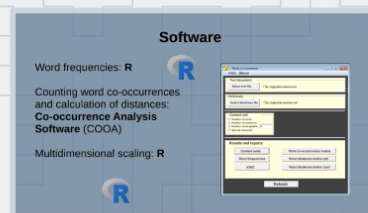


Quantitative Textual Analysis in Sociology II



Data & Analysis

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

Analytical procedure:

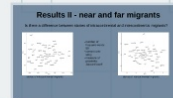
1. selection of 50 most frequent semantic words; excluding highly contextual words such as people, time, lot, month, day, don't, am etc.
2. splitting the corpus into context units (stories=paragraphs)
3. counting co-occurrences of selected words within the context units
4. creation of a standardized 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)

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.**



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 20 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 essential conditions of subjects of contention.
• the content area referring to the opposition strategies and negotiation with movement opponents.

Results I - gender differences

Interpretation guide:
• the most frequent words are displayed close to the center 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.
• differences aimed to not have substantive meaning, and the graph can be only rarely rotated and turned.



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
5. computation of a two-dimensional graph using multidimensional scaling technique (MDS)

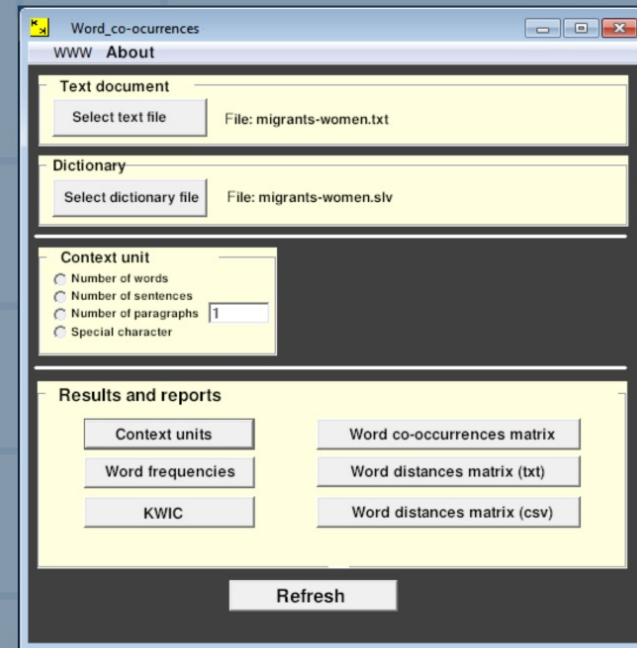
Software

Word frequencies: **R**



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

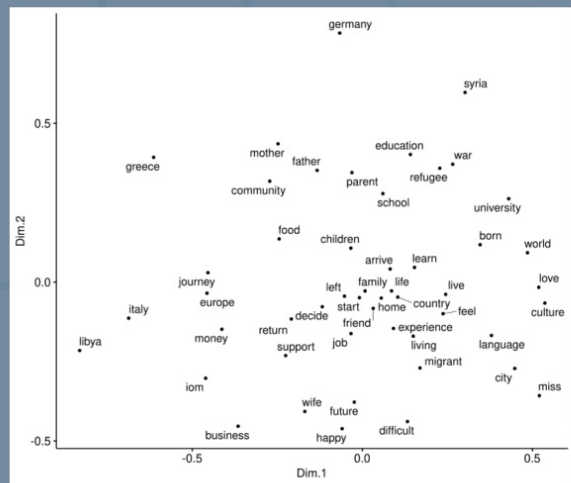
Multidimensional scaling: **R**

A screenshot of a web-based software interface titled "Word_co-occurrences". The interface has a menu bar with "WWW" and "About". It is divided into several sections: "Text document" with a "Select text file" button and the filename "migrants-women.txt"; "Dictionary" with a "Select dictionary file" button and the filename "migrants-women.slv"; "Context unit" with radio buttons for "Number of words", "Number of sentences", "Number of paragraphs" (selected), and "Special character", with a text input field containing "1"; and "Results and reports" with buttons for "Context units", "Word frequencies", "KWIC", "Word co-occurrences matrix", "Word distances matrix (txt)", and "Word distances matrix (csv)". A "Refresh" button is located at the bottom of the interface.

Results I - gender differences

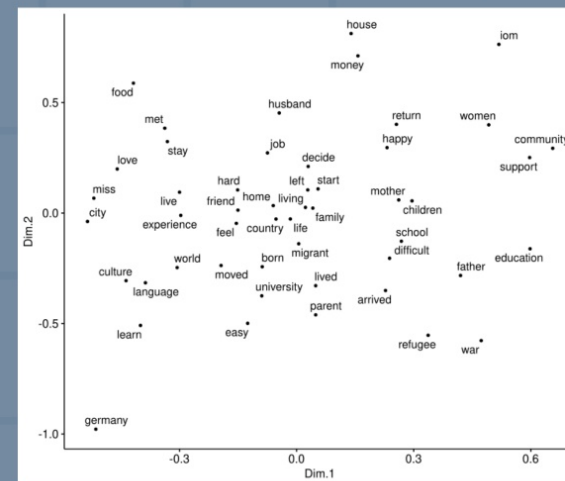
Interpretation guide:

- 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.

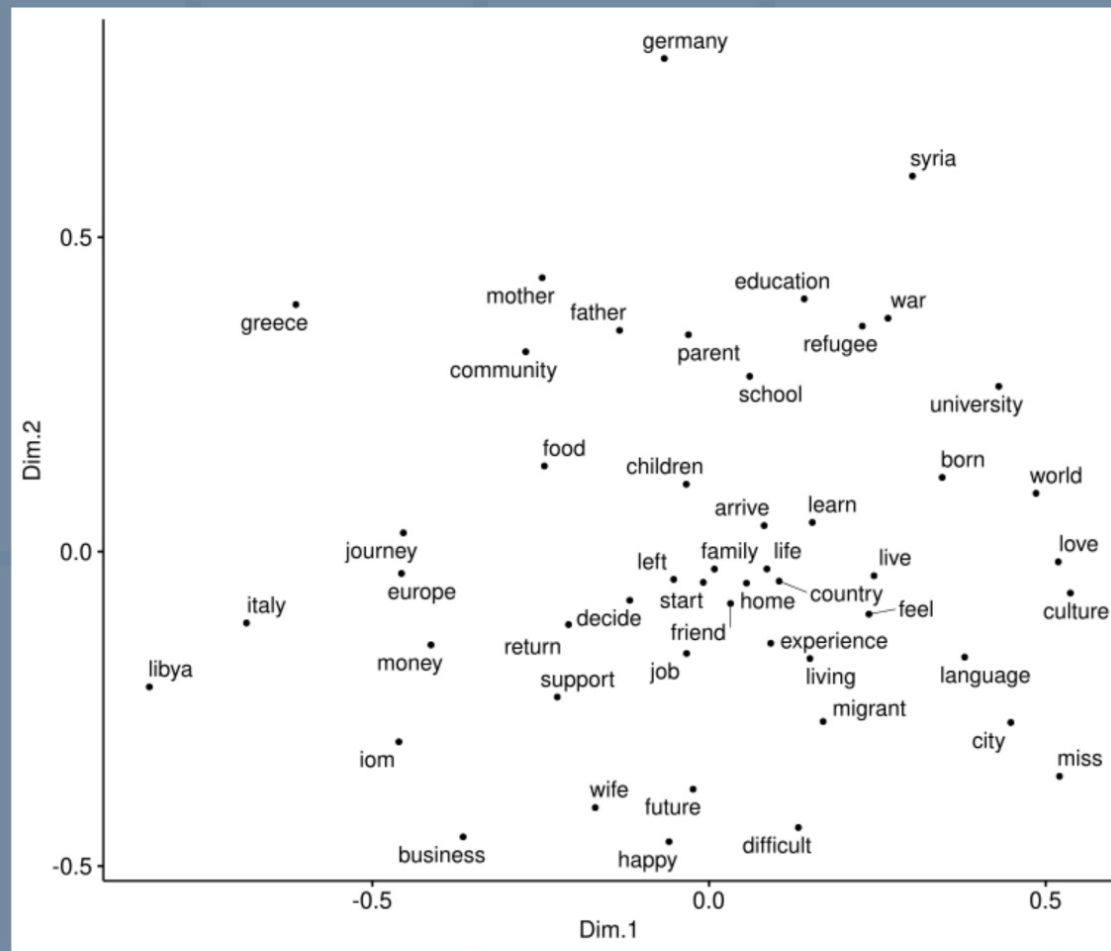


migrant men's stories

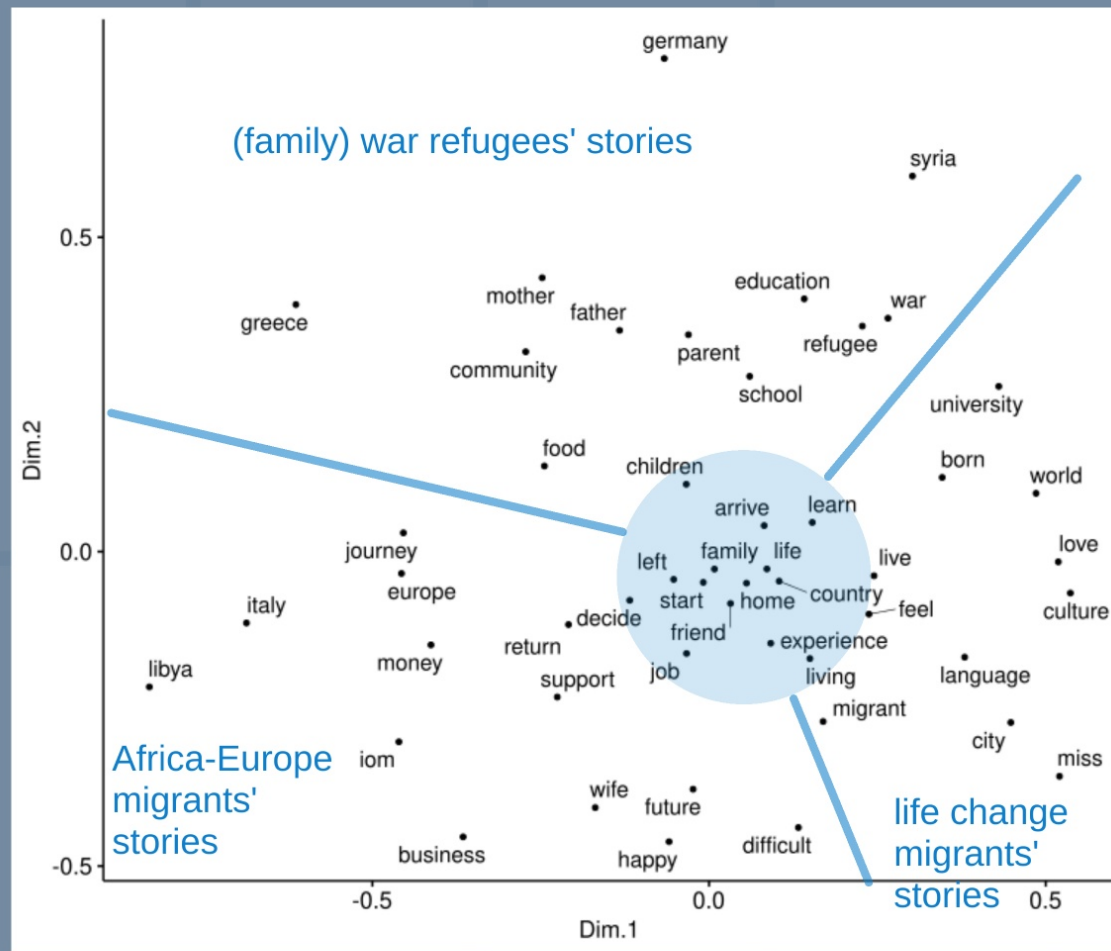
- number of frequent words: 50
- context unit: story
- measure of proximity: Jaccard coeff.



migrant women's stories

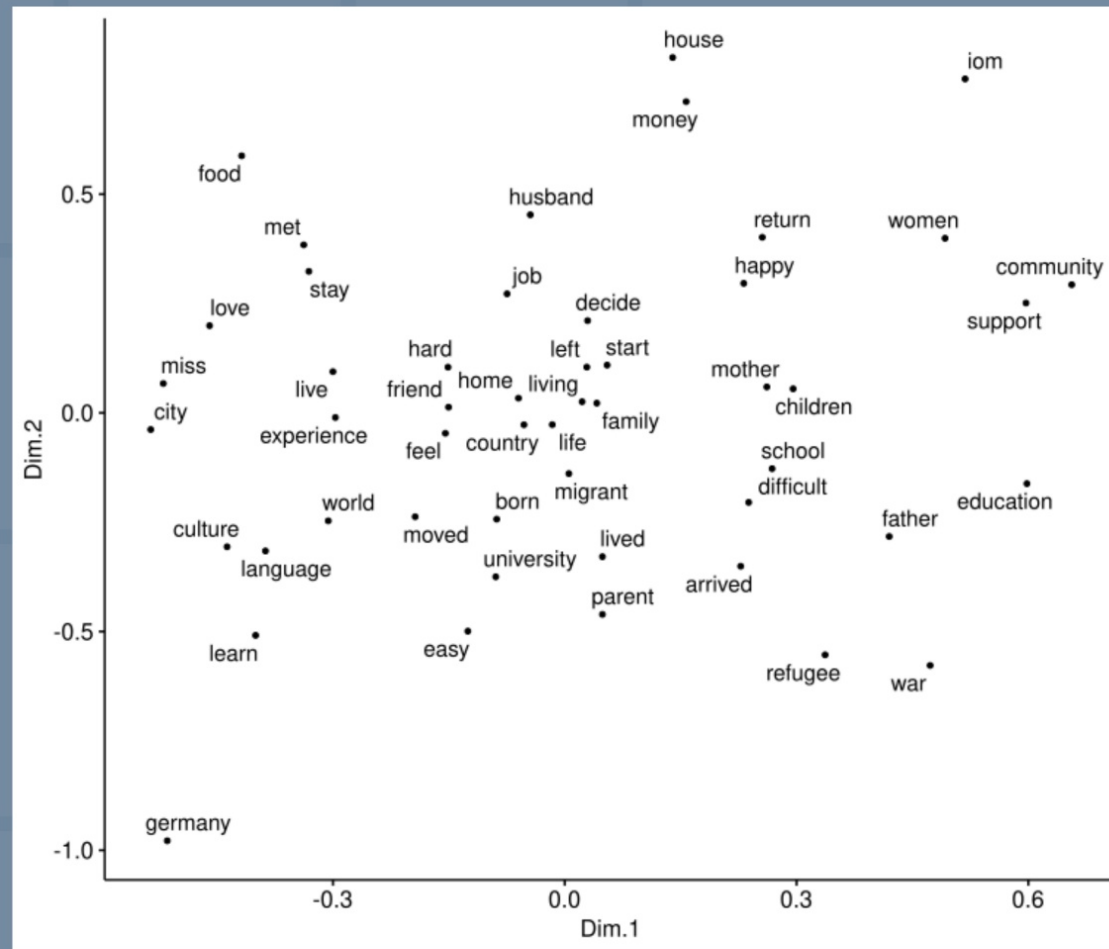


- number
frequent
50
- context
story
- measure
proximity
Jaccard



- number
frequent
50
- context
story
- measure
proximity
Jaccard

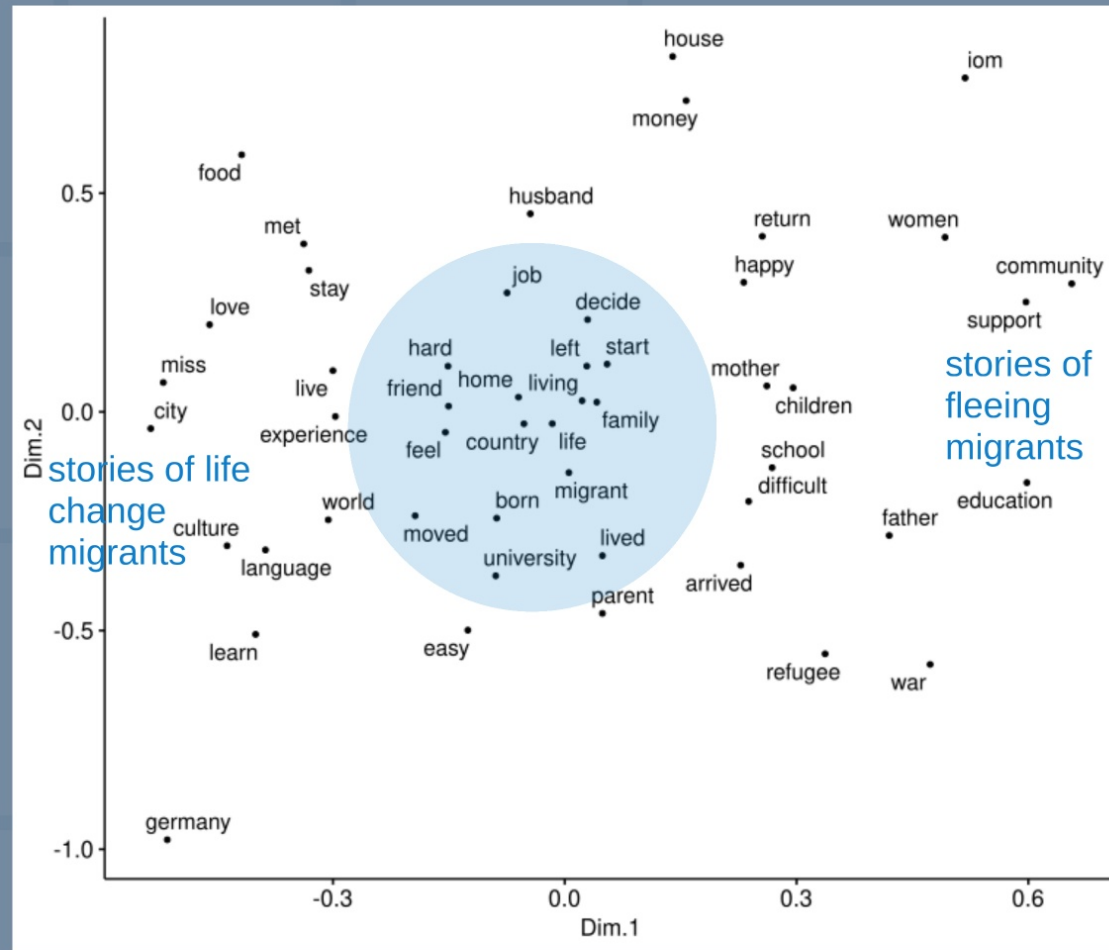
re of
ity:
d coeff.



