Operationalising the Hermeneutic Grouping Process in Corpus-assisted Discourse Studies

Philipp Heinrich and Stephanie Evert

Chair of Computational Corpus Linguistics
Friedrich-Alexander-Universität Erlangen-Nürnberg
Bismarckstr. 6, 91054 Erlangen, Germany

philipp.heinrich@fau.de stephanie.evert@fau.de

Abstract

We propose a framework for quantitative-qualitative research in corpus-assisted discourse studies (CADS), which operationalises the central process of manually forming groups of related words and phrases in terms of "discoursemes" and their constellations. We introduce an open-source implementation of this framework in the form of a REST API based on Corpus Workbench. Going through the workflow of a collocation analysis for *fleeing* and related terms in the German Federal Parliament, the paper gives details about the underlying algorithms, with available parameters and further possible choices. We also address multi-word units (which are often disregarded by CADS tools), a semantic map visualisation of collocations, and how to compute assocations between discoursemes.

1 Introduction and Related Work

Corpus-assisted discourse studies (CADS) (Baker, 2006; Baker et al., 2008; Mautner, 2009) are a highly effective approach for exploring and understanding socio-political discourse, often building on a theoretical background rooted in critical discourse analysis (Fairclough, 2015). CADS research focuses on interpreting, explaining, and critiquing discourses surrounding socially contentious issues, intricate historical phenomena, and dominant narratives (Wodak and Meyer, 2015, 11). Typical examples of the themes explored in CADS include socio-economic concerns like austerity (Griebel et al., 2020), global challenges such as climate change (Grundmann and Krishnamurthy, 2010; Wang and Huan, 2023), and political ideologies such as right-wing or nationalistic perspectives (Baker and McEnery, 2005; Gabrielatos and Baker, 2008; Wodak, 2015, 2018).

CADS research usually relies on "low-level" corpus-linguistic techniques such as concordancing as well as keyword and collocation analyses (Baker, 2006). They are complemented by a hermeneutic interpretation of the observations that takes the wider socio-pragmatic context into account, but which is also influenced (more or less explicitly) by the intuitions and preconceptions of researchers. The use of corpora aims to mitigate such biases and the cherry-picking of examples that support them. A typical CADS investigation starts with a detailed examination of keywords and collocates (Baker, 2006; Baker et al., 2008). Keywords are lemmata[1] that occur with significantly higher frequency in a target corpus than in a reference corpus and indicate either important topics of the discourse (for a target corpus related to the theme of the study) or characteristic framings used by certain groups of actors (e.g. for right-wing vs. left-wing newspapers). Collocates are lemmata that are statistically associated (i.e. tend to co-occur) with a particular node lemma (or set of lemmata). They might indicate, e.g., the salient framings and evaluations associated with a certain topic indicated by the node lemmata (e.g. *refugee, displaced person*). Tentative interpretations obtained from this "distant reading" of the corpora are then confirmed and refined by "close reading" of concordances for individual lemmata, displaying their corpus occurrences in a compact tabular format with left and right context.

Relevant methodological research in corpus linguistics has focused on identifying suitable association measures and other parameter settings for the identification and ranking of keyword and collocation candidates (Stubbs, 1995; Hardie, 2014; Evert et al., 2017; Evert, 2022). However, it has been

[1] Analyses are typically carried out on the basis of lemmata rather than word forms. In most European languages beside English, different inflected forms of the same lemma are often selected due to syntactic constraints and do not to indicate different discourse-specific meanings. We thus refer to lemmata throughout our contribution; analyses can of course also be carried out on the basis of word forms or other annotation layers (such as POS-disambiguated lemmata or semantic tags).

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 33–44

established that there is no single "best" measure (Evert, 2008), leading researchers to advocate for the integration of multiple perspectives provided by different algorithms and parameter configurations (Gries, 2019, 2021).

A crucial step in CADS is the manual grouping of related keywords and collocations, which are then interpreted in terms of discursive patterns (topics, discursive strategies, positions or fragments (cf. Jäger, 2015, 80)). This "meso level" of discourse analysis (Fairclough, 2015, 58) thus forms the bridge between linguistic and discursive patterns. Most CADS research relies on off-the-shelf concordancing tools (such as CQPweb, AntConc, and #LancsBox) or SaaS platforms (such as SketchEngine and english-corpora.org), which are limited in the parameters of quantitative analysis such as choice of association measure (depending on the specific tool used) and present keywords and collocations as tables ranked by association score (making it difficult to recognise discursive patterns among them). The grouping process invariably happens outside the concordancing tools, using spreadsheet software or pen and paper.

Our aim is to improve the quantitative-qualitative interface in CADS research by (i) introducing an operationalisation of the grouping process in terms of "discoursemes" (see Section 2) and (ii) providing better software tools that integrate discoursemes into quantitative corpus analysis. We thus stay very close to established and successful practice in CADS, which at its core induces groupings and discursive patterns from the observed data in a corpus-driven fashion. This is markedly different from other ongoing research that might also contribute to the future of CADS. One strand focuses on machine learning techniques leveraging human "ground truth" annotations to detect functional properties of texts (or text segments) such as emotion (Wegge and Klinger, 2023) or sarcasm (Plepi et al., 2023). With the advent of large language models, another line of research is now concerned with zero-shot detection of topics (Navarretta and Hansen, 2023) or narratives (Heinrich et al., 2024).

In this contribution, we show both the possibilities of a discourseme-based operationalisation of CADS analyses and the technical challenges that come along with it, together with recommendations for best practices. Since "design and capabilities" of tools are essential to making sense of linguistic data (Anthony, 2013, 141), and off-the-shelf concordancing tools such as CQPweb (Hardie, 2012) do not provide any reasonable functionality to support the grouping process, we offer an open-source REST API for CADS research implemented in Python[2] with a corresponding OpenAPI Specification[3]. It builds on CWB (Evert and Hardie, 2011) for corpus storage, whose corpus query processor CQP (Evert and The CWB Development Team, 2022) allows efficient querying of large tokenised corpora, retrieving pairs of corpus positions for match and matchend of the query, respectively.

