MigParl (v1.0.1-rc)

Corpus Documentation

Andreas Blaette (andreas.blaette@uni-due.de), Christoph Leonhardt (christoph.leonhardt@uni-due.de)

2020-01-27

Abstract

This paper introduces the MigParl corpus. We outline available data, the data preparation process for preparing corpora of parliamentary debates as well as the sampling strategy used to obtain a thematically coherent corpus of debates concerned with migration and integration.

A Corpus of Plenary Protocols

MigParl is a corpus of plenary debates in the German regional states (Bundesländer) which are concerned with topics of migration and integration. It follows in the footsteps of GermaParl corpus, a corpus of plenary protocols of the German Bundestag, which, just as MigParl, was developed in the PolMine Project. As such, it shares GermaParl's motivation and general purpose (see Blätte and Blessing 2018: 810). As the use cases and requirements of a corpus of plenary protocols is described in depth in Blätte and Blessing (2018), the main focus of this paper is the description of the specificities of MigParl itself.

The MigParl corpus is a processed thematic subset of the plenary protocols published by the German regional states from (mostly) January 2000 to December 2018. While 15 of the 16 regional states do provide data for roughly this time period, the Saarland does not provide protocols for the time period before September 2004. As a thematically specialized corpus, it does not contain all debates, but only those speeches which are relevant for migration and integration research. See section "Sampling" for the sampling strategy used.

MigParl is made available as a R data package called "MigParl". The package includes the functionality to download a linguistically annotated, indexed and consolidated version of the corpus that has been imported into the Corpus Workbench (CWB).

Along with the package, a small sample corpus ("MigParlMini") is provided. MigParlMini is annotated and indexed the same way as MigParl, however only contains about 1% of the data the full corpus does provide.

The R data package is designed to work smoothly with the analytical tools offered by the R package `polmineR`.

The corpus is versioned by means of a build date in the corpus registry file.

Most of the data was prepared from pdf documents issued by the state parliaments. These pdf documents have then been turned into plain text files with `trickypdf`, an R package which has been developed within the PolMine project and is freely available on GitHub.\footnote{https://github.com/PolMine/trickypdf) The corpus preparation workflow is described below.

Corpus Preparation

The preparation of the TEI version of MigParl implements the following workflow:

- **Preprocessing**: Prepare consolidated UTF-8 plain text documents (ensuring uniformity of encodings, conversion of pdf to txt if necessary);
- **XMLification**: Turn the plain text documents into TEI format: Extraction of metadata, annotation of speakers etc.;
- **Consolidation**: Consolidating speaker names and enriching documents.

As this section hugely overlaps with the preparation of GermaParl, we refer to Blätte and Blessing (2018) again. The challenges remain the same, as does the motivation and aspiration to create a robust and sustainable framework for corpus preparation. Nevertheless, some updates are necessary.

The R package `trickypdf` is, as mentioned above, still in use to turn pdf documents, which almost exclusively occur in a two-column layout, into plain text files. As an exception, Hesse did provide Microsoft Word instead of pdf documents for the time until 2002. These documents where converted into plain text as well.

The conversion of plain text files in structurally annotated XML documents is provided by the `framework for parsing plenary protocols` or `frappp`, which has been developed in the PolMine project. `frappp` is used to facilitate a workflow in which regular expressions are formulated for a number of items which should be identified, in particular metadata, speakers, interjections and agenda items. As with the previous approach using the corpus toolkit (`ctk`), false positives and false negatives are handled by both a list of known mismatches and preprocessing steps which cleans up some faulty input data.

Particular care has to be taken to ensure that identified speakers are consistently named. This is done with reference to external data sources of names which are complemented with a list of known aliases. The primary external data source used for the MigParl data is Wikipedia (Blätte and Blessing 2018: 813).

