*Proceedings of ParlaCLARIN II Workshop*, pages 58–65
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020


# Comparing Lexical Usage in Political Discourse across Diachronic Corpora

**Klaus Hofmann[1], Anna Marakasova[2], Andreas Baumann[1], Julia Neidhardt[2], Tanja Wissik[3]**
[1]University of Vienna, [2]TU Wien, [3]Austrian Academy of Sciences
Vienna, Austria
[1]{andreas.baumann, klaus.hofmann}@univie.ac.at,
[2]{anna.marakasova, julia.neidhardt}@tuwien.ac.at, [3]tanja.wissik@oeaw.ac.at

**Abstract**

Most diachronic studies on either lexico-semantic change or political language usage are based on individual or structurally similar corpora. In this paper, we explore ways of studying the stability (and changeability) of lexical usage in political discourse across two corpora which are substantially different in structure and size. We present a case study focusing on lexical items associated with political parties in two diachronic corpora of Austrian German, namely a diachronic media corpus (AMC) and a corpus of parliamentary records (ParlAT), and measure the cross-temporal stability of lexical usage over a period of 20 years. We conduct three sets of comparative analyses investigating a) the stability of sets of lexical items associated with the three major political parties over time, b) lexical similarity between parties, and c) the similarity between the lexical choices in parliamentary speeches by members of the parties vis-à-vis the media's reporting on the parties. We employ time series modeling using generalized additive models (GAMs) to compare the lexical similarities and differences between parties within and across corpora. The results show that changes observed in these measures can be meaningfully related to political events during that time.

**Keywords:** diachronic corpora, lexical stability, political discourse

## 1. Introduction

Lexical associations among words change over time. This is particularly evident for the lexical contexts associated with words denoting named entities, such as political parties in public discourse. Various approaches have been developed to make contextual (or semantic) drift quantitatively tangible (Kim et al., 2014; Hamilton et al., 2016a; Hilpert and Correia Saavedra, 2017). However, most of the research in this area has been limited to studies based on single diachronic corpora. The same is true of studies on political language usage, which either use single or structurally comparable corpora. In this paper we explore ways of comparing lexical contexts associated with named entities, viz. political parties, across two corpora with substantially different structures and text types, representing the language of the Austrian media and the Austrian parliament, respectively.

Our approach is motivated by a socio-linguistic interest in how different domains of text production (media vs. parliament) shape political discourse. An additional aim is to investigate to what extent discourse is sensitive to political events, such as elections, and the changing representational roles of political parties that elections entail. While much research on these topics is carried out by close reading of relevant primary texts (Fairclough, 1995a; Wodak, 2010), we demonstrate how such qualitative analyses can be guided and complemented by quantitative methods that are both transparent and relatively simple. To that end, we analyze two diachronic corpora of Austrian German, namely a diachronic media corpus (AMC) and a corpus of parliamentary records (ParlAT, Section 3.). We focus on the lexical contexts associated with political parties in both corpora and measure their cross-temporal stability. Since the two corpora show substantial differences with respect to their structure and size, one of the methodological challenges consists in extracting data from the datasets that allow for meaningful comparison.

In what follows, we discuss our data and the methods used to analyze them in more detail. We present the analytical results from the cross-corpus comparisons and interpret them in relation to Austria's political history over the past 20 years.

## 2. Related Work

Most diachronic studies on either lexico-semantic change or political language usage are based on individual or more or less comparable corpora. Many of the recent computer linguistic advances in the area of semantic change tracking and detection have been based on the large Google books corpus or a genre-controlled sub-sample from it (Hamilton et al., 2016b; Dubossarsky et al., 2017; Rosenfeld and Erk, 2018). While the great advantage of using this resource lies in its unmatched size, it is not a balanced linguistic corpus in the strict sense. For that reason, the *Corpus of Historical American English* (COHA) is also often used for studying semantic change (Hamilton et al., 2016a; Eger and Mehler, 2016). In either case, however, the internal structure(s) of the corpora have been of minor relevance for these studies, which are mostly interested in global linguistic mechanisms and trends regarding lexical semantic change and are therefore not primarily content-focused.

In contrast, content and internal structure are of critical importance for studies of political language usage. Studies approaching political discourse from a qualitative perspective often exploit one specific type of political texts, such as parliamentary records (Ilie, 2010; Sealey and Bates, 2016; Archer, 2018; Truan, 2019; Waddle et al., 2019). Qualitative analyses comparing parliamentary records to other resources are much rarer (Ilie, 2004; Archakis and Tsakona, 2010). Similarly, quantitative approaches to political language are usually confined to one source of politically relevant texts (Huang et al., 2019).

Systematic comparative research of political language usage across structurally different corpora, particularly employing quantitative methods, is still outstanding, not least because of the challenges that such an approach faces. The present contribution explores some avenues towards that goal.