The API provides an extensive set of features designed to facilitate CADS analyses, including:

- classic CADS features such as CQP queries, concordancing (including various filtering and sorting techniques), query breakdown (including distribution across meta data), meta data management (using information stored in structural attributes in CWB), subcorpus creation, and collocation and keyword analysis;
- visualisation of collocation and keyword profiles via semantic maps (cf. Figure 1); and
- endpoints for managing discoursemes and discourseme constellations.

Here, we concentrate on the workflow of defining discoursemes via the result table of a collocation analysis and the technical challenges of its implementation (Section 3). Some reasonable discoursemes are given as illustrative examples. Note that the API can be accessed via HTTPS and thus allows analysts to combine an interactive graphical user interface (GUI) with low-level API calls when forming discoursemes, then use other tools such as R for further quantitative analyses of the discoursemes and their constellations (see Section 4 for a brief discussion).

2 Discoursemes and Constellations

Let us start from the example of a collocation analysis in a CADS investigation. In order to define the node of the collocation analysis, a researcher will manually select a set of lemmata and/or lemma sequences that identify a topic of interest such as refugees (e.g. *refugee* and *displaced person*). They will then scan the table of collocations (or multiple tables obtained with different parameter settings) to spot groups of related words that reflect common

[2] https://github.com/ausgerechnet/cwb-cads

[3] See the interactive documentation on our own production server at https://corpora.linguistik.uni-erlangen.de/cwb-cads/docs.

discursive patterns associated with the topic. For example, collocates like *Syria* and *Lybia* indicate debates about the refugees' origin, while *displacement*, *expulsion*, and *famine* indicate "push factors" of migration. The collocates in a group tend to be semantically related, but this is not always the case. The key criterion is whether they express the same meaning aspect within the discourse (as the example of *famine* shows).

Our approach to overcoming the current limitations of CADS practice rests on understanding this central grouping step as the formation of *discoursemes*, which we define as (minimal) units of lexical meaning in the context of a given discourse. Our goal here is to provide an operational concept that has a clear hermeneutic definition (unit of meaning in the context of a discourse) but can also be approximated via lists of lemmata and thus identified automatically in corpora, forming a link between qualitative and quantitative methods. We enclose references to discoursemes in angle brackets, e.g. ⟨origin⟩ and ⟨push factors⟩ for the groups mentioned above. The node of the collocation analysis is also understood as a discourseme ⟨refugees⟩, which just happens to be defined a priori by the researcher rather than via grouping collocations.

It is worth pointing out that not all occurrences of a lemma 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 the discourseme. Our operationalisation of discoursemes as manually formed groups of lemmata must thus be considered an approximation, since there will be false positives (occurrences of these items that do not in fact belong to the discourseme) and false negatives (occurrences of the discourseme that are realised through other linguistic expressions that are not frequent enough to show up among the keywords and collocations).

Our approach also recognises explicitly that discursive patterns do not arise from individual discoursemes (as the qualitative interpretation in traditional CADS might suggest), but rather from *constellations* of discoursemes. The discourseme ⟨flood of people⟩ mentioned above might comprise lemmata like *flood*, *surge*, or *pour into*, but they only evoke the discursive pattern "migrants as a flood of people" when used in conjunction with ⟨refugees⟩, ⟨migration⟩ or a similar discourseme. Such constellations are often implicit in CADS studies: e.g. groups of collocates form discoursemes that co-occur in a constellation with the node discourseme of the collocation analysis. We make this explicit in our approach, where the node of a collocation analysis is always a discourseme. 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 important advantages for future CADS research:

1. The quantitative-qualitative bridge at the meso level of discourse analysis becomes more formalised and reproducible. Listing discoursemes (as sets of lemmata and lemma sequences) and their constellations can be regarded as a form of research documentation.
2. Discoursemes can be fed back into quantitative analyses and visualisations. We exemplify the usefulness of this in our case study below.
3. Discoursemes can be used as a starting point for further analysis steps, e.g. as node of a collocation analysis.
4. 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.
5. Statistical distributions of discoursemes can be determined (mostly) automatically, giving useful indications of the statistical distribution of discursive patterns (indicated by discourseme constellations).

3 Working with Discoursemes

As a running example, we will look into GermaParl[4], a corpus of all debates of the German federal parliament. Our goal is to describe discoursemes (via lists of lemmata) and to combine them into constellations that approximate discursive patterns, e.g. the framing of refugees as human beings in need for protection or questioning the legitimacy of seeking asylum.

Discoursemes can be created from the results of a collocation analysis, which puts them in a con-

[4] https://zenodo.org/records/10421773

stellation with the node discourseme of the analysis. We focus on the discourse around the discourseme ⟨fleeing⟩ (our "topic discourseme") in legislative period 19 (LP19), and on the parliamentary groups *Bündins90/Die Grünen* (GRUENE, a left-leaning environmentalist party) and *Alternative für Deutschland* (AfD, a right-wing populist party). However, we understand discourseme formation as an iterative process in which (i) different parameter settings for the same analysis can be used (e.g. different association measures or context definitions), and (ii) multiple analyses can be carried out (e.g. for different node discoursemes or (sub-)corpora). We concentrate on the formation of discoursemes via collocation analysis here. The API also supports an approach via keyword analysis (with a somewhat easier implementation). Of course, both approaches can be combined in a single study.

As mentioned above, discourseme descriptions for a given corpus are usually obtained by manually selecting lemmata from an n-best list of keywords or collocations, but they can also include multi-word units. Frequency counts for discoursemes are obtained in the same way as for individual lemmata, i.e. by counting all their occurrences in the corpus; some special precautions are necessary if a discourseme contains multi-word units (see Section 3.2). Such frequency counts are the basis for discourseme assocations (Section 3.6) as well as for further quantiative analyses.

The topic discourseme plays a special role in that it has to be defined a priori, and researchers have to take care not to miss relevant lemmata (or introduce false positives). For the example at hand, a manually curated list of lemmata is used based on the CQP query

```
[lemma=".*flucht.*" %cd]
```
[5]

Additional candidates can be suggested via semantic similarity search in word embeddings (Mikolov et al., 2013), which is supported by the API.

