Proceedings of ParlaCLARIN II Workshop, pages 66–74
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

The Europeanization of Parliamentary Debates on Migration in Austria, France, Germany, and the Netherlands

Andreas Blätte, Simon Gehlhar, Christoph Leonhardt
University of Duisburg-Essen, University of Duisburg-Essen, University of Duisburg-Essen
andreas.blaette@uni-due.de, simon.gehlhar@uni-due.de, christoph.leonhardt@uni-due.de

Abstract

Corpora of plenary debates in national parliaments are available for many European states. For comparative research on political discourse, a persisting problem is that the periods covered by corpora differ and that a lack of standardization of data formats inhibits the integration of corpora into a single analytical framework. The solution we pursue is a 'Framework for Parsing Plenary Protocols' (frappp), which has been used to prepare corpora of the Assemblée Nationale ("ParisParl"), the German Bundestag ("GermaParl"), the Tweede Kamer of the Netherlands ("TweedeTwee"), and the Austrian Nationalrat ("AustroParl") for the first two decades of the 21st century (2000-2019). To demonstrate the usefulness of the data gained, we investigate the Europeanization of migration debates in these Western European countries of immigration, i.e. references to a European dimension of policy-making in speeches on migration and integration. Based on a segmentation of the corpora into speeches, the method we use is topic modeling, and the analysis of joint occurrences of topics indicating migration and European affairs, respectively. A major finding is that after 2015, we see an increasing Europeanization of migration debates in the small EU member states in our sample (Austria and the Netherlands), and a regression of respective Europeanization in France and – more notably – in Germany.

Keywords: corpus creation, parliamentary debates, topic modeling, Europeanization, migration

1. Introduction: Migration and Europeanization

European politics have been challenged profoundly by the large inflow of refugees in 2015 and successive years.[1] Migration has moved to the top of the political agenda of Europe and has become a highly controversial issue – with a huge impact on electoral politics, coalition formation and parliamentary proceedings. For instance, in the Netherlands, a dispute over basic care for rejected asylum seekers in 2015 almost provoked the premature end of the governing coalition. In Germany, the governing coalition of Christian Democrats (CDU), the Christian Social Union (CSU) and the Social Democratic Party (SPD) has been on the verge of collapse due to migration disputes between the coalition parties in 2018. In France, the Asylum and Immigration Act passed in July 2018 triggered a fierce parliamentary dispute. In Austria, Sebastian Kurz from the Austrian People's Party (ÖVP) won the Federal Chancellery in October 2017, having moved the party to a more restrictive stance towards immigration.

Migration has become one of the most challenging issues for the future of the European Union (EU). In the 2019 campaign for the elections for the European Parliament, migration affairs were center stage. However, it is important to gain a comprehensive understanding, whether a European perspective is systematic or episodic in policy debates. A common ground and a common view of key challenges in the European political landscape is deemed to be necessary for European policy-making. Yet if the prospect of European politics depends on the commonality of perceptions, the question arises, whether issues are generally contextualized in a European manner or whether perceptions remain being defined exclusively from the national context. This is why the Europeanization of migration debates in the national parliaments of EU countries is relevant for the outlook of European politics.

Of course, debates on migration and integration are not at all limited to parliamentary debates. Discursive hotspots will often be located somewhere else, in the digital realm amongst others. However, the parliamentary arena has a central role for the agenda of the political-administrative system. Also, one advantage is that plenary debates provide comparative data. Still, comparing parliamentary debates is not an easy road. Whereas pdf documents are almost universally accessible and in the public domain, standardized and machine-readable versions of parliamentary debates covering the same period under investigation remain a goal yet to be reached when working towards the aim of safeguarding a "data-rich" future for social science research (King, 2011).

For this purpose, we develop a "Framework for Parsing Plenary Protocols", or "frappp" in short. This framework defines a generic workflow for preparing corpora of plenary protocols, limiting the marginal cost for preparing a corpus for an additional parliament to defining regular expressions and the development of supplementary data to consolidate the corpus.

Previously, the *frappp*-approach has been used to prepare corpora of all German regional parliaments and of the UN General Assembly. To explore and demonstrate the advances that may result from an improved data preparation workflow, we started to apply the procedure to a limited number of parliaments across Europe. Based on the theoretical consideration that Western European states have experienced a comparable political development, including a similar history of immigration (Messina, 2007), this study focuses on Germany, France, the Netherlands and Austria. The corpora prepared for the German

[1] We gratefully acknowledge funding of the Stiftung Mercator for the Mercator Forum Migration and Democracy (MIDEM) that has been instrumental for this research. MIDEM is a research center of the Technische Universität Dresden in cooperation with the University of Duisburg-Essen, funded by Stiftung Mercator.

Bundestag, the French Assemblée Nationale, the Tweede Kamer of the Netherlands and the Austrian Nationalrat are called “GermaParl”, “ParisParl”, “TweedeTwee” and “AustroParl”.

