Topic Ontologies of the Slovene Blogosphere: A Gender Perspective

Iza Škrjanec[1], Senja Pollak[2]
[1]Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
[2]Jožef Stefan Institute, Ljubljana, Slovenia
skrjanec.iza@gmail.com, senja.pollak@ijs.si

Abstract

In the past years, blogs have become an increasingly popular genre for publishing content on the Web. Blogs are also one of the five genres of the Janes corpus of Slovene user-generated content. The aim of this paper is to explore the topics of the blog subcorpus of Janes using OntoGen, a semi-automatic and data-driven ontology editor. In addition to the construction of the topic ontology of the blogs from two Slovene blog portals, special focus is placed on the topical variation in entries by male and female bloggers. First, the keywords of selected topics differentiating male and female blog entries are analysed. Next, we present two topic ontologies, one based on blog entries by private female and the other by private male users, and contrast them against each other. The analysis has shown that both groups write about politics, family, romance and sexuality, environment, and nutrition. Men seem to blog more about spectator sports, music and literature, the Roman-Catholic Church, the refugee crisis, and biology; in contrast, female authors discuss religion, emotions and social politics.

Keywords: blogs, topic ontologies, gender, keyword analysis

1. Introduction

In corpus linguistics, corpora serve as the main resource for either testing various hypotheses or developing linguistic theories based on the corpus data. This is why it is important to learn about the properties of the corpus we are working with, e.g. recognizing frequent topics by observing keywords (Kilgarriff, 2012).

For Slovene, the topics of blogs in particular have not yet been studied; however, Logar Berginc and Ljubešić (2013) contrasted two Slovene corpora of various genres against each other: the crawled slWaC[1] corpus and the reference Gigafida[2] corpus. Using the LDA topic modelling method, a number of n topics for each of the corpora was constructed. When comparing the topics, Logar Berginc and Ljubešić found that some topics appeared in both corpora (domestic policy, team sports, finance, war, terrorism, publications and culture, local politics, health and law). The slWaC corpus contains more documents on film and music, travelling and tourism, foreign affairs and classified ads. In contrast, the following topics are more prominent in Gigafida: cities, street traffic, public events, television and radio programs, individual sports, and work. Some differences between the reference corpus and the Janes corpus including blog entries (but also tweets, news comments, forum posts) have been identified through collocation analysis in Pollak (2015).

In this paper, we focus on topical variation between male and female bloggers. For English there have been some studies on how the content in social networks posts or spoken language correlates with the demographic factors of users, such as gender and age. Using data mining techniques, Argamon et al. (2007) found that male bloggers tend to write about religion, politics, business and the Internet more frequently, while female bloggers blog about conversation, domestic environment, fun, romance, and swearing more than men. Schmid (2003) carried out a comparable study on the spoken part of the BNC corpus. He conducted a list of words typical for 14 different topics. Using relative frequencies, he observed which topics are more dominant in the corpus of female and male speakers. An overrepresentation of female speakers was detected in topics dealing with clothing, basic colors, home, food and drink, body and health, and people. In contrast, the domains of work, computing, sports, and public affairs were considered more typical of the male subcorpus. The domains on swearing and car and traffic occurred equally in the speech of both groups.

In comparison to the topic keyword analysis as for example in Logar Berginc and Ljubešić (2013), the approach selected for this study results in hierarchical ontologies which allow the identification of subtopics for each topic, enables the user to be involved in the process of ontology construction, and provides the visualization of the constructed ontologies. In addition to the understanding of the topics of the Slovene blogosphere, the main contribution of our paper is the research of gender and the Slovene language in social media. We thus wish to contribute to existing studies, e.g. on the use of emoticons and expressive punctuation in tweets (Osrajnik et al., 2015) and the discourses about women and men (Škrjanec et al., 2016).

The rest of the paper is structured as follows. In Section 2, the blog subcorpus and the text preparation are presented. The OntoGen tool and the ontology construction process are described in Section 3, and discussed in Section 4. In Section 5, we conclude the paper and suggest further work.