3.1 Collocations

Our first step is a collcation analysis for the topic discourseme ⟨fleeing⟩ as node. Co-occurrences are determined for all unigram lemmata in the specified context around the node discourseme; see Appendix A for a discussion of context types and their definition. In our API, we allow context specification by a mix of surface and textual co-occurrence. For the case study at hand we include all corpus positions up to $w = 10$ tokens around the node discourseme, but only in the same sentence.

Evert (2004, 68: fn. 23) recommends that the node itself should be removed from the co-occurrence context, as each of its instances would count as a co-occurrence with itself, inevitably leading to a very high (and spurious) association score. However, we argue here that the situation is different in CADS because the same lemma can belong to multiple discoursemes. Removing all the occurrences of all lemmata of the node discourseme from the context might inadvertently discard instances of other discoursemes. It is thus better, and technically easier, to work with the full context including the node. In order not to confuse analysts, the API masks the lemmata of the node discourseme by default, so they are not displayed in the semantic map visualisation (cf. Section 3.4).

Following contingency table notation (see Table 4 in Appendix A), we refer to the number of instances of a collocate within the context as O_{11} (with R_1 being the number of corpus positions in the context) and to the number of instances outside of the context as O_{21} (with R_2 the number of corpus positions in the remainder of the corpus). Note that this directly translates to keyword analyses, where O_{11} corresponds to the number of occurrences in the target corpus and O_{21} to the number of occurrences in the reference corpus. Since there is no "best" association measure, the API offers selection from a wide range of association measures.[6] We recommend starting with a measure that combines statistical significance with effect size, such as a log-likelihood-filtered odds-ratio or conservative log ratio (LRC) (Evert, 2022); see Appendix A for more details.

3.2 Multi-word units

The API allows the manual definition of multi-word units (MWUs) as lemma sequences.[7] MWUs can either form discoursemes by themselves or be included in a discourseme alongside unigrams and other MWUs. MWU matches span several corpus positions and may thus (partially) overlap with cor-

[5] This CQP query uses a regular expression to find all lemmata that contain the substring *flucht*; `%cd` tells CQP to perform a case-insensitive search and ignore diacratics.

[6] As implemented in the Python module `https://pypi.org/project/association-measures/`.

[7] As with the suggestion of similar items, the API can easily be extended to automatically suggest MWU candidates, e.g. by means of named-entity recognition.

pus positions of other lemmata within the same or in other discoursemes. Internally, all discourseme descriptions are translated into CQP queries and we set CQP's matching strategy to *longest* in order to count corpus positions at most once. As an example, consider the *Bundesamt für Migration und Flüchtlinge* (the German Federal Office for Migration and Refugees, BAMF). If this MWU were to be included in a discourseme description also comprising the unigram *Flüchtlinge*, only occurrences of *Flüchtlinge* that are not included in the MWU would be considered as additional matches.

Furthermore, MWUs can overlap partially with the context. For co-occurrence counts of discoursemes, we thus have to define how partial overlaps are counted. A simple approach would be to assume a co-occurrence for any partial overlap, i.e. if at least the start or the end token of the match span is included in the context (and we do in fact retrieve concordance lines for all these cases when discourseme constellations are inspected). However, to ensure mathematical consistency, we only count discoursemes as co-occurrences (towards O_{11}) if they are completely within the context, and all other occurrences as outside the context (towards O_{21}).

Alternatively, the API also allows to count partial overlaps as co-occurrences. To ensure mathematical consistency in this case, we have to obtain counts on token level, which means that a single occurrence of a MWU increases the frequency count (O_{11} or O_{21}) usually by more than one. This makes MWUs more sensitive to detection by association measures based on statistical significance, but leaves effect-size measures such as odds-ratio and the recommended LRC largely unaffected.

3.3 The choice of reference frequencies

Typically, reference frequencies O_{21} and O_{22} for collocation analyses are gained from the remainder of the corpus, i.e. all corpus positions that are not included in the context. Other approaches are possible, however, and these alternatives are especially important when working with subcorpora.

Table 1 shows association scores for collocates of the discourseme ⟨fleeing⟩ in the subcorpus of debates by the AfD in LP19, subject to different reference frequencies. It lists the top-20 candidates when compared against the entire remaining corpus (column "cf. GermaParl") and displays their reference frequency counts (O_{21}), association scores, and ranks when compared to other reference frequencies.[8]

Which frequency comparison is the most reasonable one? The three comparisons answer slightly different questions about the discourse around ⟨fleeing⟩ of AfD in LP19:

1. a comparison with the full corpus yields the collocation profile of ⟨fleeing⟩ as used by AfD in LP19,
2. a comparison with LP19 yields collocations of ⟨fleeing⟩ as used by AfD, against the background of the general discourse in LP19, and
3. a comparison with AfD in LP19 yields collocations of ⟨fleeing⟩ against the background of the overall AfD discourse in LP 19.[9]

As our goal is a collocation analysis for the discourseme ⟨fleeing⟩ (rather than, say, a keyword analysis for AfD or LP19), it may seem straightforward to prefer the third option. However, the very telling label *Mittelmeermigrant* ('mediterranean migrant') was coined by AfD in LP19 in the context of ⟨fleeing⟩ and does not occur anywhere else in the corpus. Due to its low frequency ($O_{11} = 2$) its association to ⟨fleeing⟩ is not significant within the small AfD-LP19 subcorpus, but is much more remarkable when compared against the entire corpus. For this reason, our API allows users to choose between the first and the third option.

3.4 Visualising collocation profiles

As has been pointed out above, the choice of association measure has a profound impact on collocational profiles (see Appendix A for a brief discussion). Although it is convenient to rely on a single measure, different association measures often provide complementary perspectives that need to be combined in order to capture the full picture. Some researchers have thus argued for a multidimensional visualisation of collocation profiles (Gries, 2019, 397ff), similar to the topographic maps in Figures 2 and 3 (Appendix A). Such maps can aid in understanding the different properties of association measures and provide a visual representation of the statistical profiles of collocates. However, the main task in CADS analyses is grouping

[8] We do not remove "stop words" from collocation profiles because punctuation marks, prepositions, etc. can be important for certain discourses. Our approach via semantic maps makes it easy for analysts to ignore such stop words.

[9] We do not include a comparison with the complete AfD subcorpus (across all periods) in this list for the simple reason that AfD only entered the federal parliament in LP19. Newer LP are not included in our version of GermaParl, so subcorpus AfD would be identical to AfD-LP19.

	cf. GermaParl				cf. LP19				cf. AfD-19			
item	rank	O11	O21	LRC	rank	O11	O21	LRC	rank	O11	O21	LRC
Deutschsprachförderung	1	5	17	6.95	4	5	15	3.50	10	5	6	1.38
Globale	2	18	669	6.89	1	18	298	4.43	2	18	52	3.65
Migration	3	31	2781	6.25	3	31	1390	3.64	3	31	262	2.86
Migrant	4	16	2423	4.70	8	16	781	2.72	25	16	378	0.61
BAMF	5	8	718	4.04	28	8	453	1.09	19	8	56	0.88
Gesundheitsfonds	6	9	1047	3.97	2	9	91	3.82	6	9	28	2.24
UNRWA	7	4	60	3.94	17	4	21	1.75	–	4	11	**0**
berufsbezogen	8	5	170	3.83	7	5	22	3.02	14	5	7	1.23
2015	9	18	6777	3.56	11	18	1368	2.26	12	18	291	1.32
sogenannter	10	54	44544	3.48	5	54	4105	3.31	5	54	902	2.34
Asylbewerber	11	13	4151	3.28	6	13	347	3.24	11	13	144	1.33
Mittelmeermigrant	12	2	0	3.17	–	2	0	**0**	–	2	0	**0**
Wirtschaftsmigrant	13	3	24	3.05	–	3	19	**0**	–	3	17	**0**
syrisch	14	9	1998	3.04	19	9	435	1.62	17	9	77	0.91
Bundesamt	15	16	9439	2.74	10	16	921	2.48	4	16	87	2.65
Pakt	16	13	6289	2.68	13	13	863	1.94	9	13	131	1.46
Erdogan	17	7	1348	2.55	54	7	642	0.02	–	7	146	**0**
Heimatland	18	8	2119	2.48	12	8	241	1.99	18	8	55	0.90
“	19	48	75571	2.47	49	48	31305	0.14	–	48	5063	**0**
Aufnahmegesellschaft	20	3	40	2.36	30	3	7	0.98	–	3	4	**0**

Table 1: Excerpt of collocation rankings of discourseme ⟨fleeing⟩ in AfD-19 subject to different reference frequencies. There are $R_1 = 13{,}344$ tokens in the context W (⟨fleeing⟩) in AfD-19; the reference corpora (excluding W) contain $R_2 = 271{,}064{,}105$ (GermaParl), $R_2 = 22{,}274{,}643$ (LP19), and $R_2 = 2{,}531{,}322$ (AfD-19) tokens, respectively. The table lists the top-20 collocates cf. GermaParl as ranked by conservative log-ratio (LRC).

collocates based on their discourse-specific semantics, not according to similarities in their frequency distribution or contingency tables.

A better way of supporting the manual grouping step is to visualise collocates in a semantic map, i.e. a two-dimensional projection that arranges collocates by their semantic similarity according to high-dimensional word embeddings. Although general semantic similarity is not the only criterion that discoursemes are based on (cf. the example of ⟨push factors⟩ above, which includes both *famine* and *expulsion*), most of the lemmata in a discourseme tend to be semantically similar in practice. A semantic map is therefore an excellent starting point for the grouping process. Our API combines the semantic map coordinates with the score of the selected association measure, which can be visualised by font size or other means (cf. Figure 1).

We use embeddings trained out-of-domain on German Wikipedia here. In principle, we could train embeddings on the corpus at hand to increase representativeness of the target domain. However, GermaParl is comparatively small with ca. 270 million tokens (compared to billions of tokens of Wikipedia) and CADS analyses are often carried out on much smaller corproa. To our knowledge there is no well-established way to fine-tune pre-trained embeddings on an in-domain corpus (except to train from scratch on the combined data). A further alternative is the use of context-sensitive embeddings, yielding a different representation for each occurrence of the same lemma depending on its context. Since the semantic map is a type-level visualisation, a global representation would have to be obtained, e.g. by averaging over all individual token embeddings in the target corpus. Note that administrators can easily prepare and deploy such global context-sensitive embeddings, giving a high degree of flexibility to the API.

For the two-dimensional projection, we use t-distributed stochastic neighbour embedding (van der Maaten and Hinton, 2008) by default, but other techniques can also be selected; the API e.g. offers uniform manifold approximation and projection as an alternative (McInnes et al., 2018).

3.5 Comparing collocation profiles

The semantic map in Figure 1 also allows for a qualitative comparison of collocation profiles. We can e.g. observe on the right-hand side of Figure 1 that GRUENE talks about the ⟨risk⟩ that refugees are taking (*Lebensgefahr*, ‘risk of death’) whereas *angeblich* (‘alleged(ly)’), *sogenannt* (‘so-called’), and the use of quotation marks indicates that the