The data covers two decades (1999-2019). From a technical point of view, our investigation begins at a time when parliaments started to offer ”born digital” versions of parliamentary proceedings. Furthermore, our timeframe is defined by important events regarding migration policy in Europe. The Tampere Summit of October 1999 was an important milestone for the closer cooperation of European countries on migration policy, including far-reaching steps towards the harmonization of European asylum law (Trauner, 2016). The May 2019 European elections were dominated by the migration issue and can be seen as a preliminary endpoint of a long period with a high salience for the issue. Methodologically, we use topic modeling (Blei et al., 2003) as a technique to detect the thematic focal points of debates and the co-occurrence of migration-related topics and topics indicating a European perspective. Our paper demonstrates that using computer-assisted text analysis in combination with large-scale textual data is a highly efficient approach to gain findings about the degree of Europeanization of debates on migration in national parliaments. These findings would be utterly tedious to obtain otherwise. But before we turn to data, methodology and results in more detail, we develop the theoretical vanguard more precisely.

2. Theory: Europeanization as a Matter of Attention

Do policies become more similar across the European countries? The assumption of an increasing convergence of policies among EU countries is deeply embedded in the European integration project. In the social sciences, convergence was initially understood as a legislative harmonization process among countries. Researchers have identified various factors leading to convergence. At a systemic level, increasing interdependencies of nation states and the continuous expansion of international organizations (e.g. the EU), were expected to bring about convergence. Interdependence results in legal and normative obligations that should, at least theoretically, lead to legislative convergence of EU countries (Holzinger et al., 2007).

A large number of policy areas can be the subject of convergence processes, including migration policy. Indeed, the ability of EU countries to regulate migration has been shaped profoundly by European integration. The establishment of the European Single Market and the free movement of EU citizens has imposed a set of important restrictions on EU countries to regulate migration at the national level. In addition, some issues have shifted to the European level. The EU has gained relevance for asylum policy as well as security and border control (Trauner, 2016). Thus, the comparative analysis of the convergence of the policy output in migration policy-making is a well-justified and interesting research perspective.

Indeed, convergence studies have a strong focus on policy output (Nordbeck, 2013). In this context, convergence is defined as “any increase in the similarity between one or more characteristics of a certain policy (e. g. policy objectives, policy instruments, policy settings) across a given set of political jurisdictions (supranational institutions, states, regions, local authorities) over a given period of time” (Knill, 2005).

But convergence can be understood more broadly. Kerr (1983) defines convergence as a “tendency of societies to grow more alike, to develop similarities in structures, processes and performances”. Accordingly, convergence does not necessarily mean a congruent or identical reaction to a certain problem, but rather it refers to a gradual approximation, for example in the choice of policies (Scholz, 2012). Theories of convergence entail the empirical necessity to measure similarities across political systems. Accordingly, the focus on specific policies or political outcomes is just one option. A focus on the discursive and communicative patterns of parliaments is a viable alternative. With regard to the question of a common European perception and contextualization of specific problems, we draw on the literature on Europeanization and on the emergence of a European public sphere.

Studies on the European public sphere argue that the EU depends on a common frame of reference shared by citizens of EU countries (Trenz, 2015; Lingenberg, 2010). A common approach taken by these studies is to identify a European public sphere based on the salience of issues in national media systems (Trenz, 2015). The public sphere is defined “as a site where public discourses and popular identities are framed” (Trenz, 2015). At the heart of this approach is the conviction that the mass media constitute the public sphere. In this research, Europeanisation is measured by the “general level of attention the media pays to political news from the EU” (Trenz, 2015). The visibility of European events, actors and issues is the empirical hallmark of this approach.

With regard to our research, the parliamentary arena is no less important to understand Europeanization and convergence, similar to media system analysis. A focus on the frame of reference of parliamentary attention has important methodological consequences. Our interest in the larger trends concerning migration and European affairs implies that an in-depth analysis of the speeches is not necessary. Distant reading rather than close reading is required (Moretti, 2013). Statements about parliamentary attention at a higher level of abstraction make parliamentary discourse comparable and indicate using text mining techniques. The focus on attention structures is furthermore supported by the methodology of the Comparative Agendas Project (CAP).

The CAP monitors policy processes by tracking government activity in response to the challenges they face. These activities can take a variety of forms, including holding hearings or giving speeches (Baumgartner et al., 2019). Bevan (2019) argues that measuring attention is important because every policy change assumes that the “policy is first attended to”. The project has established a comprehensive database recording the “date as well as a minimum of additional information about each issue” (Baumgartner et al., 2019). Baumgartner el al. (2019) argue: “If the key issue is how much attention is being directed at an issue, and if the attention reflects enthusiasm or criticism, then traditional 'deep reading' of the text was not needed".

The CAP methodology justifies why an abstract measurement of attention in parliamentary discourse may provide important insights. This is the starting point of our research: Speeches can be classified with the help of computer-based procedures (topic models). The aim is to identify attention for two relevant issues in speeches: First, speeches with a migration policy reference and second, speeches with a European policy reference. When we know which speeches address migration policy, and which speeches refer to the European level of policy-making, we can obtain statements on the overlap of categories. It is assumed that the appearance of both categories in one speech is an indicator that the migration issue was discussed in the European context, in the sense that a Europeanization of the topic is taking place.