## 3. Data

Parliamentary records are a prime source for studying political discourse. They are published periodically according to a stable procedure, which makes them particularly valuable for diachronic investigations, and even though they usually undergo some amount of editing, their almost verbatim character renders them closer to spoken discourse than related sources (Winters, 2017). A second type of texts that is also commonly used for studying both political discourse and language change is newspapers and media publications more broadly (Böhning, 2017; Gloning, 2017). Usually, these source types are used independently of each other. It is our aim to explore ways of studying them together.

### 3.1. Austrian Media Corpus

The Austrian Media Corpus (AMC) (Ransmayr et al., 2013) is a diachronic text corpus containing Austrian newspapers, magazines, press releases, transcribed television interviews, news stories from television etc. from the last 30 years. It was created as part of a public-private cooperation between the Austria Press Agency (APA) and the Austrian Centre for Digital Humanities (ACDH) at the Austrian Academy of Sciences (ÖAW). With over 44 million articles, it is one of the largest text corpora for German and definitely the largest for Austrian German. As it is a monitor corpus, new material is being processed and added continuously. The linguistic data has been tokenized, lemmatized and part-of-speech tagged. In all, it contains 10.5 billion tokens representing 40 million word forms and 33 million lemmas. Even though the AMC includes data from a longer time span, we only use data covering the years 1997 to 2016 in the present analysis, which coincides with the duration of six successive Austrian governments. We also restricted the data set to the newspaper sub-corpus which has 5.5 billion tokens.

### 3.2. Corpus of Austrian Parliamentary Records

The Corpus of Austrian Parliamentary Records (ParlAT) contains the parliamentary records of the National Chamber (*Nationalrat*) – one of two chambers of the Austrian parliament. At present, ParlAT covers the official transcripts (from shorthand) from the XXth to the XXVth legislative periods (1996–2017) (Wissik and Pirker, 2018). Besides being tokenized, part-of-speech tagged and lemmatized, ParlAT also contains special TEI markup in accordance with the Parla-CLARIN guidelines (Erjavec and Pančur, 2019). All speeches delivered by members of parliament (as well as unauthorized interjections by members) are marked up as utterances `<u>` and each speaker is identified and marked up, accordingly. Thus, every utterance can be linked to a specific speaker. Additional comments and notes supplied by the stenographers are also encoded (e.g. applause etc.). The corpus consists of approximately 75 million tokens representing over 600 000 word forms and 400 000 lemmas. Again, for the present study we only use the years 1997 to 2016.

### 3.3. Data preprocessing

Our basic aim is to compare the language used to talk about the parties in the media to the language used by party members themselves in parliament. This leads to a fundamental problem regarding the comparability of the data: one of the corpora (ParlAT) is made up of texts by individual speakers, whose party affiliations are relevant for our purposes, while the other (AMC) is made up of texts whose authorship is irrelevant. Thus, we had to preprocess our data in a specific way in order to make them amenable to comparative study.

First it was necessary to determine which units of linguistic analysis were to represent political discourse. Based on the assumption that political topics and concepts are most emblematically represented by common nouns (such as *Arbeit* 'work, employment', *Marktwirtschaft* 'market economy' or *Nation* 'nation'), we limited our selection to this word class. It has been shown that nouns are most sensitive to semantic changes caused by cultural shifts (Hamilton et al., 2016a). We extracted all common nouns (by their lemmas) from the two corpora and applied stop words filtering. The list of stop words included numerals, the names of months and days of the week as well as the titles of officials (i.e. councillor, president, etc.), which were considered to be uninformative.

Next, we created subcorpora for each political party per year (from 1997 to 2016). The following political parties were included in the analysis: the Austrian People's Party (ÖVP), the Social Democratic Party of Austria (SPÖ) and the Freedom Party of Austria (FPÖ/BZÖ). The latter covers both the original Freedom Party as well as a splinter group – the Alliance for the Future of Austria – which formed in 2005 and took over the FPÖ's role in government. Because of personal and thematic continuities between the two, we decided not to separate them in the current study. Moreover, the Austrian Green Party (Die Grünen) were excluded due to potential confusion in the AMC with a German party of the same name.

Due to the different annotation structures in the two corpora and the fact that they represent markedly different types of texts in general, we had to define our notion of 'lexical contexts' in different ways. For ParlAT, it was the lexical items that politicians actually used in their speeches that we were primarily interested in, so 'lexical context' in this case denotes the set of common nouns that occurred in the party members' speeches. The process of linking speech to party was rather straightforward, since speaker IDs for every utterance can be linked to metadata including the speakers' party affiliations. Only speeches by elected representatives were included, whereas procedural utterances (e.g. by the chair) as well as interjections were omitted.

In order to obtain comparable subcorpora for the AMC, representing discourse *about* rather than *by* the respective parties, we extracted context windows around the party names (*SPÖ*, *ÖVP*, *FPÖ*, *BZÖ*) as they occurred in the text material. A window length of 20 words (10 nouns preceding and 10 nouns following a party name) was chosen, which is

analogous to the median length (19 words) of the selected utterances in ParlAT. In sum, we compiled 120 subcorpora (one for each party in each year for each corpus), which we take to represent the lexical contexts of the parties across the 20-year investigation period.

