From Linguistic to Discursive Patterns: Introducing Discoursemes as a Basic Unit of Discourse Analysis

Critical Approaches to Discourse Analysis across Disciplines
www.cadaadjournal.com
Vol 16 (2): 87 – 111

PHILIPP HEINRICH
Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
philipp.heinrich@fau.de

ANDREAS BLOMBACH
Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

NATHAN DYKES
Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

STEPHANIE EVERT
Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

TAMARA FUCHS
Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

LINDA HAVENSTEIN
Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

FABIAN SCHÄFER
Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

Abstract
Corpus-assisted discourse studies (CADS) combine corpus linguistics and (critical) discourse analysis to explore how language reflects and shapes discourse. CADS researchers derive macro-level social explanations based on micro-level description of texts. Interpretation (the meso level of discourse analysis) is mostly accomplished by deriving discursive patterns from linguistic patterns observed across many texts. CADS blends close reading (examining individual examples in concordances) and distant reading (looking at results of keyword and collocation analyses). While effective, we argue that CADS lacks a true integration of qualitative and quantitative techniques, which results

in a unidirectional workflow where qualitative-hermeneutic interpretation is detached from quantitative analysis. To bridge this gap, we propose operationalising the grouping of linguistic surface realisations in terms of discoursemes – building blocks for discourse analysis. Discursive patterns can then be approximated by co-occurrences of discoursemes. We demonstrate the usefulness of our approach by means of a case study analysing the discourse related to refugees in the German federal parliament during two salient moments in Germany's history. The case study is carried out using a new open-source software toolkit that facilitates the construction of a consistent database of discoursemes and overcomes some of the technical limitations faced by most CADS studies.

Key words: *Corpus-assisted Discourse Studies (CADS); Critical Discourse Analysis (CDA); German parliamentary debates (GermaParl); Corpus Workbench (CWB); discoursemes*

1. Introduction

Corpora have been an important addition to discourse studies over the last decades. Combining discourse analysis with the methods of corpus linguistics makes studies more systematic and improves their reproducibility. Since a corpus-linguistic approach enables researchers to analyse large text collections in full, they do not have to face the common criticism of using only 'cherry-pick[ed] small and unrepresentative data samples in order to suit [their] preconceived notions about hidden ideological meanings' (Mautner, 2009, p. 34; see also Baker et al., 2008; Bubenhofer et al., 2019; Rheindorf & Wodak, 2020 Wodak, 2015a). Consequently, the field of corpus-assisted discourse studies (CADS) aims to combine quantitative methods from corpus linguistics with the qualitative-hermeneutic interpretation process of Critical Discourse Analysis (CDA, see Appendix 1 for a list of abbreviations). Although CDA comprises a plethora of definitions, executions, and methodologies, it is generally agreed to be an interdisciplinary and multi-method approach to discourse analysis, focusing on the interrelation of societal power and language use (Fairclough et al., 2013; Reisigl & Wodak, 2009; van Dijk, 2015).

While quantitative methods of CADS operate on words and *linguistic patterns* (the 'micro level'), the underlying research questions usually address the interplay of discourse and society (the 'macro level'), far removed from individual words and even texts. The quantitative-qualitative combination in CADS happens primarily at a level in between (the 'meso level'), where analysts aim to derive *discursive patterns* (see Section 2) from the linguistic patterns observed in corpus data (by means of corpus-linguistic techniques such as reading concordances and comparing frequencies, see Section 3). An additional quantitative component is to determine the statistical distribution of these discursive patterns (across time, political parties, different newspapers, individual speakers, etc.). However, hermeneutic interpretation remains a central pillar in the process; close familiarity with the data as well as relating quantitative and qualitative levels of the analysis to each other are considered essential ingredients.

In this paper, we propose to operationalise discursive patterns in terms of *constellations* of *discoursemes*. We define discoursemes as (minimal) discursive units of lexical meaning in the context of a specific discourse. We argue that a harmonisation of the numerous methodological approaches on the meso level can be accomplished through a discourseme-based approach, which provides an operational connection between the qualitative and quantitative

aspects of discourse analysis. We have also developed an open-source toolkit as an integrated software environment for discourseme-based CADS.

2. Discourse and Critical Discourse Analysis

2.1 Methodological Pluralism

The very meaning of the word 'discourse' is multifaceted and varies between disciplines. In this paper, we deal with the methodological current of discourse analysis that sees itself in the tradition of Michel Foucault. This approach gains its coherence from the shared assumptions that

a) discourses form the objects of which 'they speak' (Foucault, 1972, p. 49) (without going so far as to equate the former and the latter),
b) discourse/knowledge and power are inextricably interconnected (CDA therefore critically untangles the power structures created by categories such as race, class, and gender),
c) researchers contribute to the construction of the discourse they study by selecting and analysing source materials.

Despite these basic commonalities, individual proponents of CDA use varying definitions of the term 'discourse'. According to Link, for example, discourses are 'institutionalised, regulated ways of speaking as spaces of possible statements that are linked to actions'[1] (Link, 2016, p. 121). Jäger (2015), on the other hand, describes discourses as 'trans-subjective producers of social reality and socio-cultural patterns of interpretation'[2] (p. 27), namely a 'flow of "knowledge" or social stock of knowledge through time'[3] (Jäger & Jäger, 2007, p. 23). The most radical position regarding the relationship of discourse and reality is taken by Laclau and Mouffe (2001), arguing against orthodox materialist Marxism that society (i.e. 'class') does not exist as the comprehensive, prediscursive material basis of all social processes, but is itself discursively constructed. Hence, despite building on Foucault, they reject his differentiation of discursive and non-discursive practices.

In contrast, the distinction between discursive and non-discursive practices is essential to Fairclough and Chouliaraki (cf. Chouliaraki & Fairclough, 1999; Fairclough, 2013, 2015). They define discourses as the 'semiotic elements of social practices', including 'language [...], nonverbal communication [...] and visual images [...]' (Chouliaraki & Fairclough, 1999, p. 38). Jäger uses Foucault's term 'dispositive' to encompass both discursive and non-discursive practices, namely the 'procedural interrelatedness of knowledge'[4] embedded into acts of 'speaking/thinking – acting – objectification'[5] (Jäger, 2015, p. 113). Wodak and Reisigl (2009) consider discourse as 'a cluster of context-dependent semiotic practices that are situated within specific fields of social action' (p. 89).

Despite these many conceptual differences, all CDA approaches share the goal of studying language in order to understand how language and discourse shape and reflect social power, relationships, and ideologies.

2.2 Defining the Meso Level of CDA

Beyond conceptual variations in defining discourse, proponents of CDA have also put forth different methodological approaches, resulting in substantial methodological pluralism regarding the operationalisability of Foucault's

meandering thought. At the risk of generalising, we can at least say that CDA deals with three different – albeit interconnected – analytical levels, the *micro*, *meso*, and *macro* level. Fairclough describes them as (a) description (micro level), (b) interpretation (meso level), and (c) explanation (macro level) (cf. Fairclough, 2015, pp. 58–59).[6] The micro level is concerned with language, namely the formal properties of an individual text and their relation to other texts and contexts (cf. Behnam & Mahmoudy, 2013; Fairclough, 2015). This includes structure, grammar, vocabulary, intertextuality, and rhetorical or literary devices (Johnson & McLean, 2020, p. 380). The macro level of social structures, namely the level of knowledge/power, refers to the 'relationship between discourse, ideology, and the sociomaterial world', including 'implicit and explicit rules, norms, or mores governing discourse and society' (Johnson & McLean, 2020, p. 380; cf. also Behnam & Mahmoudy, 2013; Fairclough, 2013, 2015). In their analyses, most discourse-analytical studies focus on the meso level, functioning as the connective tissue between the micro level of text (and its linguistic properties) and the macro level of discourse and non-discursive practices. Therefore, we will proceed to briefly summarise how the aforementioned authors have defined the meso level in order to develop an applicable methodology from the body of Foucault's work and his exegetes. This will lay the methodological groundwork for our approach to be fleshed out below.

The meso level of discourse analysis can be further divided into two sub-levels, examining different units of interpretation. Meso level 1 pertains to 'discursive fragments' (*Diskursfragmente*) that can accumulate to 'discursive strands' (*Diskursstränge*) (Jäger & Jäger, 2007, pp. 25–27; Jäger, 2015, p. 80), which may be described as topic-specific instances or segments of a discourse or text. Both can be sub-discourses of a larger discourse or partial overlaps between discourses. Meso level 2 is concerned with more fine-grained analysis. It targets discursive strategies, namely intra- and inter-discursive interpretations and contextualisations. This can include 'discursive strategies' and 'topics' (cf. Reisigl & Wodak, 2009), 'collective symbols' (*Kollektivsymbole*) (such as the word *flood*, framing immigrants as a potential threat), namely 'images of meaning collectively embedded in a culture' (Link, 2009, p. 42), or so-called 'nodal points', which 'articulate' (i.e. connect) discursive moments via what Laclau (1996) has called 'empty signifiers' (i.e. words or concepts that lack a fixed meaning, deriving their significance from their usage context, such as the term *the people* in populist discourses). All of these units can be referred to with the umbrella term of *discursive patterns* or, more broadly, as discursive positions (*Diskursposition*) (cf. Jäger 2015), which encompass framings of a concept, specific claims or attitudes towards a concept. Following Fairclough, Mulderrig and Wodak's (2013) use of 'semiosis' as an alternative term for discourse, namely as 'an analytical category describing the vast array of meaning-making resources available to us' (encompassing 'words, pictures, symbols, design, colour, gesture, and so forth'), we introduce the term 'discourseme' to describe these minimal discursive units of lexical meaning in the context of a specific discourse (see below).