3. Data

3.1. A Framework for Parsing Plenary Protocols

In line with our research interest, we prepared and augmented four corpora of parliamentary debates, from Austria's Nationalrat ("AustroParl"), the French Assemblée Nationale ("ParisParl"), Germany's Bundestag ("GermaParl") and the Dutch Tweede Kamer ("TweedeTwee").[2]

The raw data for building the corpora was obtained from the parliaments' websites. While most of the necessary data is provided as pdf documents,[3] this format is not apt for technically advanced analyses. A toolchain of several R packages developed in the context of the PolMine Project was used to transform the raw data into a more format suitable for corpus analysis: The first of those, `trickypdf`, processes pdf files with challenging layouts, providing a convenient workflow to extract text from pdf documents with more complex layouts featuring two columns as well as text on the margins.[4]

In the next step, the plain text output of `trickypdf` needs to be scanned for structural information and annotated accordingly. To ensure replicability and sustainability, plenary data should be prepared in a way which satisfies the principles of FAIR (Wilkinson, 2016). At the same time, barriers of data preparation should be minimized. With these goals in mind, instead of resorting to individual solutions for each parliament, the R package `frappp` was developed which strives "[t]o reduce necessities to re-invent the wheel in new corpus preparation projects, [and] uses techniques of object-oriented programming and offers a framework that runs the user through the corpus preparation workflow" (Blätte and Leonhardt, 2019).

To transform plain text to XML, regular expressions are used to extract relevant meta-information and to store it in the structured data format of the XML output document. Thus, corpora contain information on the legislative period and the date of a speech. They report the parliamentary group membership of a speaker as well as the role of the speaker. Interjections are also annotated. This structural annotation of the original text permits to create complex and multi-layered sub-corpora, which are the prerequisite for comparative analyses. Undoubtedly, a coherent standardization of plenary data is required. One of the most valid solutions is provided by the guidelines of the Text Encoding Initiative (TEI).[5] While being merely TEI-inspired, the XML output of `frappp` is a preliminary simplified approximation that may be an initial step towards standardization.

TEI/XML is useful for standardization and as a data exchange format. However, it is not necessarily appropriate for analysis. Changing the format is only a first step. After this stage of "XMLification", the speeches were tokenized and annotated linguistically.[6] All words were lemmatized and assigned to a part of speech. Stanford CoreNLP was used for tokenization and Part-of-speech-tagging (as well as Named Entity Recognition for the Austrian corpus) (Manning et al., 2014). The TreeTagger was used for lemmatization (Schmid, 1995). The general preparation process of the TEI files is described more in-depth for the GermaParl corpus which served as a model and prototype for the further corpora that have been prepared (Blätte and Blessing, 2018).

In a last step, the data was imported into the IMS Open Corpus Workbench (Evert and Hardie, 2011). This was done with the R package cwbtools. Thus, a data release will entail offering the TEI/XML data as well as the CWB indexed corpus. For reproducing results, the latter is the relevant basis.[7]

Only a part of the corpus we use in this analysis for the Dutch case is prepared as described above. Data before 15 September 2015 is taken from the ParlSpeech corpus by Rauh et al. (2017b) and then merged with a newly prepared corpus of Dutch protocols.

3.2. Structural annotation

Structural annotation is the key to obtain relevant research findings inside the corpora. As previously mentioned, while XML is ideal for long-term storage and interoperability, the indexed corpus version is the relevant resource for concrete research and publication projects. In the jargon of the Corpus Workbench (CWB), annotation layers are called "structural attributes". Table 1 provides an overview about available attributes, their description, possible values and the corpora they are available for.

[2] These corpora were developed experimentally at the time of writing. They shall be released in 2020.

[3] The German Bundestag switched to a thoroughly annotated XML format starting with the 19th legislative period (beginning in September 2017).

[4] The package is available at GitHub, see: `https://github.com/PolMine/trickypdf`.

[5] The Parla-CLARIN standard (Erjavec and Pančur, 2019) discussed at the 2019 ParlaFormat Workshop is the reference suggestion at this stage.

[6] The Dutch corpus was tokenized using the openNLP interface for R (Hornik, 2016) with its Dutch language model (Hornik, 2015).

[7] In the case of GermaParl, which serves as a model for future releases of the other corpora, the XML data is available via a GitHub repository (see `https://github.com/PolMine/GermaParlTEI`). The indexed corpus is deposited at Zenodo (Blaette, 2020), to be deposited at a CLARIN repository at a later stage.

Structural Attribute	**Description**	**Possible Values**	**Availability**
date	date of utterance	YYYY-MM-DD	AT, FR, GER, NL
year	year of utterance	YYYY	AT, FR, GER, NL
speaker	speaker of utterance	full name of speaker	AT, FR, GER, NL
party	party affiliation of speaker	party of speaker	AT, FR, GER, NL
session	number of session, the utterance was held in	numeric	AT, FR, GER, NL
interjection	whether utterance is interjection or not	logical, TRUE or FALSE	AT, FR, GER, NL
role	role of the speaker	presidency / mp / government	AT, FR, GER
lp	legislative period	numeric	AT, FR, GER
agenda_item	agenda item	number of the agenda item	AT, FR, GER†
agenda_item_type	type of agenda item	debate / question_time / government_declaration	AT, GER†
id	continuous number of processed plenary protocols	numeric, starting from 1	AT, FR
parliamentary_group	parliamentary group of party	parliamentary group of speaker	FR, GER

†Applies to the released version of the GermaParl corpus, not the update used in the following analysis.

