TOPIC ONTOLOGIES OF THE SLOVENE BLOGOSPHERE: A GENDER PERSPECTIVE

Iza Škrjanec[1], Senja Pollak[2]

[1]Jožef Stefan International Postgraduate School, Ljubljana

[2]Jožef Stefan Institute, Ljubljana

Faculty of Arts, Ljubljana, 28 September, 2016

Overview

- “getting to know your corpus“ (Kilgarriff 2012)
- **topic ontology**: a set of topics connected with hierarchical relations (insight)

- main objective
- the Slovene blog corpus
- the OntoGen tool
- the topic ontologies: male and female bloggers
- conclusion

Motivation: the “gender question“

- the same but different?
- gender and language use
  - **HOW?** → variation observed in the phenomena:
    - **syntactic patterns**: polite forms and hedges in speech (Schmid 2003)
    - **vocabulary and emoticons**: swearing in speech (Baker 2014), emoticons in tweets (Osrajnik et al. 2015)
    - **topics**

Slovene blog corpus

- blog = website containing **diary-like** textual documents (entries/posts)
  - comment section

- Slovene blog corpus: part of **Janes v0.4 corpus** of Slovene user-generated content (Fišer et al. 2016)
- blog entries from two online platforms:
  - **PublishWall**
  - **RTVSLO blog**
- size: over **40,000** blog entries by over **800** bloggers