2. Corpus Description and Data Preparation

The corpus of Slovene blogs used in this paper is one of the subcorpora in version 04 of the Janes corpus of user-

[1] http://nlp.ffzg.hr/resources/corpora/slwac/

[2] http://www.gigafida.net/

generated Slovene. The corpus was compiled within the Janes[3] research project and contains various genres of user-generated content: tweets, news comments, forum posts, user and talk pages from Wikipedia, and blog entries and blog comments. In this paper, the focus is placed on the compilation and properties of the blog subcorpus, for which two Slovene blog portals were crawled: *publishwall.si* and *rtvslo.si* (Fišer et al., to appear).

The blog entries of the Janes corpus were contributed by over 800 users, which we annotated for their account type (private or corporate) and gender [4] (female, male and undefined). Corporate accounts belong to different companies or journalists, the rest are private. The gender was manually assigned based on the use of grammatical gender when referring to self; the profile picture and username. If we were not able to identify the user as male or female, the tag "neutral" (meaning "undefined") was used.

For our study, we selected the blog posts of male and female private users. Private users wrote over 29,000 entries altogether (female: 9,056; male: 20,105). For the ontology construction, blog entries in Slovene were taken into consideration.

Disregarding the gender and account type, the average length of blog entries and comments in the entire blog subcorpus is about 70.16 words (85.42 tokens) per entry. Since clustering algorithms perform better on longer texts than on shorter ones, blog entries with minimum of 100 full words (no stop words) were used for ontology construction (9,039 entries by male and 3,771 by female users).

2.1.1 Text Preparation

The original vertical file of the Janes blog subcorpus was parsed into a format supported by OntoGen, in which each blog entry is represented with a single line containing the blog entry ID, category (female or male), and the lemma form of each token. All preprocessing steps were carried out with a simple Python programme. Stop words were removed. OntoGen cannot process diacritics and other special characters, so these were replaced with character sequences that enable the reconstruction of the original form.

3. Topic Ontology Construction

In this section, the OntoGen tool and the construction of three topic ontologies are presented.

3.1 The OntoGen Tool

For the construction of Slovene blog topic ontologies, we used the OntoGen tool[5], which is a semi-automatic data-driven ontology editor that combines text mining techniques with a fairly simple user interface (Fortuna et al., 2007). OntoGen is based on Bag-of-Words (BoW) vector representations of documents, weighted by the Term Frequency-Inverse Document Frequency weights. The tool provides subtopic suggestions based on the k-means clustering algorithm, with the parameter k being set by the user. The user then decides whether to add the clusters to the ontology. The user can also manually move the documents and provide labels for the clusters (topics). Additionally, if the input documents are pre-categorized, a method for grouping the instances according to the labels is also supported.

The user can influence the division into subtopics by employing the Active learning functionality that is based on the SVM (Support Vector Machines) active learning method. The user provides a term or a set of keywords that represent a new subtopic to be added to the ontology. This action is followed by iterative model refinement through user interaction by answering to the question whether a