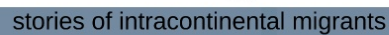
r of
nt words:

t unit:

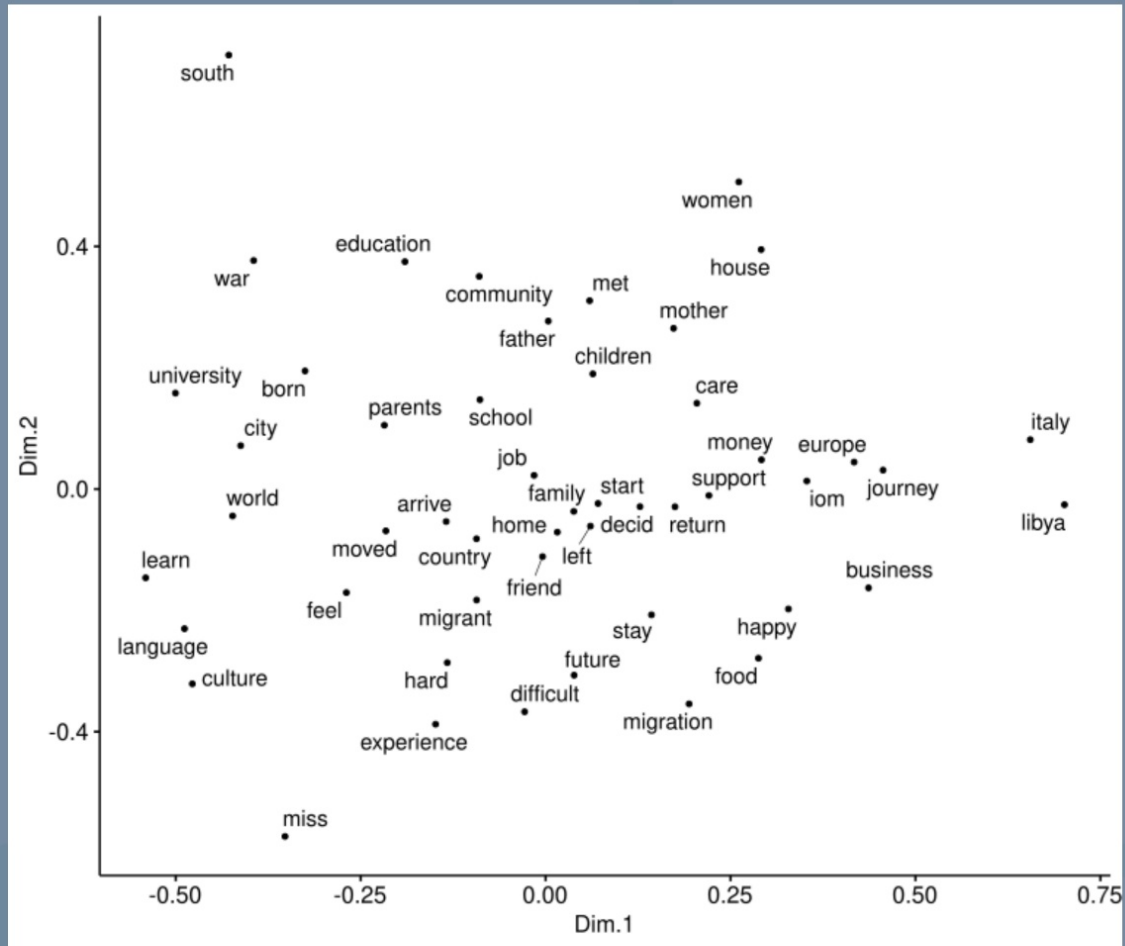
re of
ity:
d coeff.