3. Structuring Discourse in CADS

3.1 The Hermeneutic Grouping Process

One particularly successful way in which CADS studies have brought together quantitative and qualitative analysis is through the manual grouping of keywords and collocations, as discussed in-depth by Baker (2004). Both keywords and collocations refer to words (or other lexical items) that are statistically salient because of their high frequency: keywords are significantly more frequent in a given target corpus than in a reference corpus, whereas collocations are salient because of their high co-occurrence frequency with a given node word (often defining the topic of a discourse). Keywords and collocations thus highlight vocabulary that is strongly associated with a discourse or with certain actors within the discourse. After keywords or collocations have been calculated with the chosen parameter settings, the key hermeneutic contribution of the researcher lies in grouping them according to their shared discursive functions. This approach is based on the premise (shared with other semi-automatic approaches such as topic modelling) that discursive patterns can be characterised by words and other surface realisations (observed at the micro level). Thus, the grouping process operates on the *meso level* by relating these surface items to more general narratives (the macro level).

As a fairly typical example, consider a researcher who wants to study discourses on migration in a particular corpus (cf. our case study below). They may start by identifying a list of words or multiword units that are used to talk *about* migration (thus defining the topic of the discourse): *(im)migration*, *(im)migrant*, etc. They may then carry out a collocation analysis using this word list as a node (provided their software tool is not limited to single words as nodes) in order to identify salient words in the *context* of migration, and manually form groups of related collocates: e.g. one group comprising words such as *underage* and *unaccompanied*; another one comprising *displacement* and *expulsion*; etc. These groups are meant to reflect common discursive patterns: *underage* and *unaccompanied* both indicate vulnerable groups; *displacement* and *expulsion* indicate forced migration ('push factors' of migration).

The words in these groups are usually paradigmatically related to one another, i.e. they occur in similar linguistic contexts. In some cases, they are (near-)synonyms (e.g. *displacement* and *expulsion*). More generally, they often belong to the same lexical or semantic field – thus they may share similar meanings or be conceptually related to each other (Trier, 1931; Geeraerts, 2010). However, a key aspect of the hermeneutic grouping process is that analysts are not restricted to pre-defined semantic fields or to only grouping together words with common lexical semantic properties. Instead, as Baker (2004) points out, categories resulting from this analysis highlight the discursive function of the items in question as opposed to their general lexical semantics (p. 352). For instance, in his case study on a target corpus of erotic fiction stories, he identifies a category that he labels 'hypermasculinity' and which includes words like *football*, *beer*, and *army*. While these words clearly do not have similar referential meanings, in this specific target corpus, they all help to construct an archetype of a certain traditionally masculine gender role.

In a purely qualitative phase of the analysis that finally links to the macro level, the manually formed groups are interpreted as indicators of certain discursive patterns, which often necessitates a much broader perspective than what is expressed by the lexical items themselves. This is often supported by the inspection of concordance lines for each group or close reading of individual examples. For example, *underage* typically references a vulnerable group when combined with the topic of migration (which is implicit in the collocation analysis) but might refer to a legal category in other contexts within a wider migration discourse. Similarly, a keyword might only indicate a specific discursive pattern in the particular target corpus from which it was obtained. Labelling groups of collocates (or keywords) thus brings in many unspoken assumptions and generalisations, and the interpretations may only be valid for the particular collocation or keyword analysis at hand. A broader picture only emerges at the purely qualitative macro level.

Sometimes additional quantitative analyses look at the statistical distributions of certain words, often those seen as particularly evocative of a discursive pattern. However, this step is usually entirely separate from the keyword or collocation analysis and the manual formation of groups representing discursive patterns. The CADS process sketched here thus involves a combination of four *methodological paradigms*:

(a) empirical methods at the micro level (keywords, collocations, and concordances),
(b) qualitative-hermeneutic interpretation on the meso level (grouping of surface realisations and identification of discursive patterns),
(c) purely qualitative discussion of the broader discourse at the macro level, and
(d) quantitative analysis (statistical distribution of words indicative of discursive patterns) and corresponding macro-level explanations.

A key (and largely unsolved) challenge is how to combine sophisticated quantitative methods (for both (a) and (d)) effectively with the qualitative interpretation ((b) and (c)). Most existing work – not only in CADS but also in Digital Humanities and Applied Corpus Linguistics – leaves much to be desired in this respect: some studies are simply a collocation or keyword analysis (e.g. Grundmann & Krishnamurthy, 2010; Fox, 2006) others start by grouping related collocations and keywords but then develop discursive patterns through a purely qualitative interpretation (e.g. Poole, 2016; Partington, 2006). In fact, the same holds true for non-computational CDA, since 'the methodical implementation of the approach often turns into a deductionist, interpretative process, remaining vague with regard to its specific procedure' (Keller, 2011, p. 165).[7]

CADS studies often have to rely on subjective impressions or single observations (brought forth as illustrative examples in the paper) because the discursive patterns have not been operationalised in a way that would allow them easily to be searched in the corpus. Moreover, different settings or parameter choices in collocation and keyword analyses can only be explored via separate analyses that are brought together in a comparative qualitative interpretation, as the manual grouping process is carried out separately for each analysis (necessarily so, as explained above). A related problem is that the traditional workflow of moving between spreadsheets and concordances limits the interaction between the quantitative and qualitative perspectives. In the output that is typically produced by current corpus-linguistic software,

keywords and collocates are displayed as a table ranked by some statistical measure. In order to perform the manual grouping efficiently, researchers need to export this table from the concordancer programme and load it into a spreadsheet, where it can conveniently be filtered and reordered. At the same time, concordance displays are only available through the concordancer itself. Finally, CADS studies often fail to include multiword units such as *asylum seeker* when manually forming groups, simply because most tools for collocation and keyword analysis operate on single words for technical reasons.

The evolution of CADS research is thus held back by a lack of true integration of the quantitative and the qualitative realm. This also manifests in the fact that statistical distributions are obtained separately for selected words rather than for the discursive patterns that have been identified (as they should be). In summary, current CADS research lacks:

(i) an operationalisation of the interpretative process at the meso level (which would allow its results to be used for further quantitative analysis),
(ii) a feedback loop between qualitative and quantitative approaches, and
(iii) appropriate software tools supporting an integrated analysis.

3.2 Operationalising CADS: Introducing Discoursemes

Our approach to overcoming these limitations rests on conceptualising the central quantitative-qualitative step of CADS as the formation of *discoursemes*, which we define as (minimal) discursive units of lexical meaning in the context of a given discourse. A discourseme combines (i) an intensional meaning description at the qualitative level, (ii) an operationalisation in terms of lexical items at the empirical-quantitative level, and (iii) an extensional realisation (its instances across different corpora) as a basis for further quantitative analysis. It thus ties together all four of the methodological paradigms of CADS identified above. Our goal here is to provide an operational concept that has a clear hermeneutic definition (discursive unit of meaning in the context of a discourse) but can also be approximated via lexical items and thus identified automatically in corpora, forming the required link between qualitative and quantitative methods. We enclose references to discoursemes in bold vertical bars throughout this text; examples include the |hypermasculinity| discourseme mentioned above, topic-discoursemes (such as |migration|), metaphors (such as a |flood of people|), certain groups of agents (such as |political parties| or |right-wing parties|), and evaluation with regards to a certain aspect (such as |lazy|).

The lexical items forming the operationalisation of a discourseme are usually lemmata or word forms but may also include multiword expressions such as *asylum seeker* and *bring to bear*, multiword named entities such as *European Union*, or other linguistic constructions (Touileb & Salway, 2014). It is worth pointing out that not all occurrences of a lexical item will always belong to the corresponding discourseme. For instance, the lemma *flood* will typically be assigned to the metaphor discourseme |flood of people| in a migration context, but its occurrence in *displaced families are uprooted again by severe floods* does not belong to this discourseme. The operationalisation of a discourseme in terms of manually created groups of lexical items (usually obtained from a collocation or keyword analysis) must thus be considered an *approximation*, since there will be both false positives (instances of these items that do not in

fact belong to the discourseme) and false negatives (e.g. occurrences of the discourseme that are realised through pronominal anaphora or coreference). Sometimes, partial disambiguation can be achieved through additional corpus annotation such as word-sense disambiguation (Word-Sense Disambiguation [WSD], e.g., between *displacement* of people and the *displacement* of a car engine). However, WSD tools are still unreliable and do not take into account the discourse-specific meaning of words. In practice, analysts will have to carry out at least a cursory concordance analysis of each lexical item to be included in a discourseme, in order to ensure that a majority of its occurrences have the intended discourse-specific meaning (and, in the case of collocation analysis, that its co-occurrence with the node is not coincidental).