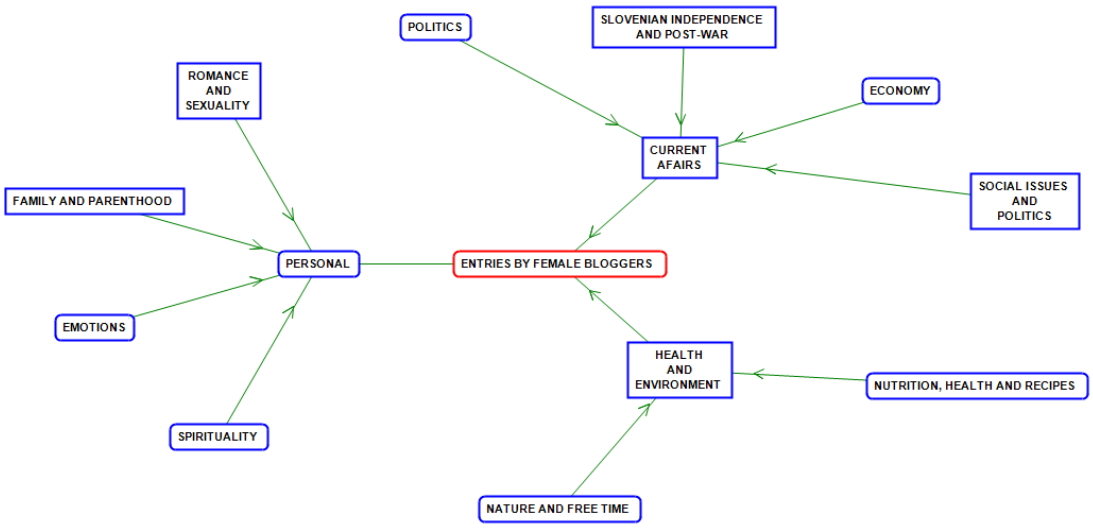

Figure 1: Topic ontology of entries by female bloggers.

[3] The Janes project webpage: http://nl.ijs.si/janes/

[4] The bloggers themselves were not contacted and asked about their self-identification. Thus, the claims on their gender are based solely on the use of grammatical gender.

[5] The OntoGen tool: http://ontogen.ijs.si/

particular document belongs to the topic or not. The user can decide when to stop the active learning process and the new (sub)topic is added to the ontology.

For each topic, OntoGen provides a list of keywords, which are the words that are the most descriptive for the content of cluster, i.e. words with the highest weights in the document centroid vectors (ibid.). Another view is gained by inspecting SVM keywords, which are the words most distinctive for the selected concept with regard to its sibling concepts in the hierarchy (e.g. words contrasting male and female entries categorized in a selected topic).

3.2 The Construction of Three Topic Ontologies

The dataset with entries by private male and female bloggers was imported into OntoGen. The ontology was built using k-means for topic suggestions, and the active learning functionality, as well as by manually arranging the ontology. Because the entries were pre-categorized according to the user gender, we could examine the keywords and SVM keywords of topics, whereby the topics with a more or less comparable number of entries by female and male bloggers were selected for analysis. Two topics (*Romance and sexuality*; *Political system*) and their keywords are presented in Table 1. In addition, we constructed a topic ontology for entries by female (Figure 1) and male (Figure 2) users[6].

	Female		Male	
	Keywords	SVM k.	Keywords	SVM k.
Romance and sexuality	moški, ženska, film, želeti, partner, življenje, ljubezen, prijatelj, ženski, odnos #431	moški, želeti, partner, čutiti, strah, razmišljati, potrebovati, fb, spolnost, telo	ženska, moški, film, sex, prijatelj, ženski, žena, rak, dekle, želeti #329	sex, ženska, žena, film, mati, bivši_žena, obraz, zgodbica, brada, punca
Political system	družba, sistem, obstajati, lasten, ego, narod, zavest, vrednota, človeški, življenje #129	želeti, obstajati, narod, telo, izkušnja, lasten, ego, sposoben, različen, zavest	družba, sistem, kapitalizem, družben, problem, človeški, življenje, planet, znanost, vrednota #503	družba, sistem, bitje, sodoben, demokracija, ideja, planet, materialen, svoboda, stoletje

Table 1: Keywords for the topics *Romance and sexuality*, and *Political system*.