One remaining challenge is the time-dependency of these external checks. Sometimes the information in the protocol and the information in the external data might differ. A member of parliament might change parties within a legislative period. In this case, our approach will label this member's party affiliation with the one found in the external data source. The same is true with names. If the protocol says "Ulla Schmidt" and the external data source "Ursula Schmidt", then the external information is used.

This external data is stored in a Git repository.

Sampling Strategy

MigParl is a thematic subset of plenary protocols relevant for migration and integration research. As such, a sampling strategy was necessary to determine this relevance from a base population of all protocols of the German regional state. We follow a two-pronged sampling approach.

Topic Model-Based Sampling

- *topic modelling*: a topic model for each of the regional state parliament corpora was calculated and the 100 most relevant terms per topic retrieved
- *dictionary approach*: A number of core terms conveying the concept of migration and integration were theoretically derived by literature review
- *topic identification*: the dictionary and the 100 most relevant terms per topic were matched against each other. A topic with more than five hits from the dictionary was deemed relevant for migration and integration
- *document identification*: for each speech, the probability to belong to one of the identified topics was calculated. If the sum probability exceeded a threshold, the speech was considered relevant
- these relevant speeches were included in the `MigParl` corpus

Dictionary-Based Sampling

- the MigPress dictionary was used to identify speeches in which at least five instances of the dictionary terms occur.
- this differs from the threshold of one occurrence used in the creation of the `MigPress` corpus. This is due to difference in language use in parliamentary settings and the greater average length of speeches compared to newspaper articles.
    - these relevant speeches were included in the `MigParl` corpus as well

Annotation

Linguistic Annotation

The XML/TEI version of the initial plenary protocols is taken through a pipeline of standard Natural Language Processing (NLP) tasks. Stanford CoreNLP is used for tokenization, part-of-speech (POS) and named-entity (NE) annotation. To add lemmata to the corpus, TreeTagger is used.

This linguistic annotation is part of the corpus as so-called positional attributes (p-attributes). The following table provides short explanations of the p-attributes in the MigParl corpus.

p-attribute	description	values
word	the word as it occrs in speech	word

p-attribute	description	values
pos	the part-of-speech-tag of the word according to the Stuttgart-Tübingen Tagset	for example ADJA, NN, VVINF
lemma	the lemmatized form of the word	lemma
ner	Named Entities	O, ORGANIZATION, PERSON, LOCATION, MISC

In the so-called token stream the linguistic annotation looks like this:

word	pos	lemma	ner
Frau	NN	Frau	O
Präsidentin	NE	Präsidentin	O
,	$,	,	O
meine	PPOSAT	mein	O
Damen	NN	Dame	O
und	KON	und	O
Herren	NN	Herr	O
!	$.	!	O

Structural Annotation (Metadata)

In the XML/TEI data format, all passages of uninterrupted speech are tagged with metadata, or so-called structural attributes (s-attributes). For instance, parliamentary speeches are often interrupted by interjections - the information whether an utterance is an interjection or an actual speech is maintained in the corpus. The legislative period, session, date, name of a speaker and his/her party are included, among others. The structural annotation is the basis for all kinds of diachronic or synchronic comparisons users may want to perform.

The following table provides short explanations of the s-attributes which are present in the MigParl corpus.

s-attribute	description	values
lp	legislative period	3 to 21 (dependend on regional state)
session	session/protocol number	1 to 161
agenda_item	agenda item	number of the agenda item
agenda_item_type	type of agenda item	debate/question_time/government_declaration/.
date	date of the session	YYYY-MM-TT (e.g. ‘2013-06-28’)
calendar_week	calendar week derived from date accoring to ISO 8601	YYYY-Week (e.g. 2001-01
year	year of the session	2000 to 2018
interjection	whether contribution is interjection	TRUE/FALSE
role	role of the speaker	presidency/mp/government
speaker	Name	speaker name
party	Party	party of the speaker
regional_state	regional state the debate is held	regional state abbreviation
speech	individual speech within a debate	combination of speaker name, date and number of utterance
migration_integration		

s-attribute	description	values
_probability	sum probability for migration and integration relevance	numeric between 0 and 1
url	the url of the source file	url
src	the type of the source file	pdf or doc
source_dict	whether speech is sampled by dictionary approach	TRUE/FALSE
source_topic_model	whether speech is sampled by topic modelling approach	TRUE/FALSE

Using the MigParl corpus

Getting started - installing MigParl

The MigParl data package that includes the CWB indexed version of the corpus is hosted at a private CRAN-style package repository on the Web-Server of the PolMine Project. The polmineR package offers a convenient installation mechanism.

```
library(polmineR)
if ("drat" %in% rownames(available.packages()) == FALSE) install.packages("drat")
drat::addRepo("polmine") # lowercase necessary in this case
if ("MigParl" %in% rownames(available.packages()) == FALSE){
  install.packages("MigParl")
}
```

After installing the MigParl package, the package only includes a small subset of the MigParl corpus. The subset serves as sample data and for running package tests. To download the full corpus, use a function to download the full corpus from an external webspace (provided by the University of Duisburg-Essen, for the time being):

```
library(MigParl)
migparl_download_corpus()
```

To check whether the installation has been successful, run the following commands. For further instructions, see the documentation of the polmineR package.

```
library(polmineR)
use("MigParl") # to activate the corpus in the data package
```

```
## ... activating corpus: MIGPARL
```

```
## ... activating corpus: MIGPARLMINI
```

```
corpus() # to see whether the MIGPARL corpus is listed
```

```
##           corpus     size template
## 1 GERMAPARLMINI   222201     TRUE
## 2       MIGPARL 51470707     TRUE
## 3   MIGPARLMINI   268552     TRUE
## 4       REUTERS     4050     TRUE
```

```
if ("MIGPARL" %in% corpus()[["corpus"]]) size("MIGPARL") # corpus size
```

```
## [1] 51470707
```

A very brief tutorial

The CWB indexed version of MigParl can be used with the CWB itself, or with any tool that uses the CWB as a back end (such as CQPweb). However, most technical decisions during corpus preparation had in mind to optimise using the MigParl corpus in combination with the polmineR package. Please consult the documentation of the polmineR package (README, vignette, manual) to learn how to use polmineR for working with MigParl. Here, we can only offer a very brief tutorial for basic commands. Note that in the following examples, we will use MIGPARLMINI to illustrate functions, as the full MIGPARL corpus will only be available after the full installation of the corpus.

First, you may want to learn about the s-attributes (structural attributes), and the p-attributes (positional attributes) that are available.

```
s_attributes("MIGPARLMINI")
```

```
##  [1] "id"                                "speaker"
##  [3] "party"                             "role"
##  [5] "lp"                                "session"
##  [7] "date"                              "src"
##  [9] "url"                               "regional_state"
## [11] "interjection"                      "year"
## [13] "agenda_item"                       "agenda_item_type"
## [15] "speech"                            "migration_integration_probability"
```

```
p_attributes("MIGPARLMINI")
```

```
## [1] "word"  "pos"   "lemma" "ner"
```

To learn about the values of s-attributes, specify the param s_attribute:

```
s_attributes("MIGPARLMINI", "date")
```

```
##  [1] "2002-04-17" "2007-01-25" "2009-10-21" "2013-06-06" "2004-03-18"
##  [6] "2007-02-22" "2013-10-09" "2004-07-21" "2005-07-20" "2013-06-04"
## [11] "2016-02-02" "2017-12-07" "2000-01-25" "2002-02-26" "2005-11-23"
## [16] "2012-06-28" "2013-06-26" "2016-07-12" "2002-08-30" "2016-09-30"
```

```
s_attributes("MIGPARLMINI", "party")
```

```
##  [1] "CDU"            "PDS"            "DVU"            "LINKE"
##  [5] "SPD"           "FDP"           "GRUENE"        "CSU"
##  [9] "FREIE WAEHLER" "NA"            "GAL"           "NPD"
## [13] "parteilos"     "AfD"
```

To inspect keywords-in-context (KWIC), use the kwic-method:

```
K <- kwic("MIGPARLMINI", query = "Integration", left = 3, right = 3)
if (interactive()){
  K
} else {
  knitr::kable(K@stat[1:10,], format = "pandoc")
}
```

match_id	left	node	right
1	verdeutlichen , dass	Integration	nicht nur eine
2	als Mittel der	Integration	– diesen wertevermittelnden
3	CDU-Fraktion “ Schulische	Integration	und Förderung von
4	Förderung der schulischen	Integration	junger Ausländer und
5	und Herren ,	Integration	kann nur gelingen
6	Beitrag zur schulischen	Integration	leistet . (
7	dass die soziale	Integration	nicht nach der
8	zum Prinzip der	Integration	und auch zu
9	den Prozess der	Integration	einzutreten . Das
10	Voraussetzungen für eine	Integration	weithin eher verschlechtert

The count-method is used for counting. You can supply one or multiple queries:

```
count("MIGPARLMINI", query = "Integration")
```

```
##          query count         freq
## 1: Integration   179 0.0006665376
```

```
count("MIGPARLMINI", query = c("Integration", "Flucht", "Abschiebung"))
```

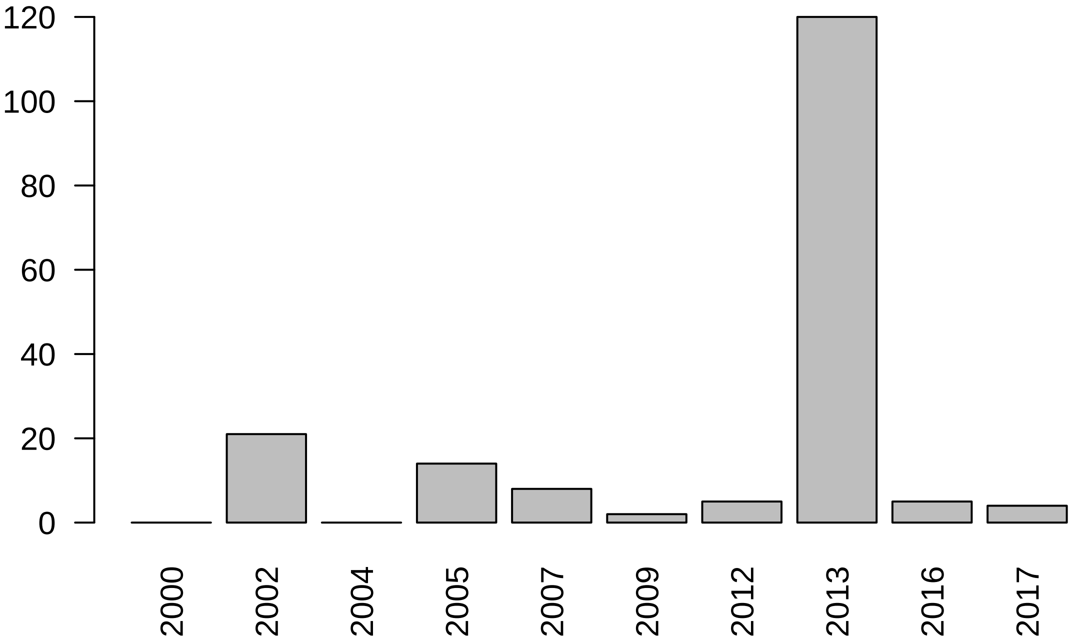
```
##          query count         freq
## 1: Integration   179 6.665376e-04
## 2:      Flucht     4 1.489469e-05
## 3: Abschiebung    17 6.330245e-05
```

To get the dispersion of a query, use the dispersion-method.

```
D <- dispersion("MIGPARLMINI", query = "Integration", s_attribute = "year")
```

Visualise the result as a barplot...

```
barplot(D[["count"]], names.arg = D[["year"]], las = 2)
```



The cooccurrences-method will get you words which do occur more frequently together with the query term than statistically expected.

```
C <- cooccurrences("MIGPARLMINI", query = "Wir")
C@stat[1:5]
dotplot(C)
```

These are some of the core functions, applied to the whole corpus. The whole point of the structural annotation of the corpus (s-attributes) is to facilitate the creation of subcorpora / partitions. So every method that has been introduced can be applied to a partition.

```
year2016 <- partition("MIGPARL", year = 2016)
count(year2016, query = c("Asyl", "Flucht", "Abschiebung"))
dispersion(year2016, query = "Flüchtlinge", s_attribute = "regional_state")
```

Finally, note that the methods of polmineR can also be used with the pipe functionality offered by the magrittr package.

```
cooccurrences("MIGPARLMINI", query = "Europa") %>%
  subset(!word %in% c(tm::stopwords("de"), ",", ".")) %>%
  subset(count_coi >= 3) %>%
  dotplot()
```

This is just a short glimpse into the analytical opportunities of using MigParl in combination with polmineR. One of the most important aspects that cannot be explained here is the possibility to use the syntax of the Corpus Query Processor (CQP) that comes with the CWB back-end. The as.TermDocumentMatrix method will prepare data structures efficiently needed for more advanced analytical techniques such as topic modelling. Consult the vignette of the polmineR package to learn more!

Some caveats

A set of general remarks may help to avoid pitfalls when working with MigParl:

- Plenary protocols meticulously report interjections. To maintain the integrity of the original documents, interjections are annotated in the corpus. By using the s-attribute 'interjection' that assumes the values 'TRUE' or 'FALSE', you can limit your analysis to speech or interjections.
- In contrast to GermaParl, a distinction between party affiliation and parliamentary group is not included in the current version of MigParl. This has mainly the practical reason that there are not too many differences between party and parliamentary group on regional level as there is no "CDU/CSU" parliamentary group as in the German Bundestag.
- For users working with previous versions of `MigParl` (> 1.0.1-RC) it might be necessary to subset the corpus so that it only comprises speeches which are sampled with the topic modelling approach (source_dict == FALSE)

Conclusion

MigParl, a thematic subset of debates of the German regional states, is available as a linguistically annotated and indexed version in form of an R data package. The data collection, spanning 18 years of German parliamentary debates, is complete and an extension is not planned immediately. As with GermaParl, however, "[t]he data is intended to be open, versioned, reproducible, accessible and sustainable, with a focus on successively improving data quality" (Blätte and Blessing 2018: 816). As MigParl followed to footstep of previous endeavours such as GermaParl, by making the workflows, resources and methodological considerations publicly available along with the data, we hope that the data can contribute to the furthering of a public digital archive of democracy.

Annex

Corpus data (by year)

year	size
2000	2353640
2001	1963334
2002	1926908
2003	1321108
2004	1897378
2005	1804022
2006	2145901
2007	1799229
2008	1914093

year	size
2009	1707252
2010	2118794
2011	2338515
2012	2270699
2013	2401247
2014	3203446
2015	6325062
2016	5221455
2017	4410779
2018	4347845

Corpus data (by regional state)

regional_state	size
BB	1909575
BE	3109974
BW	3399109
BY	4091021
HB	3073804
HE	5077787
HH	2601390
MV	3082076
NI	4269479
NW	4236904
RP	2543312
SH	2610605
SL	1045692
SN	3525393
ST	2251090
TH	4643496

References

Blätte, Andreas, and Andre Blessing. 2018. "The GermaParl Corpus of Parliamentary Protocols." In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (Lrec 2018)*, edited by (Conference chair)Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, et al. Miyazaki, Japan: European Language Resources Association (ELRA).