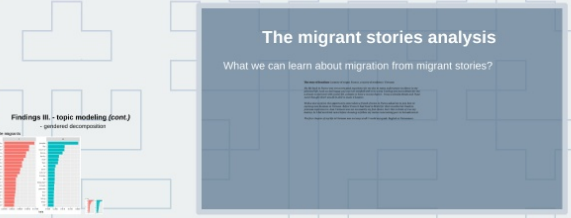
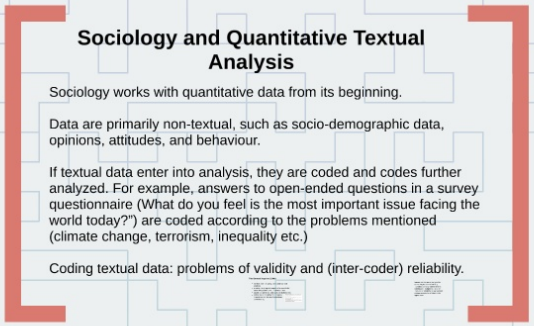
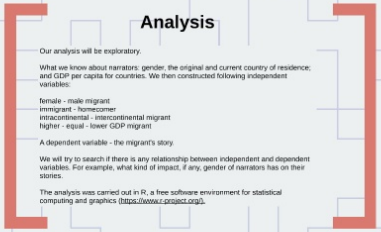
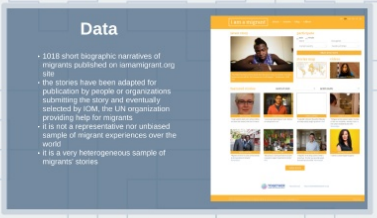
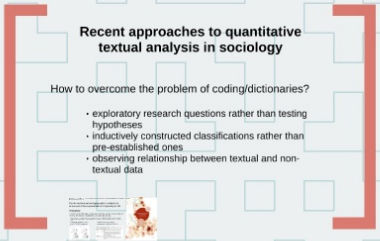
Quantitative Textual Analysis in Sociology I

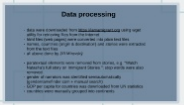


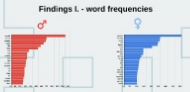












Data Analytics for Students of Social Studies and Humanities. Tuesdays 10:40-12:10
https://ufal.mff.cuni.cz/courses/npfl134

Martin Hájek (martin.hajek@fsv.cuni.cz)

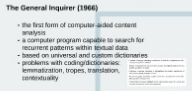
Sociology and Quantitative Textual Analysis

Sociology works with quantitative data from its beginning.

Data are primarily non-textual, such as socio-demographic data, opinions, attitudes, and behaviour.

If textual data enter into analysis, they are coded and codes further analyzed. For example, answers to open-ended questions in a survey questionnaire (What do you feel is the most important issue facing the world today?") are coded according to the problems mentioned (climate change, terrorism, inequality etc.)

Coding textual data: problems of validity and (inter-coder) reliability.



Validity

The General Inquirer (1966)

- the first form of computer-aided content analysis
- a computer program capable to search for recurrent patterns within textual data
- based on universal and custom dictionaries
- problems with coding/dictionaries: lemmatization, tropes, translation, contextuality

Certainty: Language indicating resoluteness, inflexibility, completeness, and
tendency to speak ex cathedra

Activity: Language featuring movement, change, the implementation of ideas
and the avoidance of inertia

Optimism: Language endorsing or highlighting the positive entailments o
some person, group, concept, or event

Realism: Language describing tangible, immediate, recognizable matters tha
affect people's everyday lives

Commonality: Language highlighting the agreed-upon values of a group an
rejecting idiosyncratic modes of engagement

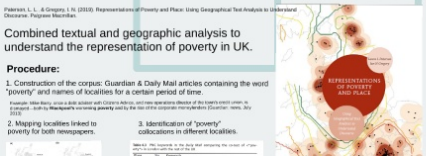
reliability:

- human coders are employed to solve the problem of validity
- it creates a new problem of time, training and budget, but also of inter-coder reliability or agreement
- various measures of inter-coder agreement

Recent approaches to quantitative textual analysis in sociology

How to overcome the problem of coding/dictionaries?

- exploratory research questions rather than testing hypotheses
- inductively constructed classifications rather than pre-established ones
- observing relationship between textual and non-textual data



Paterson, L. L., & Gregory, I. N. (2019). Representations of Poverty and Place: Using Geographical Text Analysis to Understand Discourse. Palgrave Macmillan.

Combined textual and geographic analysis to understand the representation of poverty in UK.

Procedure:

1. Construction of the corpus: Guardian & Daily Mail articles containing the word "poverty" and names of localities for a certain period of time.

Example: Mike Barry, once a debt adviser with Citizens Advice, and now operations director of the town's credit union, is dismayed—both by **Blackpool's** worsening **poverty** and by the rise of the corporate moneylenders (Guardian: news, July 2013)

2. Mapping localities linked to poverty for both newspapers.

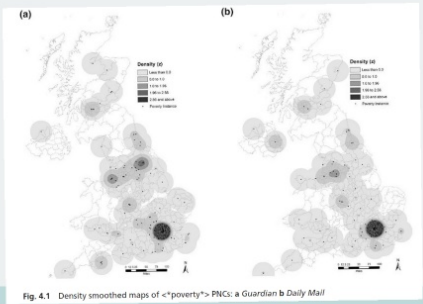
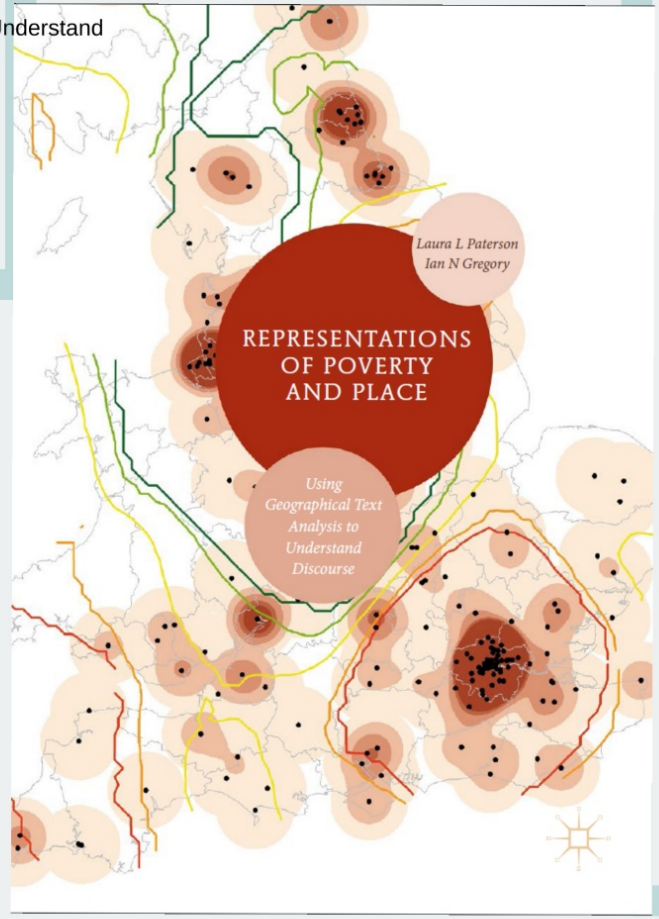


Fig. 4.1 Density smoothed maps of <*poverty*> PNCs: a *Guardian* b *Daily Mail*

3. Identification of "poverty" collocations in different localities.

Table 4.3 PNC keywords in the *Daily Mail* comparing the co-text of <*poverty*> in London with the rest of the UK

Place	Sig.	Keywords
London	<.01	two, one, rate, reached, country, protest, he, Victorian, years
London	<.05	slum, caf, says, reason, cereals, imported, walked, Keely, poorest, improved, great, compared, depression, Facebook, home, become, during, high, rose, class, byword, nostalgia, despite, situation, capital, broadly, miles, men, walk, show
Rest of UK	<.01	too, risks, lowest, people, left, care, we
Rest of UK	<.05	level, making, road, citing, wherever, small, fuel, herself, find, families, local, her, poorer, five, here, taking, miss, ignoring



The migrant stories analysis

What we can learn about migration from migrant stories?

The story of Jonathan (country of origin: France, country of residence: Vietnam

My life back in France was not a very good experience for me due to many unfortunate incidents in my personal life. I did not feel happy and was not satisfied with it in a way. Leaving was just a dream for me, a dream to start over with a new life, a dream to have a second chance. It was a distant dream and I had never thought that I would be able to make it happen.

With a nice surprise, the opportunity came when a friend of mine in France asked me to join him in opening new business in Vietnam. Before France I had lived in Brazil for three months but I had no previous experience in Asia. Vietnam was not necessarily my first choice, but I like to think of it as my destiny. So I did not think twice before choosing to follow my instinct and taking part in this adventure.

The first chapter of my life in Vietnam was not easy at all. I could not speak English or Vietnamese...

The migrant stories analysis

What we can learn about migration from migrant stories?

The story of Jonathan (country of origin: France, country of residence: Vietnam

My life back in France was not a very good experience for me due to many unfortunate incidents in my personal life. I did not feel happy and was not satisfied with it in a way. Leaving was just a dream for me, a dream to start over with a new life, a dream to have a second chance. It was a distant dream and I had never thought that I would be able to make it happen.

