EurLex Fun

Paula & Robert

2024-07-18

Research Angle

- Interested in looking at how different policy narratives/ positions/priorities may change or evolve over time
- Interested in EU politics and EU institutions
- Substantive policy focus: migration (Paula), digitization (Rob)

Initial Research Question

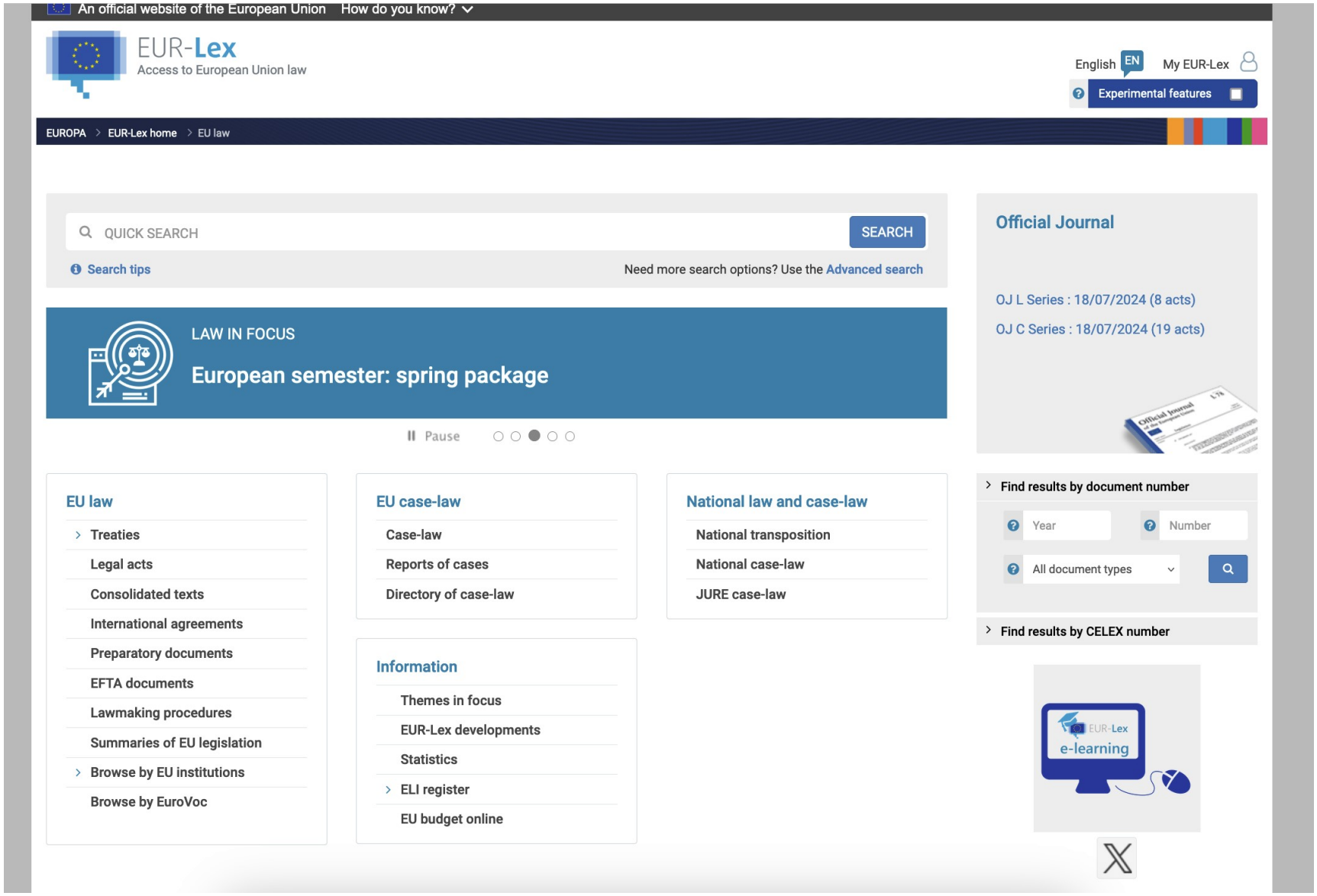
- The EU is often portrayed in certain academic + policy discourses as a ‘rights’ driven actor
  - But the reality is more complicated, and in recent years complex politics driving a shift away from ‘rights’ and more towards ‘security’ on certain issues
- Q: Can we use text-as-data approaches to see substantive and/or discursive changes in EU policy documents?

Data Collection

Potential Universe of Textual Material

- EU Press Corner (press releases, speeches, ‘statements’, policy papers)
- EurLex (EU legislative documents, legal documents, a range of non-legeslative documents)
- Academic datasets (EU speech corpus, EU ParlLawSpeech)

EurLex


An official website of the European Union How do you know?
EUR-Lex
Access to European Union law
English EN
My EUR-Lex
Experimental features
EUROPA > EUR-Lex home > EU law
QUICK SEARCH
SEARCH
Search tips
Need more search options? Use the Advanced search
LAW IN FOCUS
European semester: spring package
Pause
Official Journal
OJ L Series : 18/07/2024 (8 acts)
OJ C Series : 18/07/2024 (19 acts)
EU law
Treaties
Legal acts
Consolidated texts
International agreements
Preparatory documents
EFTA documents
Lawmaking procedures
Summaries of EU legislation
Browse by EU institutions
Browse by EuroVoc
EU case-law
Case-law
Reports of cases
Directory of case-law
National law and case-law
National transposition
National case-law
JURE case-law
Information
Themes in focus
EUR-Lex developments
Statistics
ELI register
EU budget online
Find results by document number
Year
Number
All document types
Find results by CELEX number
EUR-Lex
e-learning