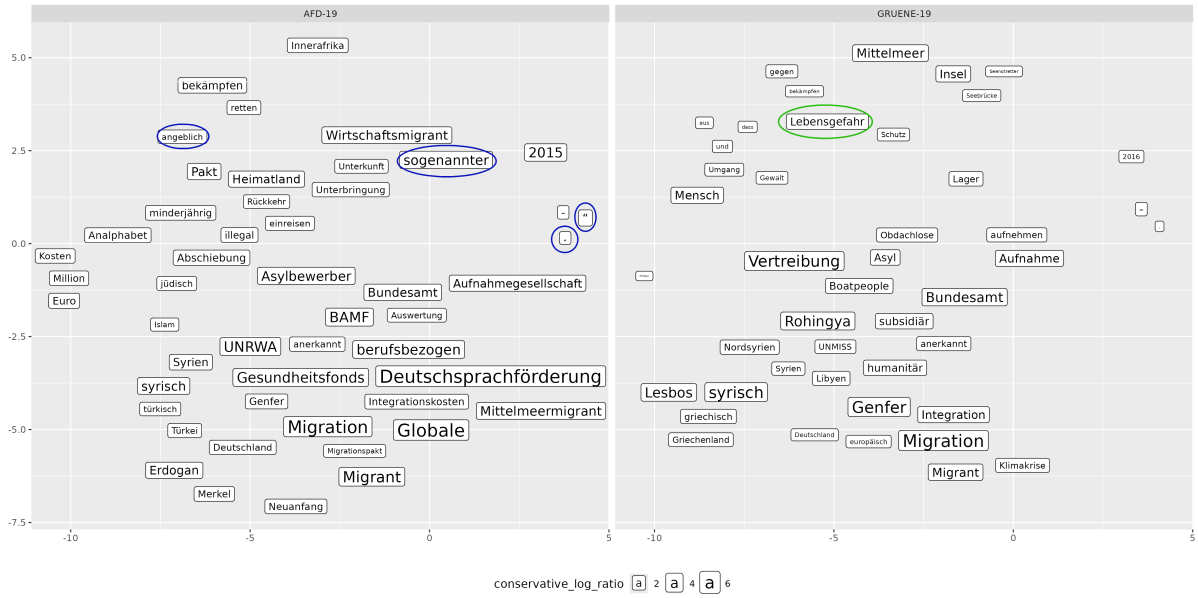

Figure 1: Semantic map visualisation of the collocation profiles of discourseme ⟨fleeing⟩ in two subcorpora (left panel: AfD in LP19, right panel: GRUENE in LP19). Both profiles are cf. GermaParl.

AfD is doubtful about the official narrative. Such qualitative comparisons of collocation profiles are often very fruitful in CADS studies. We argue that semantic maps are highly effective for this purpose: collocates appear at the same coordinates in both panels of Figure 1 rather than at entirely different ranks in two n-best lists, aiding in a direct visual comparison.

For a quantitative comparison of two given collocation profiles (or keyword lists), several approaches are available. One possibility is rank-biased overlap (Webber et al., 2010):

$$\mathrm{rbo}(P_1, P_2; p) = (1-p)\sum_{d=1}^{\infty} p^{d-1} A(d)$$

where $A(d)$ is the proportion of shared lemmata in the top d ranks of profiles P_1 and P_2, and the sensitivity parameter p controls the depth of the comparison. We are planning to support such quantitative comparisons in a future version of the API.

Furthermore, we plan to include a quantitative method for analysing collocation profiles over time that aims to find disruptions in the usage of words ("usage fluctuation analysis", UFA) (McEnery et al., 2019, 418). In UFA, a corpus must be partioned into overlapping sliding windows across time; subsequently, the resulting profiles are iteratively compared providing a scalar value, and finally a statistical regression model is estimated to detect outliers in the corresponding time series.[10]

3.6 Discourseme associations

In order to identify discourseme constellations, it is quite straightforward to look at pairwise co-occurrences of discoursemes and their assocation strength. For this purpose, one discourseme is taken as the node of a collocation analysis and the co-occurrences of all lemmata and lemma sequences from the other discourseme are added up. This is reasonable because items from the same discourseme do not overlap, which is ensured by our query-based matching (cf. Section 3.2). Focussing iteratively on each discourseme in the database, pairwise associations between all discoursemes can be calculated, yielding a network structure with discoursemes as nodes and discourseme associations as edges.

As an illustration, Table 2 shows association scores between ⟨fleeing⟩ and a tentative (and incomplete) set of discoursemes created from its collocation profiles in the subcorpora GRUENE-LP19 and AfD-LP19 (cf. GermaParl). Because of the way the discoursemes were formed, all entries in the table are relevant constellations with ⟨fleeing⟩. The API yields both a global association score for each discourseme and individual scores for its lem-

[10] McEnery et al. (2019) use Gwet's AC1 (Gwet, 2001) to compare profiles. The formula is similar to the Cohen's Kappa but incorporates a different method for estimating the probability of chance agreement, which helps mitigate the issues associated with marginal imbalances.