Note that in our case study below, we form discoursemes by grouping together lemmata rather than inflected word forms. This is first and foremost an opportunistic choice: especially in inflecting languages such as German, forming manual groups is more convenient on the basis of lemmata because the different inflected forms will often be assigned to the same discourseme. Moreover, lemmatisation often improves the statistical analysis of collocates and keywords: pooling all inflected forms reduces data sparseness and shows associations that might remain hidden when looking at individual word forms (Evert, 2005, p. 35, fn. 3). By contrast, the definition of topic-discoursemes via lemmata is merely a matter of convenience (since all inflected forms could simply be listed).

Our approach also recognises explicitly that discursive patterns do not arise from individual discoursemes in most cases (as the qualitative interpretation in traditional CADS might suggest), but rather from *constellations* of discoursemes. The discourseme |flood of people| might comprise words like *flood*, *surge*, or *pour into*, but they only evoke the discursive pattern 'migrants as a flood of people' when used in conjunction with |migration| or a similar topic-discourseme. Such constellations are often implicit in CADS studies: e.g. groups of collocates form discoursemes that co-occur in a constellation with the topic-discourseme represented by the node of the collocation analysis. We make this explicit in our approach: the node of a collocation analysis is always a discourseme, often defined a priori by the researcher. It is noteworthy that discourseme constellations provide a partial solution to the lack of (discourse-specific) word sense disambiguation discussed above, due to the mutual disambiguation of discoursemes within a constellation (e.g. *displacement* is unlikely to refer to a car engine when used in conjunction with the discourseme |migration|).

Our proposed operationalisation in terms of discoursemes and discourseme constellations offers several advantages for future CADS research:

(i) The quantitative-qualitative bridge at the meso level of discourse analysis becomes more formalised and reproducible. Listing discoursemes, their operationalisation (as sets of lexical items) and their constellations can be regarded as a form of research documentation.

(ii) Discoursemes can be fed back into quantitative analyses and visualisations. We exemplify the usefulness of this in our case study below.

(iii) Discoursemes can be used as a starting point for further analysis steps, e.g. as node of a follow-up collocation analysis.

(iv) Discoursemes need not be based on a single keyword/collocation analysis, but can incrementally grow during a study, taking different corpora and perspectives into account.
(v) Statistical distributions of discoursemes and their constellations can be determined automatically across different corpora and sub-corpora, giving useful indications of the statistical distribution of discursive patterns (also exemplified in our case study).

We have implemented a software toolkit for integrated CADS analyses based on our approach.[8] This software allows the incremental creation of a database of discoursemes that can be refined and reused across multiple keyword and collocation analyses. It visualises keywords and collocations in the form of interactive semantic maps, and can display concordances for lemmata, discoursemes and constellations using IMS Open Corpus Workbench (CWB, https://cwb.sourceforge.io) as a back end – the main additional feature here is the visual highlighting of all words that belong to any of the discoursemes in a constellation. Occurrences of discoursemes and their constellations can automatically be identified in corpora for further statistical analysis. Besides the obvious advantages of such an integrated system (which eradicates the need to switch between spreadsheets, a concordancer, and data analysis software such as R), the toolkit allows analysts to include multi-word entities (MWEs) in the definition of discoursemes, thus overcoming a major technical limitation of previous work in CADS. Our toolkit provides both an experimental web-based interactive user interface and a well-documented REST API that can be accessed from data analysis tools such as R or Python.

4. Case Study

In our case study, we focus on migration-related discoursemes in parliamentary debates at two decisive moments of Germany's post-Cold War history: (a) 1993/94, a period that saw the migration of refugees from the Balkans at the time of the Yugoslav War, and (b) 2015/16, when escalating conflicts in the Middle East, particularly in Syria, led to what was then labelled a European 'refugee crisis' (*Flüchtlingskrise*) of even larger proportions (Griebel & Vollmann, 2019; Herbert, 2014; Wodak, 2009, 2015b). Comparing such closely related discourses is particularly fruitful for illustrating the applicability of our discourseme-based CADS approach, because they revolve around the same topic, namely flight and migration, which ensures direct comparability. However, they also carry several distinctions which mandate their comparison: (a) the political setting (a coalition of Christian conservatives CDU/CSU and liberal FDP with Germany's second largest party, the social democrats (SPD), in the opposition in 1993/94, and a grand coalition of CDU/CSU and SPD in 2015/16), and (b) public opinion and political impact, namely a transpartisan agreement on a so-called 'asylum compromise' (*Asylkompromiss*) and preceding discursive normalisation of derogatory terms such as 'asylum abuse' (*Asylmissbrauch*) (Weinzierl, 2009) by heated-up conservative mass media outlets and political actors in the 1990s, and a highly polarised debate about chancellor Angela Merkel's open door policy (often described as 'migration crisis' (*Migrationskrise*) in public and political discourse), with the years of 2015/16 becoming a focal point for the rise of the right-wing populist party AfD in Germany.[9]

In contrast to most CADS/CDA research, our case study does not provide any (social) critique but aims to show the usefulness of using discoursemes to ensure reproducibility and transferability between (sub-)corpora. Our discoursemes and all quantitative analyses are available online.[10]

4.1 Data

We used GermaParl v2.0.0 beta.1,[11] a data set that covers all plenary protocols of the German federal parliament from 1949 to 2021. This corresponds to legislative periods (LPs) 1 through 19, i.e., the first 19 terms of the German parliament. We applied some custom post-processing on the CWB-indexed corpus, especially improved lemmatisation. Our final data set comprises 271,077,449 tokens in 4340 plenary protocols and 950,028 speeches. Here, we focus on LP12 (1990–1994) and LP18 (2013–2017) and five parliamentary groups: CDU/CSU, SPD, Die Grünen, PDS, and Die Linke. Note that Die Linke was founded as the result of a merger of PDS and WASG; we will refer to it as PDS / Die Linke where applicable. Table 1 shows the number of speeches and tokens across parliamentary groups and LPs.

	LP12		**LP18**	
parl. group	**# speeches**	**# tokens**	**# speeches**	**# tokens**
CDU/CSU	6,871	3,750,530	7,182	5,673,912
SPD	10,015	4,342,204	5,008	3,801,588
Die Grünen	1,408	822,312	5,958	2,731,466
PDS / Die Linke	2,047	1,184,823	4,488	2,435,579
total	**20,431**	**10,099,869**	**22,636**	**14,642,545**

Table 1: Amount of text for different parliamentary groups across legislative periods in GermaParl

4.2 Topic-Discourseme

As a starting point for our analysis (i.e. the topic-discourseme used as node in the following collocation analysis), we chose a set of lemmata indicating fleeing and refugees. This topic features in immigration discourses in both periods (LP 12 and 18). Other concepts such as 'asylum (seeker)', 'migrant', or 'migration' are excluded because they are substantially more frequent in one of the two periods than in the other. To include a broad variety of words, we manually compiled a list of German search terms based on an initial CQP query for lemmata containing the strings *flucht* or *fliehen* (case-insensitive and ignoring diacritics). Besides the literal strings, we mostly found noun compounds (e.g. *Fluchtursache* [cause of flight], *Sowjetzonenflüchtling* [Soviet zone refugee], *Flüchtlingslager* [refugee camp], …).

Our final corpus query includes almost all lemmata that contain one of the two initial search strings, except for some false positives unrelated to our

research question (e.g. *Steuerflucht* [tax evasion], *Kapitalflucht* [capital flight], *Ausflucht* [excuse, tergiversation]). Moreover, the inspection of random concordance lines (and an initial collocation analysis) revealed some tokenisation errors, which we circumvented by including custom MWEs like *Flücht Linge* in the topic-discourseme. Table 2 contains the overall frequency breakdown in the whole corpus, where 'other' represents the aggregated information for all lemmata beyond the 15th rank.

lemma	**translation**	**freq**	**ipm**	**ratio**
Flüchtling	[refugee]	20,097	74.14	0.44
other		10,516	38.96	0.23
Flucht	[flight]	3,791	13.98	0.08
fliehen	[(to) flee]	1,557	5.74	0.03
flüchten	[(to) flee]	1,282	4.73	0.03
geflüchtet	[fled]	1,270	4.69	0.03
Fluchtursache	[cause of flight]	1,257	4.64	0.03
Sowjetzonenflüchtling	[Soviet zone refugee]	809	2.98	0.02
Flüchtlingslager	[refugee camp]	777	2.87	0.02
Zuflucht	[refuge]	760	2.8	0.02
Flüchtlingspolitik	[refugee policy]	741	2.73	0.02
Flüchtlingsstrom	[stream of refugees]	715	2.64	0.02
Flüchtlingskonvention	[refugee convention]	616	2.27	0.01
Bürgerkriegsflüchtling	[civil war refugee]	599	2.21	0.01
flüchtig	[fleeting, ephemeral, volatile]	381	1.41	0.01
Flüchtlingskrise	[refugee crisis]	376	1.39	0.01
total		**45,649**	**168.2**	**1**

Table 2: Frequency breakdown of the topic-discourseme in the whole corpus; we report absolute frequencies (column 'freq') as well as the proportion of all query matches ('ratio') and instances per million tokens ('ipm')

Table 2 shows that the discourse is dominated by the lemma *Flüchtling* (more than 20,000 occurrences, i.e. 44% of all instances, with 74 instances per million tokens), followed by *Flucht* (14 ipm), *fliehen* (6 ipm) and *flüchten* (5 ipm). Figure 1 shows the topic breakdown across parliamentary groups and recent LPs. As expected, the topic is very prominent in LP12 (1990–1994) and LP18 (2013–2017).