## 4. Methods

The majority of studies on computational detection of diachronic change in word usage and meaning make use of a distributional semantics approach, and in particular, prediction-based word embedding models (Kutuzov et al., 2018; Tahmasebi et al., 2018). However, state-of-the-art word embedding models are rather sensitive to the amount of the data used for training. Apart from the fact that various subcorpora from our dataset are not sufficiently large to train a word embedding model, the specific ways in which we preprocessed our data, as determined by our comparative research interest, makes the application of word embedding models problematic. This is particularly so, since the already relatively small ParlAT corpus needs to be split into year-wise subcorpora in order to make diachronic comparisons possible.

Thus, we opted for a simpler but at the same time more accessible approach for investigating the lexical stability in the contexts of target words (in this case, party names) over time and across domains. Since it has been shown that semantic shifts can be usefully quantified by means of the Jaccard index (Jaccard, 1912) (i.e. the size of the intersection of two sets divided by the size of their union), we used it as a measure of similarity between two sets of words (Buntinx et al., 2017; Rodina et al., 2019) representing either different years, parties or corpora. Furthermore, we employed statistical modeling of time series to analyze the diachronic dynamics of the resulting Jaccard index values.

### 4.1. Word set statistics

In order to address the issue of the small sizes of the subcorpora for each party per year and their uneven distribution, we applied Jaccard distance to equally sized sets of words. Thus, for each year and political party under consideration we created a set of distinctive words, which we take to be characteristic of that party in that particular year. We examined several statistical measures (pointwise mutual information word co-occurrence matrix counts, logistic regression coefficients, cosine similarity of count-based word vectors, etc.) to obtain these characteristic word sets for each subcorpus. However, only two of these were found to be reliable and useful with regard to our research interest, namely word frequency and a $\chi^2$-based keyword measure. All other measures under considerations yielded small intersections of word sets in diachronic comparisons, making similarity estimates unreliable.

The former statistic simply consists of the N most frequent words in a subcorpus. The $\chi^2$-based keyword metric is calculated as follows: First, to measure distinctiveness of words in a subcorpus we ran a $\chi^2$-test for all the word frequencies in the party subcorpus for the year X against the aggregated word frequencies for the remaining parties' subcorpora for the year X; then, we filtered the resulting statistics based on the p-value ($p < 0.05$) as well as on positive/negative distinctiveness of the words, i.e. we only included words with a positive $\chi^2$ statistic, representing words with a significantly higher occurrence likelihood for a given party compared to the other parties; and, finally, we sorted words by their $\chi^2$ value and took N words with the highest value. Set size was chosen in such a way that noise is minimized. Sets of 200 words were found most informative and methodologically robust. Smaller sets were found to be overly sensitive to year-wise fluctuations, often producing values close to zero for any given year, while larger sets did not substantially alter the results. We conducted comparative analyses of Jaccard similarity in three different ways. First, in order to detect changes in the lexical sets for each party over time, we calculated the Jaccard similarity between the word set for any given year and the very first year to see to what extent the lexical sets had shifted. Second, for each year we computed pairwise similarity values between the word sets of the three parties to see to what extent their lexical usage overlapped. Third, for each party and year we examined the similarity between the word sets from the two corpora.

### 4.2. Time series modeling

Time series of similarity measures were modeled by means of generalized additive models (GAM) (Wood, 2017), in which time was implemented as a smooth predictor term. This allows a more fine-grained inspection of successive patterns of convergence and divergence between the word sets compared to standard linear regression models. In a graph represention of the model, the non-linear dependencies between variables become visible as curves. The number of knots in the smooth term (i.e. how flexible we allowed the curves to be) was optimized based on the model's Akaike Information Criterion (AIC), a measure of a model's goodness of fit that also considers complexity. This retains maximal informativity of the model while avoiding undue sensitivity to individiual data points. Autocorrelation in the time series was accounted for through autoregressive modeling (AR(1)) (Akaike, 1969). For computations, the R libraries `mgcv` (Wood, 2011) and `itsadug` (van Rij et al., 2017) were used.

## 5. Results

We conducted three sets of comparative analyses with the nominal word sets extracted from AMC and ParlAT (see Section 3.). The purpose was to establish a) how stable or changeable the noun vocabularies associated with the three major parties were in the two discourse domains during the 20 years under investigation, b) how similar the vocabularies linked to the individual parties were to one another, c) how much similarity there was between the vocabularies used by the parties in parliamentary speeches on the one hand and by the media in their reporting on the parties on the other. Additionally we asked whether any changes observed in these measures could be related to political events during that time.[1]

[1] All Jaccard indices can be found at https://drive.google.com/drive/folders/1m9Nuv1M6lac81aijiE-QXEPBJCj8J8T0?usp=sharing.

## 5.1. Lexical stability per party

First, we measured lexical stability both in the media's coverage of the three main political parties in Austria and in the parties' own language as used in parliament. The time series displayed in Figures 1 and 2 trace Jaccard indices (JI) of the $\chi^2$-based keyword sets for each party, where the JI for any specific year represents the amount of lexical overlap between the set for that year and the set for the very first year of the investigated period (i.e. 1997). In essence, this measure gauges to what extent discourses by and about parties moved away from the point of departure.