4. Discussion

A keyword list can tell us something more about the main ideas and concepts users blog about concerning a particular topic. After constructing a common topic ontology of entries by both groups of users, we observed the entries on romance and sexuality to compare the keywords and SVM keywords of both groups. From keywords in Table 1, we can learn that male and female bloggers use similar keywords ("woman", "man", "want") with some variation. Observing the SVM keywords, which point out the differences, it is evident that female bloggers use more verbs ("feel", "think", "need"), while male bloggers focus more on the participants ("ex-wife", "girlfriend", "mother") and appearance ("face", "beard"). The keyword "crab"/"cancer" suggests that the topic is still somewhat noisy. The keywords for the topic *Political system* also reveal similarity between entries by men and women ("society", "system", "life", "human"), whereas in entries by female bloggers the topic of nation comes forward. In the entries by male bloggers, terms like "capitalism",

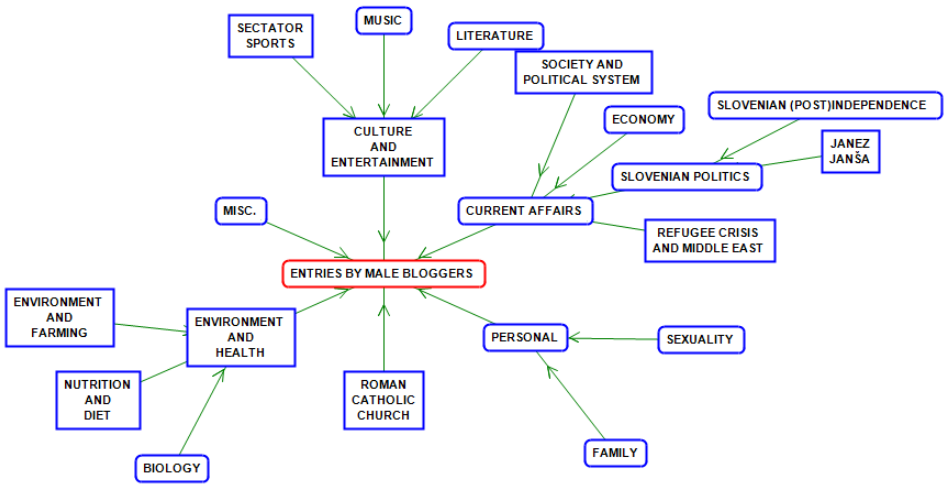

Figure 2: Topic ontology of entries by male bloggers.

[6] For the common ontology, lemmas of uni- and bigrams with the minimum frequency of 20, and for the gender specific ontology, the minimum frequency was set to 10.

"democracy" and "freedom" indicate a specific political issue.

A topical comparison of blog entries by female and male bloggers (Figures 1 and 2) shows some interesting similarities and differences. Both groups seem to write about the environment, nutrition, family and parenthood, sexuality, and politics, in particular the subtopic on Slovenian politics and the (post)independence era (the Independence War, post-war killings and the role of former communists today). Another common topic is economy (mostly Slovene and EU). One of the more prominent topics of male bloggers is that on the Slovene politician Janez Janša, mostly concerning his 2013–2015 trial for corruption. An evident topic on current affairs is also that of the refugee crisis in the male ontology. In contrast to female bloggers, male authors contributed a significant number of entries on biology, spectator sports, music and literature. They also discuss the role of the Roman Catholic Church. In turn, female bloggers write more about spirituality in connection to various religious beliefs, and nature. Emotions are also a prevalent blog topic of female users; additionally, they pay special attention to social politics and issues, such as handicapped people and their social rights.

5. Conclusion

In the paper, we described the process of topic ontology construction and keyword analysis of blog entries from two Slovene blog portals. The goal was to contrast the topics covered by female and male bloggers.

To avoid over-generalization on gendered topics, it is important to take into account the distribution of blog entries among bloggers. Some topics are heavily dominated by a very small number of bloggers (*Biology*, *Social issues and politics*), but this is not visible in the ontology. When using quantitative methods to explore gender and language use, it seems the tendency is to favour differences, while backgrounding similarities, what Baker (2014) calls the "difference mindset". The findings of studies such as this one may suggest and show mostly the differences. However, the language and topics of a single gendered group is not homogenous, which is what Baker (ibid.) discovered when he contrasted "same-sex" parts of the BNC spoken among each other using Manhattan Distance for a list of keywords. He found that some pairs of "same-sex" parts vary more than pairs of "mixed-sex" combinations.