Table 1: Structural Attributes of Corpora

Structural attributes are named in an intuitive way. Yet the distinction between party and parliamentary group needs to be explained: The attribute *party* denotes the party affiliation of a speaker. This may be different from the parliamentary group of the speaker, as indicated by the attribute *parliamentary_group*. This distinction is particularly decisive. For instance, government actors are often members of a party, but do not necessarily adhere to a parliamentary group. It also happens that politicians from different parties or with no party affiliation join a common parliamentary group.

The annotation of interjections is another particular feature: Interjections are not part of a speech itself but are "infused‘" by other participants of the debate. For example, applause during a speech would be annotated as an interjection (AustroParl: "Allgemeiner Beifall", ParisParl: "Applaudissements sur divers bancs", GermaParl: "Beifall bei der CDU/CSU und der SPD", TweedeTwee: "Applaus"), as would be laughter (AustroParl: "Allgemeine Heiterkeit", ParisParl: "Rires"), interjections by individual speakers (GermaParl: "Speaker [Parliamentary Group]: Das ist ja unglaublich!") and context information such as the closing of the session (TweedeTwee: "Sluiting 22.22 uur").

Finally, *agenda_item* and *agenda_item_type* describe the agenda item of a debate as identified in the protocol. Whereas *agenda_item* provides a running number of agenda items by protocol, *agenda_item_type* provides a categorization of the agenda item call.

At this stage, not all structural attributes are available for all corpora. TweedeTwee is sparsely annotated by comparison. This is due to the fact that we use the Dutch ParlSpeech corpus as a basis and adopt the annotation provided there (Rauh et al., 2017a). In addition, as Rauh et al. (2017a) explain, the attribute for session is not available before January 2011 and is thus identical with the date for Dutch data. Explicit information about interjections are also only available in TweedeTwee after September 2015. Furthermore, the parliamentary group is only annotated in the GermaParl and the ParisParl corpus. Yet due to the dynamics of the French party system (at least when it comes to party names), this attribute is annotated less reliably in the French corpus than in its German counterpart.

Finally, the difficulty to achieve a reliable annotation agenda items is substantial. For instance, small variations in the language used by a parliament's presidency when calling a new agenda item may cause regular expressions to fail. These limitations need to be kept in mind when working with large and diverse data that has been prepared in an automated process. Given the workflow we used, Austrian and German protocols are rather similar and easier to process than the French and Dutch data. For both AustroParl and GermaParl, documents were available digitally born for the entire period of interest. Both interjections and speakers could be detected in a reliable fashion in the text. For TweedeTwee, we addressed issues of the limited data availability (in a format we wanted to work with at least) by using data previously prepared by Rauh et al. (2017a). ParisParl presented particular challenges: Interjections were presented as very short speeches. Speakers were annotated with a variation of patterns that challenged the approach of `frappp` that is based on regular expressions. This was addressed by adjusting the extraction pipeline to include further text formatting information to identify speaker calls and by employing a rather large collection of external data (parliamentary data from Wikipedia, see Blätte and Blessing (2018)) to check for speaker mismatches. To conclude: We do acknowledge that every new corpus preparation project has its own intricacies. Still, while some data specific adjustments to the pipeline are still required, the framework `frappp` enhanced the efficiency of the data preparation process substantially and was a prerequisite to obtain a congruent dataset for the four countries under investigation.

3.3. Descriptive Statistics

The following descriptive statistics present essential information on the subsets of the corpora that have been used. The period of investigation we defined covers the period between the Tampere summit in October 1999 to the elections to the European Parliament in May 2019 (always including the full month). The corpora cover a broader time span, yet with variations, making the temporal standardization necessary. Once the consistency of coverage is ensured, AustroParl comprises about 62 million tokens. ParisParl has a size of about 203 million tokens. The subset of GermaParl examined here is 97 million tokens. Finally, TweedeTwee comprises of about 135 million tokens. To supplement this initial overview over the data, figure 1 reports the number of tokens in the four corpora per year.