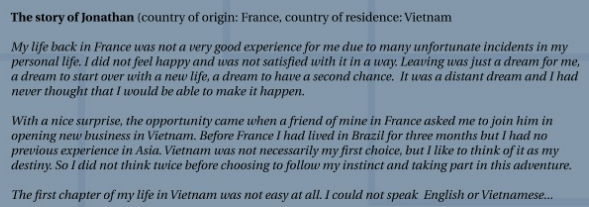
With a nice surprise, the opportunity came when a friend of mine in France asked me to join him in opening new business in Vietnam. Before France I had lived in Brazil for three months but I had no previous experience in Asia. Vietnam was not necessarily my first choice, but I like to think of it as my destiny. So I did not think twice before choosing to follow my instinct and taking part in this adventure.

The first chapter of my life in Vietnam was not easy at all. I could not speak English or Vietnamese...

- we can try to understand migrant's experience from individual stories = a hermeneutic (qualitative) approach

The migrant stories analysis

What we can learn about migration from migrant stories?



The story of Jonathan (country of origin: France, country of residence: Vietnam

My life back in France was not a very good experience for me due to many unfortunate incidents in my personal life. I did not feel happy and was not satisfied with it in a way. Leaving was just a dream for me, a dream to start over with a new life, a dream to have a second chance. It was a distant dream and I had never thought that I would be able to make it happen.

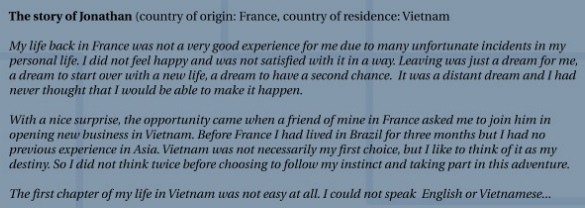
With a nice surprise, the opportunity came when a friend of mine in France asked me to join him in opening new business in Vietnam. Before France I had lived in Brazil for three months but I had no previous experience in Asia. Vietnam was not necessarily my first choice, but I like to think of it as my destiny. So I did not think twice before choosing to follow my instinct and taking part in this adventure.

The first chapter of my life in Vietnam was not easy at all. I could not speak English or Vietnamese...

- we can try to understand migrant's experience from individual stories = a hermeneutic (qualitative) approach
- we can compare stories of different categories of migrants, observe similarities and differences in what they told and make inferences about migration (how is narrated and experienced) = quantitative textual approach

The migrant stories analysis

What we can learn about migration from migrant stories?



The story of Jonathan (country of origin: France, country of residence: Vietnam

My life back in France was not a very good experience for me due to many unfortunate incidents in my personal life. I did not feel happy and was not satisfied with it in a way. Leaving was just a dream for me, a dream to start over with a new life, a dream to have a second chance. It was a distant dream and I had never thought that I would be able to make it happen.

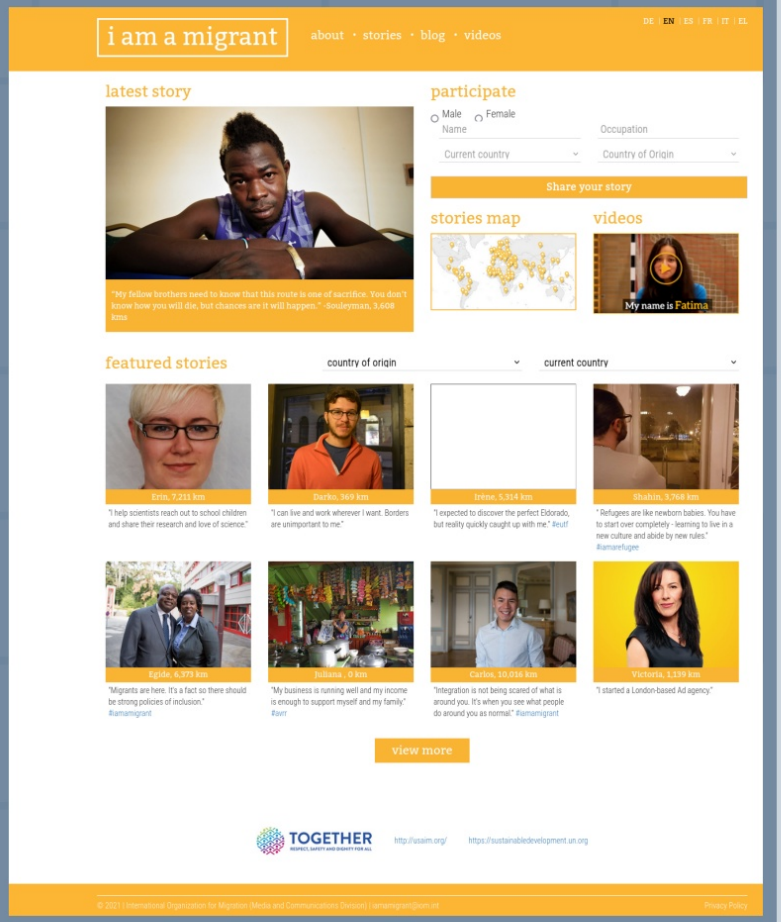
With a nice surprise, the opportunity came when a friend of mine in France asked me to join him in opening new business in Vietnam. Before France I had lived in Brazil for three months but I had no previous experience in Asia. Vietnam was not necessarily my first choice, but I like to think of it as my destiny. So I did not think twice before choosing to follow my instinct and taking part in this adventure.

The first chapter of my life in Vietnam was not easy at all. I could not speak English or Vietnamese...

- we can compare stories of different categories of migrants, observe similarities and differences in what they told and make inferences about migration (how is narrated and experienced) = quantitative textual approach

Data

- 1018 short biographic narratives of migrants published on iamamigrant.org site
- the stories have been adapted for publication by people or organizations submitting the story and eventually selected by IOM, the UN organization providing help for migrants
- it is not a representative nor unbiased sample of migrant experiences over the world
- it is a very heterogeneous sample of migrants' stories



Data processing

- data were downloaded from https://iamamigrant.org using wget utility for retrieving files from the Internet
- html files (web pages) were converted into plain text files
- names, countries (origin & destination) and stories were extracted from the text files
- all above done by Jiří Mírovský

- paratextual elements were removed from stories, e.g. "Watch Natasha's full story on Immigrant Stories."; stop words were also removed
- gender of narrators was identified semiautomatically (gendernamefinder.com + manual search)
- GDP per capita for countries was downloaded from UN statistics
- countries were manually grouped into continents

Analysis

Our analysis will be exploratory.

What we know about narrators: gender, the original and current country of residence; and GDP per capita for countries. We then constructed following independent variables:

female - male migrant
immigrant - homecomer
intracontinental - intercontinental migrant
higher - equal - lower GDP migrant

A dependent variable - the migrant's story.

We will try to search if there is any relationship between independent and dependent variables. For example, what kind of impact, if any, gender of narrators has on their stories.

The analysis was carried out in R, a free software environment for statistical computing and graphics (https://www.r-project.org/).

Text analytical techniques used in the study

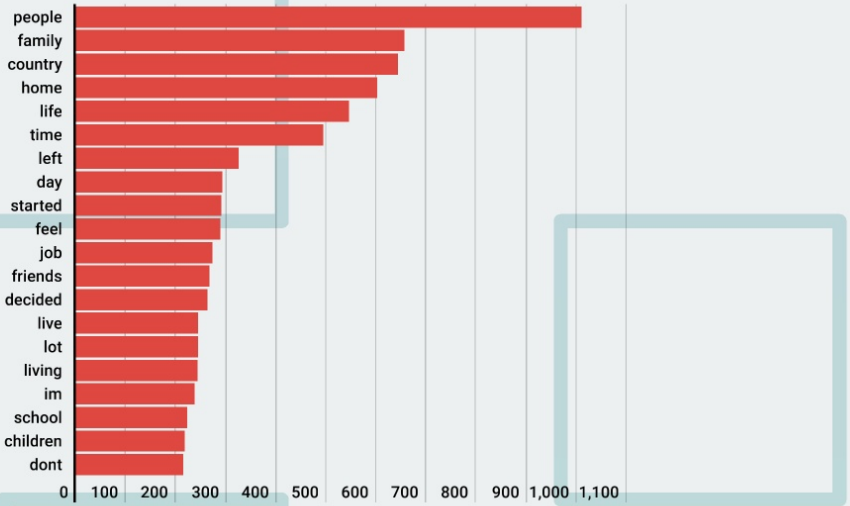
Word frequencies: calculation of most frequent semantic words; it gives us basic information about words through which the migrant experience is expressed

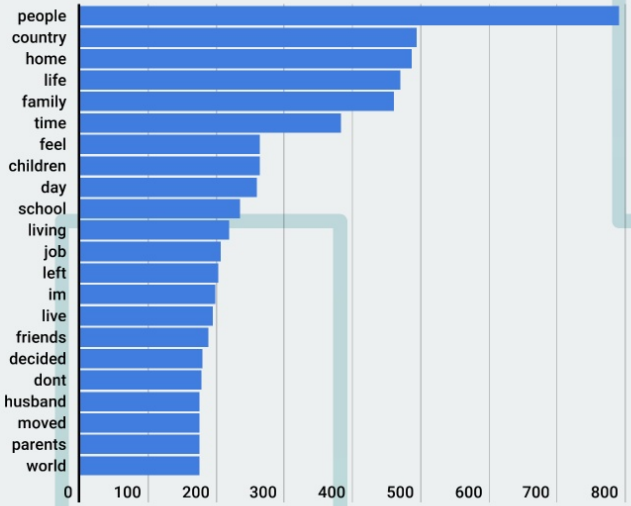
Bigrams: identification of frequent pairs (collocations) of semantic words occurring in the narratives; it informs us about typical entities, such as "primary school" or "speak English"