		GRUENE-19			AfD-19		
discourseme	lemma (sequence)	O11	O21	LRC	O11	O21	LRC
⟨BAMF⟩		19	1235	7.44	24	1230	7.50
—"—	Bundesamt für Migration und Flüchtling	16	512	8.08	16	512	7.64
—"—	BAMF	3	723	1.62	8	718	5.25
⟨migration⟩		36	5215	6.68	47	5204	6.75
—"—	Migration	28	2784	7.02	31	2781	6.79
—"—	Migrant	8	2431	4.17	16	2423	5.55
⟨origin⟩		39	18268	5.03	44	18263	4.82
—"—	syrisch	14	1993	5.73	9	1998	4.04
—"—	Boatpeople	2	4	5.66	0	6	0.00
—"—	Rohingya	5	208	5.65	1	212	0.00
—"—	Nordsyrien	3	120	3.94	0	123	0.00
—"—	Libyen	6	2938	2.56	1	2943	0.00
—"—	Syrien	7	6066	2.06	11	6062	3.03
—"—	Heimatland	0	2127	0.00	8	2119	3.58
—"—	Innerafrika	0	5	0.00	2	3	5.51
—"—	Islam	0	1439	0.00	5	1434	2.44
—"—	jüdisch	2	3368	0.00	7	3363	2.47
⟨push⟩		17	2939	5.95	8	2948	3.60
—"—	Vertreibung	17	2912	5.87	5	2924	1.91
—"—	Wirtschaftsmigrant	0	27	0.00	3	24	6.53
⟨route⟩		32	31146	3.88	31	31147	3.37
—"—	Lesbos	4	98	5.87	0	102	0.00
—"—	Mittelmeer	8	2246	4.02	3	2251	0.00
—"—	Lager	7	4475	2.59	2	4480	0.00
—"—	Griechenland	8	8095	2.17	7	8096	1.29
—"—	Mittelmeermigrant	0	2	0.00	2	0	7.44
—"—	Türkei	5	15155	0.00	12	15148	2.01
—"—	einreisen	0	1075	0.00	5	1070	2.97
⟨asylum⟩		8	7131	2.77	16	7123	4.09
—"—	Asyl	7	2968	3.48	3	2972	0.00
—"—	Asylbewerber	1	4163	0.00	13	4151	4.28
⟨accommodation⟩		46	60400	3.62	40	60406	2.91
—"—	Aufnahme	12	8419	3.32	8	8423	1.70
—"—	Integration	15	14882	3.08	8	14889	0.88
—"—	aufnehmen	15	31725	1.98	8	31732	0.00
—"—	Aufnahmegesellschaft	0	43	0.00	3	40	5.29
—"—	Unterbringung	2	3011	0.00	7	3006	2.76
—"—	Unterkunft	2	2320	0.00	6	2316	2.60
⟨collaboration⟩		6	13962	0.93	36	13932	4.83
—"—	Seebrücke	2	14	4.32	0	16	0.00
—"—	Erdogan	1	1354	0.00	7	1348	3.84
—"—	Globale Flüchtlingsforum	0	16	0.00	1	15	0.00
—"—	Migrationspakt	0	138	0.00	3	135	3.45
—"—	Pakt	0	6302	0.00	13	6289	3.47
—"—	UNRWA	0	64	0.00	4	60	6.06
—"—	türkisch	3	6074	0.00	8	6069	2.11
⟨help⟩		26	49256	2.80	26	49256	2.36
—"—	Seenotretter	2	16	4.24	0	18	0.00
—"—	subsidiär	5	823	3.79	2	826	0.00
—"—	UNMISS	4	673	3.12	0	677	0.00
—"—	Schutz	14	42054	1.38	11	42057	0.30
—"—	berufsbezogen Deutschsprachförderung	0	20	0.00	5	15	8.89
—"—	retten	1	5670	0.00	8	5663	2.24
⟨risk⟩		4	340	4.84	–	–	–
—"—	Lebensgefahr	4	340	4.84	–	–	–
⟨doubt⟩		–	–	–	65	58177	3.86
—"—	angeblich	–	–	–	11	13633	2.13
—"—	sogenannter	–	–	–	54	44544	3.87

Table 2: Tentative (and incomplete) discourseme formation for collocations of ⟨fleeing⟩ in two subcorpora (R_1 = 9,830, R_2 = 271,067,619 in GRUENE-19 and R_1 = 13,344, R_2 = 271,064,105 in AfD-19). Both scores of individual lemmata and lemma sequences and global discourseme scores are provided.

mata and lemma sequences. The discourseme associations provide a bird's-eye view on the distribution across subcorpora: we can e.g. see that the ⟨BAMF⟩ (the German Federal Office for Migration and Refugees) plays a role for both GRUENE and AfD, whereas ⟨collaboration⟩ is clearly only associated with AfD. Associations for individual items give a more detailed view, revealing e.g. that although ⟨accommodation⟩ is associated with both parliamentary groups, AfD has a particularly high association for *Aufnahmegesellschaft* ('receiving society'), which might prompt us to reconsider the inclusion of this lemma in the discourseme.

In our approach, discoursemes are usually created and extended iteratively by working with different parameter settings and in different subcorpora. This also has an impact on discourseme associations, cf. Table 3. Working solely on a collocation profile of ⟨fleeing⟩ in the subcorpus of GRUENE in LP19, for instance, would not have brought up the lemmata *Unterbringung* ('accommodation') und *Unterkunft* ('lodging') for the discourseme ⟨accommodation⟩. Inclusion of these lemmata does however change its association with ⟨fleeing⟩, increasing the LRC score from 3.00 to 3.11, even though the two additional lemmata are not significant by themselves (with an LRC of 0).

item	O11	O21	LRC
⟨accommodation⟩	27	40144	3.00
. . . Aufnahme	12	8419	3.32
. . . aufnehmen	15	31725	1.98
⟨accommodation⟩	31	45475	3.11
. . . Aufnahme	12	8419	3.32
. . . aufnehmen	15	31725	1.98
. . . Unterbringung	2	3011	0.00
. . . Unterkunft	2	2320	0.00

Table 3: Two alternative definitions for discourseme ⟨accommodation⟩. Frequencies taken from subcorpus GRUENE in LP19 cf. GermaParl ($R_1 = 9{,}830$, $R_2 = 271{,}067{,}619$).

4 Working with the API

As outlined in the introduction, interaction with the API is possible both from dedicated GUIs and through low-level API calls e.g. from widespread languages such as Python or R. Both interaction methods operate on the same discourseme database, ensuring consistency across tools while giving users the freedom to select the most convenient option for their analysis.

Typically, a graphical frontend is ideal for tasks such as defining the topic discourseme, forming discoursemes based on collocation profiles or keyords, and examining concordance lines of discoursemes or individual lemmata. We have already experimented with a prototype frontend[11] and are currently developing an improved version, which can be found in the `cwb-cads` repository linked above. Development of the new frontend, implemented in React, focusses on flexible selection of semantic maps, more straightforward definitions of MWUs, and efficient recalculations.

For operations such as analysing the distribution of discoursemes or discourseme networks across metadata variables, or exporting discourseme descriptions and individual concordance lines for research documentation, API calls are more suitable. Manuals for working with the API are available in the `cwb-cads` repository.

5 Conclusion and Future Work

We have presented a CWB-based REST API that aims to provide convenient methods for CADS researchers, offering a variety of parameter choices to enable customised and comprehensive research. We have outlined the available parameters and provided guidelines on making reasonable selections.

A significant contribution of our work is the conceptual and technical framework for working with manually defined semantic groups ("discoursemes"). The paper includes details on the calculation of discourseme scores and how to tackle the challenges associated with multi-word units (MWUs) and overlapping discoursemes.

It is worth mentioning that our approach necessarily shares the same limited perspective on discourse as classic CADS. Working on word or lemma types means neglecting word-sense or discourse-specific disambiguation. However, this is somewhat mitigated by the fact that discoursemes are mutually disambiguated within constellations.