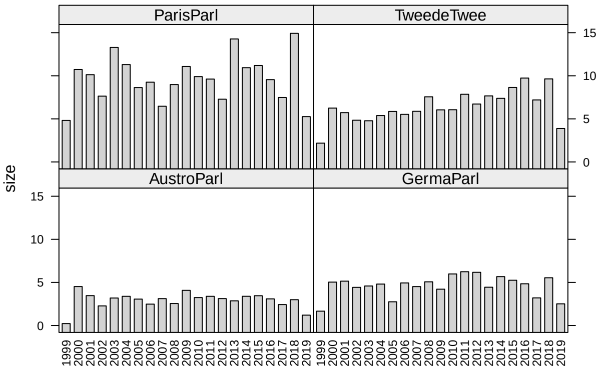

Figure 1: Corpus size per year

As presented in figure 1, there are substantial differences among parliaments in terms of plenary productivity. This observation is particularly true when shifting our attention from a blunt token count to a more substantial identification of speeches. Indeed, parliamentary proceedings have a specific logic. The notion of *speeches* in our data set is derived from a technical definition. A “speech” is defined as a coherent set of utterances of an individual speaker on a single day. Since it is reasonable to assume that a speaker can present more than one speech per day, the following heuristic is used: If two utterances of the same speaker on the same day are interrupted by more than 500 tokens of another speaker, these two utterances are assumed to be two separate speeches. If they are interrupted by less than 500 tokens, they are assumed to be one speech merely interrupted by interjections or organizational interventions. As Rauh et al. (2017a) noted, the number of speeches differs between countries due to different parliamentary settings and understandings. This also applies to the annotation of interjections – which differs as well (Rauh et al., 2017a). We can confirm this statement beyond the corpora examined by Rauh.

The procedure to identify speeches results in an initial distinction of speeches that does not assume a minimum required length for considering an utterance a speech. As illustrated by 2, a histogram of the lengths of (unfiltered) speeches for four parliamentary corpora, there is a substantial variation of the lengths of speeches.

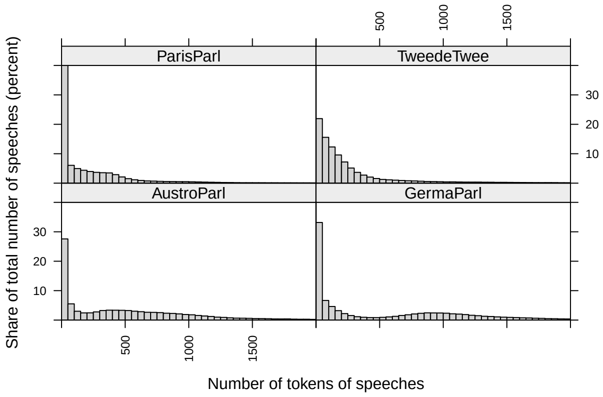

Figure 2: Length of speeches - histogram

An interesting insight conveyed by the histograms is that the distribution of the length of speeches is very different between Germany and Austria on the one side, and the Netherlands and France on the other side. Differences in parliamentary culture may explain this variation.[8] The histograms also indicate that the computational heuristic to detect speeches results in many very short contributions to parliamentary debates. These stumps are very unlikely to be speeches in a substantial sense. For our analysis, we assumed that contributions of a speaker need to surmount 100 tokens to qualify as a speech. This kind of threshold is also an appropriate requirement that the topic modeling technique will work well.

Assuming that at least 100 words are required to make a speech, figure 3 conveys the number of speeches given in the Assemblée Nationale, the Tweede Kamer, the Nationalrat and the Bundestag per year. The plot conveys that speeches are not evenly distributed across time. There is a notable fluctuation between the years that is easily explained for the fringe years: The period of investigation starts with the Tampere summit (October 1999) and ends with the May 2019 European election. The number of speeches in the initial and trailing year is unsurprisingly curtailed. Furthermore, parliaments are subject to cyclical fluctuations. There is a decline of the number of speeches during election years. Notably, this is much clearer in Germany, Austria and France than in the Dutch parliament.

[8] Exploring these differences between the corpora is beyond the scope this paper, but deserves further investigation.

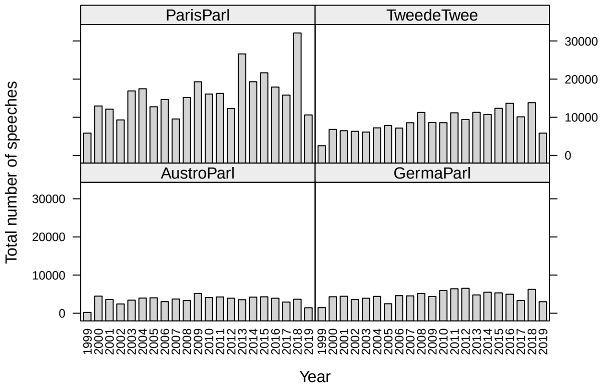

Figure 3: Number of speeches per year

To sum up core findings on speeches in the four corpora, table 2 presents the total number of speeches computationally detected, the number of speeches with at least 100 tokens and the share of speeches that are stumps (less than 100 tokens). Excluding stumps from the analysis has a mitigating effect on the number of speeches analyzed, but there is a remaining substantial variation of the extent of plenary speech-making to be considered. Speeches technically detected that are not stumps (more than 100 words) were that basis for the topic modeling described in the next section.

corpus	speeches (all)	speeches (min. 100)	stumps (per cent)
GermaParl	158 537	95 537	39.7
ParisParl	785 707	334 627	57.4
TweedeTwee	296 184	185 912	37.2
AustroParl	110 198	73 794	33.0

Table 2: Summary of core features of the corpora

4. Methodology: Measuring Europeanization Using Topic Models

We are interested in shifts in the combined attention targeting both migration and European affairs. To process a very substantive amount of data, a method classifying data in an efficient and reproducible way is needed: The parliamentary discourse during this period comprises millions of words and several hundred thousand speeches. The data-driven method of topic modeling is a useful solution. Topic models are methods for determining thematic structures in large unstructured text collections (Bock et al., 2016) which have been applied successfully to corpora of parliamentary speech in previous studies (Greene and Cross, 2017). Latent Dirichlet Allocation (LDA) is a classic procedure in topic modeling (Blei et al., 2003) and is still considered to be a state-of-the-art solution for probabilistic topic models (Rahimi et al., 2016).