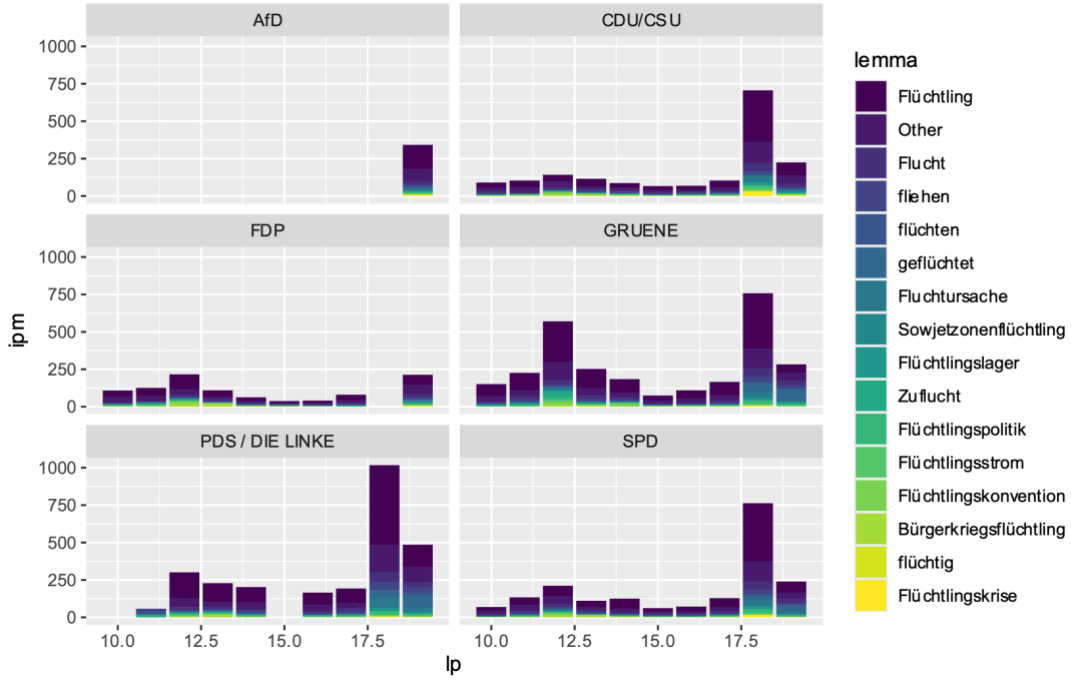

Figure 1: Frequency breakdown of the topic-discourseme across parliamentary groups and LPs 12–19

4.3 Collocation Analyses

We started with an unrestricted collocation analysis in the whole corpus, with a context window of 10 tokens to the left and to the right of the node discourseme (limited by sentence boundaries). The collocation profile for the whole corpus is shown in Figure 2 in the form of a *semantic map* (cf. Evert & Heinrich, 2019), in which the positions of the lemmata are based on word embeddings. Since these are high-dimensional and thus cannot be directly shown in a two-dimensional plot, an algorithm to project the embeddings onto a two-dimensional plane must be used. Here, we opt for *t*-distributed stochastic neighbour embedding (van der Maaten & Hinton, 2008). Text size is proportional to association strength. This visualisation technique is especially helpful for manual clustering of lemmata, since the most salient lemmata directly catch the researchers' eyes and semantically similar words are close to one another. Our toolkit allows researchers to form discoursemes by drag & drop and automatically shows the corresponding concordance lines when selecting an item.

Note that the parameters chosen for this plot (association measure, window, context break, cut-off) are somewhat arbitrary: there is no general 'best practice' (see e.g., Haider, 2019). For most of the present study, we opt for conservative log-ratio (LRC) as association measure, which combines two key

dimensions of association strength (effect size and statistical significance) in a single value (cf. Evert, 2022).[12] LRC has not come into wide-spread use yet, since it is relatively new. However, many studies use similar measures, such as log-likelihood-filtered Log Ratio (binary logarithm of relative risk) or Mutual Information (MI).

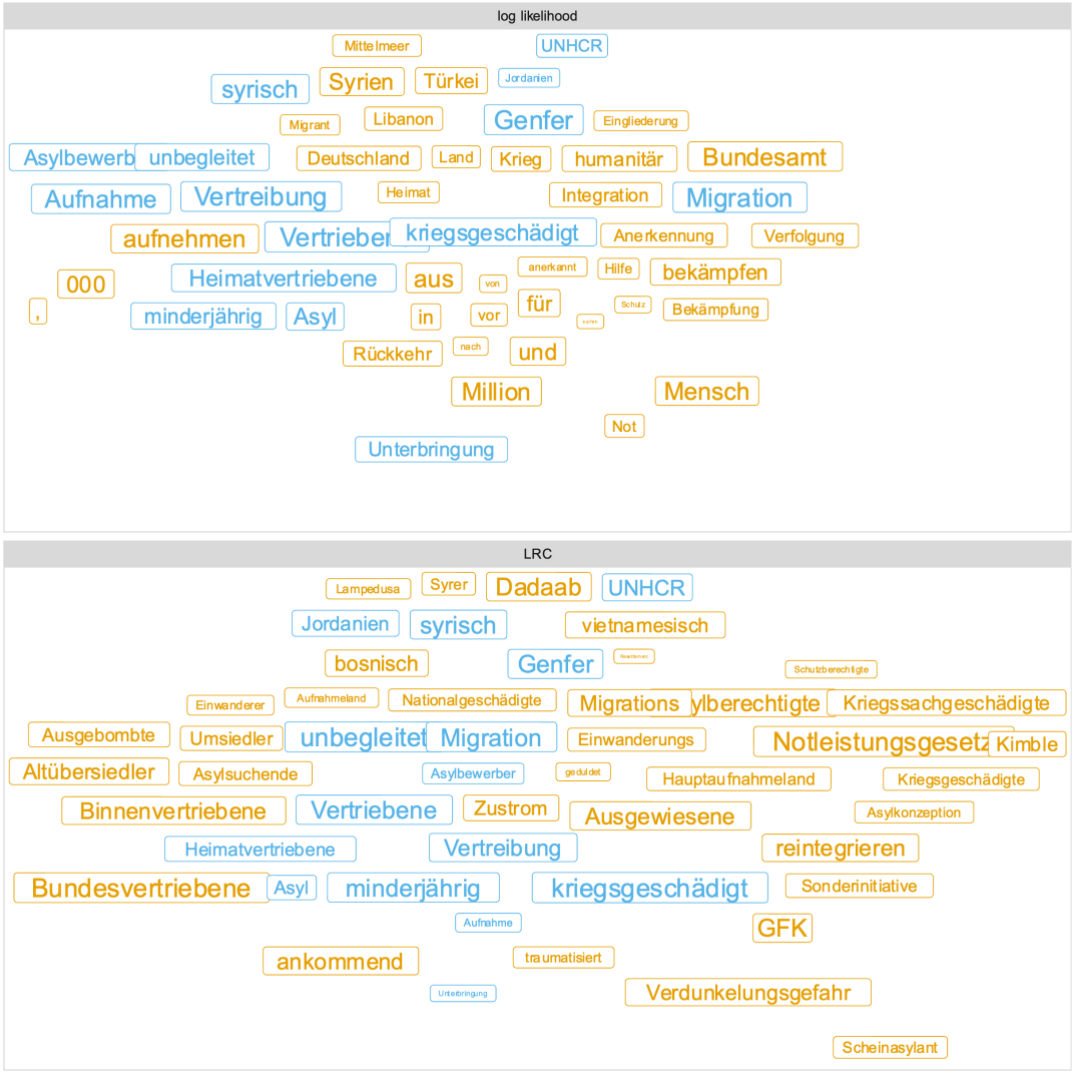

Figure 2: Semantic map of the collocation profile of the topic-discourseme in the whole corpus. Top: using log-likelihood ratio as association measure, bottom: LRC. Collocates that occur in both profiles are displayed in blue, others in orange.

