CLARIN

ParlaCLARIN Workshop: Creating and Using Parliamentary Corpora

Miyazaki, 7 May 2018

A Corpus of Grand National Assembly of Turkish Parliament

Onur Güngör
Boğaziçi University, Istanbul, Turkey
Huawei R&D Center, Istanbul, Turkey

onurgu@boun.edu.tr
onur.gungor@huawei.com

Mert Tiftikçi
Boğaziçi University, Istanbul, Turkey

mert.tiftikci@boun.edu.tr

Çağıl Sönmez
Boğaziçi University, Istanbul, Turkey
cagil.ulusahin@boun.edu.tr

Outline

- Motivation
- Building
  - Crawling
  - Processing
- Analysis
- Conclusions

Motivation

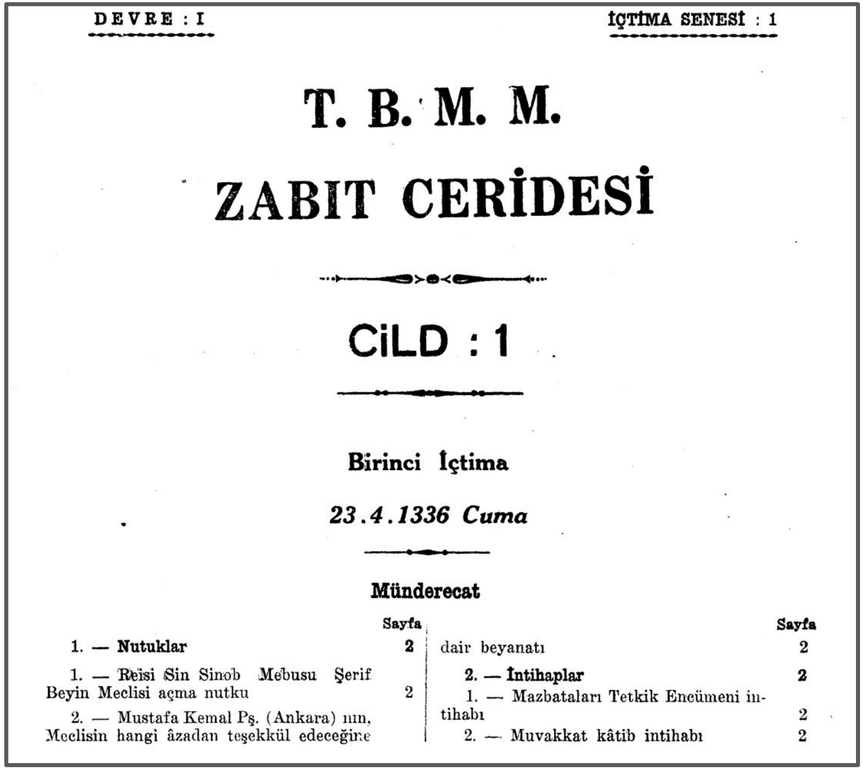
- Parliaments
  - discussions
- Decisions
  - people
  - the country
  - and the world

Motivation

- Transcriptions of these discussions are important
  - Political perspective
    - Historians
    - Political scientists
  - Language perspective
    - Linguists
    - Computational linguists

Problem

- Parliaments provide these transcriptions as
    - HTML
    - PDF
    - or through online search interfaces
- Problematic to access for scientific analysis