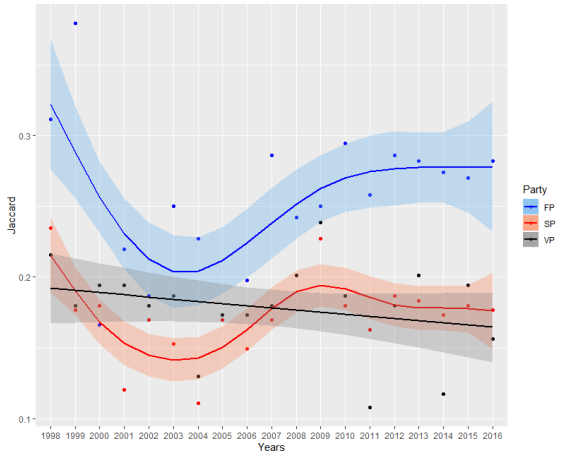


Figure 1: AMC, Jaccard index per party, 1997–2016, lexical similarity to first year, lexical sets based on $\chi^2$-tests (n = 200 per party per year)

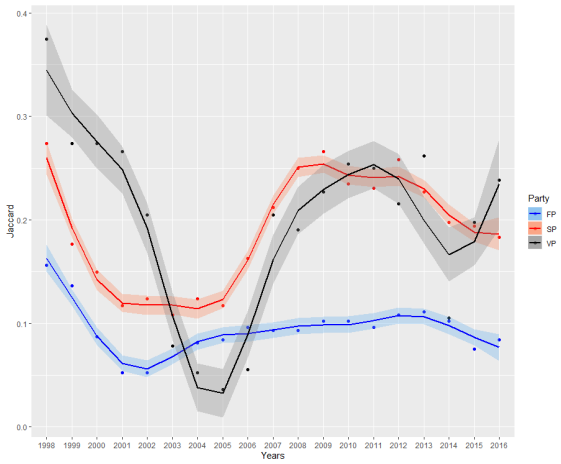


Figure 2: ParlAT, Jaccard index per party, 1997–2016, lexical similarity to first year, lexical sets based on $\chi^2$-tests (n = 200 per party per year)

In the AMC, the keyword sets for the right-wing FPÖ/BZÖ and the centre-left SPÖ undergo significant changes in the first half of the period, represented by a significant drop in JI values. In the second half, the sets regain similarity with the keyword sets of the first year. The JIs for the two parties vary between 0.11 and 0.38, i.e. about 20% to 55% of the 200 keywords is shared between the years. For the centre-right ÖVP, no significant changes can be detected, the JIs hovering around 0.18, i.e. roughly 30% overlap.

In ParlAT, similar patterns emerge for FPÖ/BZÖ and SPÖ, as both parties witness significant drops in lexical similarity to the first year, and again partly revert to the original keyword sets during the second half of the period. Here, the ÖVP also sees significant changes paralleling those of the other parties. JIs for all three parties oscillate between 0.04 and 0.38 (i.e. between 8% and 55% overlap). These findings indicate that the media discourse related to the ÖVP (as found in the AMC) is generally less variable compared to the other parties, even though all parties exhibit substantial variability in ParlAT. It is also worth noting that in the AMC the discourse surrounding the FPÖ/BZÖ remains relatively more faithful to its initial state compared to the other parties, while in ParlAT FPÖ/BZÖ generally exhibits lower values.

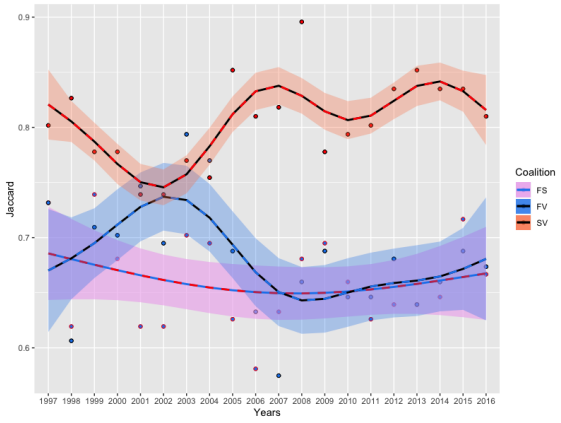


Figure 3: AMC, Jaccard index per party, 1997–2016, lexical similarity between parties, lexical sets based on frequency of occurrence (n = 200 per party per year)