In Figure 2, we present the top 50 collocates for both LRC (bottom panel) and the well-established log-likelihood ratio (LLR, top panel). This comparison shows that although there is relatively little overlap between the specific collocates identified by the two measures, they indicate similar discursive patterns. In our terminology, they often identify different collocates that belong to the same discourseme (LRC e.g., retrieves *Syrer* [Syrian] among the top 50 collocates, whereas LLR retrieves *Syrien* [Syria]). Further visualisations omitted here for reasons of brevity make clear that other measures yield similar discoursemes, with the usual caveats: log-likelihood ratio retrieves highly frequent words (such as prepositions), Log Ratio is biased towards low-

frequency words that are often not statistically significant. As for context settings, we remark that the smaller the context window, the more collocates tend to be part of MWEs around the node lemmata rather than the kind of linguistic patterns relevant for CADS.

The semantic maps in Figure 2 are rather crowded – as mentioned above, the task of CADS researchers is to categorise lexical items into discoursemes. Two directly apparent discoursemes are traces of the related discourses of |asylum| (*Asyl* [asylum], *Asylbewerber* [asylum seeker], *Asylrecht* [asylum law], ...) and |migration| (*Migration* [migration], *Migrant* [migrant], *Einwanderer* [immigrant], ...). Another obvious cluster indicates |push| factors (*Vertreibung* [displacement], *Vertriebene* [displaced person], ...). Two lemmata indicate some refugees' |need for protection| (*unbegleitet* [unaccompanied], *minderjährig* [underage]). Several demonyms, adjectives, and names indicate the |origin| of refugees (*syrisch* [Syrian], *Syrer* [Syrian], ...) as well as |third countries| (*Jordanien* [Jordan]) and the refugee |route| (*Lampedusa*). We can also find traces of the discourse about victims of World War II (|WW2|: *Kriegssachgeschädigte* [war-damage claimant], *Kriegsgeschädigte* [war-damaged], *Nationalgeschädigte* [person affected by national policies]) and the collapse of the German Democratic Republic (|GDR|: *Altübersiedler* [people who fled or immigrated from the GDR to the Federal Republic of Germany before 1989]). Similarly, the *Notleistungsgesetz* [emergency benefit act] was a special |(draft) law| with regards to the accommodation of refugees from the part of Germany occupied by the Soviet Union after World War II. Less salient, but still visible, is a lemma indicating the |trauma| experienced by refugees due to their flight (*traumatisiert* [traumatised]). Lemmata such as *Aufnahmeland* [host country] indicate discourse about the refugees' |accommodation|, whereas *Ausgewiesene* [expelled people] indicates their |rejection|. Finally, we also see the |United Nations| High Commissioner for Refugees (*UNHCR*) being mentioned. Figure 3 shows the collocation profile after these discoursemes have been formed and only the remaining collocates are shown individually. It represents an intermediate step in the manual grouping process.

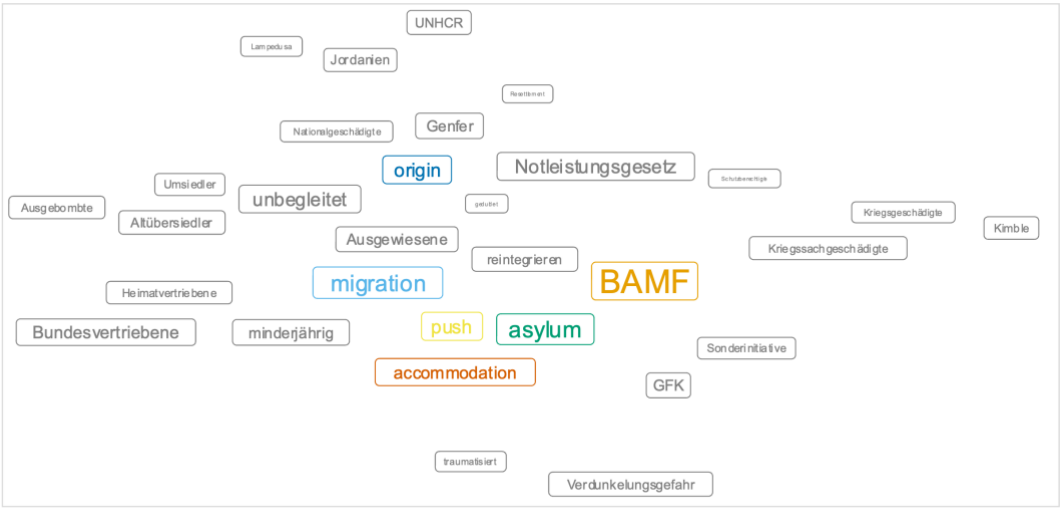

Figure 3: Semantic map of the collocation profile after partial discourseme formation; discoursemes are coloured to distinguish from regular lemmata

Note that the collocations profiles only show single words (unigrams) due to technical limitations shared with most collocation analysis tools. Some of them may be part of larger MWEs, e.g. *kriegsgeschädigt* [war-damaged], which belongs to an MWE indicating the predecessor of the Federal Office of Migration and Refugees (*BAMF*): *Bundesministerium für Vertriebene, Flüchtling[e] und [K]riegsgeschädigt[e]*[15]; only its occurrences outside this MWE belong to the discourseme |WW2|. Our toolkit allows us to include this MWE as well as the corresponding minister (*Bundesminister für Vertriebene , Flüchtling[e] und [K]riegsgeschädigt[e]*) in the discourseme |BAMF|.[14] Further discoursemes with MWEs are |human rights|, which includes the Convention Relating to the Status of Refugees (|GFK|: *Genfer Flüchtlingskonvention*), as well as German and European |initiatives| (*Sonderinitiative* [special initiative], *Fluchtursache[n] bekämpfen* [(to) fight the reason(s) for fleeing]*, Flüchtling[e] reintegrieren* [re-integrate refugees]).

Last but not least, it is worth mentioning that our broadly defined topic-discourseme also yields some collocates that we are not interested in, most notably *Verdunkelungsgefahr* [danger of collusion] (a |crime|) from the MWE *Flucht- und Verdunkelungsgefahr* [risk of flight and suppression of evidence]. Further false positives are *flüchtig[e] organisch[e] Verbindung* [volatile organic compounds] (important in chemistry or |biology|) and the US-American TV series *Kimble Auf der Flucht* ['The Fugitive'] (|Kimble|).

4.3.1 Collocation analyses in legislative periods 12 and 18

Since we are particularly interested in LP12 and 18, we repeat the collocation analysis in these subcorpora. In contrast to the approach adopted by CQPweb[15], we use marginal frequencies from the whole corpus to compute association scores in the subcorpora. We believe that this approach is most suitable for our research question, as it shows how strongly collocates are attracted to the combination of the topic and respective LP (and political party in latter analyses). Figure 4 shows a full set of discoursemes formed from the top 50 collocates in these two LPs. Note that discoursemes (and their operationalisations as word lists) are the same for both panels and were formed across both collocation profiles.

It is immediately apparent that both periods show a strong association with the |migration| and |asylum| discoursemes, LP12 more so than LP18. Additionally, both LPs focus on |push| factors (*Elend* [misery], *Hunger* [hunger], *Bürgerkrieg* [civil war], *Krisengebiet* [crisis area], *Kriegsgebiet* [war area]). The |origin| discourseme is also prominent in both subcorpora, but obviously with different surface realisations: in LP12, it primarily refers to the Balkans (*Bosnien* [Bosnia], *Jugoslawien* [Yugoslavia], *Kroatien* [Croatia], *Herzegowina* [Herzegovina]), whereas |origin| in LP18 includes countries in Africa and the Middle East (*Syrien* [Syria], *Eritrea* [Eritrea], *Nordirak* [Northern Iraq], *Libanon* [Lebanon]). Similarly, |third countries| such as *Türkei* [Turkey] or *Jordanien* [Jordan] are only found in LP18. Regarding the (potential) |approval| of refugees, LP12 mentions a *right* to stay (*Bleiberecht*), while the equivalent in LP18 is merely a *perspective* of staying (*Bleibeperspektive*). The most striking discourseme unique to LP12 is, however, a cluster of numbers referring to |(draft) laws| and the corresponding legal process (*12/6852*, *12/3094*) (*Drucksache* literally means 'printed matters',

which refers to administrative acts).[16] Further discoursemes only found in LP12 are |human rights| and its counterpart, the |rejection| of refugees (the euphemistic *Repatriierung* [repatriation]).