Data Collection Strategy

Migration:

- EurLex R package, collect metadata by policy area, ‘sector tags’, and EurVoc policy tags (total n = 2877)
- Use metadata to pull full documents (EurLex package via EurLex API)

Digital:

- Used EurLex website’s advanced search function w. boolean query “child safety” AND “internet” to get metadata (n = 82)
- Eurlex package to pull text of full documents by CELEX number

Analysis Methods (1/3)

- Structural topic models
- Seeded/keyword assisted topic models
- Word embeddings
- Latent semantic scaling

Analysis Methods (2/3)

> although topic models can **explore themes of a corpus** (e.g., Roberts et al. 2014), **they do not necessarily measure specific concepts** of substantive interest. Although researchers have also relied upon topic models for measurement purposes (e.g., Bagozzi and Berliner 2018; Blaydes, Grimmer, and McQueen 2018; Barberá et al. 2019; Dietrich, Hayes, and O'brien 2019; Grimmer 2013; Martin and McCrain 2019), they acknowledge that these **fully automated models often inadvertently create multiple topics with similar content** and combine different themes into a single topic…
>
> (Eshima et al. 2024, p. 730)

Analysis Methods (3/3)

AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

Check for updates

Keyword-Assisted Topic Models

Shusei Eshima and **Kosuke Imai** Harvard University
Tomoya Sasaki Massachusetts Institute of Technology

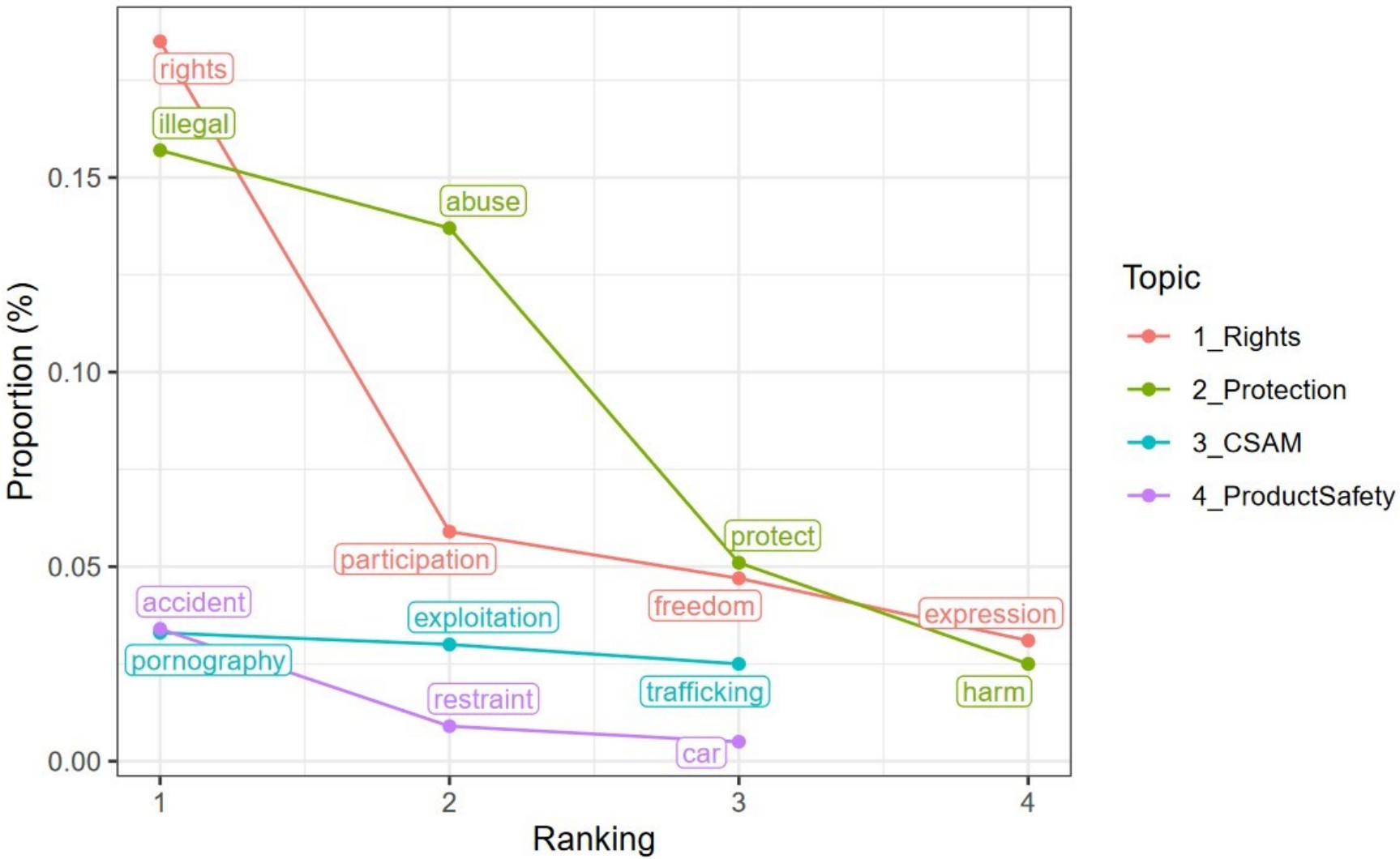
Abstract: *In recent years, fully automated content analysis based on probabilistic topic models has become popular among social scientists because of their scalability. However, researchers find that these models often fail to measure specific concepts of substantive interest by inadvertently creating multiple topics with similar content and combining distinct themes into a single topic. In this article, we empirically demonstrate that providing a small number of keywords can substantially enhance the measurement performance of topic models. An important advantage of the proposed keyword-assisted topic model (keyATM) is that the specification of keywords requires researchers to label topics prior to fitting a model to the data. This contrasts with a widespread practice of post hoc topic interpretation and adjustments that compromises the objectivity of empirical findings. In our application, we find that keyATM provides more interpretable results, has better document classification performance, and is less sensitive to the number of topics.*