The LDA makes two assumptions: First, documents consist of several topics with different weights and second, the text corpus is composed of a certain number of topics (Bock et al., 2016). An LDA model describes the probability distribution of topics over the complete corpus and indicates the share of each topic in the respective documents or speeches. It also describes the probability that specific words belong to a specific topic. In the LDA context, the term “topic” should not be equated prematurely with what is understood as a topic or issue in a social science context. Topics, within the context of topic modeling, are latent constructs that are indicated by a collection of words that are related. The thematic definition of specific topics is performed by the researcher by giving an interpretation to the most probable words contained in a topic by assigning a label to it (Wiedemann and Niekler, 2016). In research practice, some topics are unspecific and difficult to interpret, while other topics are much clearer and easier to classify. To achieve a valid classification, a close reading of texts will usually be desirable.

Knowing which mixture of topics is present in a speech, we can make statements about the joint occurrence of migration and European affairs in a speech, and on the Europeanization of migration debates. To implement this idea, an LDA model was calculated for each corpus, which was then interpreted by the members of the MIDEM research project. The coding instructions were simple and straightforward. A reference to migration issues or to European affairs was determined based on the 50 most relevant terms of the topic. The joint interpretation of the models by the team of researchers was sought to establish intersubjectivity. A number of examples shall illustrate the topics which have been selected.

- For *GermaParl*, three topics were identified as indicative for migration (152, 181 and 210). For example, the top words for topic 152 are ”Deutschland” (Germany), ”Flüchtlinge” (refugees), ”Menschen” (people), ”Asylbewerber” (asylum seekers) and ”Asyl” (asylum). Three topics were selected as indicative for a European reference (54, 71 and 179). Topic 54 is characterized by the words ”Europa” (Europe), ”Union” (union), ”Europäischen” (European), ”europäischen” (European) and ”Europäische” (European).

- In the analysis of *ParisParl*, two topics indicate a reference to migration (66 and 135). The top words for topic 66 are ”asile” (asylum), ”immigration” (immigration), ”pays” (country), ”droit” (”law”) and ”étrangers” (foreigners). Three topics were seen to convey a reference to Europe (61, 162 and 195). Topic 61 is characterized by the words ”européenne” (European), ”directive” (directive), ”européen” (European), ”Commission” (commission) and ”Union” (union).

- For *TweedeTwee*, three topics were selected for migration (187, 206 and 243). Topic 187 is described by ”asielzoekers” (asylum seekers), ”Nederland” (the Netherlands), ”mensen” (people), ”land” (country), ”IND” (probably the Dutch Immigration and Naturalisation Service). Three topics entail European references (1, 24 and 160). Top words for topic 1 are ”Europese” (European), ”Europa” (Europe), ”Unie” (union), ”lidstaten” (member states) and ”Europees” (European).

- Three migration topics were identified in *AustroParl* (41, 64 and 152). The top words for topic 41 are ”Österreich” (Austria), ”Asylbewerber” (asylum seeker), ”Asyl” (asylum), ”Verfahren” (procedure) and ”Asylverfahren” (asylum procedure). Three topics (57, 212 and 215) indicate European references. For topic 57, top words are ”Union” (union), ”Europäischen” (European), ”Europa” (Europe), ”Europäische” (European) and ”europäischen” (European).

An extensive documentation of the topics identified to pertain either to migration or European affairs is included in the Technical Annex for this paper that is available online.[9]

5. Analysis: The Europeanization of Migration Debates

The empirical strategy we pursue is to analyze co-occurrences of migration and European issues in speeches based on topic models. These co-occurrences were determined based on the five most probable topics per speech. The number of top topics considered may be chosen based on various criteria. Based on several tests, opting for the first five topics per speech was considered a suitable choice. This way we identified speeches with a migration reference (mig), speeches with a European reference (eu), and speeches with both references (mig+eu).

To generate results that are neither too rough nor too fine-grained, we aggregated the investigation period into roughly five-year periods. There is a set of reasons to be considered. First, there is the cyclical fluctuations already mentioned, i.e. the slumps of plenary activity in election years. Second, aggregation is necessary to achieve significant numbers.

For the GermaParl corpus a total of 4341 speeches with reference to migration and 6632 speeches with reference to European issues have been found. 324 of those overlap. See table 3 for a breakdown per period of interest.

mig+eu	mig	eu	rel	chi	period
63	596	1793	10.57	24.43	1999-2004
39	572	1480	6.82	1.50	2005-2009
94	1144	1910	8.22	19.49	2010-2014
128	2029	1449	6.31	5.62	2015-2019

Table 3: Topic Cooccurrences in the GermaParl corpus

While the absolute number of speeches referencing migration increases, the number of speeches referencing both migration and Europe increases less rapidly in absolute terms. Their relative share compared to all migration related speeches all in all decreases.