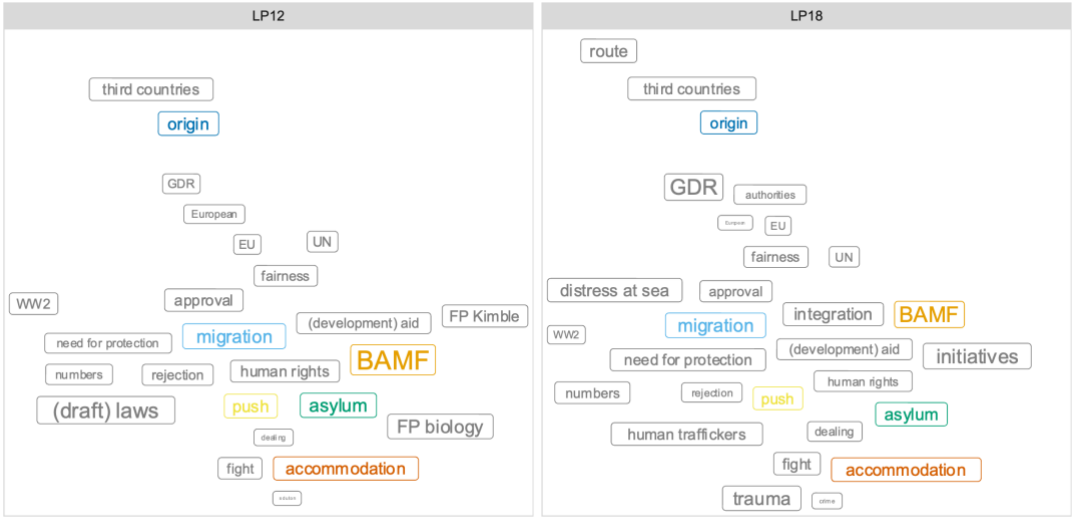

Figure 4: Semantic maps of collocation profiles of the topic-discourseme in LP12 and LP18, respectively, fully categorised into discoursemes; coloured discoursemes are the same as the ones in Figure 3

Overall, we find LP12 to display a more compact discourseme landscape, with a smaller number of discoursemes, reflecting a common ground in the parliamentary discussion towards the handling of the issue. In contrast, LP18 shows a larger number of discoursemes and thus a more diverse range of discursive patterns. Additionally, we find prominent outliers, with discoursemes describing the situation of the refugees, such as |need for protection| (*unbegleitet* [unaccompanied], *minderjährig* [underage], *schutzbedürftig* [in need of protection]) and |trauma|, but also interim measures for handling the situation, like |accommodation| (*Unterbringung* [housing], *Unterkunft* [lodging], *Aufnahme* [admission]). These discoursemes – emphasising sympathy and urgency – are almost completely absent from LP12. In LP18, we also find strong traces of the discourse about the |fight| against the influx (*Bewältigung* [overcoming], *bekämpfen* [(to) fight]) and the aforementioned (European) |initiatives|, as well as the |route| across the Mediterranean (*Ägäis* [Aegean Sea], *Mittelmeer* [Mediterranean Sea], *Lampedusa*), |distress at sea| (*Seenot* [distress at sea], *seeuntauglich* [unseaworthy], *ertrinken* [(to) drown]), and |human traffickers| (*Schleuser*). Semantically more distant and more numerous, the discoursemes prominent in LP18 seem to reflect a more contrarian and diverse parliamentary discussion on how to address the ongoing refugee movements at that time, with calls for sympathy and action, and less of a discursive consensus on established administrative frameworks, as in LP12.

4.4 Discourseme Associations

Having built up a database of discoursemes in this way, we can carry out quantitative analyses beyond the level of individual words and look at pairwise associations between discoursemes; either in subcorpora or in the whole corpus. For reasons of computational efficiency, we calculate pairwise associations based on co-occurrence of discoursemes in sentences (rather than a token-based span). We can easily visualise the resulting network structure using a force-directed graph layout algorithm, with pairwise association strengths as edge weights. One small drawback of this approach is that coordinates change for each visualisation, so figures for different subcorpora are not as easily comparable to one another as semantic maps.

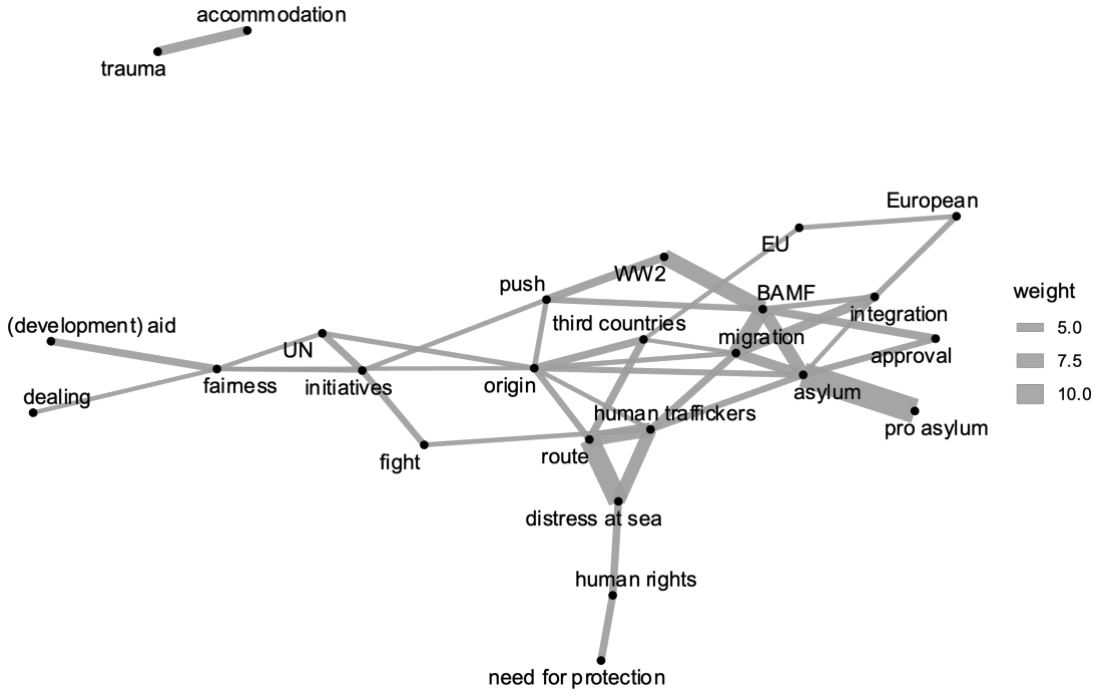

Figure 5: Visualisation of network of pairwise discourseme associations in the whole corpus, without the topic-discourseme (discoursemes as nodes and association strengths as edge weights)

Figure 5 shows the network layout created by the Fruchterman-Reingold algorithm (Fruchterman & Reingold, 1991), again using LRC as association measure. We omit the topic-discourseme (|fleeing|) – since we already know it is highly associated with all discoursemes per construction – as well as edges with LRC < 3 for reasons of clarity. Strong associations among pairs of discoursemes in this network indicate relevant discourseme constellations (which in turn are indicative of discursive patterns). For example, we can see that |human rights| and |need for protection| go hand in hand, and |human traffickers| are strongly associated with the |route| and, ultimately, the refugees' |distress at sea|.

4.4.1 Discourseme associations across parliamentary groups and legislative periods

Although Figure 5 provides a compelling overview, focussing on specific discoursemes (such as the initial topic-discourseme of our case study) can be even more fruitful. Table 3 shows all discoursemes from the analyses above and their association strengths with the topic-discourseme, computed separately for each parliamentary group and for LP12 vs. 18.

Table 3 is sorted by mean LRC across both periods. The topic-discourseme is mostly discussed in terms of the obvious discoursemes |asylum| and |migration|; the discourse also prominently features |BAMF|, Germany's central migration authority, and the refugee's |origin|. This is followed by a range of discoursemes discussing urgent issues such as |accommodation|, but also strategic and preventive measures such as the discoursemes of |fight| of the influx, reducing |push| factors, and the role of |third countries| and the |route|. Only then do we find discoursemes on the process after arrival in Germany, with focus on refugees' |human rights|, |integration|, |approval| and implemented measures, namely |initiatives|. The following discoursemes concerned with the refugee's emergency situation are partly exclusive to LP18 and therefore last, and discuss |trauma|, |distress at sea| and |people smugglers|, but also the |need for protection|, which in LP12 was only discussed by opposition parties. The |rejection| of refugees is also only ever addressed by opposition parties in both legislative periods. The discourseme |(development) aid| can be interpreted as either in relation to the previous cluster of discoursemes or as a measure of prevention of causes of flight. Mainly in LP18, the distribution and legal frameworks of the |EU| was a point of discussion. In summary, we can assert that the parliamentary discourse in both legislative periods and across parliamentary groups is first concerned with the administrative handling of the issue, secondly urgent but also preventive measures, third, with the process after arrival in Germany, and then humanitarian concerns.

Insight into collocation profiles for each parliamentary group in both legislative periods also allows for tracing shifts in discursive positions over time and political fractions. Unsurprisingly, the |asylum| and |migration| discoursemes are prominent across all subcorpora. Similarly, the |origin| of refugees and |push| factors are not associated with any legislative period or parliamentary group in particular; recall however that the actual realisations of these discoursemes differ between parliamentary groups. The |BAMF| discourseme is only indicative of the discourse of the governing parliamentary group CDU/CSU in LP12 (and less so of the SPD) and in large parts of the governing grand coalition of CDU/CSU and SPD in LP18, which could be interpreted as a discourse defending implemented administrative measures and mechanism to address the issue. The same might be true for the discourseme |initiatives|, which is exclusive to LP18 and only is referred to by the governing parties, however all the more. On the contrary, the process and fate after denial of status reflected in the discourseme |rejection| is only addressed by opposition parties and never governing parties, which is an observation warranting further investigation.