Verification Materials: The data and materials required to verify the computational reproducibility of the results, procedures and analyses in this article are available on the *American Journal of Political Science* Dataverse within the Harvard Dataverse Network, at: https://doi.org/10.7910/DVN/RKNNVL.

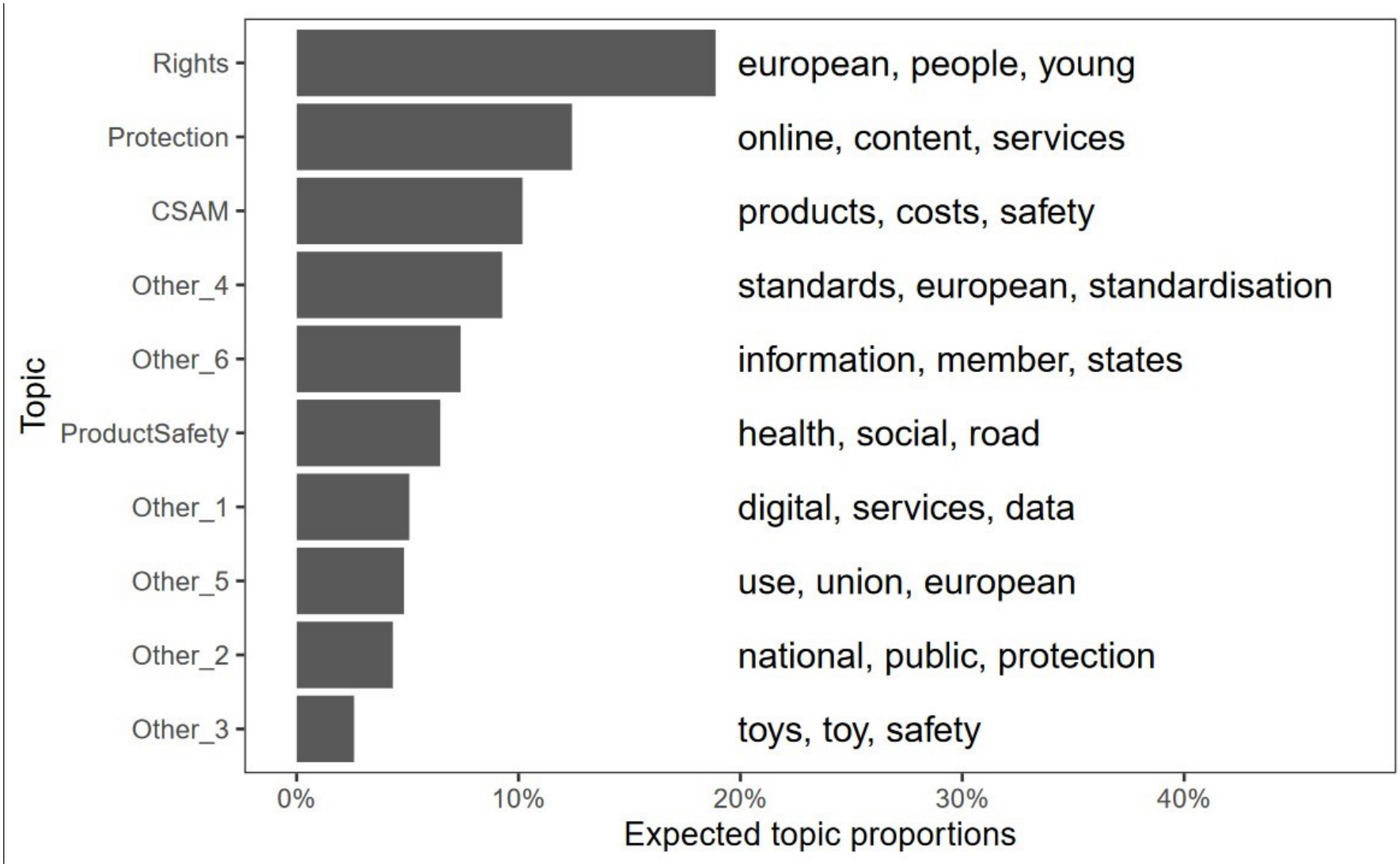
Results (Rob 1/3)

- start with selecting keywords (exploratory, based on domain knowledge) around 4 topics
  - Human Rights
  - Child Protection
  - Child Abuse (CSAM)
  - Product Safety

Results (Rob 2/3)