In spite of considering this issue, the analysis has shown that some topics (*Refugee crisis*, *Janez Janša, Biology*, *Spectator sports*, *Music and literature*) seem more prominent in entries by male bloggers, while female bloggers typically contribute to topics like *Religion*, *Nature*, *Emotions*, *Social politics*. When writing about mutual topics (*Romance and sexuality*, *Political system*), female and male bloggers discuss them from different perspectives. Our methodology can be applied to explore the topics of the entire blog subcorpus, including corporate users and those undefined in terms of gender. The information on the predominant topic of the entry could enrich the existing blog metadata: user gender, account type, the linguistic and technical standardness and sentiment of the text. In the future, the automatization of the topic labelling could be performed by combining clustering and terminology extraction as shown in Fortuna et al. (2008). Adding the topic to the metadata enables a more fine-grained analysis of discursive strategies for the same topic with regard to the gender of the user, which is something we plan to carry out in the future.

6. Acknowledgements

The work described in this paper was funded by the Slovenian Research Agency within the national basic research project "Resources, Tools and Methods for the Research of Nonstandard Internet Slovene" (J6-6842, 2014-2017).

7. References

Argamon, Shlomo, Moshe Koppel, James W. Pennebaker and Johnatan Schler (2007). Mining the Blogosphere: Age, gender and the varieties of self-expression. *First Monday*, 12(9).

Baker, P. (2014). *Using Corpora to Analyze Gender*. London: Bloomsbury.

Fišer, D., Erjavec, T., Ljubešić, N. (to appear): Janes v0.4: korpus slovenskih spletnih uporabniških vsebin. *Slovenščina 2.0 – Special Issue*.

Fortuna, B., Grobelnik, M., Mladenić, D. (2007). OntoGen: Semi-automatic Ontology Editor. *HCI International 2007*, July 2007, Beijing. 309–318.

Fortuna, B., Lavrač, N., Velardi, P. (2008). Advancing Topic Ontology Learning through Term Extraction. In Ho, T., Zhou, Z. (eds), *Proceedings of PRICAI 2008: Trends in Artificial Intelligence*. Hanoi, Vietnam, December 15-19, 2008. 626–635.

Kilgarriff, A. (2012). Getting to know your corpus. In Sojka, P., Horak, A., Kopecek, I. (eds), *Proceedings of the 15th International Conference on Text, Speech and Dialogue* (TSD2012), pages 3-15. Brno, Czech Republic: Springer.

Logar Berginc, N., Ljubešić, N. (2013). Gigafida in slWaC: tematska primerjava. *Slovenščina 2.0*, 1 (1): 78–110.

Osrajnik, E., Fišer, D., Popič, D. (2015). Primerjava rabe ekspresivnih ločil v tvitih slovenskih uporabnikov in uporabnic. Fišer, D. (ed), *Zbornik konference Slovenščina na spletu in v novih medijih*. Ljubljana: Znanstvena založba Filozofske fakultete, 50–74.

Pollak, S. (2015). Identifikacija spletno specifičnih kolokacij pogostega besedišča. Fišer, D. (ed), *Zbornik konference Slovenščina na spletu in v novih medijih*. Ljubljana: Znanstvena založba FF UL, 57–62.

Schmid, H. J. (2003). Do men and women really live in different cultures? Evidence from the BNC. In: Wilson, A., Rayson, R. and McEnery, T. (eds), *Corpus Linguistics by the Lune. Lódź Studies in Language 8*. Frankfurt: Peter Lang. 185-221.

Škrjanec, I., Sobočan, A. M., Pollak, S. (2016). The lexical environments of woman and man in the corpus of Internet Slovene. In: Granić, J., Kecskes, I. (eds), *Proceeding of the 7th INPRA Conference*. 10-12 June 2016, Split, Croatia. 161.