As can be seen in table 4 In the ParisParl corpus, there were significantly more speeches with both migration and European references. A total of 10751 speeches with reference to migration could be found compared to 20070 speeches with reference to European issues. 1123 of those overlap. Table 4 illustrates the breakdown by period of interest.

mig+eu	mig	eu	rel	chi	period
151	1583	4758	9.54	87.86	1999-2004
164	1829	4030	8.97	81.38	2005-2009
343	2770	5496	12.38	329.44	2010-2014
465	4569	5786	10.18	294.12	2015-2019

Table 4: Topic Cooccurrences in the ParisParl corpus

Similar to GermaParl, the absolute number of speeches concerning migration increases in ParisParl. The number of speeches referring to Europe remains relatively stable. The same applies to the relative share of speeches which are both referencing migration and Europe which, after a slight peak in the period of 2010 to 2014 almost returns to its initial value.

In the Netherlands, shown in table 5, both the migration issue and the European issue also received a considerable amount of attention. Although significantly smaller than the French or German parliaments, a total of 9162 speeches were given on migration policy. In comparison, 14789 speeches were delivered on European policy issues. The overlap between the two topics was 870. See table 5 for the description period by period. A look at the relative share shows a moderate increase from 7.50 percent in the first period under study to 12.88 percent in the fourth period.

mig+eu	mig	eu	rel	chi	period
148	1974	3025	7.50	5.48	1999-2004
161	2168	3439	7.43	3.72	2005-2009
181	2070	4034	8.74	19.31	2010-2014
380	2950	4291	12.88	206.23	2015-2019

Table 5: Topic Cooccurrences in TweedeTwee corpus

In the Dutch corpus, the number of speeches about both migration and Europe increases while the relative share of migration speeches also referencing Europe sees an uptick in the final period of investigation.

mig+eu	mig	eu	rel	chi	period
36	480	1282	7.50	6.47	1999-2004
69	770	1582	8.96	6.64	2005-2009
72	682	1488	10.56	24.92	2010-2014
303	1505	1245	20.13	508.39	2015-2019

Table 6: Topic Cooccurrences in the AustroParl corpus

The number of speeches on migration policy issues in Austria increased strongly during the period under study, while the number of speeches on European policy first increased and then decreased again. A total of 3437 migration policy and 5597 European policy speeches were found. Table 6 provides an overview over the development by period.

Resulting from a rise in the number of migration speeches and a slight decrease in the number of European policy speeches, a substantial increase of the relative frequency of Europeanized migration speeches was observed.

The development of the relative share of the number of

[9] https://polmine.github.io/ParlaCLARIN2020/TechnicalAnnex.html.

speeches concerned with migration which also refer to European topics compared to all migration related speeches is comparable to the TweedeTwee corpus. It is increasing steadily.

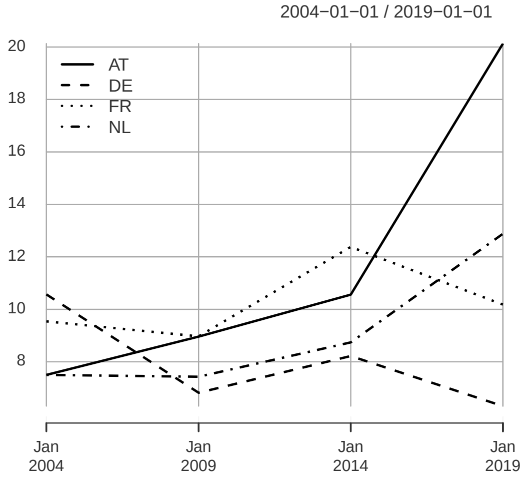

Figure 4: Shares of Europeanized Migration Speeches

The final plot 4 combines the data on the share of migration speeches that include a co-occurrence with a European reference obtained for the individual parliaments. The operationalization of Europeanization as the presence of a European reference in speeches on migration affairs shows a noteworthy trend: The four parliaments examined have witnessed a considerable absolute increase of speeches that address migration affairs. But there is a difference between debates in the national parliaments of the large and smaller EU countries. In France and Germany, the share of speeches on migration that entailed a European dimension decreased, indicating a more nation-centric and inward-looking perspective rather than a perspective that is Europeanized and takes Europe into account. The shared trend notwithstanding, France and Germany are still different: The level of Europeanization is significantly higher in France during the periods examined; the disappearance of the European point of reference is a much more a specifically German phenomenon. In juxtaposition to that, parliamentary debates in Austria and the Netherlands – two smaller EU countries – gain a stronger European orientation and get more Europeanized, as the challenges they face make the European point of view more relevant when the stakes are high.

