EUI DEPARTMENT OF POLITICAL AND SOCIAL SCIENCES

Computational Text Analysis and the Politics of Education

University of Marburg

Milan Thies

06.2025

www.eui.eu

Recap

Co-funded by the Erasmus+ Programme of the European Union

What are dictionaries?

- **Keys** that stand for a theorized concept
- **Values** that represent the keys empirically

- Example: Food (key); bread, apple, banana, chocolate, potatoes (values)

- How to measure the prevalence of a concept in a corpus?
  - We count the frequency of dictionary features (values)

Co-funded by the Erasmus+ Programme of the European Union

Dictionary creation: Best practice

1. Creating a top-features list
   - From data
2. Deductively selecting features
   - Which of these explain the theoretical concept well (unambiguously)
3. Manual validation
   - Human coders determine whether feature should be included or not

POLITICAL COMMUNICATION

Political Communication

ISSN: 1058-4609 (Print) 1091-7675 (Online) Journal homepage: https://www.tandfonline.com/loi/upcp20

(Re)Claiming Our Expertise: Parsing Large Text Corpora With Manually Validated and Organic Dictionaries

Ashley Muddiman, Shannon C. McGregor & Natalie Jomini Stroud

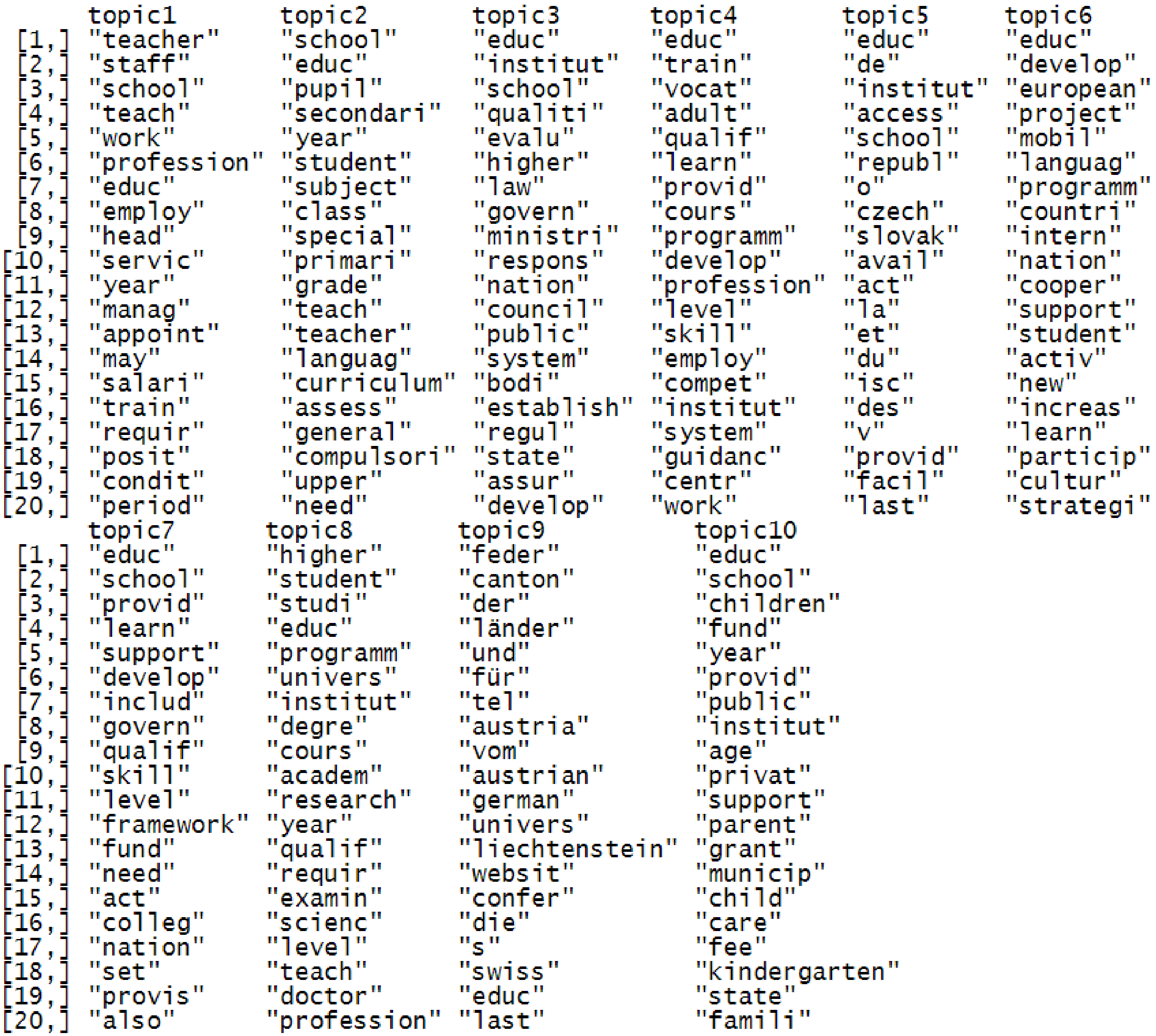
What methods does Martin (2018) use?