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

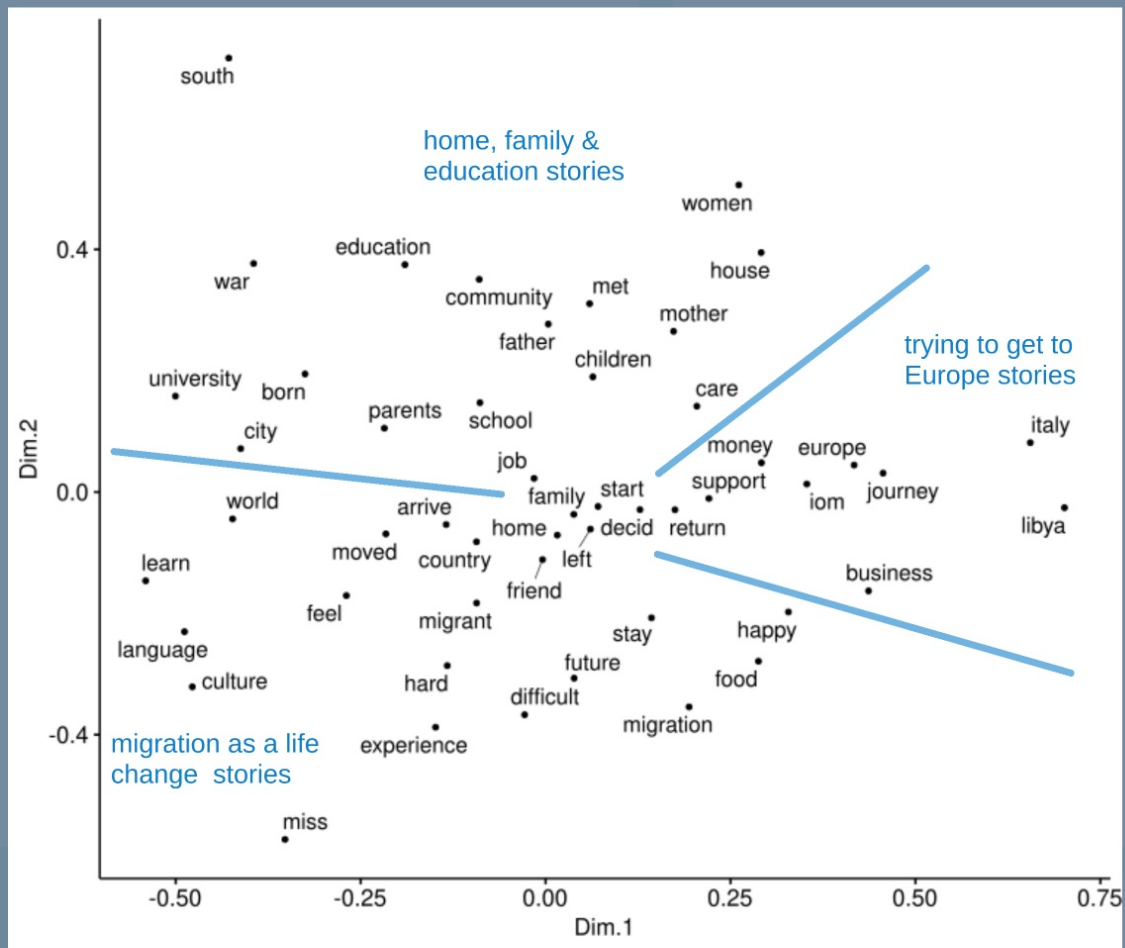


- stories of intercontinental migrants



stories of intracontinental migrants

- num
- frequ
