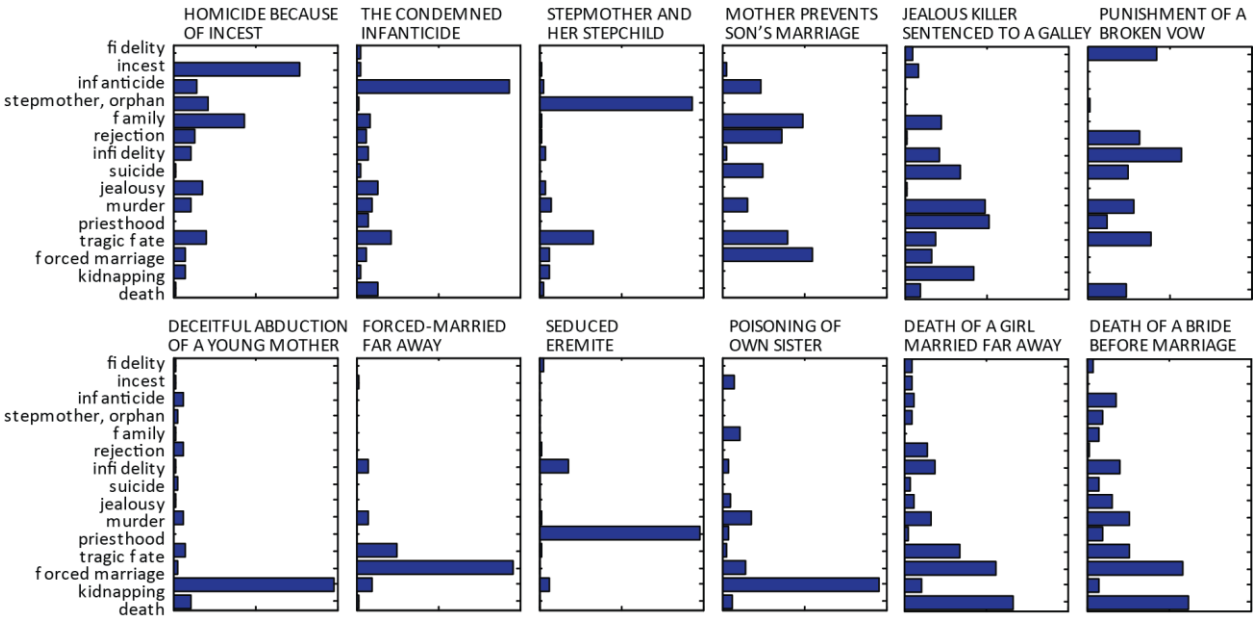


**Figure 1.** Topic distribution for selected variant types. 15 annotated topics from top down: fidelity, incest, infanticide, stepmother and orphan, family, rejection, infidelity, suicide, jealousy, murder, priesthood, tragic fate, forced marriage, kidnapping and death.

tertextuality, stories about murder, for example, prevail throughout the whole corpus, whereas stories about kidnapping, which typically involves "evil" Turks, are being represented by only a few variants. The fact that in our choice of tf-idf weighting for topic modeling, the significance of a word is inversely related to its frequency, is another factor weighing in on the topic distribution in favor of dominant, but not globally represented words. Hence, when considering 'murder' and 'kidnapping' in the same context, 'kidnapping' will often be taken as more salient. Nevertheless, LLDA has correctly assigned appropriate main topics for all selected variant types in Figure 1.

## 6. CONCLUSION

Our aim in the above experiments was to replicate some of the real world scenarios, similar to folklorist's approach to analysis and classification of topics and uncovering of latent semantic structure of a folk song. Experiment 2 has shown LDA could be used in classification of large folkloristic corpora, as it is able to disambiguate between major family types despite strong intertextuality of the corpus. Experiment 3, on the other hand, has shown that based on a set of few annotated examples we are able to infer general thematic structure and prevalent topics for different variant types in the corpus. Of course, results of these preliminary analyses should be further examined, but they nevertheless show LDA can uncover typical characteristics of individual variant type (e.g., compare variant type names and corresponding topics detected by LLDA in Figure 1). And, the ability of the LDA to detect multiple topics can further help us discover general relationships (similarities and differences) in the corpus, as shown by the three experiments.

NLP methods should be chosen with care. Our comparative analysis has shown that there are different representational structures generated by statistical and probabilistic models (see Experiment 1). These representations significantly differ in their composition, with LSA generated space having in general more 'unbalanced' distribution compared to LDA. This difference is especially evident in the Voronoi tessellation of the semantic space, where salience of individual regions is highly disproportional (e.g. dimensions overlap) given the topical distributions of the corpus. This is in line with the results from previous studies, (Blei et al., 2003; Steyvers & Griffiths, 2007) that show that probabilistic models generally output more discriminative and hence interpretable semantic structures compared to statistical similarity-space models.

In our future work we plan to enrich the studied materials with other song families and use the described techniques to visualize and explore the obtained semantic spaces. Also, we plan to study interrelations of lyric spaces to melodic spaces obtained by analyzing relationships between song melodies.

## 7. REFERENCES


Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res., 3*, 993-1022.

Golež Kaučič, M., Kumer, Z., Terseglav, M., & Vrčon, R. (1998). *Slovenske ljudske pesmi IV: Ljubezenske pripovedne pesmi*. Ljubljana: Slovenska matica.

Golež Kaučič, M., Kumer, Z., Šivic, U., Terseglav, M., & Vrčon, R. (2007). *Slovenske ljudske pesmi V: Družinske pripovednepesmi*. Ljubljana: Založba ZRC, ZRC SAZU.

Grčar, M., Krek, S., & Dobrovoljc, K. (2012). *Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik*. Paper presented at the Osma konferenca jezikovnih tehnologij, Ljubljana.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *104*(2), 211-240.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. In T. Landauer, D. S. McNamara, S. Dennis & K. W. (Eds.), *Handbook of Latent Semantic Analysis* (pp. 424–440). Hillsdale, NJ: Erlbaum.

Strle, G., & Marolt, M. (2012). The EthnoMuse digital library: conceptual representation and annotation of ethnomusicological materials. *International Journal on Digital Libraries, 12*(2-3), 105-119.