IMAGINE ALL THE PEOPLE

Literature, Society, and Cross-National Variation in Education Systems

By CATHIE JO MARTIN*

- Topic 1 seems to be about the **teaching profession**
- Topic 2 seems to be about **school education** but there is room for ambiguity
- Topic 3 is about **governance**
- Topic 4 about the education as means to create **human capital on the labor market**
- Topic 6 seems to be **European integration**
- Topic 8 is about **higher education**.
- Topic 9 is about federalism
- Topic 10 seems to be about the perspective of **families** and early childhood education and care (**ECEC**)
- Topic 5 and 7 hard to interpret because of language diversity in data set – lots of gibberish and very frequent words. Common denominator: It is about **education**. More cleaning could help.



```
      topic1       topic2        topic3        topic4        topic5       topic6
 [1,] "teacher"    "school"      "educ"        "educ"        "educ"       "educ"
 [2,] "staff"      "educ"        "institut"    "train"       "de"         "develop"
 [3,] "school"     "pupil"       "school"      "vocat"       "institut"   "european"
 [4,] "teach"      "secondari"   "qualiti"     "adult"       "access"     "project"
 [5,] "work"       "year"        "evalu"       "qualif"      "school"     "mobil"
 [6,] "profession" "student"     "higher"      "learn"       "republ"     "languag"
 [7,] "educ"       "subject"     "law"         "provid"      "o"          "programm"
 [8,] "employ"     "class"       "govern"      "cours"       "czech"      "countri"
 [9,] "head"       "special"     "ministri"    "programm"    "slovak"     "intern"
[10,] "servic"     "primari"     "respons"     "develop"     "avail"      "nation"
[11,] "year"       "grade"       "nation"      "profession"  "act"        "cooper"
[12,] "manag"      "teach"       "council"     "level"       "la"         "support"
[13,] "appoint"    "teacher"     "public"      "skill"       "et"         "student"
[14,] "may"        "languag"     "system"      "employ"      "du"         "activ"
[15,] "salari"     "curriculum"  "bodi"        "compet"      "isc"        "new"
[16,] "train"      "assess"      "establish"   "institut"    "des"        "increas"
[17,] "requir"     "general"     "regul"       "system"      "v"          "learn"
[18,] "posit"      "compulsori"  "state"       "guidanc"     "provid"     "particip"
[19,] "condit"     "upper"       "assur"       "centr"       "facil"      "cultur"
[20,] "period"     "need"        "develop"     "work"        "last"       "strategi"
      topic7      topic8       topic9          topic10
 [1,] "educ"      "higher"     "feder"         "educ"
 [2,] "school"    "student"    "canton"        "school"
 [3,] "provid"    "studi"      "der"           "children"
 [4,] "learn"     "educ"       "länder"        "fund"
 [5,] "support"   "programm"   "und"           "year"
 [6,] "develop"   "univers"    "für"           "provid"
 [7,] "includ"    "institut"   "tel"           "public"
 [8,] "govern"    "degre"      "austria"       "institut"
 [9,] "qualif"    "cours"      "vom"           "age"
[10,] "skill"     "academ"     "austrian"      "privat"
[11,] "level"     "research"   "german"        "support"
[12,] "framework" "year"       "univers"       "parent"
[13,] "fund"      "qualif"     "liechtenstein" "grant"
[14,] "need"      "requir"     "websit"        "municip"
[15,] "act"       "examin"     "confer"        "child"
[16,] "colleg"    "scienc"     "die"           "care"
[17,] "nation"    "level"      "s"             "fee"
[18,] "set"       "teach"      "swiss"         "kindergarten"
[19,] "provis"    "doctor"     "educ"          "state"
[20,] "also"      "profession" "last"          "famili"
```

What is seeded LDA topic modelling?

Co-funded by the
Erasmus+ Programme
of the European Union

Seeded LDA Topic Models

- In seeded LDA topic modelling, you can predefine topics using a dictionary of “seed words”.
- But we only nudge the model: Jagarlamudi et. al 2012, 205:

"importantly, we only encourage the model to follow the seed sets and do not force it. So if it has compelling evidence in the data to overcome the seed information then it still has the freedom to do so.”

- More on the theory of seeded LDA: [Watanabe et al (2020: Theory-Driven Analysis of Large Corpora: Semisupervised Topic Classification of the UN Speeches]()

Co-funded by the
Erasmus+ Programme
of the European Union

In-Class Project

- This could be part of your bachelor thesis / master thesis / project studies
- … or it could be just to practice what you have learned.
- The general idea: Noone can really teach you the method theoretically. You learn it by applying it.
- You can bring your own data or use the data I provide
- Do not worry of you have not understood everything yet. You have a whole day and my assistance.

Co-funded by the Erasmus+ Programme of the European Union

Your task:

- This is not just about coding. This is about applying new methods in a research context. Therefore

1. Define a research question that has a theoretical foundation (remember the first day or remember why you are interested in your topic).
2. Import, clean and organize your data. This might be the biggest challenge for those with own data.
3. Apply those methods you think are helpful to answer your research question.
4. Interpret the findings: What have you learned?
5. Outline the limitations of your analysis and describe how you could improve the analysis.
6. Visualize your findings in a nicely knit markdown file. Prepare a short presentation of 5 minutes guiding us through the html file.

Co-funded by the Erasmus+ Programme of the European Union

Data to work with

- Your own data
- UN General Assembly Speeches:
  - Corpus of texts of General Debate statements from 1970 (Session 25) to 2016 (Session 71). Docvars include country, year and session
- ParlSpeech:
  - Full text corpora of parliamentary speeches in Austria, the Czech Republic, Germany, Denmark, the Netherlands, New Zealand, Spain, Sweden, and the United Kingdom

Co-funded by the Erasmus+ Programme of the European Union