Topic modeling: identifying clusters or recurring patterns of co-occurring words (called “topics”); it provides information about potential themes or topics in the narratives

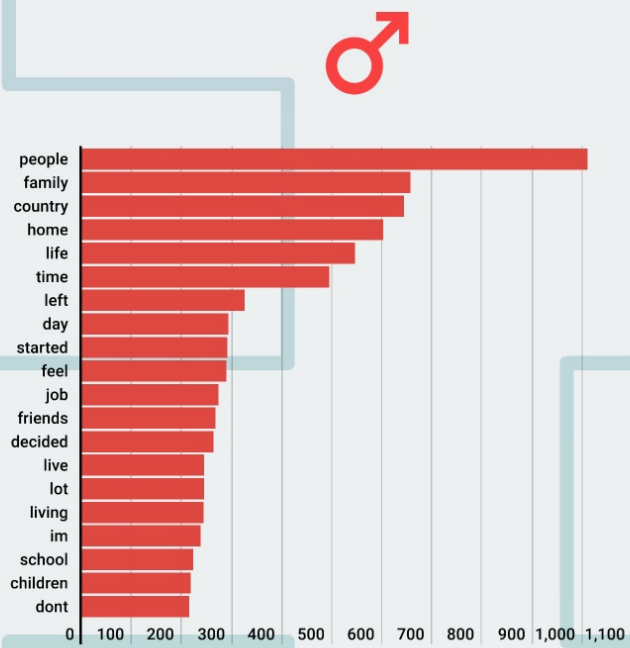
increase in analytical complexity

Findings I. - word frequencies





Findings I. - word frequencies



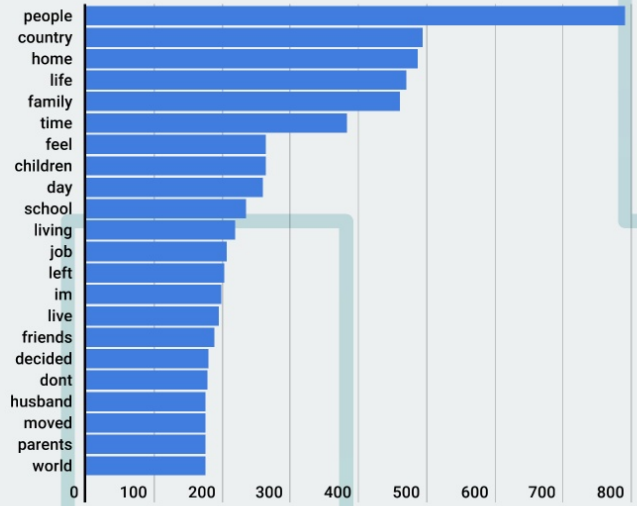
people

---

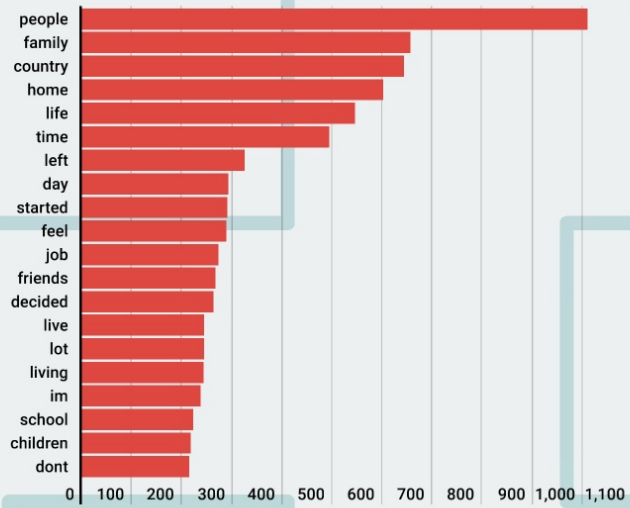
family
country
home
life
time

---

left	children
job	school
friends	living
living	job
school	left
children	friends

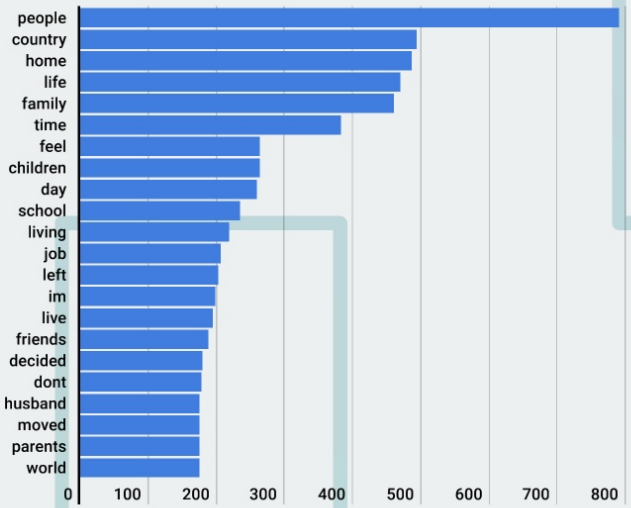


Findings I. - word frequencies



people
---
family
country
home
life
time
---

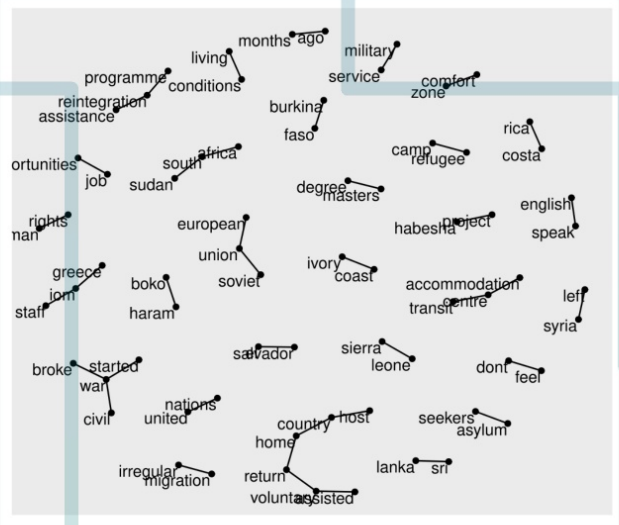
left	children
job	school
friends	living
living	job
school	left
children	friends



Results: There are no surprising differences between male and female migrants' stories on the level of words. It seems that migration is linked to families; migrants of both genders express it through similar words (-- share similar experiences?).

