Language Technologies in Humanities: Computational Semantic Analysis in Folkloristics

Gregor Strle,* Matija Marolt†

* Institute of Ethnomusicology ZRC SAZU
Novi trg 5, 1000 Ljubljana
gregor.strle@zrc-sazu.si
† University of Ljubljana, Faculty of Computer and Information Science
Večna pot 113, 1000 Ljubljana
matija.marolt@fri.uni-lj.si

1. Introduction

The paper discusses computational methods for natural language processing (NLP) and possibilities they offer to folkloristics.[1] As folkloristic materials are very challenging for NLP, due to their specific semantic-syntactic structure, inherent dialectical diversity and strong intertextuality, a robust NLP method is needed that can account for topical distribution, detect general heterogeneity, and context. The focus of this paper is on computational semantic analysis (such as word-sense disambiguation, topic recognition) and its ability to uncover latent semantic structure of folkloristic corpora.

2. Objectives

Our aim is to determine the appropriatness of NLP for analysing general patterns and relationships on the level of song types and genres, and also specify a general procedure (necessary steps) for conducting computational analysis of folkloristic materials. Two main approaches are presented and compared on the large-scale corpus of Slovenian folk songs: a statistical associative approach using Latent Semantic Analysis (LSA) and a probabilistic topic modelling approach using Latent Dirichlet Allocation (LDA). Emphasis is being placed on the practical applications of LSA and LDA models for analysing folkloristic corpora.

3. Methods

3.1. Latent semantic analysis (LSA) and Latent Dirichlet allocation (LDA)

Latent semantic analysis (LSA; Landauer and Dumais, 1997) and *Latent Dirichlet allocation* (LDA; Blei et al., 2003) are two of the most known methods used in NLP. They differ in theory and their approach to the semantic analysis. LSA does computations on high-dimensional similarity-space representations of associations between words, extracting the 'meaning' of individual words based on their proximity to other words in the semantic space. LDA, on the other hand, is a generative probabilistic topic model – it tries to uncover the blend of latent topics as distributions over documents and words. The central question for topic modeling approach is, 'what is the hidden structure behind these documents?'

The analysis and visualization of song types was made by Matlab topic modeling toolbox (TMT). In LDA, the analysis is an optimization process initialized randomly for the number of topics given as an input parameter. Consequently, multiple calculations yield slightly different results. The input parameter for the number of topics impacts the representation of semantic structure – smaller number of topics results in a more general overview of topic distributions in the corpus, whereas higher number of topics gives greater segmentation and detail.

3.2. Corpus

Both methods were tested on a large corpus of 3449 folk song variants from the collection of Slovenian folk songs (SLP I-V)[2]. The songs date back to the 19th century, with some variant types represented by only one variant and others consisting of up to 180 variants. Moreover, strong intertextuality is present throughout the corpus, which reflects characteristic phenomenon of Slovenian folk song: traveling of verses, motifs, and

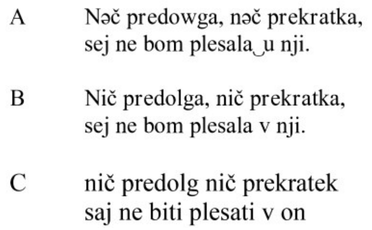
[1] This work discusses previous research on semantic analysis of folkloristic materials (Strle and Marolt, 2014)

[2] Part of the EthnoMuse multimedia archive (Institute of Ethnomusicology, ZRC SAZU): www.ethnomuse.info

thematic patterns from one song to the other, within and across variant types. This consequently affects the results, as most occurring topics and motifs dominate over lesser ones.

3.3. Document preprocessing

Lemmatization of the documents was performed in two steps (Figure 1). First, special characters used for encoding characteristics of dialect groups (such as semivowels, diphthongization, pitch accent etc.) are replaced by their grammatical equivalents. A dialect dictionary containing over 18.000 entries, specifically built from the folk song corpus, was used to translate the resulting words into literary language. In the second step, we used the statistical morphosyntactic tagger for the Slovenian language Obeliks (Grčar et al., 2012) to lemmatize the text.



A Nəč predowga, nəč prekratka,
sej ne bom plesala‿u nji.

B Nič predolga, nič prekratka,
sej ne bom plesala v nji.

C nič predolg nič prekratek
saj ne biti plesati v on

Figure 1: Lemmatization: A shows the original text, B the text after removal of a dialect and C the lemmatized text

4. Results

Two cases of using NLP analysis in folkloristics are presented: the analysis of general semantic structure of the corpus by the respective approach and the hierarchical clustering of folk song variants into folk song family types (the latter by using LDA only). Both cases give insight into thematic and semantic relationships, with the LDA clustering of folk song variants proving especially useful for building folk song typologies on the fly.

The results of general topic analysis show differences in the semantic structure of the corpus generated by LSA and LDA. There is a significant difference between both methods in their ability to the detect topics across the corpus and within various song families. Due to its simple design, LSA can only detect more prevalent topics across the semantic space. Arguably, this is the main limitation of LSA. As LSA model cannot account for topic distribution, it has difficulty detecting heterogeneity and the resulting semantic space repeatedly generalizes towards most salient (frequent) topics of the corpus. LDA is better in detecting the heterogeneity of the corpus and provides a more balanced representation of the semantic space (see Figure 2). Furthermore, the topic clusters generated by LDA correlate with the division of songs into song families and the general typology of the corpus.