DEVRE : I

İÇTİMA SENESİ : 1

T. B. M. M.

ZABIT CERİDESİ

CİLD : 1

Birinci İçtima

23.4.1336 Cuma

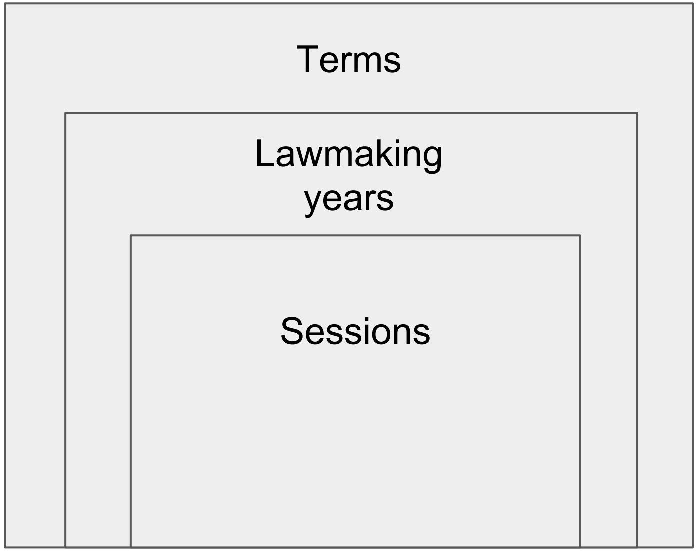
Münderecat

Our contribution

- Compiling the transcriptions of Turkish Grand National Assembly sessions
  - From 1920
  - Until 2015
- Providing an offline system for
  - Querying
  - Analyzing

The parliament

- Founded on 23 April 1920
- The members: ~400-550 elected every **term**
  - Five years, between 1920-2007
  - Four years, after 2007
- Each **lawmaking year** has **sessions**
- Sessions are transcribed by clerks
- Published in the official "Journal of Minutes of Meeting (JMM)"[1]



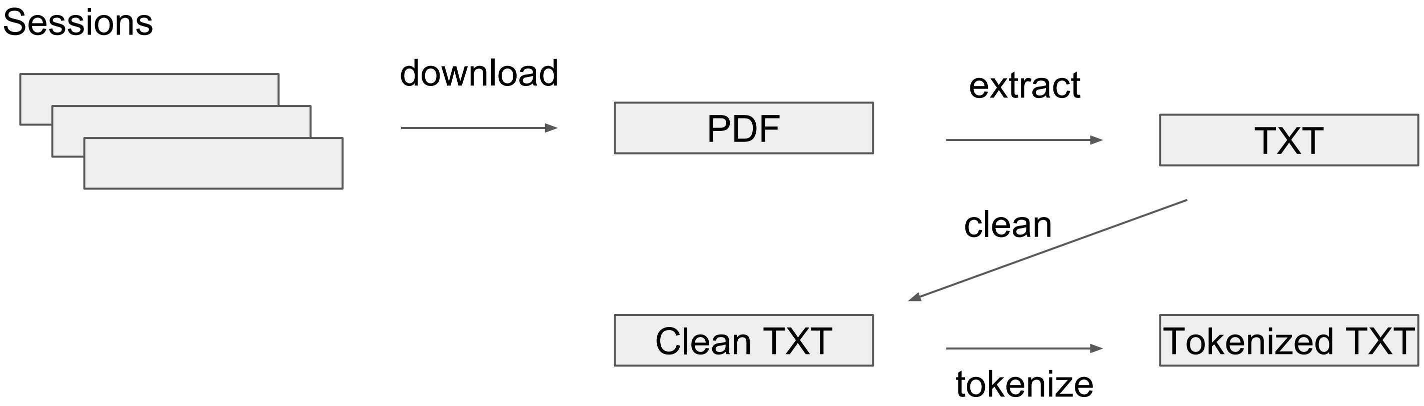
Terms

Lawmaking years

Sessions

[1] "Tutanak Dergisi" in Turkish.

Building the corpus

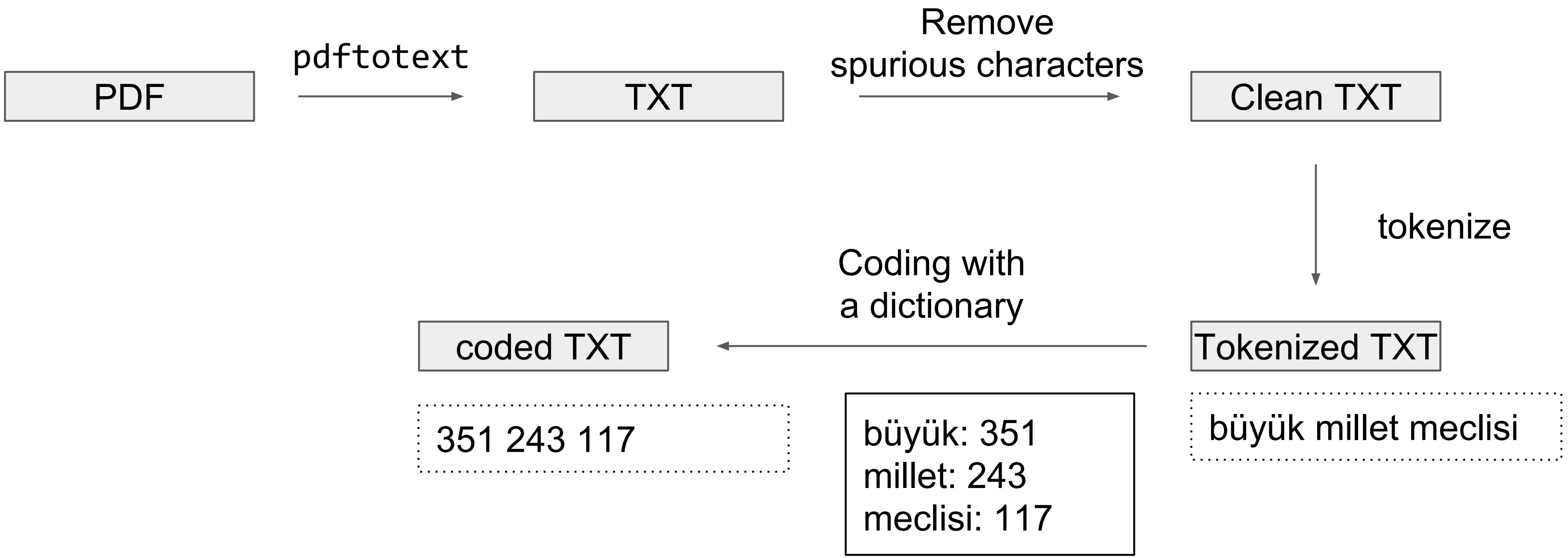


Crawling

- Session transcriptions are provided as scanned images of `JMM` in PDF format
  - Manually obtained the URL addresses of these
  - Then downloaded the files with a script

Processing (1)

- `pdftotext` was used to extract the text from the scanned images



Processing (2)

- We also record the day and the year of the session
- We do not explicitly mark
  - any person names, including the speakers
  - the political party of any member mentioned
  - the subject of the session

Analysis

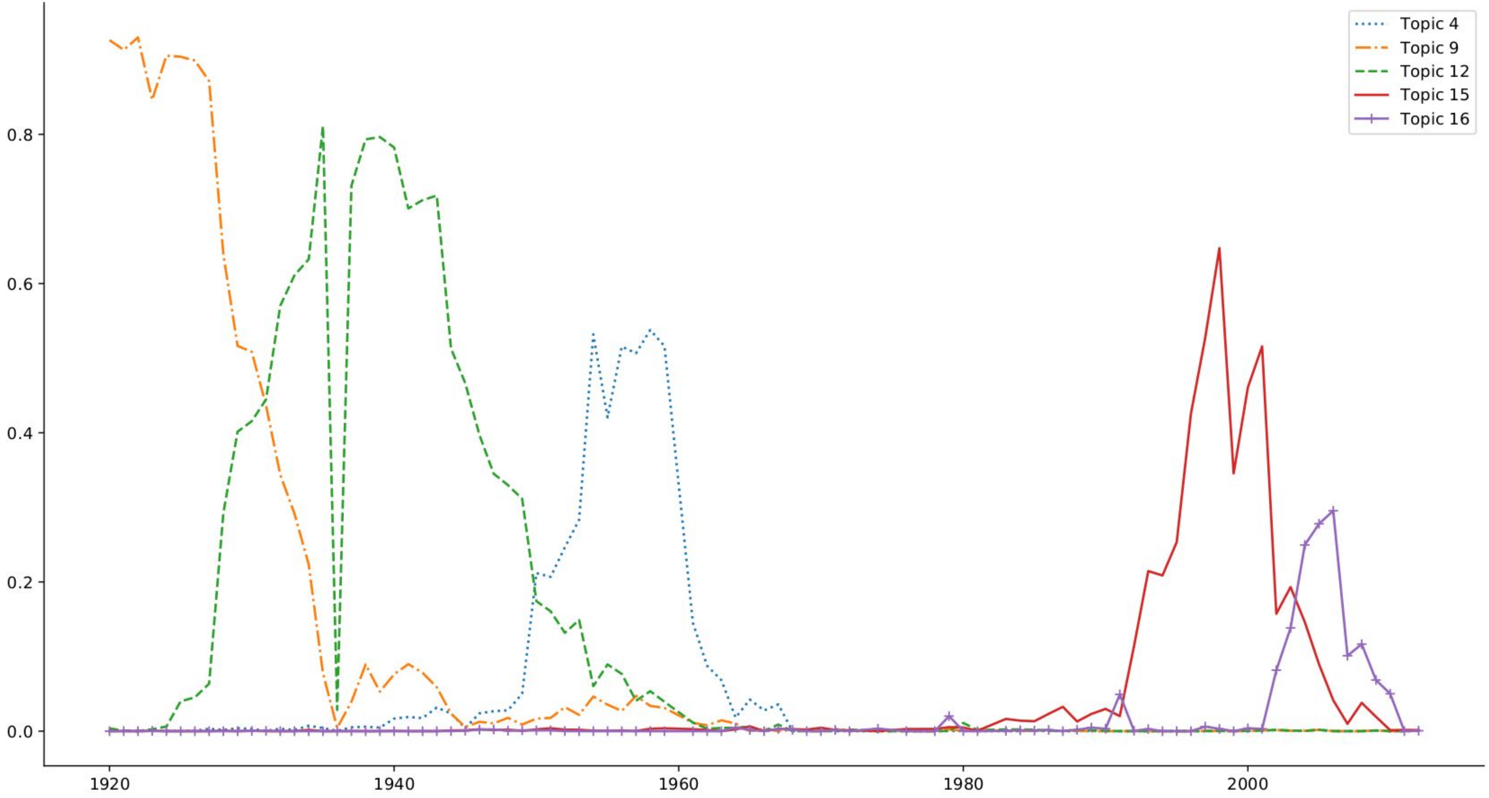
- 12645 sessions
- 208 million words
- 619 thousand unique words
- Testing the coverage
  - If we discarded unique words appearing less than 10 times
    - 318 thousand unique words
  - Remove the inflectional suffixes
  - Check if this remaining portion is present in a Turkish dictionary
  - 70% are found

Analysis: Topic distributions (1)

- latent Dirichlet allocation (LDA)
- 20 topics

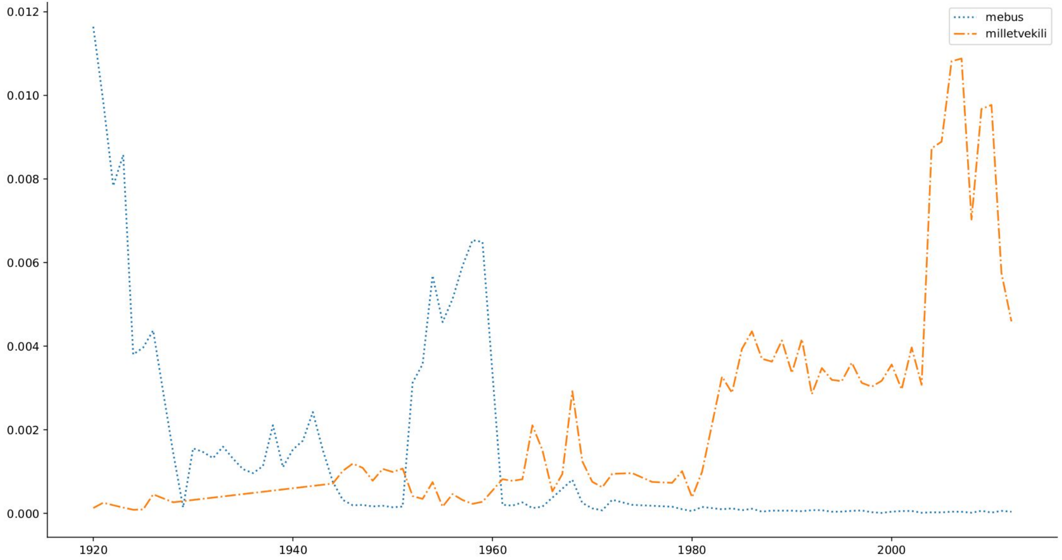
older Turkish	Topic 4	mebus	umumi	lâhiyası	muvakkat
	Topic 9	reis	mazbata	dahiliye	rey
	Topic 12	mucibince	tâbi	müdafaa	tahsisat
current Turkish	Topic 15	başbakan	milletvekili	geliş	ilgi
	Topic 16	kovuşturma	dokunulmazlık	rapor	karma

Analysis: Topic distributions (2)



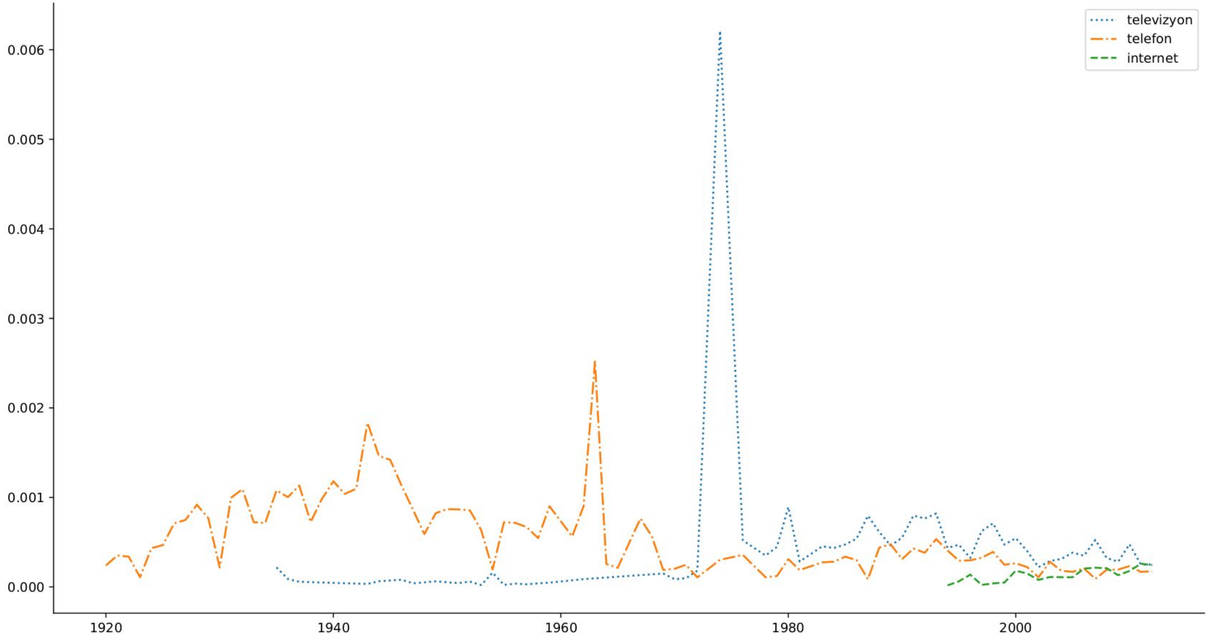
Analysis: Word usage frequencies
- ‘mebus’ vs. ‘milletvekili’

- Mebus
  - ‘Member of the parliament’
  - Originally Arabic
- Current Turkish speakers do not use
  - Frequent in the beginning of the 20th century



Analysis: Word usage frequencies
- Technological terms

- 'Televizyon' → TV
- 'Telefon' → landline
- 'internet'



Access

- the resulting file is 1.2 GB
- we share the corpus publicly[1]
- along with the offline interactive interface

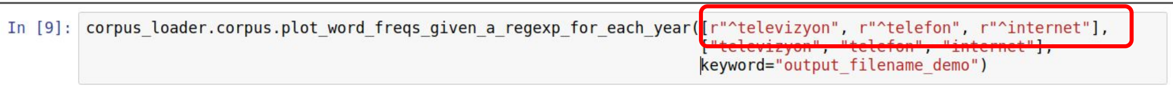
[1] https://github.com/onurgu/turkish-parliament-texts

Plotting word frequencies

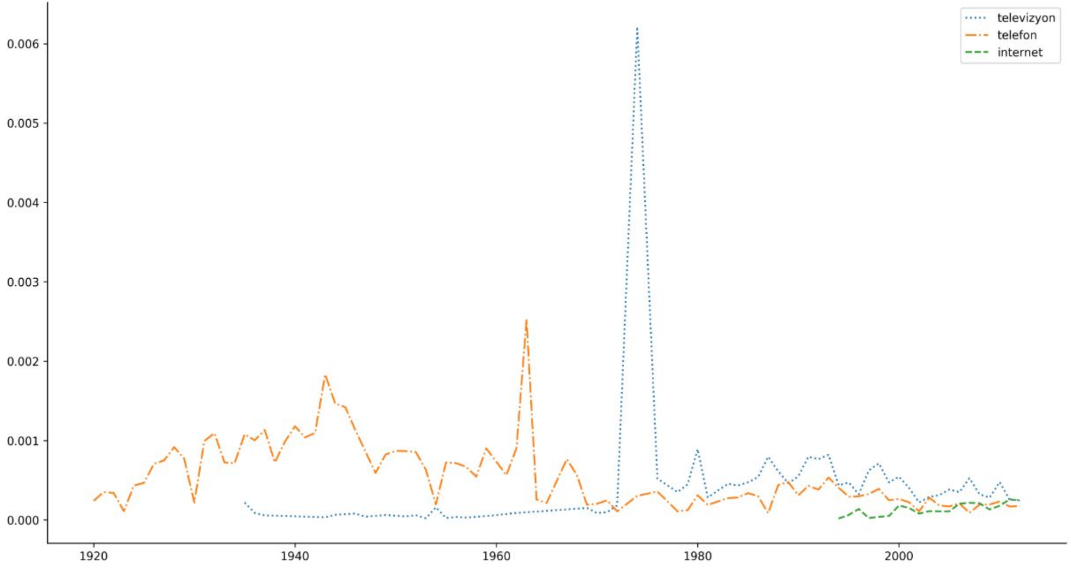
(Re)construct the corpus

```
python construct_vocab.py --command construct_corpus --corpus_filename
corpus-v0.3/tbmm_corpus --max_documents 0
```

Plot



```
In [9]: corpus_loader.corpus.plot_word_freqs_given_a_regexp_for_each_year([r"^televizyon", r"^telefon", r"^internet"],
                                                                  ~~["televizyon", "telefon", "internet"],~~
                                                                  keyword="output_filename_demo")
```



Conclusions

- Turkish parliament’s session transcriptions
  - 95 years, between 1920 and 2015
- Suitable environment for statistical research
- Interface for researchers without extensive programming skills

Future work

- Extraction of
  - The structure of the dialogue
  - Named entities: locations, person names, political parties
  - The sessions between 2015 and today
- Removing spelling errors introduced during digitization

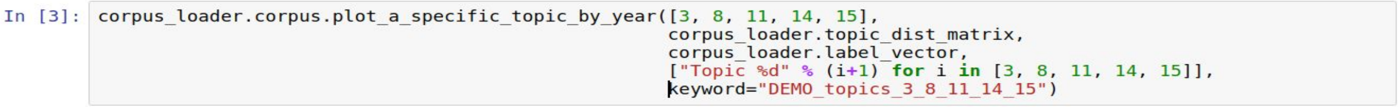
Thanks

Plotting LDA topics

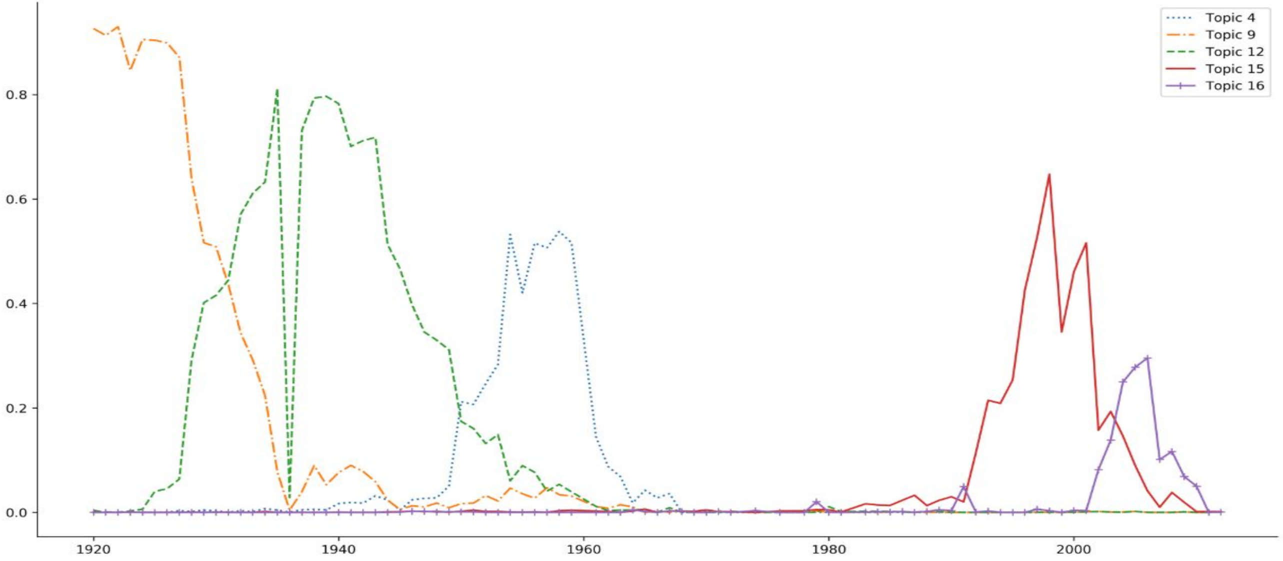
(Re)allocate topics

```
python construct_vocab.py --command construct_corpus --corpus_filename
corpus-v0.3/tbmm_corpus --max_documents 0 --train_lda
```

Plot



```
In [3]: corpus_loader.corpus.plot_a_specific_topic_by_year([3, 8, 11, 14, 15],
                                                           corpus_loader.topic_dist_matrix,
                                                           corpus_loader.label_vector,
                                                           ["Topic %d" % (i+1) for i in [3, 8, 11, 14, 15]],
                                                           keyword="DEMO_topics_3_8_11_14_15")
```



Topic 4
Topic 9
Topic 12
Topic 15
Topic 16