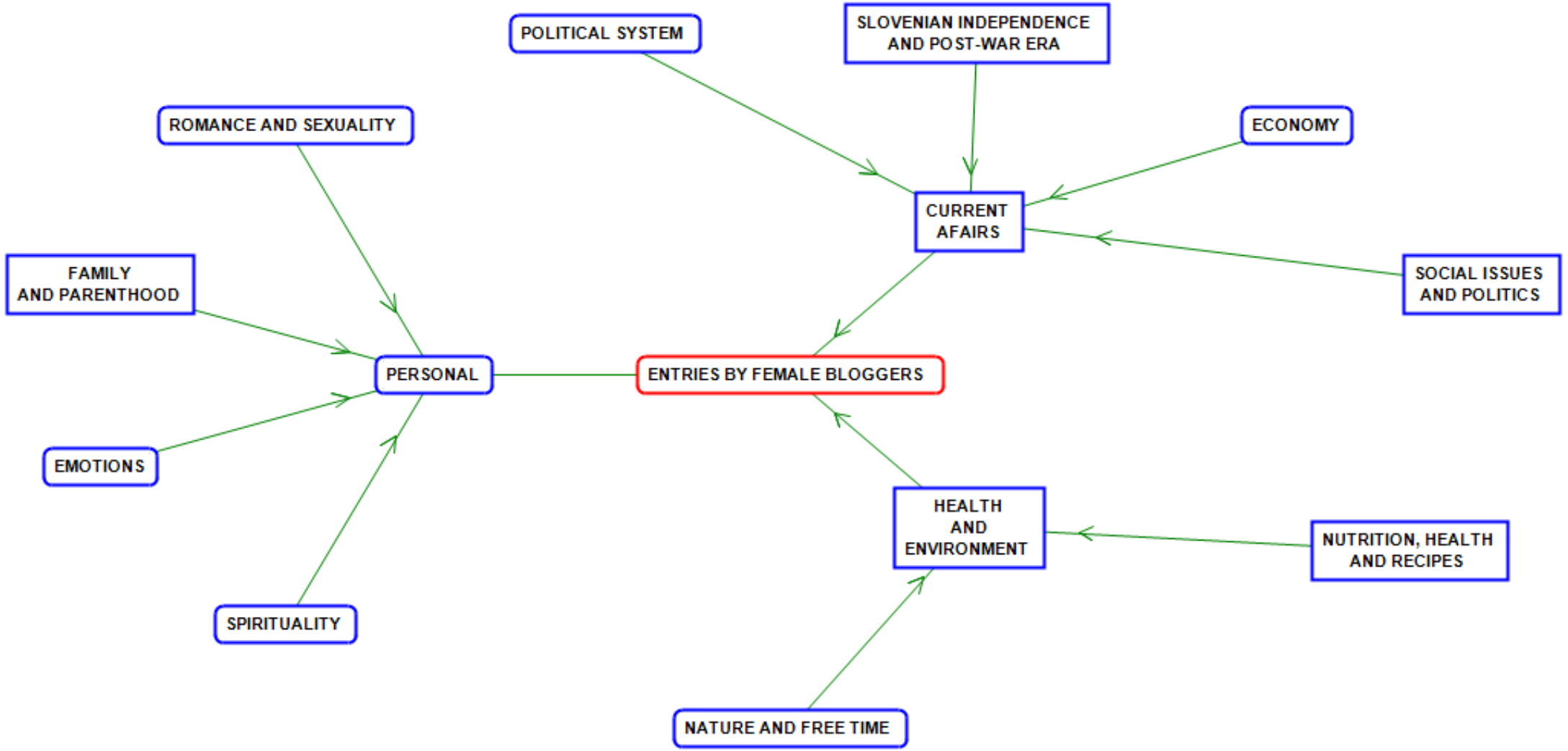
Tools and data preprocessing

- minimize manual labour and avoid subjectivity
- the **OntoGen** tool (Fortuna et al., 2007):
  - semi-automatic ontology editor
  - **k-means** clustering (**BoW** representation, **TF-IDF** weighting, **cosine** similarity)
- data preprocessing:
  - annotation of **user gender** (female/male/undefined) and **account type** (private/corporate)
  - entries by female and male private bloggers (in Slovene only)
  - **minimal length** (100 full words): **9,039** entries by male, and **3,771** by female users