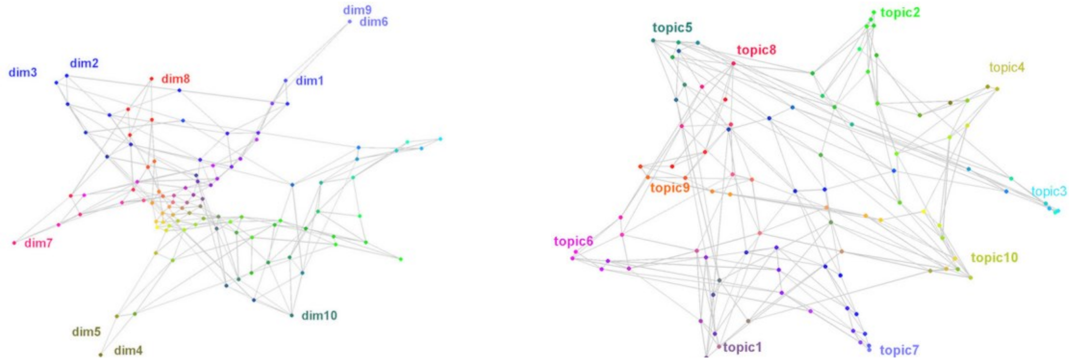


Figure 2: Comparison of semantic spaces: LSA (left) and LDA (right)

For the purpose of hierarchical clustering and visualization of song variant types only a subset of the corpus was chosen with the analysis limited to 5 clusters and two song family types, the narrative poems about fate and conflict in love and family. Our goal was to train the model on a smaller scale and investigate how it deals with the challenge of strong intertextuality present in both song types. We first calculated the average topic vector for each individual variant type by averaging the vectors of all the variants belonging to the respective variant type. The purpose of averaging is to reduce the disparity between the disproportionate number of variants representing particular variant type. The output is single (average) topic vector representing individual variant type, totalling 88 topic vectors for the song types analysed in the collection.

The method of hierarchical clustering was then used to divide variant types into clusters similar to the two folk song family types (love vs. family song type). The similarity of song types was calculated as the cosine similarity between topic vectors. Hierarchical clustering in Figure 3 shows the division of song types within the cluster, as well as the relationship between the individual clusters. Branches of the dendrogram for all five clusters are composed of 30 sub-groups, dividing all song variant types (88) into love (36) and family (52) types, with the former prevailing in clusters 2 and 5, and the latter in clusters 1, 3, and 4. This division indicates approximately 60% dominance of family narrative poems, which corresponds to the division between love and family songs in the corpus.

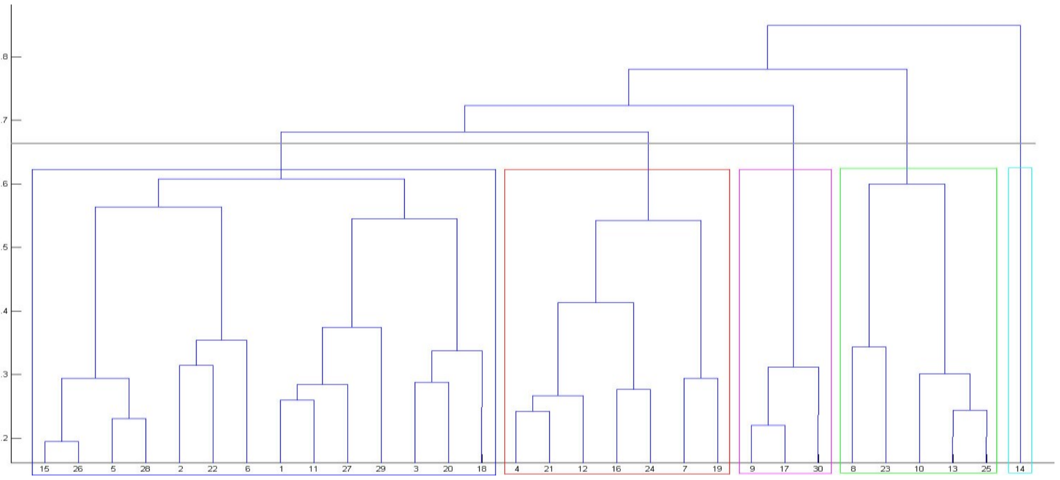


Figure 3: Hierarchical clustering of folk song poems about love and family fate. Colour boxes indicate 5 clusters: the songs about family fate prevail in clusters 1, 3, and 4, whereas the songs about love prevail in clusters 2 and 5

5. Conclusion

The main advantage of using NLP in folkloristics is the ability to analyze the semantic structure of large corpora, going beyond the limitations of traditional methods used in the office and fieldwork. Additional advantage of generative probabilistic models (such as LDA) is the ability to learn and generalize on new information, and thus expand existing analyses with new examples. This is especially handy in situations where we need to follow the semantic and typological transformations both chronologically and thematically. Future investigations will consider how computational methods can be used in folkloristics for more complex semantic and typological analyses.

6. References

Gregor Strle and Matija Marolt. 2014. New approaches: uncovering semantic structures in ethnological materials | Novi pristopi. Odkrivanje semantičnih struktur v etnoloških vsebinah. *Glasnik SED*, vol. 54/1-2, pp. 17-21.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104, pp. 211-240.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5), pp. 993-1022.

Miha Grčar, Simon Krek, and Kaja Dobrovoljc. 2012. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. In: T. Erjavec, J. Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan.