MigParl (v1.0.1-rc)

Korpusdokumentation

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

2020-01-27

Kurzdarstellung.

Dieser Artikel stellt das MigParl-Korpus vor. Es werden die verfügbaren Daten, der Datenaufbereitungsprozess zur Erstellung von Korpora bestehend aus Parlamentsdebatten sowie die Samplingstrategie zur Gewinnung eines thematisch kohärenten Korpus von Debatten zum Thema Migration und Integration beschrieben.

Ein Korpus von Plenarprotokollen

MigParl ist ein Korpus von Plenardebatten in den deutschen Bundesländern, die sich mit Themen der Migration und Integration befassen. Es knüpft an das GermaParl-Korpus an, ein Korpus von Plenarprotokollen des deutschen Bundestages, der ebenso wie MigParl im PolMine-Projekt entwickelt wurde. Als solches folgt es der Motivation und dem allgemeinen Verwendungszweck von GermaParl (siehe Blätte and Blessing 2018: 810). Da die Anwendungsfälle und Anforderungen eines Korpus von Plenarprotokollen in Blätte and Blessing (2018) ausführlich beschrieben sind, liegt der Schwerpunkt dieses Papiers auf der Beschreibung der Besonderheiten von MigParl selbst.

Das MigParl-Korpus ist ein bearbeitetes thematisches Subset der Plenarprotokolle, die von den deutschen Bundesländern von (meist) Januar 2000 bis Dezember 2018 veröffentlicht wurden. Während 15 der 16 Bundesländer Daten für etwa diesen Zeitraum zur Verfügung stellen, gibt es im Saarland keine Protokolle für den Zeitraum vor September 2004. Als thematisch spezialisiertes Korpus enthält es nicht alle Debatten, sondern nur die für die Migrations- und Integrationsforschung relevanten Beiträge. Zur verwendeten Stichprobenstrategie siehe Abschnitt “Sampling-Strategie”.

MigParl wird als R-Datenpaket mit dem Namen “MigParl” zur Verfügung gestellt. Das Paket enthält die Funktionalität, eine sprachlich annotierte, indizierte und konsolidierte Version des Korpus herunterzuladen, die in die [Corpus Workbench (CWB)] (http://cwb.sourceforge.net) importiert wurde.

Zusammen mit dem Paket wird ein kleines Beispielkorpus (“MigParlMini”) zur Verfügung gestellt. MigParlMini ist wie MigParl annotiert und indiziert, enthält aber nur etwa 1% der Daten, die das Vollkorpus liefert.

Das R-Datenpaket ist so konzipiert, dass es reibungslos mit den analytischen Werkzeugen des R-Pakets `polmineR` zusammenarbeitet.

Die Versionierung des Korpus erfolgt über ein Build-Datum in der Korpusregistrierungsdatei.

Die meisten Daten wurden aus den von den Landesparlamenten herausgegebenen PDF-Dokumenten erstellt. Diese PDF-Dokumente wurden dann mit `trickypdf`, einem R-Paket, das im Rahmen des PolMine-Projekts entwickelt wurde und auf GitHub frei verfügbar ist, in reine Textdateien umgewandelt. \footnote{https://github.com/PolMine/trickypdf). Der Workflow der Korpusaufbereitung wird im Folgenden beschrieben.

Korpusvorbereitung.

Die Erstellung der TEI-Version von MigParl setzt den folgenden Workflow um:

- **Vorbereitung**: Vorbereitung von konsolidierten UTF-8-Klartext-Dokumenten (Sicherstellung der Einheitlichkeit der Kodierungen, ggf. Konvertierung von PDF nach txt);
- **XMLifizierung**: Umwandlung der Klartextdokumente in das TEI-Format: Extraktion von Metadaten, Annotation der Sprecher etc;
- **Konsolidierung**: Konsolidierung der Sprechernamen und Anreicherung der Dokumente.

Da sich dieser Abschnitt stark mit der Aufbereitung von GermaParl überschneidet, wird nochmals auf Blätte and Blessing (2018) verwiesen. Die Herausforderungen bleiben die gleichen, ebenso wie die Motivation und der Anspruch, ein robustes und nachhaltiges Framework für die Korpusvorbereitung zu schaffen. Dennoch sind einige Aktualisierungen notwendig.

Das R-Paket `trickypdf` ist, wie oben erwähnt, immer noch im Einsatz, um PDF-Dokumente, die fast ausschließlich in einem zweispaltigen Layout vorliegen, in reine Textdateien zu verwandeln. Als Ausnahme stellte Hessen für die Zeit bis 2002 Microsoft Word anstelle von PDF-Dokumenten zur Verfügung. Diese Dokumente wurden ebenfalls in reinen Text umgewandelt.

Die Konvertierung von Klartextdateien in strukturell annotierte XML-Dokumente wird durch das im PolMine-Projekt entwickelte `Framework zum Parsen von Plenarprotokollen` oder `frappp` ermöglicht. `frappp` wird verwendet, um einen Arbeitsablauf zu erleichtern, in dem reguläre Ausdrücke für eine Reihe von Elementen formuliert werden, die identifiziert werden sollen, insbesondere Metadaten, Redner, Zwischenrufe und Tagesordnungspunkte. Wie beim bisherigen Ansatz unter Verwendung des Corpus-Toolkits (`ctk`) werden falsch positive und falsch negative Ausdrücke sowohl durch eine Liste bekannter Fehlanpassungen als auch durch Vorverarbeitungsschritte behandelt, die einige fehlerhafte Eingabedaten bereinigen.

Es wird besondere Sorgfalt darauf verwendet, dass die identifizierten Sprecher konsistent benannt werden. Dies geschieht durch Bezugnahme auf externe Datenquellen von Namen,

die durch eine Liste bekannter Aliasnamen ergänzt werden. Die primäre externe Datenquelle, die für die MigParl-Daten verwendet wird, ist Wikipedia (Blätte and Blessing 2018: 813).

Eine verbleibende Herausforderung ist die zeitliche Abhängigkeit dieser externen Prüfungen. Mitunter können sich die Informationen im Protokoll und die Informationen in den externen Daten unterscheiden. Ein Abgeordneter kann innerhalb einer Legislaturperiode die Partei wechseln. In diesem Fall wird unser Ansatz die Parteizugehörigkeit dieses Mitglieds mit der in der externen Datenquelle gefundenen kennzeichnen. Dasselbe gilt für Namen. Steht im Protokoll "Ulla Schmidt" und in der externen Datenquelle "Ursula Schmidt", dann wird die externe Information verwendet.

Diese externen Daten werden in einem Git-Repository gespeichert.

Sampling-Strategie

MigParl ist eine thematisches Subset von Plenarprotokollen, die für die Migrations- und Integrationsforschung relevant sind. Aus diesem Grund war eine Stichprobenstrategie notwendig, um diese Relevanz aus einer Grundgesamtheit aller Protokolle des deutschen Bundeslandes zu ermitteln. Wir verfolgen hier einen zweigleisigen Ansatz.

Topic Modelling-Ansatz

- *Topic Modelling*: Für jeden der Landesparlamentskorpora wurde ein eigenes Topic Model berechnet und die 100 relevantesten Begriffe pro Topic ermittelt
- *Diktionäransatz*: Eine Reihe von Kernbegriffen, die das Konzept der Migration und Integration vermitteln, wurden durch Literaturrecherche theoretisch abgeleitet
- *Themenidentifikation*: Das Wörterbuch und die 100 relevantesten Begriffe pro Topic wurden gegeneinander abgeglichen. Ein Thema mit mehr als fünf Treffern aus dem Wörterbuch wurde als relevant für Migration und Integration eingestuft
- *Dokumentenidentifikation*: Für jede Rede wurde die Wahrscheinlichkeit berechnet, zu einem der identifizierten Topics zu gehören. Wenn die summierte Wahrscheinlichkeit einen Schwellenwert überschritt, wurde die Rede als relevant betrachtet
- Die entsprechenden Reden gehen in das `MigParl` Korpus ein

Diktionärsansatz

- das MigPress-Diktionär wurde genutzt, um Reden zu identifizieren, in denen mindestens fünf Instanzen dieser Suchterme auftreten
- dies weicht von dem Schwellenwert von einem Suchterm ab, der für die Erstellung des `MigPress`-Korpus genutzt wurde. Diese Entscheidung wurde aufgrund des unterschiedlichen Sprachgebrauchs in der parlamentarischen Arena und der größeren durchschnittlichen Länge einer Rede verglichen mit einem durchschnittlichen Zeitungsartikels getroffen
- Die entsprechenden Reden gehen in das `MigParl` Korpus ein.

Annotation

Linguistische Annotation

Die XML/TEI-Version der ursprünglichen Plenarprotokolle wird durch eine Pipeline von Standardaufgaben des Natural Language Processing (NLP) geführt. Stanford CoreNLP wird für die Tokenisierung, Part-of-Speech- (POS), sowie Named-Entity- (NE) Annotation verwendet. Um Lemmata zum Korpus hinzuzufügen, wird TreeTagger verwendet.

Diese linguistische Annotation ist als sogenannte positionale Attribute (p-Attribute) Teil des Korpus. Die folgende Tabelle gibt kurze Erläuterungen zu den p-Attributen im MigParl-Korpus.

p-attribute	description	values
word	das Wort, wie es in der Rede vorkommt	word
pos	der part-of-speech-tag des Wortes nach dem Stuttgart-Tübingen-Tagset	z.B. ADJA, NN, VVINF
lemma	die lemmatisierte Form des Wortes	lemma
ner	Named Entities	O, ORGANIZATION, PERSON, LOCATION, MISC

Im sogenannten Tokenstream sieht die linguistische Annotation folgendermaßen aus:

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

Strukturelle Annotation (Metadaten)

Im XML/TEI-Datenformat werden alle Passagen ununterbrochener Sprache mit Metadaten, so genannten strukturellen Attributen (s-Attributen), versehen. Parlamentarische Reden werden häufig durch Interjektionen unterbrochen - die Information, ob es sich bei einer Äußerung um eine Interjektion oder um eine eigentliche Rede handelt, bleibt im Korpus erhalten. Hierzu gehören u.a. Legislaturperiode, Sitzungsperiode, Datum, Name des Redners und seiner Partei. Die strukturelle Annotation ist die Grundlage für alle Arten von diachronen oder synchronen Vergleichen, die die Benutzer durchführen möchten.

Die folgende Tabelle gibt kurze Erläuterungen zu den s-Attributen, die im MigParl-Korpus vorhanden sind.

s-Attribute	Beschreibung	Mögliche Werte
lp	Legislaturperiode	3 bis 21 (abhängig vom Regionalstaat)
session	Session/Protokollnummer	1 bis 161
agenda_item	Tagesordnungspunkt	Nummer des Tagesordnungspunktes
agenda_item_type	Art des Tagesordnungspunktes	Debatte / Fragezeit / Regierungserklärung / ...
date	Datum der Sitzung	YYYY-MM-TT (z.B. '2013-06-28')
calendar_week	Kalenderwoche abgeleitet vom Datum, nach ISO 8601	YYYY-Woche (e.g. 2001-01)
year	Jahr der Sitzung	2000 bis 2018
interjection	Beitrag ist eine Interjektion	TRUE/FALSE
role	Rolle des Sprechers	Vorsitz / Abgeordneter / Regierung
speaker	Name	Sprecher-Name
party	Partei	Partei des Sprechers

s-Attribute	Beschreibung	Mögliche Werte
regional_state	Bundesland, in dem die Debatte geführt wird	Bundeslandabkürzung
speech	einzelne Rede innerhalb einer Debatte	Kombination aus Sprechername, Datum und Nummer der Äußerung
migration_integration_probability	summierte Wahrscheinlichkeit für Migrations- und Integrationsrelevanz	numerisch zwischen 0 und 1
url	die URL der Quelldatei	url
src	der Typ der Quelldatei	pdf oder doc
source_dict	Die Rede wurde durch den Diktionärsansatz gesampelt	TRUE/FALSE
source_topic_model	Die Rede wurde durch den Topic Modelling-Ansatz gesampelt	TRUE/FALSE

Verwendung des MigParl-Korpus

Erste Schritte - MigParl installieren

Das MigParl-Datenpaket, das die CWB-indizierte Version des Korpus enthält, wird in einem privaten CRAN-ähnlichen Paket-Repository auf dem Web-Server des PolMine-Projekts gehostet. Das polmineR-Paket bietet einen komfortablen Installationsmechanismus.

```
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")
}
```

Nach der Installation des MigParl-Pakets enthält das Paket nur eine kleines Subset des MigParl-Korpus. Das Subset dient als Beispieldatensatz und für die Durchführung von Pakettests. Um das Gesamtkorpus herunterzuladen, verwenden Sie die folgende Funktion zum Herunterladen des Gesamtkorpus von einem externen Webspace (derzeit zur Verfügung gestellt von der Universität Duisburg-Essen):

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

Um zu überprüfen, ob die Installation erfolgreich war, führen Sie die folgenden Befehle aus. Weitere Anweisungen finden Sie in der Dokumentation des Pakets polmineR.

```
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
```

Ein sehr kurzes Tutorial

Die CWB indizierte Version von MigParl kann mit der CWB selbst oder mit jedem Tool, das die CWB als Backend nutzt (wie z.B. [CQPweb] (http://cwb.sourceforge.net/cqpweb.php)), verwendet werden. Bei den meisten technischen Entscheidungen während der Korpuserstellung wurde jedoch darauf geachtet, die Verwendung des MigParl-Korpus in Kombination mit dem polmineR-Paket zu optimieren. Bitte konsultieren Sie die Dokumentation des polmineR-Pakets (README, Vignette, Handbuch), um zu erfahren, wie Sie polmineR für die Arbeit mit MigParl nutzen können. Wir können hier nur eine sehr kurze Anleitung für die grundlegenden Befehle anbieten. Beachten Sie, dass wir in den folgenden Beispielen MIGPARLMINI zur Veranschaulichung der Funktionen verwenden, da das vollständige MIGPARL-Korpus erst nach der Komplett-Installation des Korpus zur Verfügung steht.

Zuerst möchten Sie sich vielleicht über die s-Attribute (strukturelle Attribute) und die p-Attribute (positionale Attribute) informieren, die verfügbar sind.

```
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"
```

Um etwas über die Ausprägungen von s-Attributen zu erfahren, geben Sie den Parameter s_attribute an:

```
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"
```

Um Schlagwörter und ihre Wortumfelder auszugeben, verwenden Sie die kwic-Methode (Keywords-in-Context):

```
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

Die Zählung erfolgt mittels der Count-Methode. Sie können eine oder mehrere Suchanfragen angeben:

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

```
##          query count         freq
```

```
## 1: Integration   179 0.0006665376
```

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

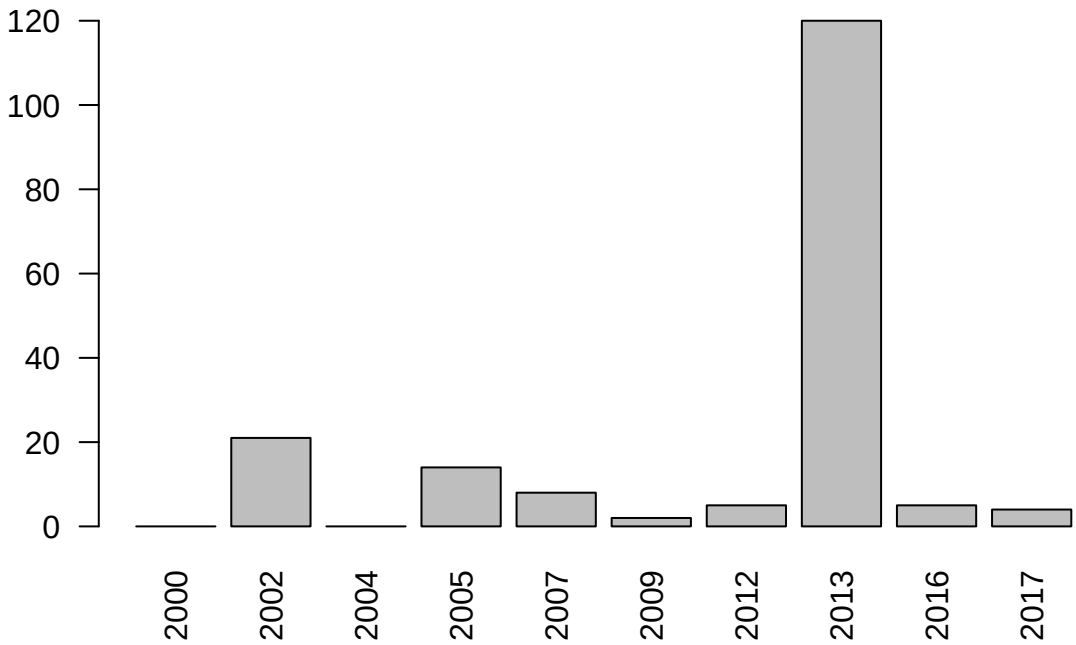
```
##          query count         freq
## 1: Integration   179 6.665376e-04
## 2:      Flucht     4 1.489469e-05
## 3: Abschiebung    17 6.330245e-05
```

Um die Streuung einer Abfrage zu erhalten, verwenden Sie die dispersion-Methode.

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

Visualisieren des Ergebnisses als Balkendiagramm:

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



Die Kookkurrenz-Methode liefert Ihnen Wörter, die zusammen mit dem Suchbegriff öfter vorkommen als statistisch erwartet.

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

Dies sind einige der Kernfunktionen, die auf den gesamten Korpus angewendet werden. Der wesentliche Punkt der strukturellen Annotation des Korpus (s-Attribute) ist es, die Erstellung von Subkorpora / Partitionen zu erleichtern. So kann jede Methode, die eingeführt wurde, auf eine Partition angewendet werden.

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

```
dispersion(year2016, query = "Flüchtlinge", s_attribute = "regional_state")
```

Schließlich ist zu beachten, dass die Methoden von polmineR auch mit der vom magrittr-Paket angebotenen Pipe-Funktionalität verwendet werden können.

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

Dies ist nur ein kurzer Einblick in die analytischen Möglichkeiten der Verwendung von MigParl in Kombination mit polmineR. Einer der wichtigsten Aspekte, der hier nicht erläutert werden kann, ist die Möglichkeit, die Syntax des Corpus Query Processors (CQP) zu verwenden, die im CWB-Backend integriert ist. Die as.TermDocumentMatrix Methode kann Datenstrukturen effizient aufbereiten, die für weitergehende analytische Techniken wie z.B. Topic Modelling benötigt werden. Bitte lesen Sie die Vignette des polmineR Pakets, um mehr zu erfahren!

Einige Vorbehalte

Eine Reihe von allgemeinen Anmerkungen soll dabei helfen, mögliche Fallstricke bei der Arbeit mit MigParl zu vermeiden:

- Die Plenarprotokolle berichten akribisch über Zwischenrufe. Um die Integrität der Originaldokumente zu erhalten, werden die Einsprüche im Korpus kommentiert. Durch die Verwendung des s-Attributs 'interjection', das die Werte 'TRUE' oder 'FALSE' annimmt, können Sie Ihre Analyse auf Sprache oder Interjektionen beschränken.
- Anders als bei GermaParl ist eine Unterscheidung zwischen Parteizugehörigkeit und Fraktion in der aktuellen Version von MigParl nicht enthalten. Dies hat hauptsächlich den praktischen Grund, dass es auf regionaler Ebene nicht allzu viele Unterschiede zwischen Partei und Fraktion gibt, da es keine "CDU/CSU"-Fraktion wie im Deutschen Bundestag gibt.
- Für Nutzer*innen, die bereits mit früheren Versionen des `MigParl` Korpus gearbeitet haben, ist es unter Umständen notwendig, ein Subset zu erstellen, das lediglich die Reden berücksichtigt, die mit dem Topic Modelling-Ansatz gesampelt wurden (source_dict == FALSE).

Fazit

MigParl, ein thematisches Subset von Debatten der deutschen Bundesländer, liegt als sprachlich annotierte und indizierte Version in Form eines R-Datenpakets vor. Die Datensammlung, die sich über 18 Jahre deutscher Parlamentsdebatten erstreckt, ist vollständig, eine Erweiterung ist kurzfristig nicht geplant. Wie bei GermaParl sollen jedoch "die Daten offen, versioniert, reproduzierbar, zugänglich und nachhaltig sein, wobei der Schwerpunkt auf der sukzessiven Verbesserung der Datenqualität liegt" (Blätte and Blessing 2018: 816). Da

MigParl in die Fußstapfen früherer Bestrebungen wie GermaParl tritt, indem es die Arbeitsabläufe, Ressourcen und methodischen Überlegungen zusammen mit den Daten öffentlich zugänglich macht, hoffen wir, dass die Daten zur Förderung eines öffentlichen digitalen Archivs der Demokratie beitragen können.

Anhang

Korpusdaten (nach Jahreszahl)

year	size
2000	2353640
2001	1963334
2002	1926908
2003	1321108
2004	1897378
2005	1804022
2006	2145901
2007	1799229
2008	1914093
2009	1707252
2010	2118794
2011	2338515
2012	2270699
2013	2401247
2014	3203446
2015	6325062
2016	5221455
2017	4410779
2018	4347845

Korpusdaten (nach Bundesland)

regional_state	size
BB	1909575
BE	3109974
BW	3399109
BY	4091021
HB	3073804
HE	5077787
HH	2601390
MV	3082076
NI	4269479

regional_state	size
NW	4236904
RP	2543312
SH	2610605
SL	1045692
SN	3525393
ST	2251090
TH	4643496

Literaturangaben

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).