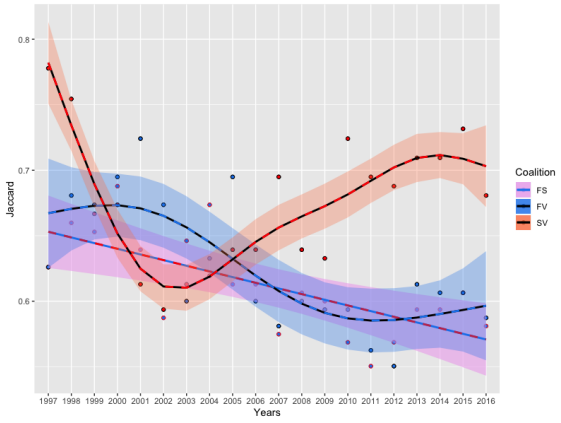


Figure 4: ParlAT, Jaccard index per party, 1997–2016, lexical similarity between parties, lexical sets based on frequency of occurrence (n = 200 per party per year)

## 5.2. Lexical similarity between parties

It is intriguing to relate the apparent slump in lexical stability during the first half of the investigated period to a major change in government: in 2000, the FPÖ (later BZÖ) entered into a coalition government with the ÖVP, which

lasted (over two legislative periods) until 2007. To further explore this observation, we next investigated whether the political vocabularies associated with two political parties exhibit higher similarity metrics during years when the parties participated in a coalition government. For this analysis, we calculated between-party JIs from annual word sets consisting of the 200 most frequent common noun lemmas associated with each party in the two corpora. For this step, simple frequency-based sets were preferred over $\chi^2$ -based sets, since the latter by definition represent party-specific distinctive vocabularies, which would depress the JI measuring lexical overlap between parties. The results in Figures 3 and 4 bear out this expectation. In both the AMC and ParlAT the similarity metrics representing the shared noun vocabulary of the 'Grand Coalition' parties (SPÖ and ÖVP) are significantly reduced in the years of the ÖVP–FPÖ/BZÖ governments, while lexical similarity between the right and centre-right parties is elevated during that time. In contrast, lexical similarity between the parties never forming a coalition government (SPÖ and FPÖ/BZÖ) seems stable at a lower level throughout the 20 years. In addition, there seems to be more lexical overlap in the AMC between SPÖ and ÖVP even during years when they did not form a government compared to the remaining non-governing party combinations. This is indicated by higher JI values generally.

We further tested the correlation between lexical similarity and participation in government by constructing a simple linear model from the same data as above, with JI as the output variable, participation in government (GOV) as a two-valued categorical predictor variable (Gov, NoGov). The (hypothetical) coalitions (COAL) were also added as an interacting predictor variable (JI∼GOV*COAL). Figure 5 and Tables 1 and 2 show that in both corpora lexical similarity between two parties is higher when they are in a coalition government together. Only in ParlAT the difference reaches statistical significance ($p < 0.001$), however, while in the AMC the difference is marginally significant ($p = 0.0582$). The models also confirm that there is a higher baseline similarity between the SPÖ and the ÖVP ($p < 0.001$) compared to other party combinations. In ParlAT, the identity of the parties does not add significantly to the predictiveness of the model.

| **Pred.** | **Levels** | **Est.** | **SE** | ***Z*** | ***p*** |
|---|---|---|---|---|---|
| Intercept | | .65 | .02 | 30.02 | <2e-16 |
| GOV | NoGov | -.04 | .02 | -1.93 | .06 |
| COAL | FP/VP | .05 | .03 | 1.92 | .06 |
| | SP/VP | .17 | .02 | 9.57 | 2.64e-13 |
| GOV× COAL | NoGov× FP/VP | .01 | .03 | .23 | .82 |

Table 1: Table AMC model, Formula: (JI∼GOV*COAL).

### 5.3. Lexical similarity across corpora

Up to this point, word sets from the two corpora have been analyzed separately, and any comparisons between them have rested on correlation tests with the corpus-specific JIs as input. In a third and final step, we addressed the question whether there is also a cross-corpus overlap between the word sets themselves and whether we could identify tendencies towards lexical convergence or divergence between the two discourse domains. For this analysis, we again relied on $\chi^2$-based keyword sets. In this case, the JIs for each party in each each year represent the amount of lexical overlap between the keywords from the AMC contexts, representing media discourse about the parties, and the keywords extracted from ParlAT, representing the parties' own use of language.

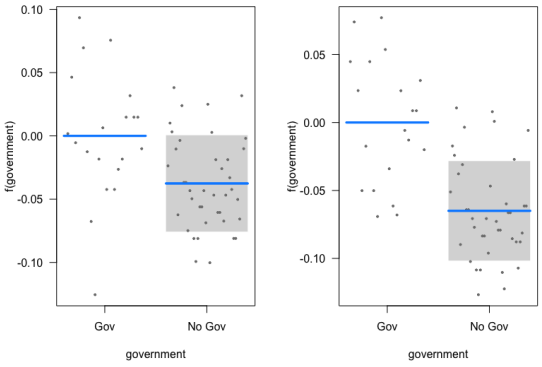


Figure 5: Linear regression models for AMC (left) and ParlAT (right), Formula: (JI∼GOV*COAL).

| **Pred.** | **Levels** | **Est.** | **SE** | ***Z*** | ***p*** |
|---|---|---|---|---|---|
| Intercept | | .68 | .02 | 32.56 | <2e-16 |
| GOV | NoGov | -.07 | 0.02 | -3.48 | .001 |
| COAL | FP/VP | -.03 | .03 | -1.06 | .29 |
| | SP/VP | .02 | 0.02 | 1.38 | 0.17 |
| GOV× COAL | NoGov× FP/VP | .02 | .03 | 0.86 | .40 |

Table 2: Table ParlAT model, Formula: (JI∼GOV*COAL).