- 50
- cont
- story
- mea
- prox
- Jacc



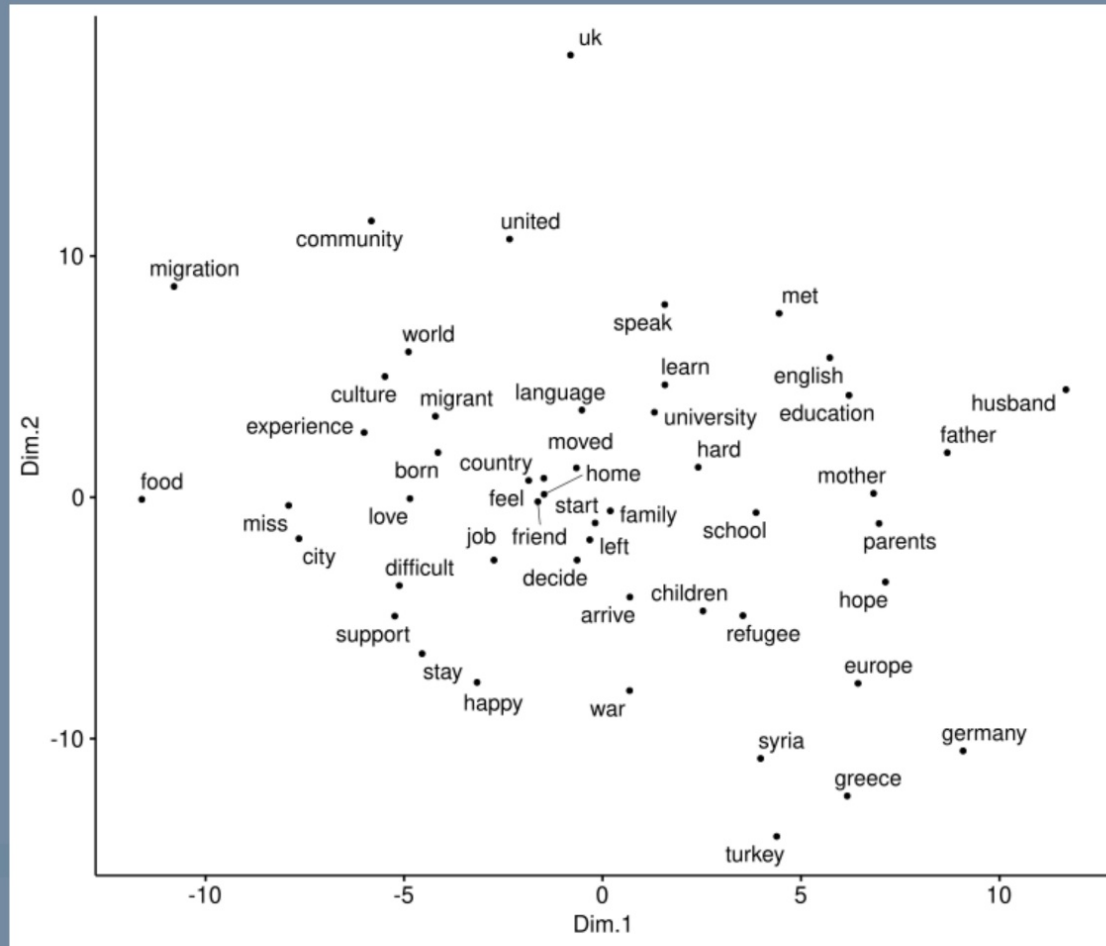
stories of intracontinental migrants

- num
- frequ
- 50
- cont
- story
- mea
- prox
- Jacc

ords:

:

eff.

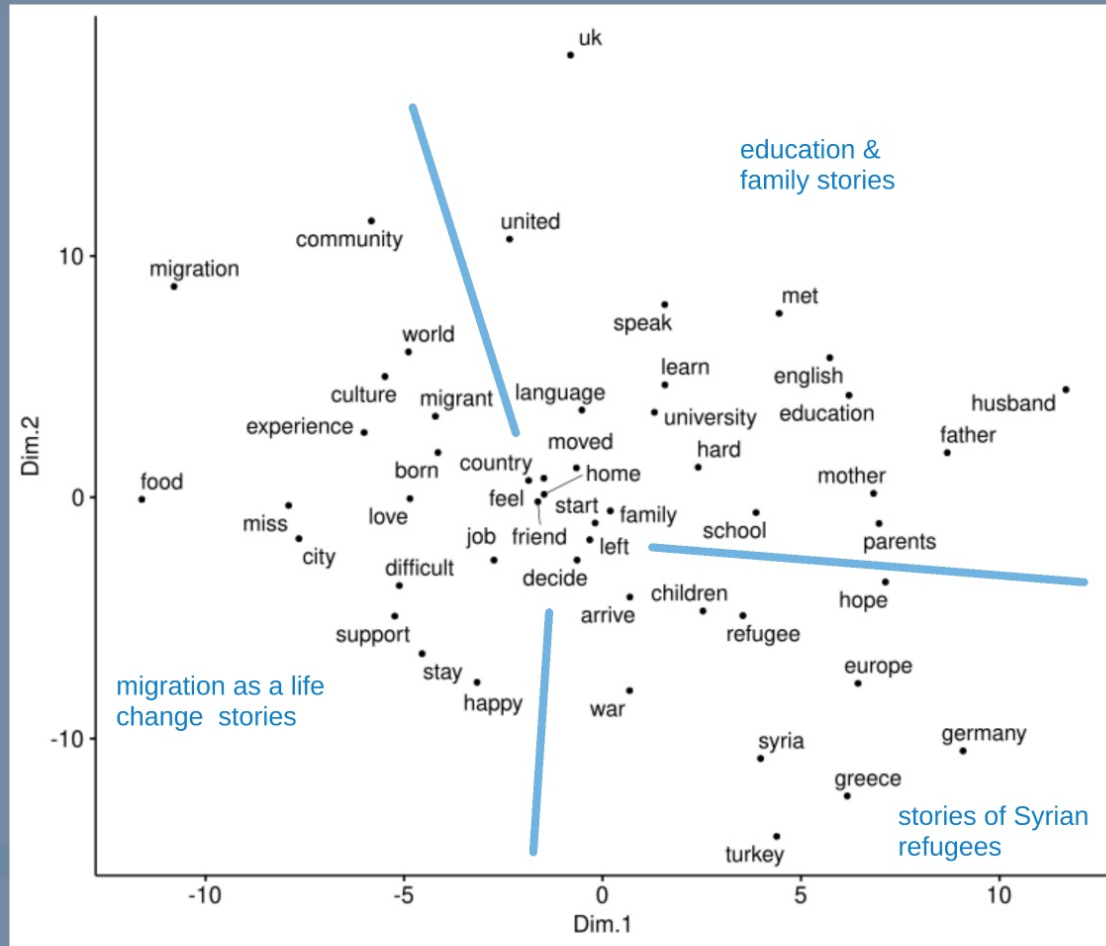


stories of intercontinental migrants

ords:

:

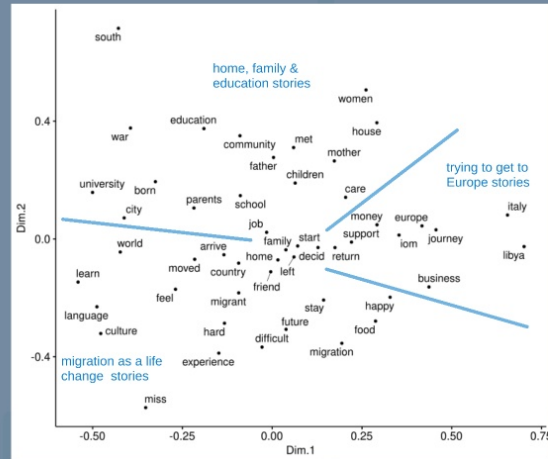
eff.



stories of intercontinental migrants

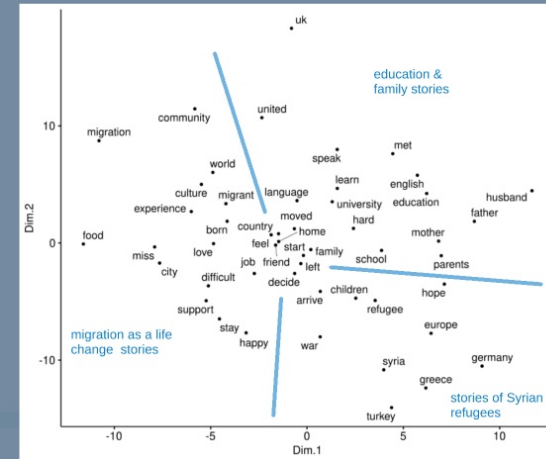
Results II - near and far migrants

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



stories of intracontinental migrants

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



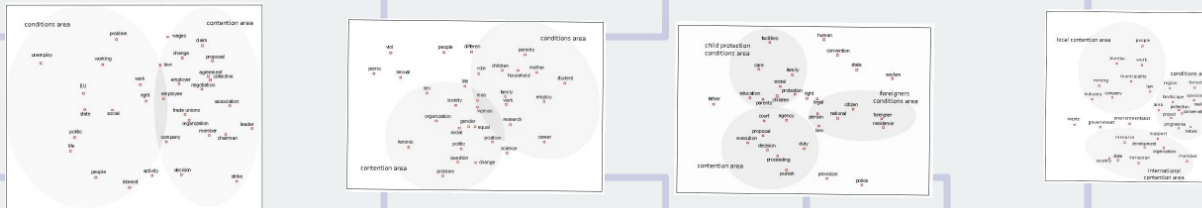
stories of intercontinental migrants

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!