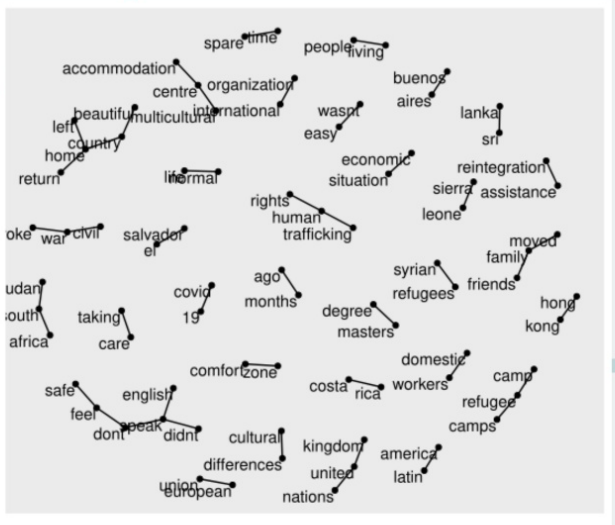
Findings II. - bigrams

men



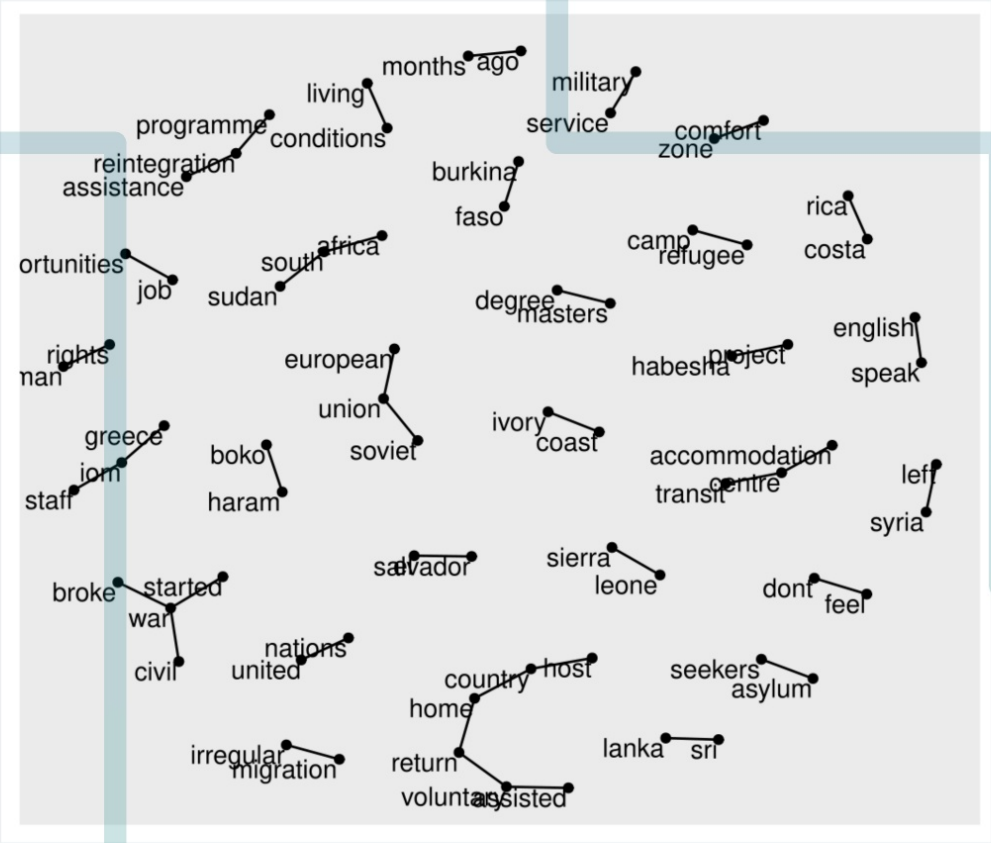
minimal occurrence = 9

women



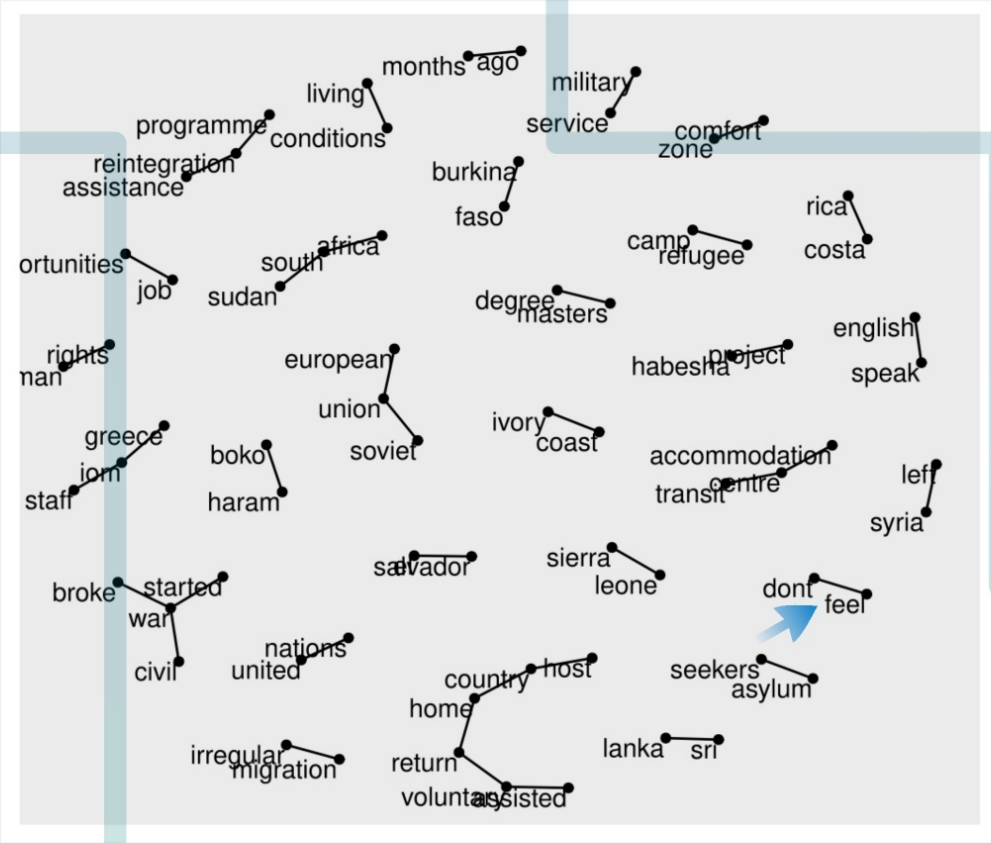
minimal occurrence = 6

men



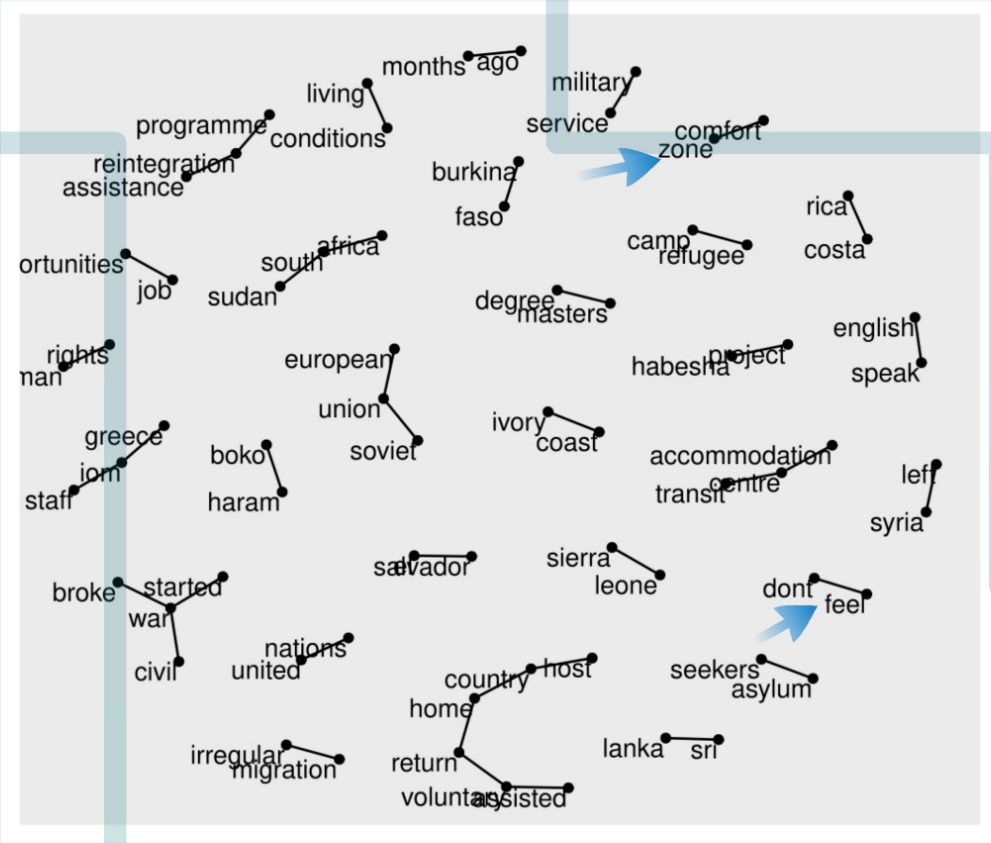
minimal occurrence = 9

men



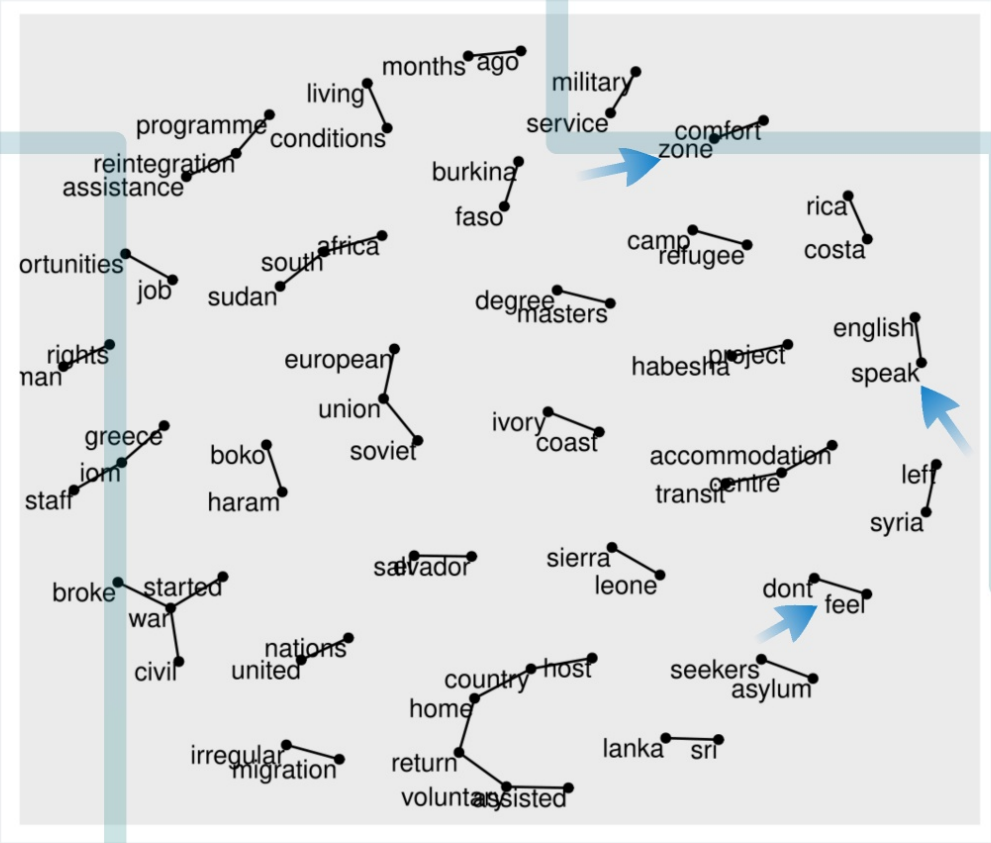
minimal occurrence = 9

men



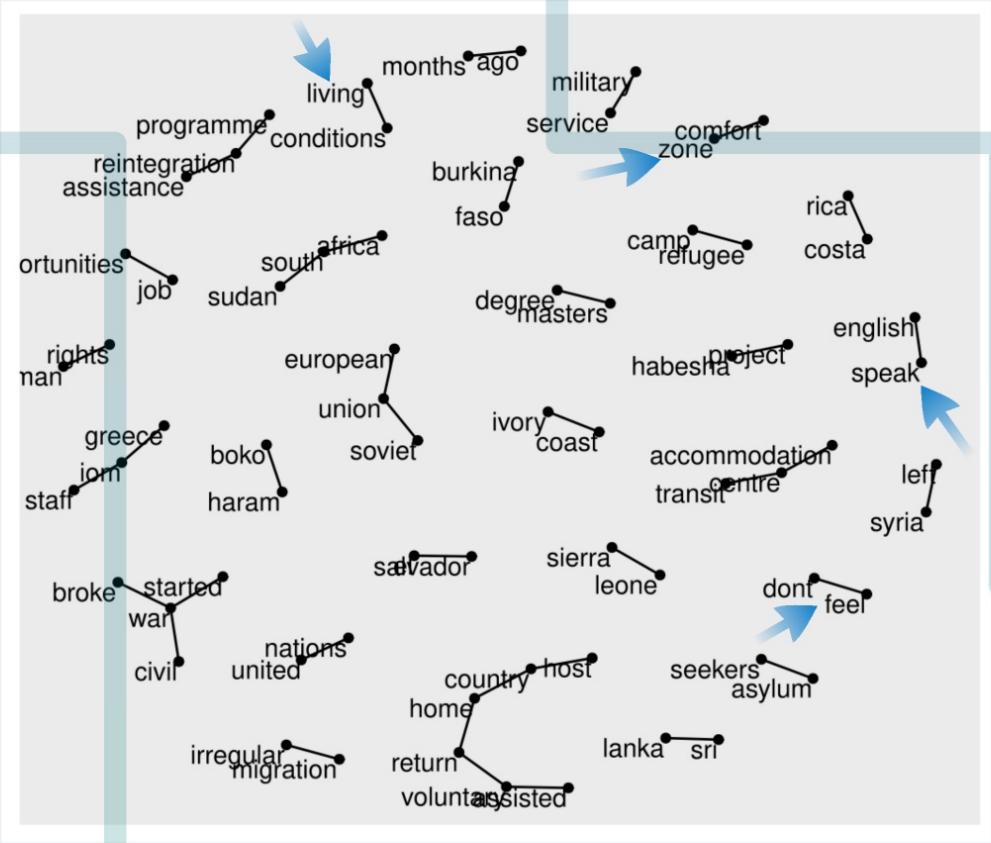
minimal occurrence = 9

men



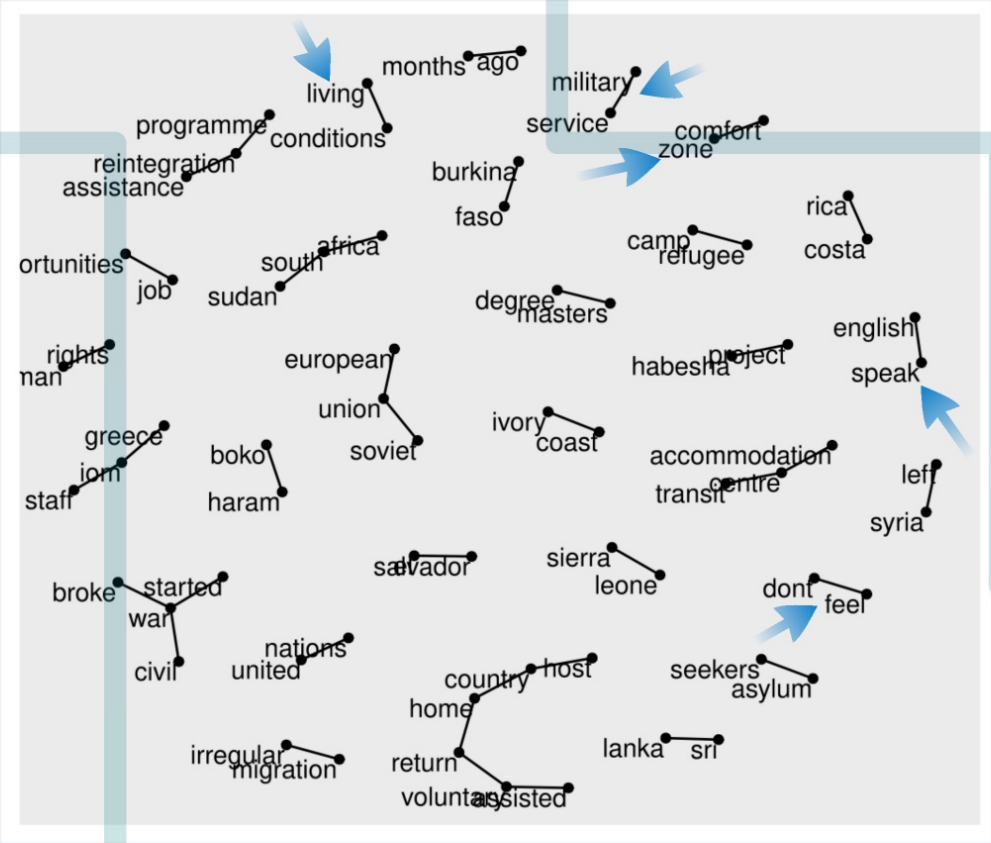
minimal occurrence = 9

men



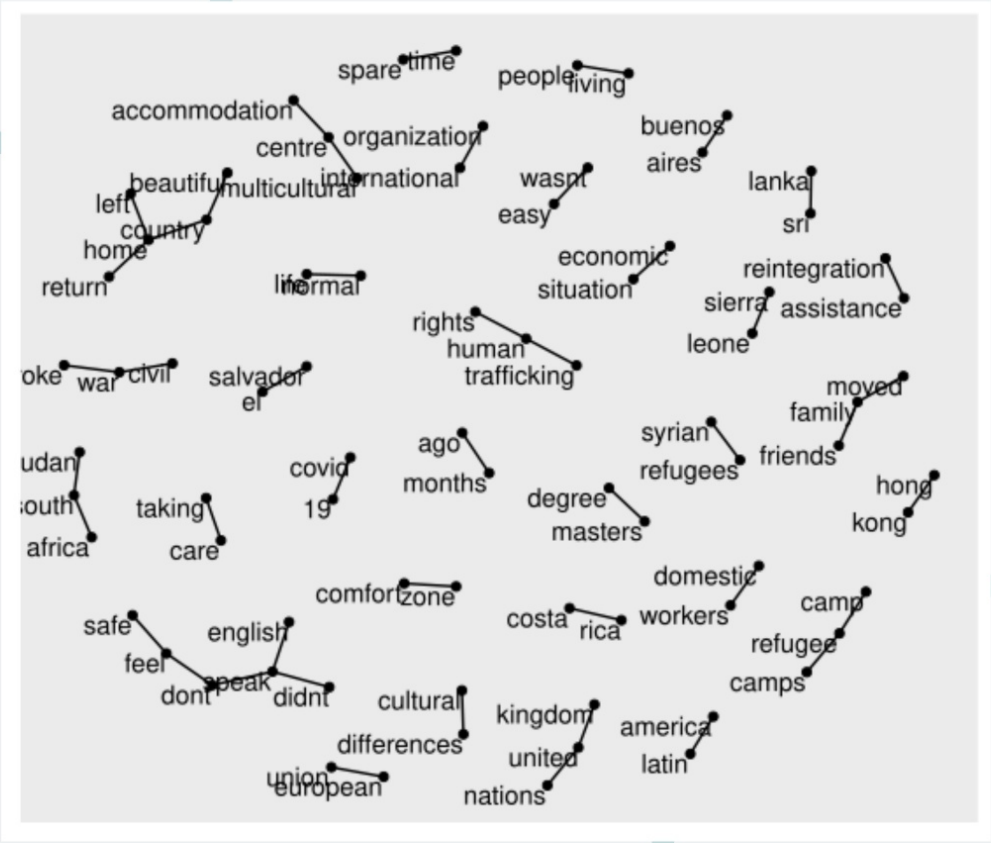
minimal occurrence = 9

men



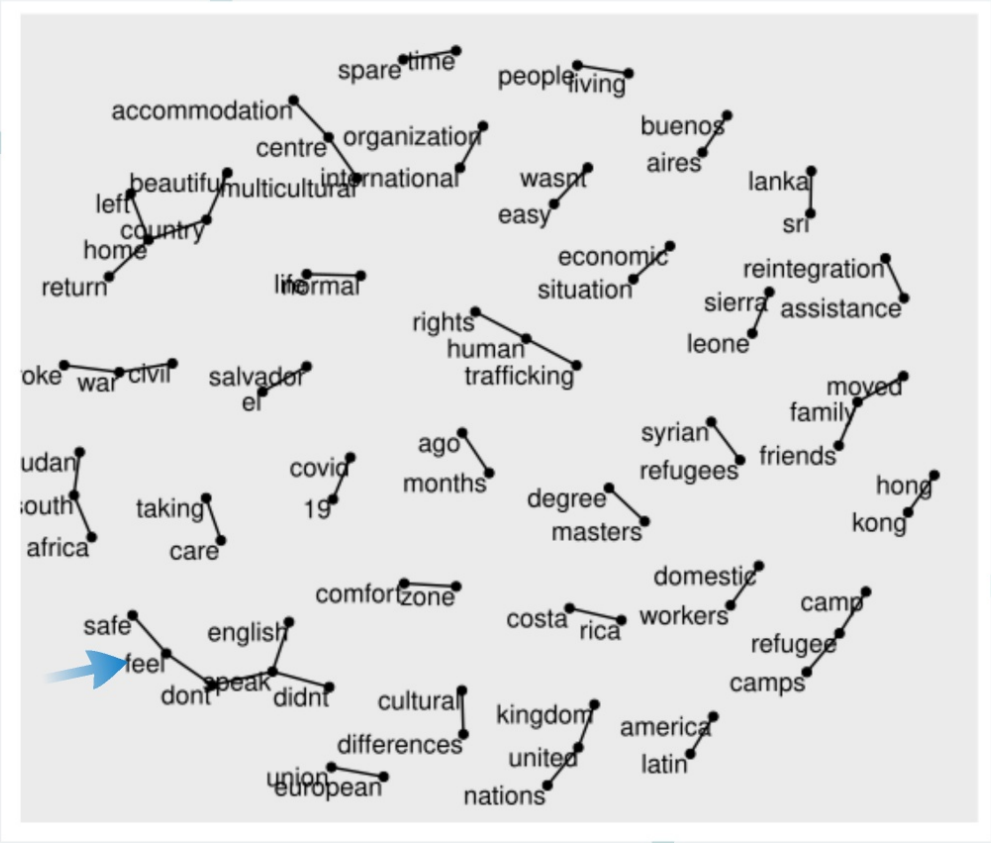
minimal occurrence = 9

women



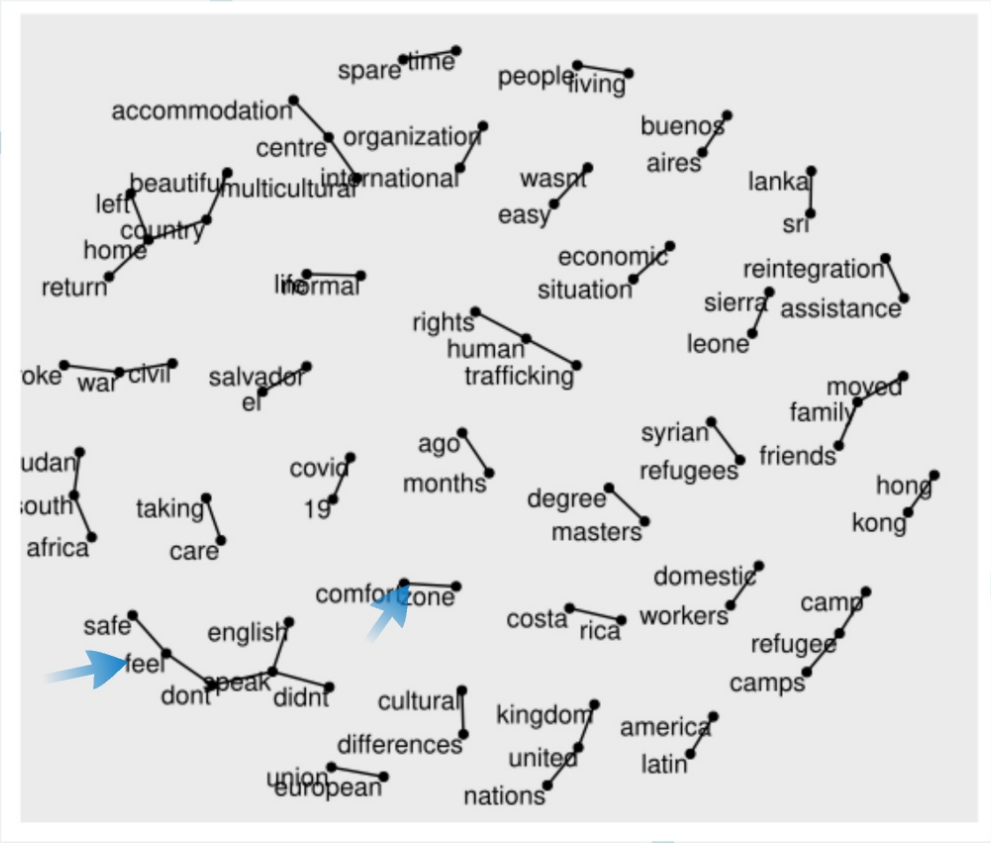
minimal occurrence = 6

women



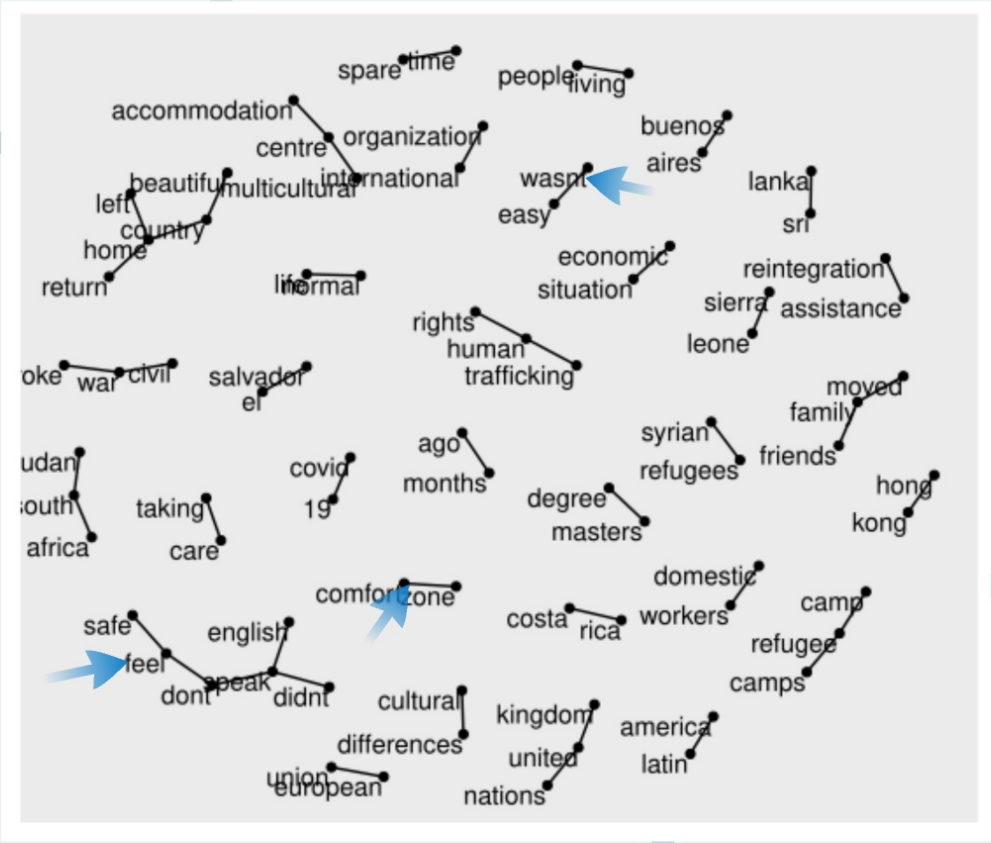
minimal occurrence = 6

women



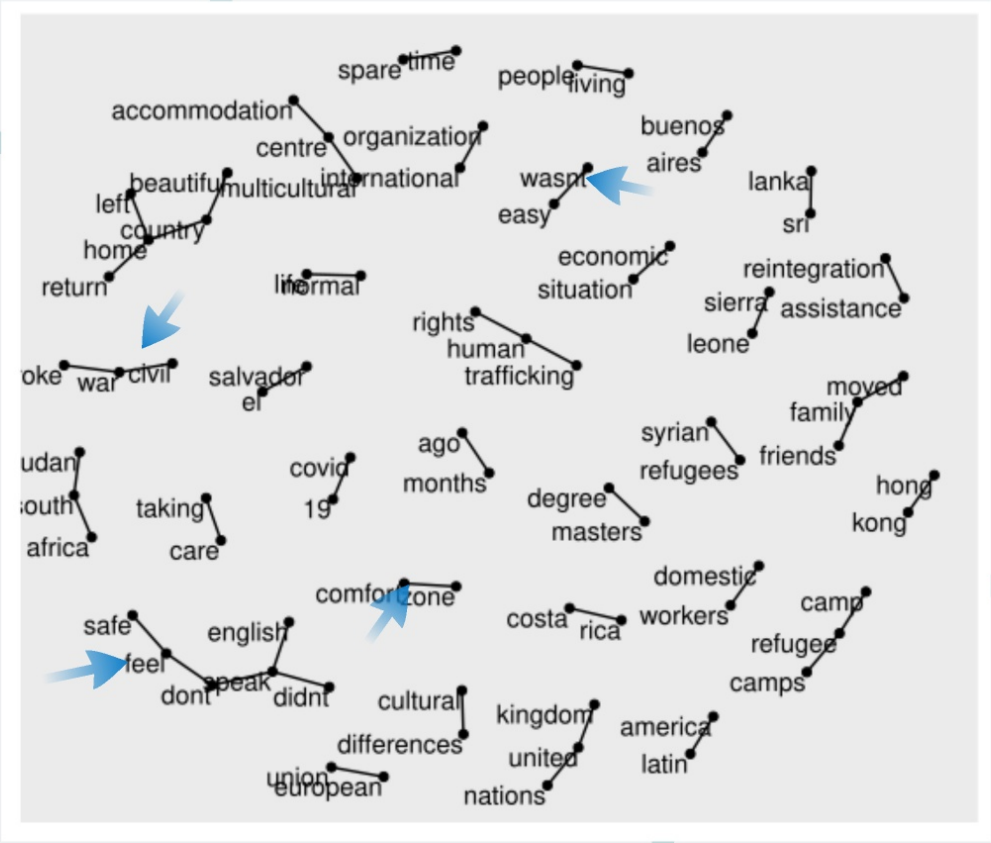
minimal occurrence = 6

women



minimal occurrence = 6

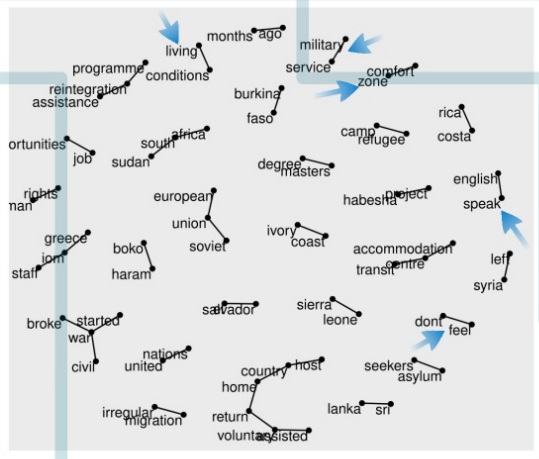
women



minimal occurrence = 6

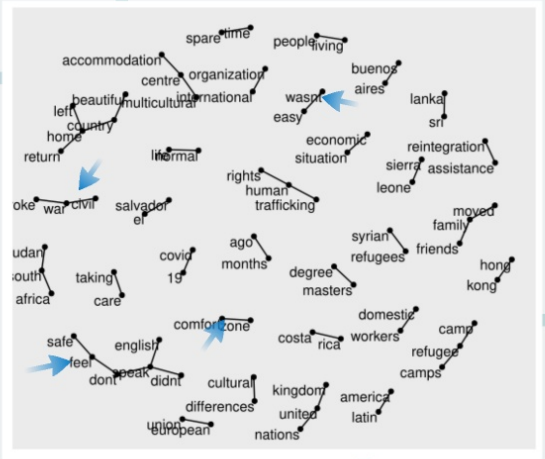
Findings II. - bigrams

men



minimal occurrence = 9

women

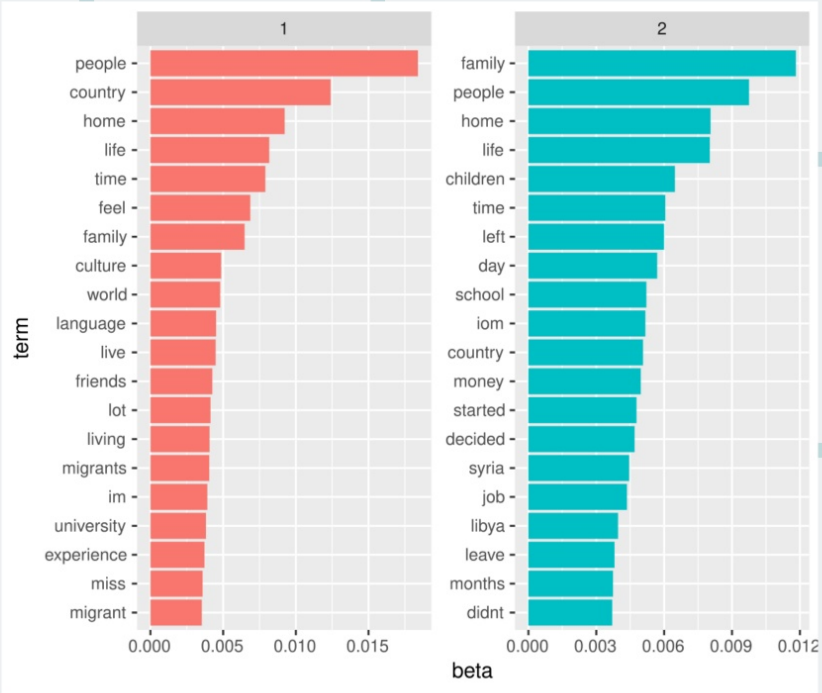


minimal occurrence = 6

Results: through the analysis of bigrams, we have got more precise information about the contexts of some frequent words in narratives; it seems that the narratives depict challenging situations and feelings rather than easy ones. It also points to some typical actors, processes, and institutions migrants encounter.

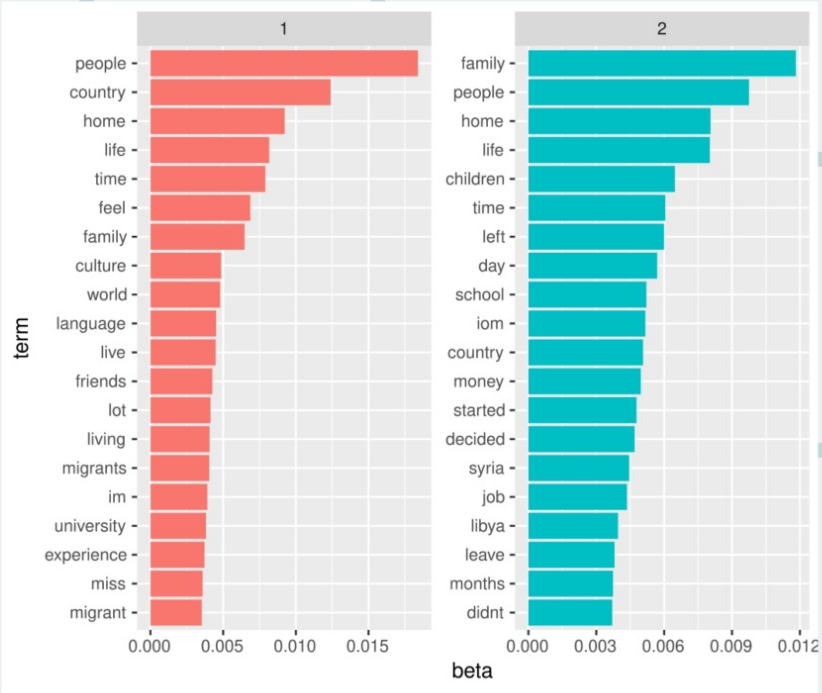
Findings III. - topic modeling

- computing macro patterns of words ("topics") in the narratives
- a single narrative can contain one or more topics
- probabilistic estimation = it can generate valid results but also artifacts



Findings III. - topic modeling

- computing macro patterns of words ("topics") in the narratives
- a single narrative can contain one or more topics
- probabilistic estimation = it can generate valid results but also artifacts



Topic 1

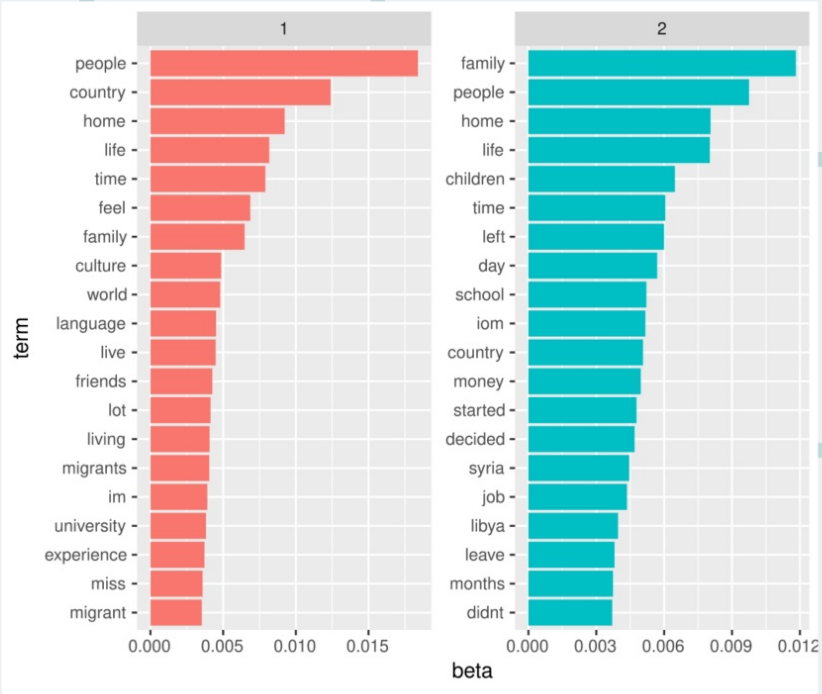
country, feel, culture, world, language, friends, migrants, uni, experience, miss

Topic 2

family, children, left, school, IOM, money, started, decided, Syria, Libya

Findings III. - topic modeling

- computing macro patterns of words ("topics") in the narratives
- a single narrative can contain one or more topics
- probabilistic estimation = it can generate valid results but also artifacts



Topic 1

country, feel, culture, world, language, friends, migrants, uni, experience, miss

Topic 2

family, children, left, school, IOM, money, started, decided, Syria, Libya

Interpretation

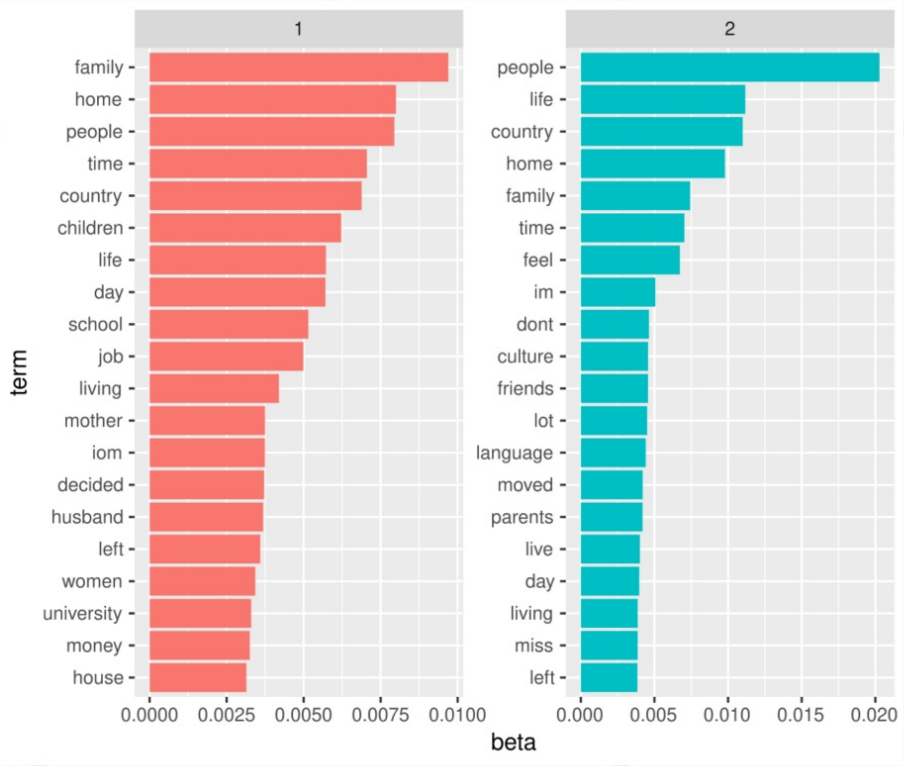
Topic 1 - general narratives of migrants about their experience of moving from one country to another, encountering different cultures, languages and feelings associated with this change

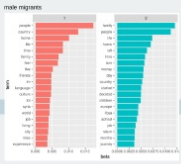
Topic 2 - a particular theme of war migrants (refugees) who fled from their homes and sought asylum elsewhere. They often come from Syria or transit through or end in Libya.

Findings III. - topic modeling *(cont.)*

- gendered decomposition

female migrants

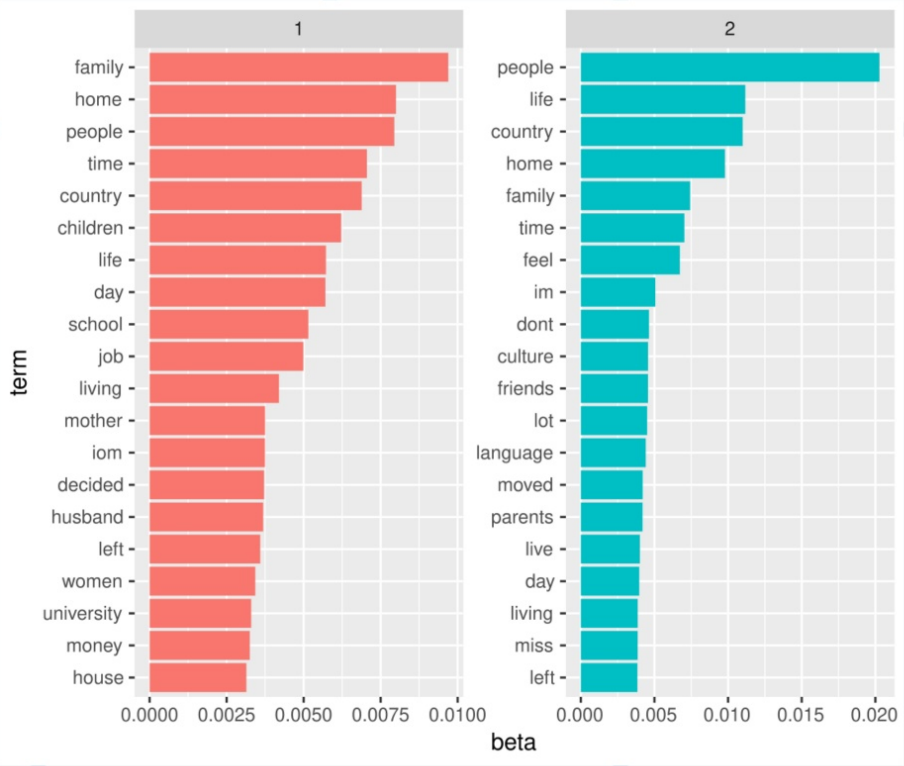




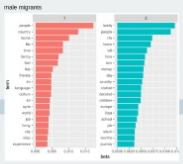
Findings III. - topic modeling *(cont.)*

- gendered decomposition

female migrants



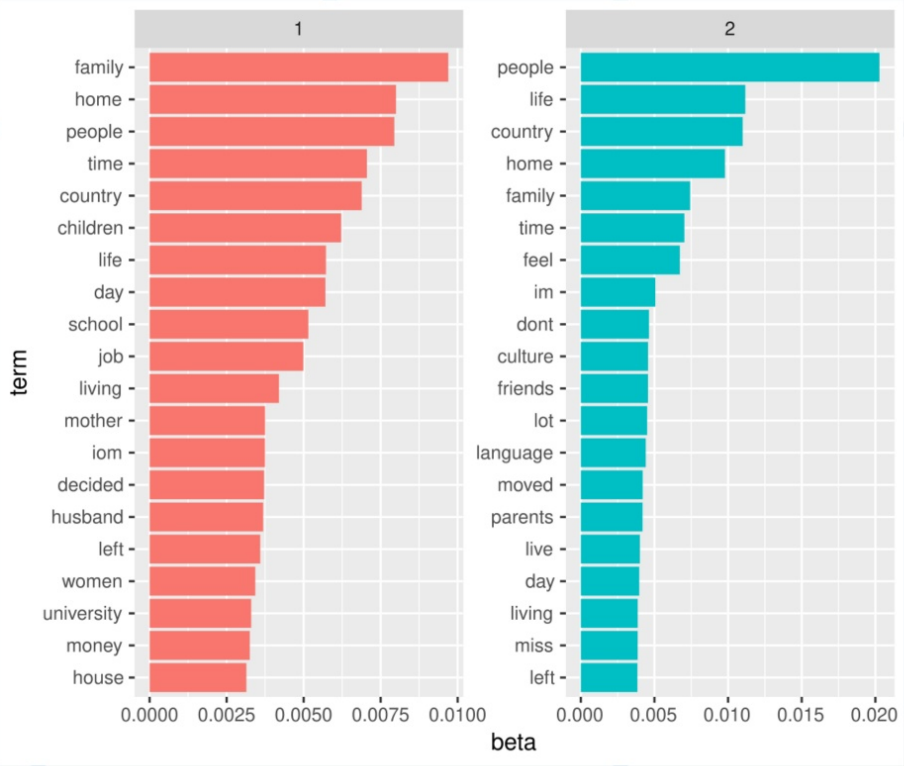
War migrant topic is dominant, general migrant experience topic is secondary.



Findings III. - topic modeling *(cont.)*

- gendered decomposition

female migrants

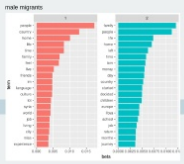


War migrant topic is dominant, general migrant experience topic is secondary.

Interpretation

A. There is more war than non-war female migrants in the sample as they are less often voluntary migrants.

B. Migration is always challenging for female migrants; therefore, they express their experience similarly to war migrants.



Homework

Carry out your own analysis on the same sample of narratives. Compare groups in any subset of narratives:

- immigrants vs. homecomers
- transcontinental vs. intracontinental
- immigrants to countries with (much) higher, equal or (much) lower GDP
- men vs. women.

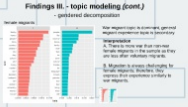
Download data from google drive https://tinyurl.com/26vpzrj6

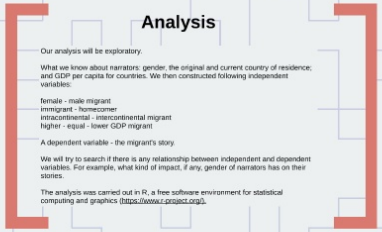
Use a web application Voyant (https://voyant-tools.org/) to explore similarities and differences between groups of stories. Play with the app and try different analytical tools.

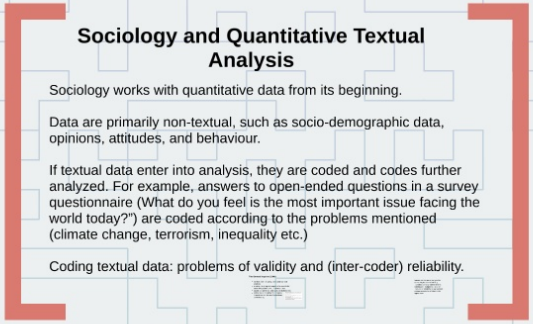
Make a report with your findings and tentative interpretation.

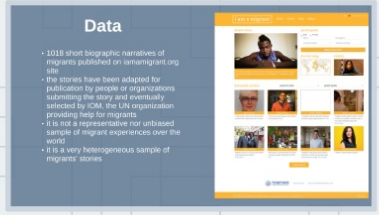
Deadline: Sunday 17.4. 2020 midnight.

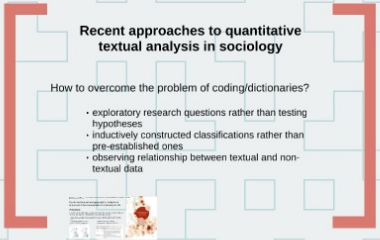
Quantitative Textual Analysis in Sociology I

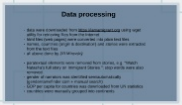












Data Analytics for Students of Social Studies and Humanities. Tuesdays 10:40-12:10
https://ufal.mff.cuni.cz/courses/npfl134

Martin Hájek (martin.hajek@fsv.cuni.cz)