As in the previous analyses, the results suggest a temporal split between the first and the second half of the investigation period, roughly corresponding to the changes in governing coalitions (Figure 6). Interestingly, two of the parties behave in an almost antithetical way: where the lexical sets from the two corpora tend towards greater convergence for the FPÖ/BZÖ, they diverge for the SPÖ, and vice versa ($r(18) = -0.86$, $p < 0.001$). The ÖVP takes an intermediate position: at first, its cross-corpus similarity metrics align more closely with those of the FPÖ/BZÖ, but after a peak during the early years of the the right/centre-right coalition government soon fall back to a trajectory that is counter-cyclical to that of the FPÖ/BZÖ and similar to that of the SPÖ (ÖVP vs. FPÖ/BZÖ: $r(18) = -0.45$, $p < 0.05$; ÖVP vs. SPÖ: $r(18) = 0.45$, $p < 0.05$). It should be noted that the JI measures in this analysis are generally smaller than those found for lexical stability per party within corpora (cf. Section 5.1.). JIs range between 0.02 and 0.21, which corresponds to between c. 4% and c. 35% shared nominal

keyword vocabulary associated with the parties across the two domains.

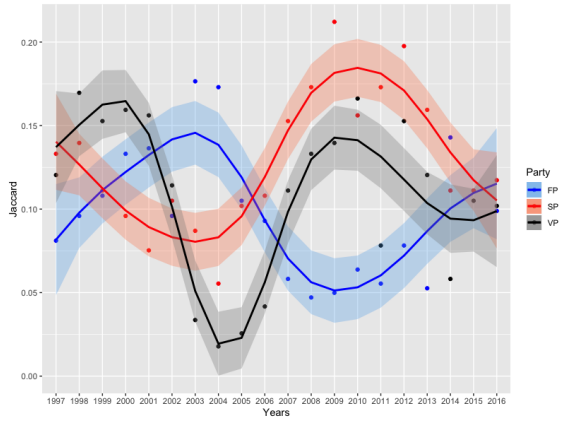


Figure 6: AMC, ParlAT, Jaccard index per party, 1997–2016, lexical similarity between corpora, lexical sets based on $\chi^2$-tests (n = 200 per party per year)

## 6. Discussion

A common theme running through all the above analyses is that lexical usage in Austrian political discourse is sensitive to changes in political realities. More importantly for present purposes, the sensitivity of language to real-world events can be traced and explored comparatively with the help of the two corpus resources presented here, the AMC and ParlAT.

Thus, it emerges from the empirical analysis that changes in governing coalitions have a visible and robust impact both on the lexical inventory used by the media to report on the three major parties and on the vocabulary used by party representatives in parliament. This is evident in the way that lexical usage adapts to a party's democratic role during a legislative period (Section 5.1.), and also in the way that the parties' speech patterns converge when they join to form a government (Section 5.2.).

These findings are not altogether unexpected, considering that the change of a party's role within the structures of a representational democracy goes hand in hand with changes in executive responsibilities and procedural matters. At the same time, a closer inspection of the keyword sets for the individual parties also suggests that lexical differences between governing and opposition roles reflect different strategies of self-representation (Gruber, 2015). When in government, the SPÖ generally uses more positive and dynamic vocabulary, including *Arbeit* 'work, employment', *Lösung* 'solution', *Möglichkeit* 'opportunity', *Maßnahme* 'measure, action', *Projekt* 'project' and *Erfolg* 'success'. In opposition, the same party's vocabulary is more antagonistic and critical, including *Forderung* 'demand, request', *Kritik* 'criticism', *Problem* 'problem', as well as a wider range of unconcealed expressions of rebuke, such as *Chaos* 'chaos', *Desaster* 'disaster', *Doppelspiel* 'duplicity', or *Ellbogengesellschaft* 'elbow society', none of which features prominently in the party's speech when in government. Some of these tendencies seem to carry over to the AMC, where *Lösung*, *Maßnahme* and *Möglichkeit* are also among the most prominent keywords associated with the SPÖ while in power. Findings such as these can serve to complement studies on how politicians defend their own record (Sealey and Bates, 2016) and negotiate differences (Harris, 2001; Archer, 2018; Waddle et al., 2019) within the confines of decorum and parliamentary rules.