6. Conclusion and Outlook

It is a well-founded suspicion rather than a consolidated research finding that an increasing salience of migration spurs very different trends with respect to Europeanization in small EU countries as compared to large EU countries. A next step is to validate the observed patterns and the explanatory thrust with a close reading of the speeches that have been classified as addressing migration and European affairs. Indeed, there are many ensuing questions that can be asked to understand and to give interpretative substance to our descriptive finding on the mixed trends of Europeanization of migration debates in the four parliaments investigated. This limitation notwithstanding, we are confident that our data and our methodology yield a result that is robust at the descriptive level.

The purpose of this paper is to demonstrate that the data basis for making statements about changing attention patterns and Europeanization can be obtained with reasonable effort, and that we do have the methodology to make statements about speech-making in the longue durée. It would be very difficult to obtain the kind of results we present without large-scale corpora and efficient techniques of corpus preparation and text analysis. The preparation of corpora of parliamentary debates is a precondition for this kind of comparative research. Using enhanced procedures for preparing corpora with limited marginal costs, such as the frappp, the "Framework for Parsing Plenary Protocols", may help to bring the vast potential of text analysis to fruition.

7. Bibliographical References

Baumgartner, F. R., Breunig, C., and Grossman, E. (2019). The Comparative Agendas Project. Intellectual Roots and Current Developments. In *Comparative Policy Agendas*, pages 3–16. Oxford University Press.

Bevan, S. (2019). Gone Fishing. The Creation of the Comparative Agendas Project Master Codebook. In *Comparative Policy Agendas*, pages 17–34. Oxford University Press.

Blaette, A. (2020). Germaparl. linguistically annotated and indexed corpus of plenary protocols of the german bundestag. CWB corpus version 1.0.6.

Blätte, A. and Blessing, A. (2018). The GermaParl Corpus of Parliamentary Protocols. In Nicoletta Calzolari (Conference chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

Blätte, A. and Leonhardt, C. (2019). The Framework For Parsing Plenary Protocols (frappp). Why parlaTEI matters. Slides presented at the ParlaFormat Workshop May 2019.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Bock, S., Du, K., Huber, M., Pernes, S., and Pielström, S. (2016). Der Einsatz quantitativer Textanalyse in den Geisteswissenschaften. *DARIAH-DE Working Papers*, 18, 01.

Erjavec, T. and Pančur, A. (2019). Parla-clarin: Tei guidelines for corpora of parliamentary proceedings.

Evert, S. and Hardie, A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium.

Greene, D. and Cross, J. P. (2017). Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach. *Political Analysis*, 25(1):77–94.

Holzinger, K., Jörgens, H., and Knill, C., (2007). *Transfer, Diffusion und Konvergenz: Konzepte und Kausalmechanismen*, pages 11–35. VS Verlag für Sozialwissenschaften, Wiesbaden.

Hornik, K., (2015). *openNLPmodels.nl: Apache OpenNLP Models for Dutch*. R package version 1.5-2.

Hornik, K., (2016). *openNLP: Apache OpenNLP Tools Interface*. R package version 0.2-6.

Kerr, C. (1983). *The Future of Industrial Societies: Convergence or Continuing Diversity?* Harvard University Press, Cambridge, Mass., 2nd edition.

King, G. (2011). Ensuring the data rich future of the social sciences. *Science*, 331(11):719–721, 2011.

Knill, C. (2005). Introduction: Cross-national policy convergence: concepts, approaches and explanatory factors. *Journal of European Public Policy*, 12(5):764–774.

Lingenberg, S., (2010). *Einleitung*, pages 13–22. VS Verlag für Sozialwissenschaften, Wiesbaden.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Messina, A. M. (2007). *The Logics and Politics of Post-WWII Migration to Western Europe*. Cambridge University Press.

Moretti, F. (2013). *Distant Reading*. Verso, London.

Nordbeck, R., (2013). *Nationale Umweltpolitik in einem internationalisierten Kontext*. Springer Fachmedien Wiesbaden, Wiesbaden.

Rahimi, M., Zahedi, M., and Mashayekhi, H. (2016). A two level probabilistic topic model. In *2016 24th Iranian Conference on Electrical Engineering (ICEE)*, pages 108–112.

Rauh, C., De Wilde, P., and Schwalbach, J. (2017a). Release note. In *The ParlSpeech data set: Annotated full-text vectors of 3.9 million plenary speeches in the key legislative chambers of seven European states*. Harvard Dataverse.

Rauh, C., De Wilde, P., and Schwalbach, J. (2017b). The ParlSpeech Data Set. Annotated Full-Text Vectors of 3.9 Million Plenary Speeches in the Key Legislative Chambers of Seven European States.

Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*.

Scholz, A., (2012). *Einleitung*. VS Verlag für Sozialwissenschaften, Wiesbaden.

Trauner, F. (2016). Wie sollen Flüchtlinge in Europa verteilt werden? Der Streit um einen Paradigmenwechsel in der EU-Asylpolitik. *integration*, 39(2):93–106.

Trenz, H.-J., (2015). *Europeanising the Public Sphere – Meaning, Mechanisms, Effects*, pages 233–251. Springer Fachmedien Wiesbaden, Wiesbaden.

Wiedemann, G. and Niekler, A. (2016). Analyse qualitativer Daten mit dem Leipzig Corpus Miner.

Wilkinson, M. D. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1).