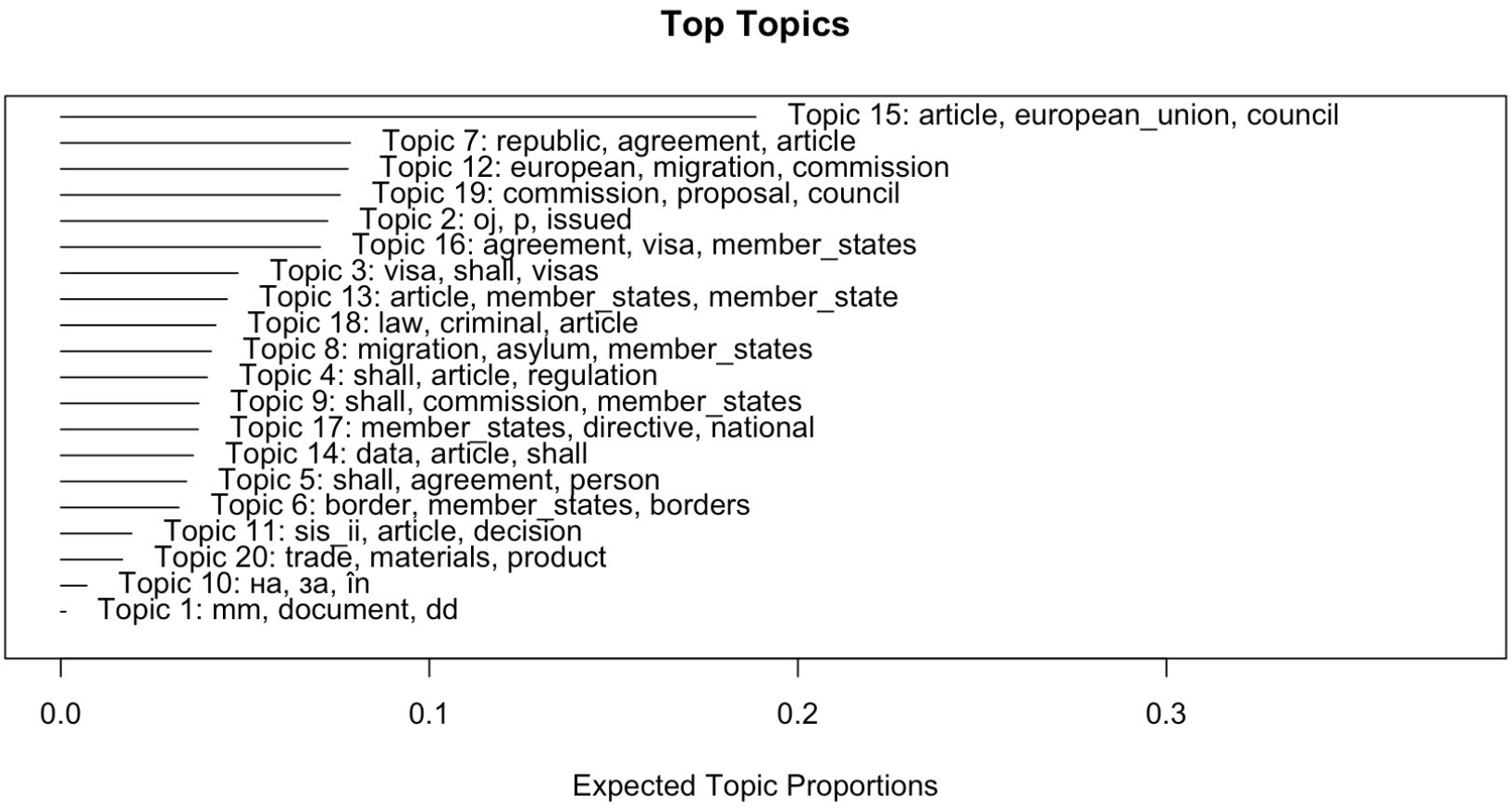


Results (Rob 3/3)