Another finding worth commenting on highlights how the lexical effects of being in government may sometimes differ between parties. As seen in Section 5.3., the lexicon associated with the FPÖ/BZÖ during the right/centre-right government showed a much greater degree of convergence between parliamentary and media discourse relative to that of its coalition partner ÖVP. This could be interpreted as evidence that the FPÖ was generally more successful in having topics or its way of speaking picked up by the media. Without a closer reading of the source materials, it is not immediately clear if this was indeed the case. It is striking, however, that many of the most widely dispersed keywords linked to the FPÖ/BZÖ in both corpora during this time designate individuals, such as *Person* 'person', *Kollege* 'colleague', *Abgeordnete* 'representative, MP', *Freiheitliche* 'member of the freedom party' and *Mitglied* 'member'. This may be a reflection of internal conflicts within the FPÖ during this time, including a party coup in 2002 (known in Austria as 'Knittelfeld' after the venue of the coup) and the eventual break-up into two parties, FPÖ and BZÖ, in 2005. These tensions and their effect on parliamentary debate may well have had a more attractive media appeal than the ÖVP's contributions, a difference that integrates well with conceptions of contemporary politics that distinguish between 'frontstage' and 'backstage' politics (Wodak, 2010).

Equally intriguing are differences between the two corpora: For example, we found little evidence to suggest that the way that the only party in power throughout the 20-year period, the ÖVP, was represented in the media changed much at all (see Section 5.1.). At least based on the JIs measuring how much of the keyword vocabulary matches the first year's vocabulary, there was little movement over time. This differs starkly from the way that lexical items characterise the speech of the ÖVP in parliament, being subject to some of the strongest fluctuations of all parties. Continuity in government may level out media coverage, but the same may not necessarily hold true for the language in parliament. Findings such as these may prompt closer investigations of disparities between what a party does in parliament and what is said about the party in the wider public discourse (Wodak, 2010).

Finally, the results provide some basis to speculate about how a party's positioning in the media discourse may differ from its role in parliament. Thus, the findings in Section 4.2. imply that the so-called 'centrist' parties (i.e. SPÖ and ÖVP) display a significantly larger amount of lexical similarity in the media compared to how much keyword vocabulary either of them shares with the right-wing FPÖ/BZÖ. Importantly, this effect is independent of whether the centrist parties formed a coalition government or not. In contrast, no such elevated baseline of lexical overlap between the centrist parties could be observed in ParlAT: here joint

participation in government turned out to be the only factor significantly influencing the amount of overlap between parties. Once again, this points to a potential disconnect between the two domains of political discourse.

Suggestive as these findings are, many of the points made above must await further study, either qualitatively by applying the analytical tools developed within the field of discourse analysis (Fairclough, 1995a; Fairclough, 1995b; Wodak and Meyer, 2001), or with a more sophisticated set of quantitative methods, including stylometric analysis of individuals or groups of speakers (Huang et al., 2019), computer-assisted content analysis and topic modeling for extracting political positions (Laver et al., 2003; Proksch and Slapin, 2010; Lauderdale and Herzog, 2016), or sentiment analysis (Taboada, 2016). Nonetheless, this study has demonstrated that a comparative analysis of two corpora with related contents but markedly different internal structures can succeed in yielding insightful and stimulating results, with great potential for the study of political discourse. Within the field of digital humanities, relatively simple and transparent methods such as the ones applied in this study can assist in identifying global trends in the compared data and point out areas of interest in the corpus data for closer scrutiny.

## 7. Conclusion

In this paper, we examined ways of comparing the stability and similarity of lexical usage across two corpora covering the same time period but otherwise exhibiting substantial differences in terms of annotation and content. We addressed these questions by means of a case study focusing on the lexical contexts associated with major Austrian political parties in two different diachronic corpora, i.e. AMC and ParlAT. We identified and discussed changes in the lexical contexts associated with political parties over time, between the parties and across the corpora. Furthermore, we were able to relate the results of the comparative analysis to real-world events.

## 8. Acknowledgements

The project *Diachronic Dynamics of Lexical Networks (DYLEN)* is funded by the ÖAW go!digital Next Generation grant (GDNG 2018-02).

## 9. Bibliographical References

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, 21(1):243–247.

Archakis, A. and Tsakona, V. (2010). The wolf wakes up inside them, grows werewolf hair and reveals all their bullying': The representation of parliamentary discourse in greek newspapers. *Journal of Pragmatics*, 42:912–923.

Archer, D. (2018). Negotiating difference in political contexts: An exploration of Hansard. *Language Sciences*, 68:22–41.

Buntinx, V., Bornet, C., and Kaplan, F. (2017). Studying linguistic changes over 200 years of newspapers through resilient words analysis. *Frontiers in Digital Humanities*, 4:2.

Böhning, H. (2017). Zeitungen und Sprachentwicklung. Beobachtungen zu den ersten eineinhalb Jahrhunderten deutscher Zeitungen. In Oliver Pfefferkorn, et al., editors, *Die Zeitung als Medium in der neuen Sprachgeschichte. Korpora - Analyse - Wirkung*, pages 7–21. Walter de Gruyter, Berlin/Boston.

Dubossarsky, H., Grossman, E., and Weinshall, D. (2017). Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1147–1156.

Eger, S. and Mehler, A. (2016). On the linearity of semantic change: Investigating meaning variation via dynamic graph models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics,*, pages 52–58.