We plan to expand the API by adding more parameters to its endpoints, further increasing flexibility. Most importantly, while the current implementation only supports pairwise associations of discoursemes, we aim to visualise these associations as discourseme networks.

[11] https://github.com/fau-klue/mmda-toolkit

Acknowledgements

This work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project no. 466328567.

References

Laurence Anthony. 2013. A critical look at software tools in corpus linguistics. *Linguistic Research*, 30:141–161.

Paul Baker. 2006. *Using Corpora in Discourse Analysis*. Continuum, London.

Paul Baker, Costas Gabrielatos, Majid KhosraviNik, Michał Krzyżanowski, Tony McEnery, and Ruth Wodak. 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.

Paul Baker and Tony McEnery. 2005. A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts. *Journal of Language and Politics*, 4(2):197–226.

Stefan Evert. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart.

Stefan Evert. 2008. Corpora and collocations. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook*, chapter 58, pages 1212–1248. Mouton de Gruyter, Berlin, New York.

Stefan Evert. 2009. 58. corpora and collocations. In Anke Lüdeling and Merja Kytö, editors, *Volume 2*, pages 1212–1248. De Gruyter Mouton, Berlin, New York.

Stefan Evert and Andrew Hardie. 2011. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 Conference*, Birmingham, UK.

Stefan Evert and The CWB Development Team. 2022. *The IMS Open Corpus Workbench (CWB) CQP Interface and Query Language Tutorial*. CWB Version 3.5.

Stefan Evert, Peter Uhrig, Sabine Bartsch, and Tobias Proisl. 2017. E-view-alation – a large-scale evaluation study of association measures for collocation identification. In *Electronic lexicography in the 21st century. Proceedings of the eLex 2017 conference*, pages 531–549, Leiden, The Netherlands.

Stephanie Evert. 2022. Measuring keyness. In *Digital Humanities 2022*, pages 202 – 205.

Norman Fairclough. 2015. *Language and Power*, 3 edition. Routledge, Oxon.

Costas Gabrielatos and Paul Baker. 2008. Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the uk press, 1996-2005. *Journal of English Linguistics*, 36(1):5–38.

Tim Griebel, Stefan Evert, and Philipp Heinrich, editors. 2020. *Multimodal Approaches to Media Discourses: Reconstructing the Age of Austerity in the United Kingdom*. Routledge, London.

Stefan Th. Gries. 2019. 15 years of collostructions: Some long overdue additions/corrections (to/of actually all sorts of corpus-linguistics measures). *International Journal of Corpus Linguistics*, 24(3):385–412.

Stefan Th. Gries. 2021. A new approach to (key) keywords analysis: Using frequency, and now also dispersion. *Research in Corpus Linguistics*, 9(2):1–33.

Reiner Grundmann and Ramesh Krishnamurthy. 2010. The discourse of climate change: a corpus-based approach. *Critical Approaches to Discourse Analysis Across Disciplines*, 4(2):125–146.

Kilem Gwet. 2001. *Handbook of Inter-Rater Reliability: How to Estimate the Level of Agreement Between Two or Multiple Raters*. STATAXIS Publishing Company, Gaithersburg, MD.

Andrew Hardie. 2012. CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3):380–409.

Andrew Hardie. 2014. A single statistical technique for keywords, lockwords, and collocations. Internal CASS working paper no. 1, unpublished.

Philipp Heinrich, Andreas Blombach, Bao Minh Doan Dang, Leonardo Zilio, Linda Havenstein, Nathan Dykes, Stephanie Evert, and Fabian Schäfer. 2024. Automatic Identification of COVID-19-Related Conspiracy Narratives in German Telegram Channels and Chats. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, page 1932–1943, Turin, Italy.

Siegfried Jäger. 2015. *Kritische Diskursanalyse: Eine Einführung*. UNRAST.

Gerlinde Mautner. 2009. Corpora and critical discourse analysis. In Paul Baker, editor, *Contemporary Corpus Linguistics*, pages 32–46. Continuum, London/New York.

Tony McEnery, Vaclav Brezina, and Helen Baker. 2019. Usage Fluctuation Analysis: A new way of analysing shifts in historical discourse. *International Journal of Corpus Linguistics*, 24(4):413–444.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *ArXiv e-prints*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Preprint*, arXiv:1301.3781.

Costanza Navarretta and Dorte H. Hansen. 2023. According to BERTopic, what do Danish parties debate on when they address energy and environment? In *Proceedings of the 3rd Workshop on Computational Linguistics for the Political and Social Sciences*, pages 59–68, Ingolstadt, Germany. Association for Computational Linguistics.

Joan Plepi, Magdalena Buski, and Lucie Flek. 2023. Personalized intended and perceived sarcasm detection on Twitter. In *Proceedings of the 3rd Workshop on Computational Linguistics for the Political and Social Sciences*, pages 8–18, Ingolstadt, Germany. Association for Computational Linguistics.

Michael Stubbs. 1995. Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language*, 1:23–55.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605.

Guofeng Wang and Changpeng Huan. 2023. Negotiating climate change in public discourse: insights from critical discourse studies. *Critical Discourse Studies*, 21(2):133–145.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28:20.

Maximilian Wegge and Roman Klinger. 2023. Automatic emotion experiencer recognition. In *Proceedings of the 3rd Workshop on Computational Linguistics for the Political and Social Sciences*, pages 1–7, Ingolstadt, Germany. Association for Computational Linguistics.

Ruth Wodak. 2015. *The Politics of Fear: What Right-Wing Populist Discourses Mean*. SAGE, London.

Ruth Wodak. 2018. *Discourse and European Integration*, volume 86 of *KFG Working Paper Series*. Freie Universität Berlin, FB Politik- und Sozialwissenschaften, Otto-Suhr-Institut für Politikwissenschaft Kolleg-Forschergruppe "The Transformative Power of Europe", Berlin.

Ruth Wodak and Michael Meyer. 2015. *Methods of Critical Discourse Studies*, chapter Critical discourse studies: history, agenda, theory and methodology. Sage.