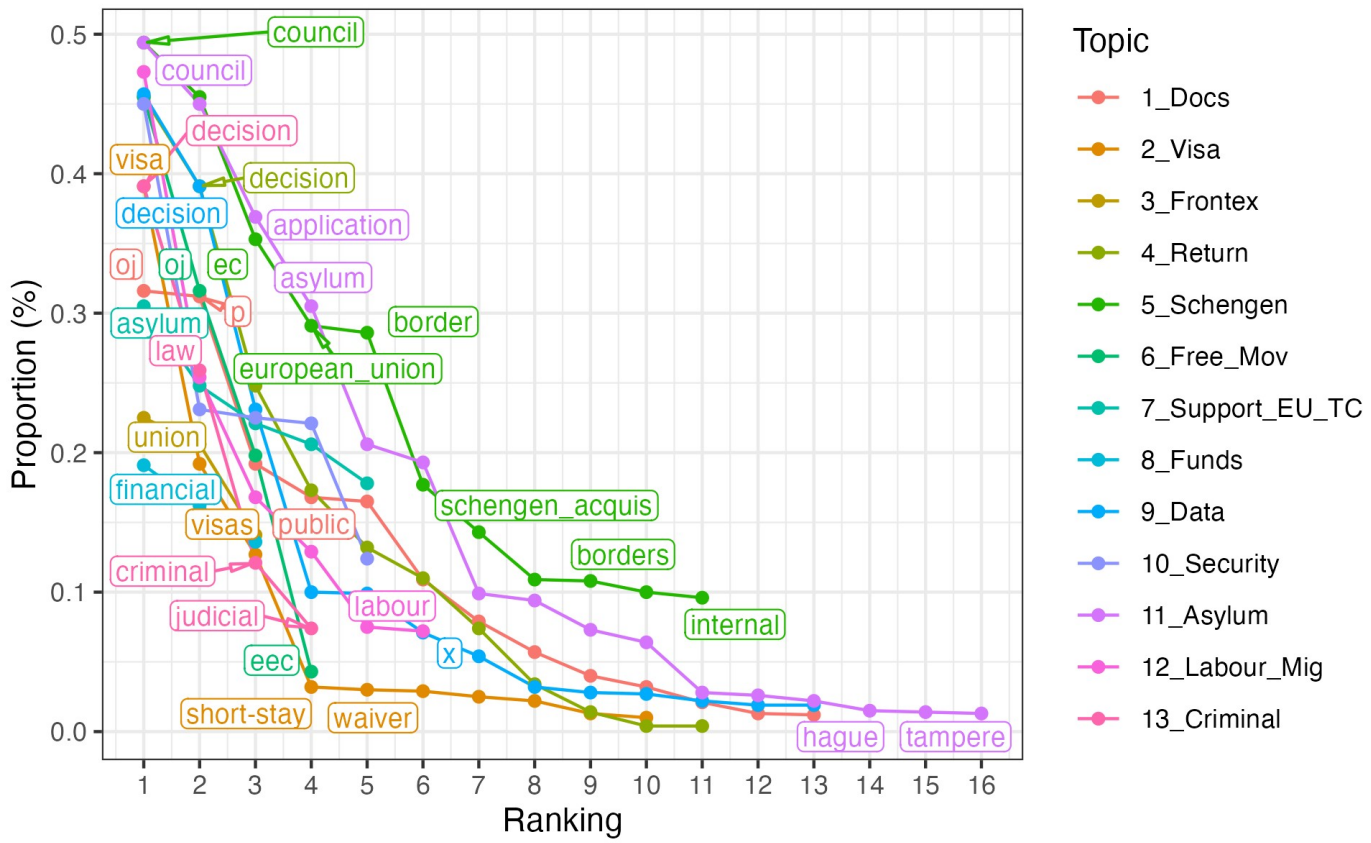
Results (Paula 1/10)

Baseline: stm with 20 topics



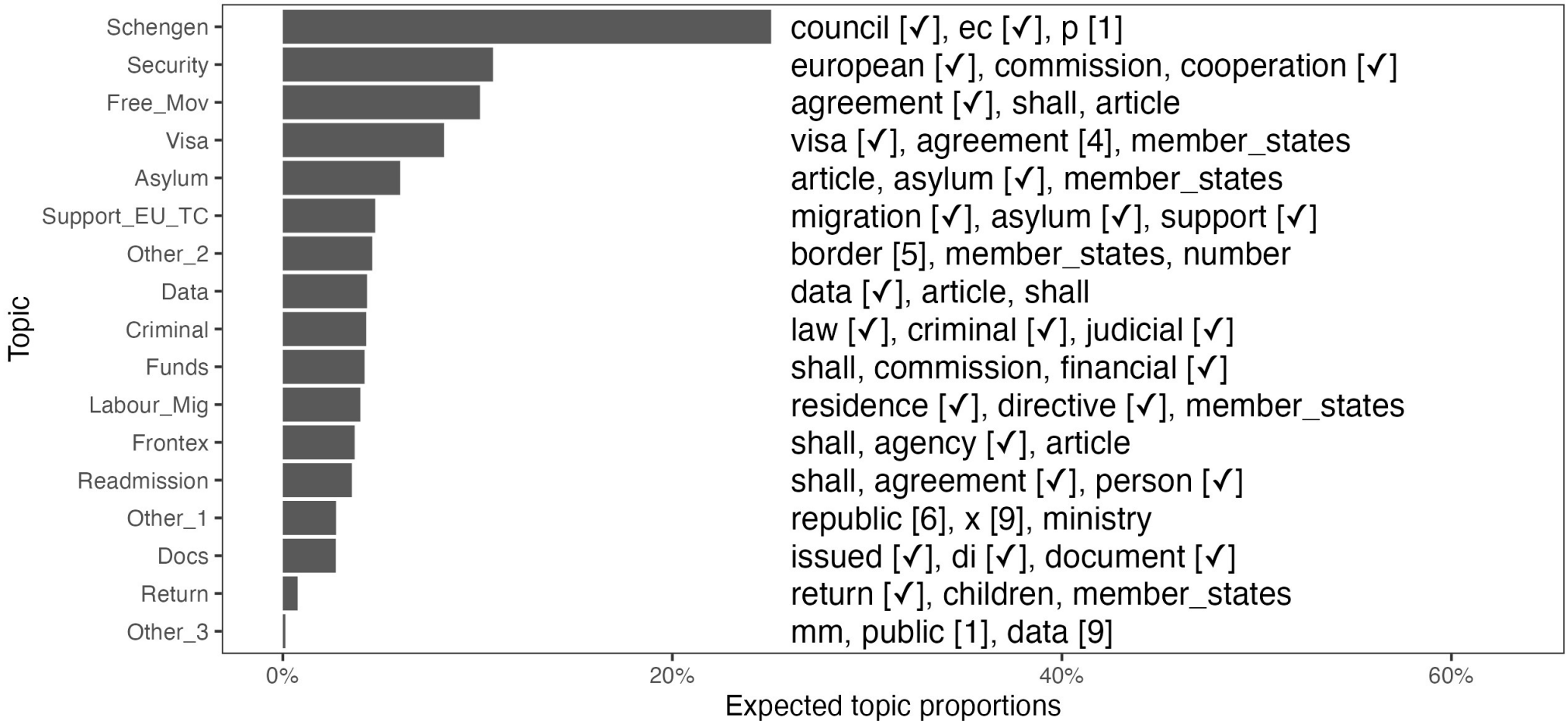
Results (Paula 2/10)