	LP12				LP18			
discourseme	**CDU/CSU**	**SPD**	**Die Grünen**	**PDS**	**CDU/CSU**	**SPD**	**Die Grünen**	**Die Linke**
|asylum|	5.77	5.62	6.08	6.63	4.73	4.51	3.84	4.03
|migration|	3.31	3.96	6.59	4.91	5.35	4.96	4.39	4.36
|BAMF|	7.10	2.81			7.18	6.85	5.43	4.59
|origin|	3.85	3.44	4.18	4.19	4.31	3.82	3.96	4.05
|push|	4.19	4.17	3.95	3.82	3.36	3.54	3.14	3.71
|accommodation|	2.90	2.54	2.76	2.11	4.04	4.31	4.17	3.53
|third countries|	3.27	2.95	1.51	1.08	4.42	4.23	4.38	4.45
|initiatives|					8.22	6.72	4.90	4.72
|route|					5.68	4.80	5.03	6.55
|fight|	1.75	2.97	1.40		4.05	3.51	2.72	3.07
|trauma|					5.37	5.23	4.18	4.24
|human rights|	2.24	2.81	2.32	0.91	2.45	2.63	2.77	2.51
|distress at sea|					5.61	2.66	2.32	6.37
|human traffickers|					4.72	4.41	2.15	4.90
|integration|					4.22	4.17	4.22	3.49
|approval|	2.88	2.10		1.82	2.10	2.24	2.54	2.32
|(development) aid|	2.51	2.18	1.63	0.71	2.37	1.98	2.16	2.10
|need for protection|			2.42	0.00	2.64	2.54	2.55	2.76
|rejection|	0.48	1.15	2.15	3.52	0.58	0.58	1.19	1.99
|EU|	0.88	0.65	1.83	0.47	1.85	1.39	1.61	1.68
|UN|	1.64	2.32			1.44	1.39	1.50	2.02
|GDR|	4.73	1.27					0.14	

Table 3: Discourseme associations with the node discourseme across parliamentary groups and legislative periods

The discourseme of |third countries|, which we interpreted as part of a discourseme cluster concerned with urgent and preventive measures, is only salient for CDU/CSU and SPD in LP12, while, in LP18, it is associated with all parliamentary groups. Also, the |accommodation| of refugees is more prominent in LP18, as well as the |fight| of the influx, as we discussed before. However, it is noteworthy that the latter unsurprisingly shows the highest association with the discourse of the conservative CDU/CSU in LP18, in contrast to LP12, when it was the discourse of the then oppositional SPD, which would generally be suspected to stress more of a humanitarian than a defensive stance. This shift in discursive position might present a worthy starting point for an in-depth analysis on how different surface realisations of the discourseme are discussed by different parliamentary groups over time. Equally noteworthy is the discourse on |integration|, which was not addressed by any parliamentary group in LP12, however the more in LP18 and across all parties, which might be in relation with the origin of refugees in LP12 and LP18 respectively – another noteworthy entry point for more analysis. The |approval| of refugees is generally associated with all parliamentary groups and LPs, save the discourse of Die Grünen in LP12. Conversely, the discourseme shows the highest association with Die Grünen in LP18, which again is a discursive shift worthy of in-depth analysis.

5. Conclusion

In the present contribution, we have argued that discoursemes are useful building blocks for combining quantitative and qualitative analysis in corpus-assisted discourse studies, regardless of specific terminology and methodological approaches. They act as a bridge between the micro and meso levels of discourse, making analyses more formalised and reproducible. A database of discoursemes can be built up incrementally across multiple analyses (e.g. for different parameters and different subcorpora), thus integrating different quantitative perspectives.

The discoursemes created for the purpose of our case study have been used to determine statistical distributions across time and parliamentary groups at a level much more closely to discursive patterns than the distributions of individual words or lemmata. In summary, the discourse on refugees and flight under the lens of a discourseme-level analysis produced results outlining the differences and commonalities of parliamentary discourse on the topic-discourseme of |fleeing| during different migration events, showing basic characteristics of discursive strands across all parliamentary groups, but also shifts in discursive positions by political fractions and highlighted worthy entry points for an in-depth analysis. Although some of these insights are 'so what'-findings (cf. e.g. Baker & Levon, 2015), they show that the proposed operationalisation of groups of lexical items in terms of discoursemes helps to abstract away from individual surface realisations – a crucial step to overcome the very lexicalised approach inherent to many CADS studies.

Despite some remaining conceptual and technical difficulties addressed in our case study, we are confident that our discourseme-based approach opens new possibilities for CADS, helping to overcome some of its current weaknesses. In particular, they provide a level of research documentation and reproducibility, a true integration of qualitative and quantitative methods, as

well as comprehensive and explainable hermeneutics. Our software toolkit provides an integrated research environment supporting this approach.

Finally, it seems self-evident that CADS can only be *supported* by computational techniques, rather than being fully automated by 'black box' AI approaches; grouping together lexical items into discoursemes depends on the analyst's research agenda and is thus inevitably subjective to a certain extent. Ultimately, we are envisioning a 'hermeneutic cyborg' (Evert, 2018), where computational methods assist human researchers in the interpretation of textual data for analysing discourse. We hope that our work offers an accessible approach to discourse analysis, which provides researchers a perspicacious view on their case studies.

Acknowledgements

Part of this research has been funded by the German Research Foundation (DFG), project no. 466328567.

Notes

1. German: 'Es sind institutionalisierte, geregelte Redeweisen als Räume möglicher Aussagen, die an Handlungen gekoppelt sind.'
2. German: 'transsubjektive Produzenten gesellschaftlicher Wirklichkeit und sozio-kultureller Deutungsmuster'
3. German: 'Fluss von "Wissen" bzw. sozialen Wissensvorräten durch die Zeit'
4. German: 'einen prozessierenden Zusammenhang von Wissen'
5. German: 'Sprechen/Denken – Tun – Vergegenständlichung'
6. 'At the micro level, the analyst is concerned with the text's syntax, metaphoric structure and certain rhetorical devices. The meso level comprised studying the text's production and consumption, concentrating on how power relations are enacted. At the macro level, the analyst considers intertextual relationships, trying to understand the broad, societal currents that are influencing the text being studied.' (Behnam & Mahmoudy 2013, p. 2196)
7. German: 'Die methodische Umsetzung des Ansatzes gerät häufig zu einem deduktionistischen, im konkreten Vorgehen unbestimmt bleibenden Interpretationsvorgang.'
8. See https://github.com/ausgerechnet/cwb-cads/. Figures and tables presented in the paper at hand were produced in R using API access to the software toolkit. The complete R scripts illustrating usage of the API for data analysis are included in the reproduction materials.
9. However, the AfD is not part of the current case study because it only entered the German federal parliament in LP19.
10. https://osf.io/84dcx/
11. See https://zenodo.org/records/10416536.
12. We use a 99.9% confidence interval (i.e. significance level $\alpha = .001$) and apply a Bonferroni correction for the number of comparisons made, i.e. the number of distinct unigram types.

13. With our analysis based on lemmata, we use square brackets to indicate the original form of the multi-word entities to ensure readability.
14. Unfortunately, it is currently not possible to exclude these instances of *kriegsgeschädigt* automatically from the discourseme |WW2|. This computationally expensive extension is under consideration for future development.
15. https://cwb.sourceforge.io/install.php#cqpweb
16. E.g. *Drucksache 12/6852: Entwurf eines Gesetzes zu dem Europäischen Übereinkommen vom 16. Oktober 1980 über den Übergang der Verantwortung für Flüchtlinge*.

References

Baker, P. (2004). Querying keywords: Questions of difference, frequency, and sense in keywords analysis. *Journal of English Linguistics*, *32*(4), 346–359. https://doi.org/10.1177/0075424204269894

Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyżanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, *19*(3), 273–306. https://doi.org/10.1177/0957926508088962

Baker, P., & Levon, E. (2015). Picking the right cherries? A comparison of corpus-based and qualitative analyses of news articles about masculinity. *Discourse & Communication*, *9*(2), 221–236. https://doi.org/10.1177/1750481314568542

Behnam, B., & Mahmoudy, B. (2013). A critical discourse analysis of the reports issued by the International Atomic Energy Agency (IAEA) Director General on Iran's nuclear program during the last decade. *Theory and Practice in Language Studies*, *3*(12), 2196–2201. https://doi.org/10.4304/tpls.3.12.2196-2201

Brinton, L. J. (2000). *The structure of modern English: A linguistic introduction*. John Benjamins.

Bubenhofer, N., Calleri, S., & Dreesen, P. (2019). Politisierung in rechtspopulistischen Medien: Wortschatzanalyse und Word Embeddings. *Osnabrücker Beiträge zur Sprachtheorie, 95*, 211–242. https://doi.org/10.17192/obst.2019.95.8658

Chouliaraki, L., & Fairclough, N. (1999). *Discourse in late modernity: Rethinking critical discourse analysis*. Edinburgh University Press. https://doi.org/10.1515/9780748610839

Evert, S. (2005). *The statistics of word cooccurrences: Word pairs and collocations* [Doctoral dissertation, University of Stuttgart]. OPUS. http://doi.org/10.18419/opus-2556

Evert, S. (2018). Distributional methods in corpus linguistics: Towards a hermeneutic cyborg. In Y. Tono & H. Isahara (Eds.), *Proceedings of the 4th Asia Pacific Corpus Linguistics Conference (APCLC2018)* (p. 010). Takamatsu, Japan.