Erjavec, T. and Pančur, A. (2019). Parla-CLARIN: TEI guidelines for corpora of parliamentary proceedings. In *Book of abstracts of the TEI2019: What is text, really? TEI and beyond*. University of Graz.

Fairclough, N. (1995a). *Media discourse*. Arnold, London/New York.

Fairclough, N. (1995b). *Critical Discourse Analysis*. Longman, London.

Gloning, T. (2017). Alte Zeitungen und historische Lexikographie. Nutzungsperspektiven, Korpora, Forschungsinfrastrukturen. In Oliver Pfefferkorn, et al., editors, *Die Zeitung als Medium in der neuen Sprachgeschichte. Korpora - Analyse - Wirkung*, pages 121–147. Walter de Gruyter, Berlin/Boston.

Gruber, H. (2015). Policy-oriented argumentation or ironic evaluation: A study of verbal quoting and positioning in Austrian politicians' parliamentary debate contributions. *Discourse Studies*, 17(6):682–702.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016a). Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121.

Hamilton, L. W., Leskovec, J., and Jurafsky, D. (2016b). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Harris, S. (2001). Being politically impolite: Extending politeness theory to adversarial political discourse. *Discourse Society*, 12(4):451–472.

Hilpert, M. and Correia Saavedra, D. (2017). Using token-based semantic vector spaces for corpus-linguistic analyses: From practical applications to tests of theoretical claims. *Corpus Linguistics and Linguistic Theory*.

Huang, L., Perry, P. O., and Spirling, A. (2019). A general model of author 'style' with application to the UK House of Commons, 1935-2018. Working Paper, https://www.nyu.edu/projects/spirling/documents/VeryBoring.pdf.

Ilie, C. (2004). Interruption patterns in british parliamentary debates and drama dialogue. In *Dialogue Analysis*

*IX: Dialogue in Literature and the Media, Part 1: Literature: Selected Papers from the 9th IADA Conference*, pages 311–326.

Ilie, C. (2010). Strategic uses of parliamentary forms of address: The case of the u.k. parliament and the swedish riksdag. *Journal of Pragmatics*, 42:885–911.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50.

Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D., and Petrov, S. (2014). Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65.

Kutuzov, A., Øvrelid, L., Szymanski, T., and Velldal, E. (2018). Diachronic word embeddings and semantic shifts: A survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397.

Lauderdale, B. and Herzog, A. (2016). Measuring political positions from legistlative speech. *Political Analysis*, 24(3):374–394.

Laver, M., Benoit, K., and Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(2):311–313.

Proksch, S.-O. and Slapin, J. B. (2010). Position taking in european parliament speeches. *British Journal of Political Science*, 40(3):587–611.

Ransmayr, J., Mörth, K., and Ďurčo, M. (2013). Linguistic variation in the Austrian Media Corpus: Dealing with the challenges of large amounts of data. In *Procedia - Social and Behavioral Sciences 95. Proceedings of the 5th International Conference on Corpus Linguistics (CILC 2013)*, pages 111–115.

Rodina, J., Bakshandaeva, D., Fomin, V., Kutuzov, A., Touileb, S., and Velldal, E. (2019). Measuring diachronic evolution of evaluative adjectives with word embeddings: the case for English, Norwegian, and Russian. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 202–209.

Rosenfeld, A. and Erk, K. (2018). Deep neural models of semantic shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 474–484.

Sealey, A. and Bates, S. (2016). Prime ministerial self-reported actions in Prime Minister's Questions 1979-2010: A corpus-assisted analysis. *Journal of Pragmatics*, 104:18–31.

Taboada, M. (2016). Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics*, 2:325–347.

Tahmasebi, N., Borin, L., and Jatowt, A. (2018). Survey of computational approaches to diachronic conceptual change. *CoRR*, abs/1811.06278.

Truan, N. (2019). Talking about, for, and to the people: Populism and representation in parliamentary debates on Europe. *Zeitschrift für Anglistik und Amerikanistik*, 67:307–337.

van Rij, J., Wieling, M., Baayen, R. H., and van Rijn, H. (2017). itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs. R package version 2.3.

Waddle, M., Bull, P., and Böhnke, J. R. (2019). 'He is just the nowhere man of British Politics': Personal attacks in Prime Minister's Questions. *Journal of Language and Social Psychology*, 38(1):61–84.

Winters, J. (2017). Tackling complexity in humanities big data: From parliamentary proceedings to the archived web. *Studies in Variation, Contacts and Change in English*, 19.

Wissik, T. and Pirker, H. (2018). ParlAT beta Corpus of Austrian Parliamentary Records. In *Proceedings of the LREC 2018 Workshop 'ParlaCLARIN: LREC2018 workshop on creating and using parliamentary corpora'*, pages 20–23.

Wodak, R. and Meyer, M. (2001). *Methods of Critical Discourse Analysis*. SAGE, London.

Wodak, R. (2010). *The discourse of politics in action: Politics as usual*. Palgrave Macmillan, Basingstoke.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3–36.

Wood, S. (2017). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2 edition.