First keyword assisted model: selected keywords based on the stm (some more pre-processing after this)



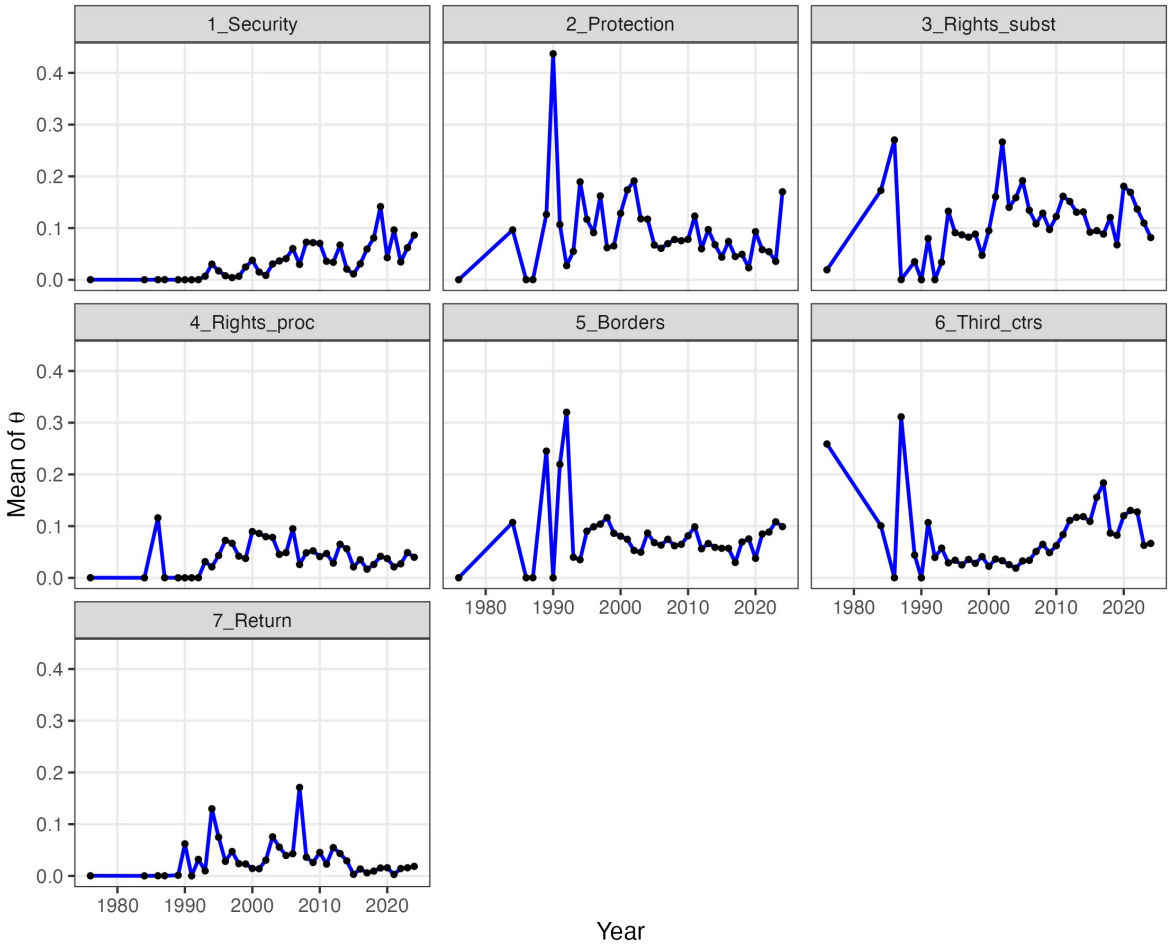
Results (Paula 3/10)

Topic proportions first model



Results (Paula 4/10)

Time trend first model



Results (Paula 5/10)

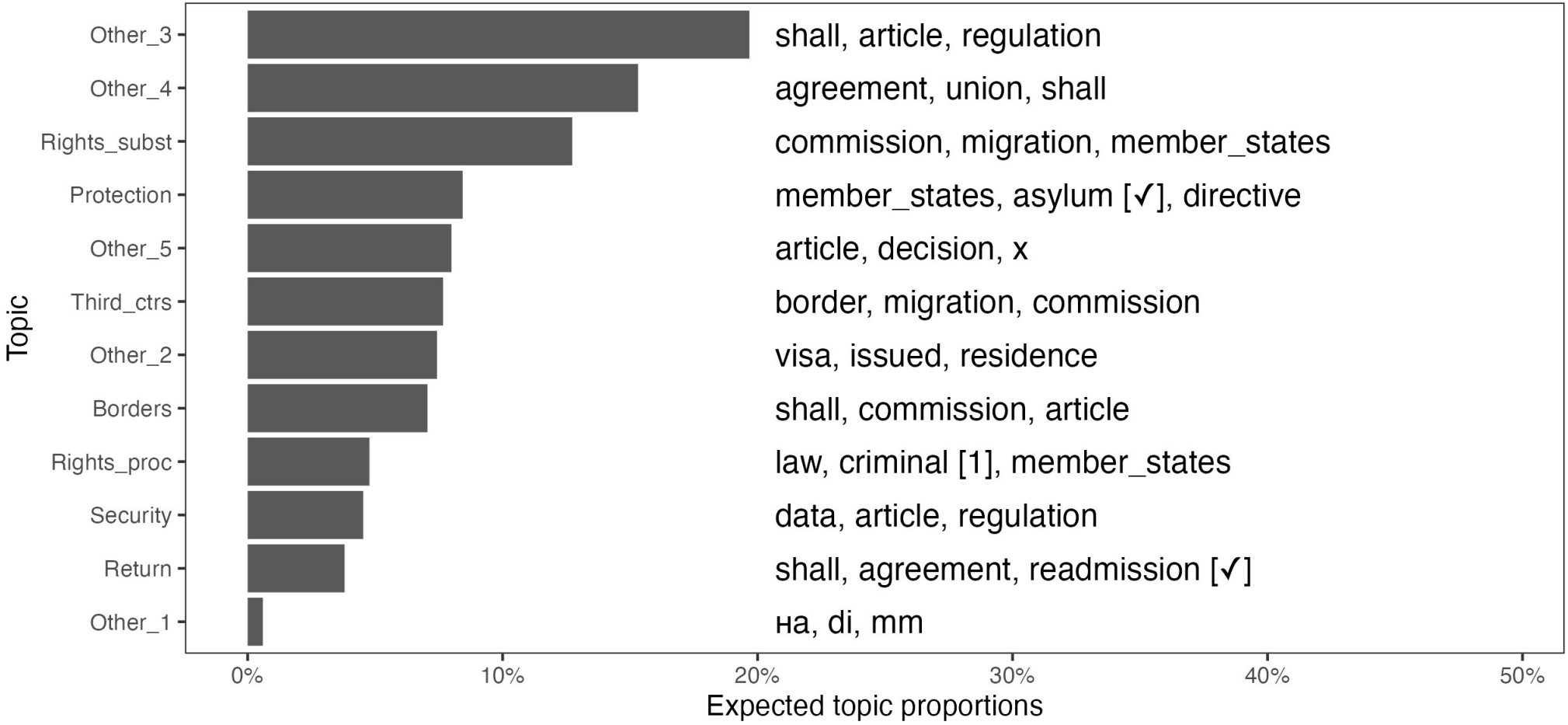
Second model: keywords selected based on research interest (rights, third countries, return)

```
keywords_free <- list(
  Security = c("security", "criminal"),#  Highest Prob T12
  Protection = c("application", "applicant",
                 "asylum","protection", "international_protection",
                 "minor",  "unaccompanied" ), #  change later to "u
  Rights_subst = c ("fundamental_rights", "human_rights"),
  Rights_proc = c( "judicial", "due", "appeal", "procedure"), # add
  Borders = c ("external_borders", "border_management", "frontex"),
  Third_ctrs = c ("third_countr", "third_countries", "third_country
  Return = c ("return", "voluntary_return", "removal", "departure",
              "return_decision", "detention",
              "return_directive",
              "readmission")) # return-related words selected based
```

Results (Paula 6/10)

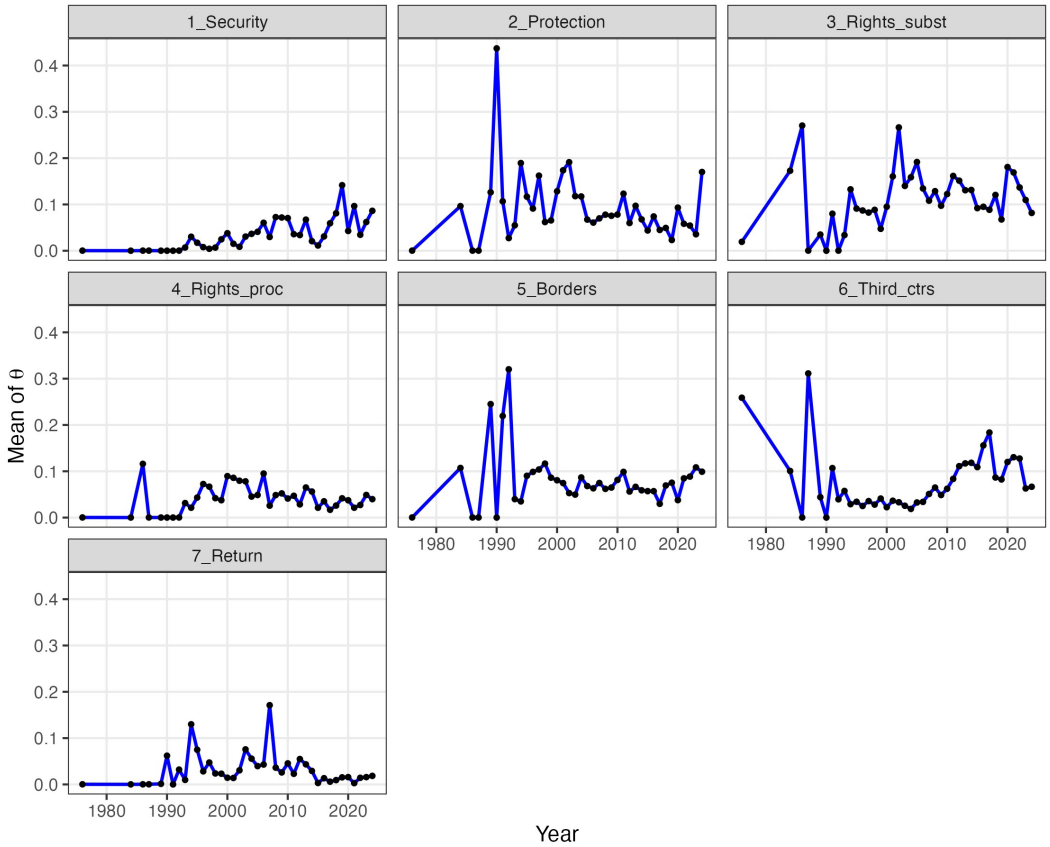
- Second model: keywords selected based on research interest (rights, third countries, return)
- subsumes “uninteresting” topics from baseline model into topics of interest
  - “Integration” keywords now part of “rights” topic