Evert, S. (2022). Measuring keyness. In Y. Wang, T. Murase, K. Nagasaki, Y. Sato, & S. Seki (Eds.), *Digital Humanities 2022: Conference abstracts* (pp. 202–205). University of Tokyo. https://dh2022.adho.org

Evert, S., & Heinrich, P. (2019, February 27–28). *Introducing MMDA: An interactive toolkit for CDA* [Conference presentation abstract]. 7th Göttingen-Hildesheim-Workshop on Computational Linguistics and Digital Humanities: Digital Methods in Political Science, Göttingen, Germany. https://www.gcdh.de/fileadmin/user_upload/HI-GÖ_Workshop_Book_of_Abstracts_Final_Version_PDF.pdf

Fairclough, N. (2013). *Critical discourse analysis: The critical study of language*. London: Routledge.

Fairclough, N. (2015). *Language and power* (3rd ed.). Routledge.

Fairclough, N., Mulderrig, J., & Wodak, R. (2013). What is discourse analysis? In T. A. van Dijk (Ed.), *Discourse studies: A multidisciplinary introduction* (pp. 357–379). SAGE Publications.

Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, *21*(11), 1129–1164. https://doi.org/10.1002/spe.4380211102

Foucault, M. (1972). *The archaeology of knowledge*. Pantheon Books.

Fox, R. (2006). Using corpus linguistics to describe corporations' ideologies. *Tourism and Hospitality Management*, *12*(2), 15–24. https://doi.org/10.20867/thm.12.2.2

Geeraerts, D. (2010). *Theories of lexical semantics*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198700302.001.0001

Griebel, T., & Vollmann, E. (2019). We can('t) do this: A corpus-assisted critical discourse analysis of migration in Germany. *Journal of Language and Politics*, *18*(5), 671–697. https://doi.org/10.1075/jlp.19006.gri

Grundmann, R., & Krishnamurthy, R. (2010). The discourse of climate change: A corpus-based approach. *Critical Approaches to Discourse Analysis across Disciplines, 4*(2), 125–146.

Haider, A. S. (2019). Using corpus linguistic techniques in (critical) discourse studies reduces but does not remove bias: Evidence from an Arabic corpus about refugees. *Poznan Studies in Contemporary Linguistics*, *55*(1). https://doi.org/10.1515/psicl-2019-0004

Herbert, U. (2014). »Asylpolitik im Rauch der Brandsätze« – der zeitgeschichtliche Kontext. In S. Luft & P. Schimany (Eds.), *20 Jahre Asylkompromiss* (pp. 87–104). transcript Verlag. https://doi.org/10.1515/transcript.9783839424872.87

Jäger, M., & Jäger, S. (2007). *Deutungskämpfe: Theorie und Praxis Kritischer Diskursanalyse*. VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-90387-3

Jäger, S. (2015). *Kritische Diskursanalyse: Eine Einführung*. Unrast.

Johnson, M. N. P., & McLean, E. (2020). Discourse analysis. In A. Kobayashi (Ed.), *International encyclopedia of human geography* (2nd ed., pp. 377–383). Elsevier. https://doi.org/10.1016/B978-0-08-102295-5.10814-5

Keller, R. (2011). *Wissenssoziologische Diskursanalyse*. VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-92058-0

Laclau, E. (1996). *Emancipation(s)*. Verso.

Laclau, E., & Mouffe, C. (2001). *Hegemony and socialist strategy: Towards a radical democratic politics* (2nd ed.). Verso.

Link, J. (2009). *Versuch über den Normalismus: Wie Normalität produziert wird* (4th ed.). Vandenhoeck & Ruprecht.

Link, J. (2016). Sprache und Kultur in der foucaultschen Diskurstheorie. In L. Jäger, W. Holly, P. Krapp, S. Weber, & S. Heekeren (Eds.), *Sprache – Kultur – Kommunikation / Language – culture – communication: Ein internationales Handbuch zu Linguistik als Kulturwissenschaft / An international handbook of linguistics as a cultural discipline* (pp. 117–126). De Gruyter Mouton. https://doi.org/10.1515/9783110224504-014

Mautner, G. (2009). Corpora and critical discourse analysis. In P. Baker (Ed.), *Contemporary corpus linguistics* (pp. 32–46). Continuum.

Partington, A. (2006). Metaphors, motifs and similes across discourse types: Corpus-Assisted Discourse Studies (CADS) at work. In A. Stefanowitsch & S. Gries (Eds.), *Corpus-based approaches to metaphor and metonymy* (pp. 258–294). De Gruyter Mouton. https://doi.org/10.1515/9783110199895.267

Poole, R. (2016). Good times, bad times: A keyword analysis of letters to shareholders of two Fortune 500 banking institutions. *International Journal of Business Communication*, 53(1), 55–73. https://doi.org/10.1177/2329488414525449

Reisigl, M., & Wodak, R. (2009). The discourse-historical approach (DHA). In R. Wodak & M. Meyer (Eds.), *Methods of critical discourse analysis* (2nd ed., pp. 87–121). SAGE Publications.

Rheindorf, M., & Wodak, R. (2020). Grenzen, Obergrenzen, Zäune: Korpuslinguistische und diskurshistorische Perspektiven auf die Normalisierung rechtspopulistischer Positionen. In K. Binner & K. Scherschel (Eds.), *Fluchtmigration und Gesellschaft* (pp. 126–148). Beltz Juventa.

Touileb, S., & Salway, A. (2014). Constructions: A new unit of analysis for corpus-based discourse analysis. In W. Aroonmanakun, P. Boonkwan, & T. Supnithi (Eds.), *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing* (pp. 634–643). https://aclanthology.org/Y14-1072.

Trier, J. (1931). *Der deutsche Wortschatz im Sinnbezirk des Verstandes: die Geschichte eines sprachlichen Feldes: Vol. 1. Von den Anfängen bis zum Beginn des 13. Jahrhunderts*. Winter.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605. http://jmlr.org/papers/v9/vandermaaten08a.html

van Dijk, T. A. (2015). Critical Discourse Analysis. In D. Tannen, H. E. Hamilton, & D. Schiffrin (Eds.), *The Handbook of Discourse Analysis* (pp. 466–485). John Wiley & Sons. https://doi.org/10.1002/9781118584194.ch22

Weinzierl, R. (2009). *Der Asylkompromiss 1993 auf dem Prüfstand: Gutachten zur Vereinbarkeit der deutschen Regelungen über sichere EU-Staaten und sichere Drittstaaten mit der Europäischen Menschenrechtskonvention, dem EU-Recht und dem Deutschen Grundgesetz*. Deutsches Institut für Menschenrechte. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-316587.

Wodak, R. (2009). The semiotics of racism: A critical discourse-historical analysis. In J. Renkema (Ed.), *Discourse, of course: An overview of research in discourse studies* (pp. 311–326). John Benjamins. https://doi.org/10.1075/z.148.29wod

Wodak, R. (2015a). 'Normalisierung nach rechts': Politischer Diskurs im Spannungsfeld von Neoliberalismus, Populismus und kritischer Öffentlichkeit. *Linguistik Online, 73*(4), 27–48. https://doi.org/10.13092/lo.73.2191

Wodak, R. (2015b). *The politics of fear: What right-wing populist discourses mean*. SAGE Publications. https://doi.org/10.4135/9781446270073

Appendix 1. List of Abbreviations

AfD	Alternative for Germany (far-right party)
BAMF	*Bundesamt für Migration und Flüchtlinge* [Federal Office of Migration and Refugees in Germany]
CADS	Corpus-assisted Discourse Studies
CDA	Critical Discourse Analysis
CDU/CSU	Christian Democratic Union of Germany (CDU) and Christian Social Union in Bavaria (CSU) (centre-right alliance)
CQP	Corpus Query Processor of CWB
CQPweb	A web app based on CWB
CWB	IMS Open Corpus Workbench (https://cwb.sourceforge.io/)
Die Grünen	Alliance 90/The Greens (centre-left party)
Die Linke	The Left (left-wing party)
FDP	Free Democratic Party (centre-right/liberal party)

GDR	German Democratic Republic
GermaParl	Corpus of plenary protocols of the German federal parliament
GFK	*Genfer Flüchtlingskonvention* [Convention Relating to the Status of Refugees]
ipm	instances per million (relative frequency)
LLR	log-likelihood ratio (association measure)
LP	legislative period (of the German federal parliament)
LRC	conservative log-ratio (association measure)
MI	mutual information (association measure)
MWE	multi-word entity
PDS	Party of Democratic Socialism (left-wing party)
REST API	an application programming interface that follows the Representational State Transfer style
SPD	Social Democratic Party of Germany (centre-left party)
WSD	word-sense disambiguation