A Parameters in Collocation Analyses

Context settings and the choice of association measure can have a huge influence on the outcome of a collocation analysis. Evert (2009) distinguishes three types of co-occurrence:

1. surface co-occurrence – where one counts up to w tokens in any direction (asymmetrical windows are obviously possible),
2. textual co-occurrence – using the whole sentence, paragraph, post, etc. as context,
3. syntactic co-occurrence – which we will ignore here because it presupposes reliable syntactic annotation and does not generalise for various parts of speech that can be included in discoursemes.

As mentioned above, the API supports a combination of surface and textual co-occurrences, defining context via a window span w and a structural context break (e.g. texts, paragraphs, or sentences). For small context windows w, collocates are e.g. often part of multi-word expressions rather than indicating discourseme constellations. Confining the context to individual texts is especially important for corpora with small "natural" text units such as tweets. A large shortcoming of CQPweb is that the context of a collocation analysis cannot be confined to individual texts (or sentences), and collocation analyses on Twitter corpora are thus oftentime misleading (since the context often includes the last or first couple of tokens of different tweets).

Given a well-defined context, all occurrences of (unigram) types can be directly classified as being inside or outside the context.[12] In "contingency table notation", these numbers are named O_{11} and O_{21}, respectively, with O_{12} the remaining number of corpus positions within the context and O_{22} the remaining number of corpus positions outside the context.

Statistical association measures allow the calculation of a single scalar value to quantify the association, either in terms of the effect size or in terms of statistical significance – or by some other heuristic (e.g. motivated by information theory). Straightforward measures are the ratio of relative frequencies (measuring the effect size) or log-likelihood ratio (measuring statistical significance). Note that effect-size measures are biased

[12] This works similar for keyword analyses, where O_{11} is the number of occurrences in the target corpus and O_{21} the number of occurrences in the reference corpus; R_1 and R_2 being the respective sizes of the corpora.

	w_2	$\neg w_2$	
$W(\langle d \rangle)$	O_{11}	O_{12}	$= R_1$
$\neg W(\langle d \rangle)$	O_{21}	O_{22}	$= R_2$
	$= C_1$	$= C_2$	$= N$

Table 4: Contingency table notation: For a focus discourseme $\langle d \rangle$ and a given lemma w_2, all lemmata in the corpus are categorised according to whether they appear within the context W of $\langle d \rangle$ or its complement (rows), and whether they are a realisation of w_2 or not (columns). The row marginals are named R_1 and R_2, the column marginals C_1 and C_2, respectively; the total number of tokens in the corpus is N.

to low-frequency terms. The association measure presented by Evert (2022) and recommended in this paper combines effect size and statistical significance: it is the binary logarithm of the lower bound of the confidence interval of relative risk.

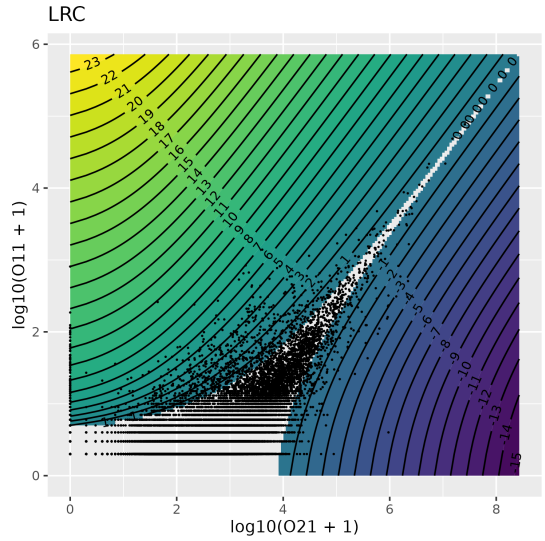

Figure 2: Topographic map for collocation profile of discourseme ⟨fleeing⟩ in GermaParl using conservative log-ratio as association measure.

As mentioned above, a wide variety of association measures are implemented in the Python package `association-measures`. This package also allows the creation of *topographic maps*, which visualise association measures in form of contour plots above a two-dimensional plane spanned by (the logarithm of) the number of occurrences in the target (the cotext) and the reference corpus (the remaining corpus). Figure 2 and Figure 3 show such topographic maps for the collocation profile of discourseme ⟨fleeing⟩ in GermaParl subject to two different association measures ($R_1 = 725{,}839$, $R_2 = 270{,}351{,}610$). Each point represents a collocation candidate of discourseme ⟨fleeing⟩ (which are identical in both figures).

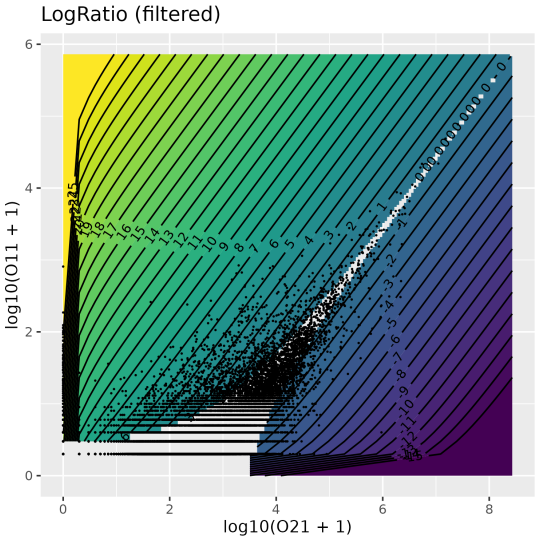

Figure 3: Topographic map for collocation profile of discourseme ⟨fleeing⟩ in GermaParl using log-ratio filtered by log-likelihood as association measure ($\alpha = 99.9\%$)

Both conservative log-ratio and log-ratio filtered by log-likelihood-ratio combine effect size and statistical significance: candidates are ranked high (upper left corner) if they appear frequently and relatively more frequently in the context than outside. Note that the two measures differ mainly in their decision for low-frequency candidates, where it is debatable how much statistical evidence is needed to support some observable effect.