Results (Paula 7/10)



Results (Paula 8/10)

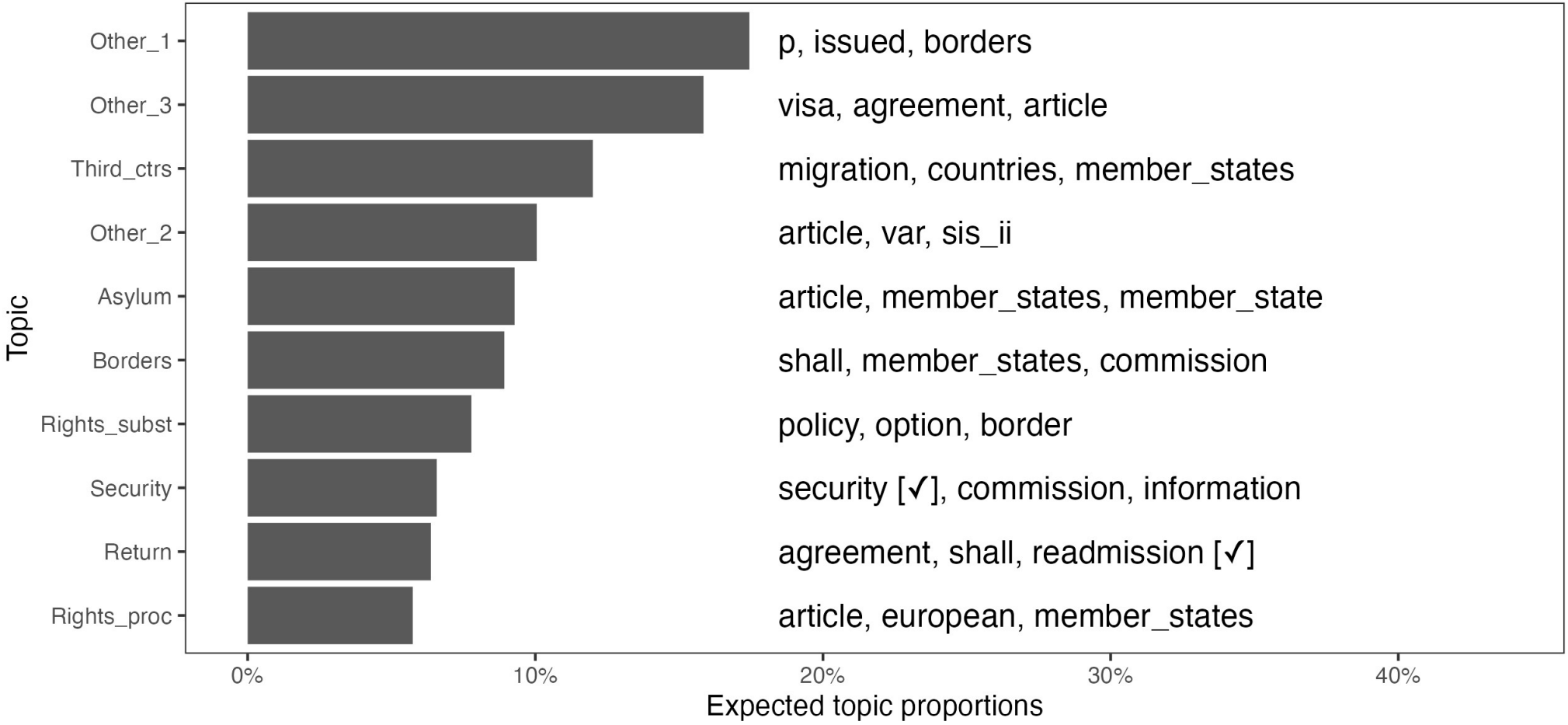
Second model: keywords selected based on research interest (rights, third countries, return)



Results (Paula 9/10)

- Third model: only “preparatory documents” (proposals, communications, white papers…) issued by the Commission; same topic selection as model 2
- Interesting: “right” topics less prevalent in Commission documents! But needs further validation (and keyword selection)

Results (Paula 10/10)



Reflections & Problems

Research Design

- Is EurLex the best source for the kinds of documents we want?
- Sampling? (diversity of documents in corpora, topic selection…)

Methods

- Difficulties in getting ‘fancier’ approaches to work (BERTopic…) - but are they better? Mixed methods?
- Keyword Selection, pre-processing questions…