Quantitative Textual Analysis in Sociology II

From topic modeling to mapping discursive areas

Topic modeling:
- Based on the assumption of semantic clustering (thematic organization)
- Produces a calculation of topics
- Produces.dominates rather than contrasts.

Suggests we are interested in distinguishing individual topics but in a more general discursive structure of textual data. For example, we know what "topics" are in negative (e.g., "leaving the home country, traveling, refugee camps") and positive (e.g., "arrival to the host country, adaptation, return").

Data & Analysis

Data: the corpus of migrant stories from Sangarwani.org; cleaned, preprocessed, and marked in male and female versions.

Analytical procedure:
1. Selection of 50 most frequent semantic words by frequency and correlation with specific words (positive/negative).
2. Creation of co-occurrence networks within the context window.
3. Accuracy of network analysis.
4. Analysis of given textual contexts of the co-occurrence networks using statistical methods.

Software

Word frequencies
- Counting word co-occurrences and calculating dissimilarity
- Multidimensional scaling

Results 1 - gender differences

Negative values: heavy traffic, poverty, inequality, unemployment.

Positive values: family, culture, adaptation.

Analysis of activist media communication

The analysis of activist media communication reveals how the construction of social identity and the representation of migrant experiences are mediated through the use of various communicative strategies and representations.
From topic modeling to mapping discursive areas

**Topic modeling:**
- based on the assumption of semantic clustering (thematic organization) of textual data
- complicated calculation of topics
- producing discontinuities rather than continuities.

Sometimes, we are not interested in distinguishing individual topics but in a more general discursive structure of textual data. For example, we know what "topics" are in migration stories: leaving the home country (- traveling - refugee camps) - arrival to the host country - adaptation (- return). We can wonder if there are more macrostructures than topics

--> analysis of frequent word co-occurrences.
Data & Analysis

Data: the corpus of migrant stories from iamamigrant.org; cleaned (paratextuals, stop-words) and divided in male and female subsets

Analytical procedure:
1. selection of 50 most frequent semantic words; excluding highly contextual words such as people, time, lot, month, day, dont, im etc.
2. splitting the corpus into context units (stories=paragraphs)
3. counting co-occurrences of selected words within the context units
4. creation of a standardized matrix of distances between words using Jaccard’s similarity coefficient
\[ J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}. \]
5. computation of a two-dimensional graph using multidimensional scaling technique (MDS)
Software

Word frequencies: R

Counting word co-occurrences and calculation of distances:
Co-occurrence Analysis Software (COOA)

Multidimensional scaling: R
Results I - gender differences

Interpretation guide:
- the most frequent words are displayed closer to the centre of the graph as they have many co-occurrences with all other words;
  less frequent words are, on the other hand, found at the periphery;
- it is essential to look at the structure as a whole, not at the exact position of a particular word;
- dimensions (axes) do not have substantive meaning, and the graph can be arbitrarily rotated and turned.

- number of frequent words: 50
- context unit: story
- measure of proximity: Jaccard coeff.

migrant men's stories

migrant women's stories
- number of frequent words
- context in story
- measure of proximity (Jaccard Index)
- number of frequent words
- context in a story
- measure of proximity (Jaccard index)
Results II - near and far migrants

Is there a difference between stories of intracontinental and intercontinental migrants?

- number of frequent words: 50
- context unit: story
- measure of proximity: Jaccard coeff.

![stories of intracontinental migrants](image1)

![stories of intercontinental migrants](image2)
stories of intracontinental migrants

- number
- frequent
- country
- story
- mean
- prox
- Jacob
stories of intracontinental migrants
words:

stories of intercontinental migrants
stories of intercontinental migrants

migration as a life change stories

education & family stories

stories of Syrian refugees
Results II - near and far migrants

Is there a difference between stories of intracontinental and intercontinental migrants?

- number of frequent words: 50
- context unit: story
- measure of proximity: Jaccard coeff.

**Interpretation:** Both categories of migrants share "family & education" and "life change" migrant stories. In this sense, distance does not matter in narrating migration. On the other hand, each category contains a specific group of migrants (Syrian war refugees and African migrants via Libya) who narrated different stories.
Analysis of activist media communication

**Data:** internet communication of 12 Czech activist organizations and interest groups representing four domains of justice claim-making: trade unionism, feminism, human rights, environmentalism

**Method:** Counting and displaying co-occurrences of 33 most frequent lexical (semantic) words in texts in each domain

**Results:** internal split of discursive space in two areas
- **the conditions area** as the discursive space for verbalizing the existential conditions of subjects of contention,
- **the contention area** referring to the opposition strategies and negotiation with movement opponents.
Conclusion

Non-coding quantitative text analysis in sociology can be used to explore various social phenomena.

A good strategy is to combine textual and nontextual (behavioral, institutional, socio-demographic) data.

The quality of data is more important than the sophistication of analytical techniques.

Interpretation is often not obvious and has a form of hypothesis proposal.

Thank you for your attention!