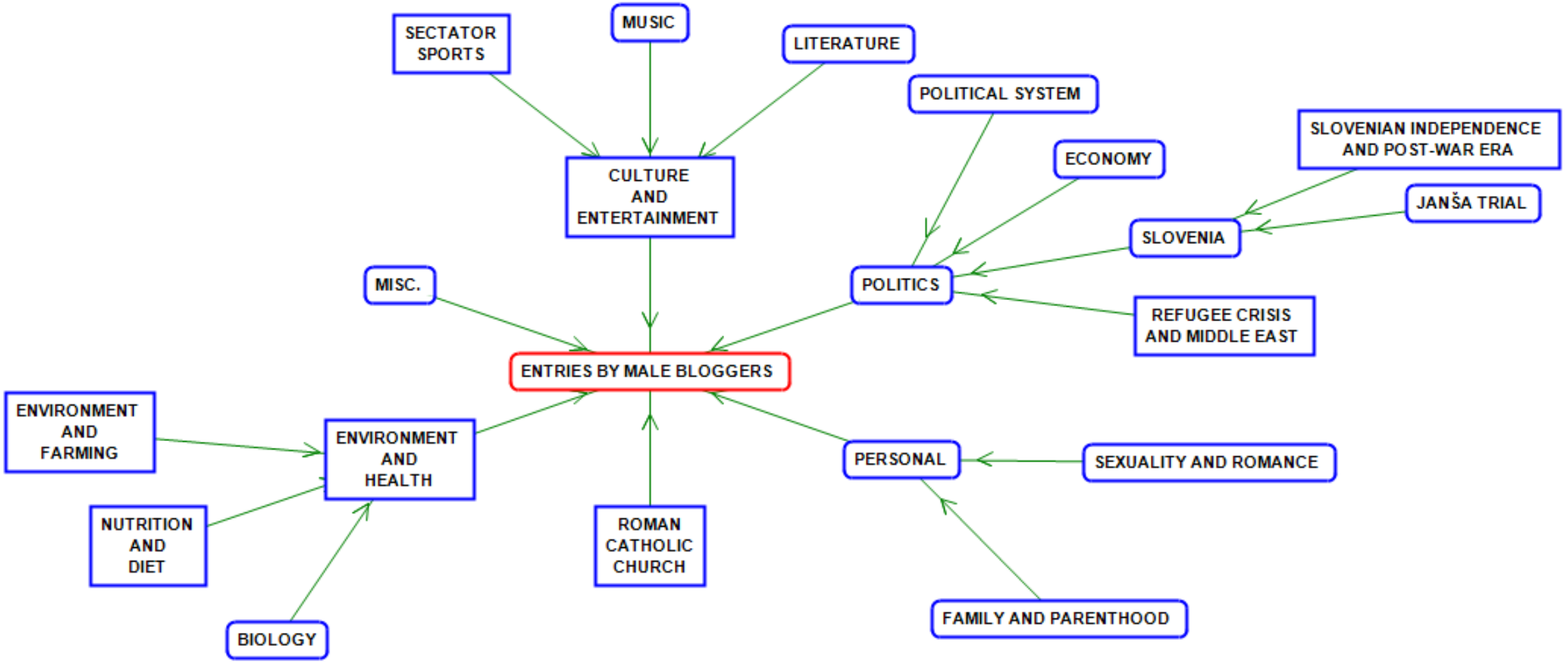
Ontology construction: semi-automatic

- automatic: k-means
- relevant parameters
  - **maximum n-gram length**: 2
  - **minimum n-gram frequency**: 10
- manual (user involvment):
  - **naming topic** and subtopics based on a list of **keywords** provided by OntoGen
  - **manually arranging** the ontology (moving the concepts)

Ontology construction: entries by private **female bloggers**



Ontology construction: entries by private **male bloggers**



Observed varieties

FEMALE BLOGGERS

- spirituality (religious beliefs)
- emotions (ljubezen/love, srce/heart, strah/fear, ljubiti/to love, čutiti/to feel, želeti/to wish)
- social politics and issues (handicapped people, social rights)

MALE BLOGGERS

- the Slovene politician Janez Janša (2013–2015 corruption trial)
- the refugee crisis
- the role of the Roman Catholic Church
- biology
- free time (spectator sports, music and literature)

COMMON TOPICS

- environment, nutrition
- family and parenthood; **sexuality**
- **politics**: Slovenian politics and the (post)independence era
- economy: Slovenian and EU

SVM keywords (distinctive)

POLITICAL SYSTEM	
Female bloggers	Male bloggers
želeti, obstajati, narod, telo, izkušnja, lasten, ego, sposoben, različen, zavest to wish, to exist, experience, own, body, ego, nation, capable, different, consciousness	družba, sistem, bitje, sodoben, demokracija, ideja, planet, materialen, svoboda, stoletje society, system, democracy, freedom, idea, planet, century, contemporary, material, being

SVM keywords (distinctive)

ROMANCE AND SEXUALITY	
Female bloggers	Male bloggers
moški, partner, strah, želeti, čutiti, razmišljati, potrebovati, fb, spolnost, telo man, partner, fear, to wish, to feel, to think, to need, fb, sexuality, body	ženska, žena, punca, bivši_žena, zgodbica, film, obraz, brada, sex woman, wife, girl, ex_wife, mother, story, film, face, beard, sex

Conclusion (1)

- several differentiating topics, but many shared by both groups
- Argamon et al. 2007 (English blogs):
  - **male bloggers**: religion, politics, business, and the Internet
  - **female bloggers**: conversation, domestic environment, fun, romance and swearing
- Schmid 2013 (BNC Spoken):
  - **male speakers**: work, computing, sports, and public affairs
  - **female speakers**: clothing, basic colors, home, food and drink, body and health, and people
- interpretation of results:
  - careful with over-generalization: **distribution matters**!
  - "the more different, the better": favouring differences, while backgrounding similarities (the "**difference mindset**", Baker 2014)

Conclusion (1)

- several differentiating topics, but many shared by both groups
- Argamon et al. 2007 (English blogs):
  - **male bloggers**: **religion**, **politics**, business, and the Internet
  - **female bloggers**: conversation, **domestic environment**, **fun**, **romance** and swearing
- Schmid 2013 (BNC Spoken):
  - **male speakers**: work, computing, **sports**, and **public affairs**
  - **female speakers**: clothing, basic colors, **home**, **food and drink**, **body and health**, and people
- interpretation of results:
  - careful with over-generalization: **distribution matters**!
  - "the more different, the better": favouring differences, while backgrounding similarities (the "**difference mindset**", Baker 2014)

Conclusion (2)

- topical overview of the corpus – further work:
  - **evaluation**: automatic (average similarity of the cluster/topic), **manual** (planned)
  - enriching the **corpus metadata**: adding the topic information (exporting topic name from ontology RDF into corpus XML)
- gender and CMC language: **a more detailed analysis** (discursive strategies)

THANK YOU!