Discourse-Level Features in Spoken and Written Communication

Hannah J. Seemann[1], Sara Shahmohammadi[2], Manfred Stede[2], Tatjana Scheffler[1]
[1] Department for German Language and Literature, Ruhr-University Bochum, Germany
[2] Department of Linguistics, University of Potsdam, Germany
hannah.seemann@rub.de shahmohammadi@uni-potsdam.de
stede@uni-potsdam.de tatjana.scheffler@rub.de

Abstract

Using PARADISE, a German corpus of thematically parallel blog posts and podcast transcripts, we test how reliably a document's original mode can be classified based on discourse-level features only. Our results show that classifying mode with a document's distribution of discourse relations as well as the frequency of discourse connectives and discourse particles is possible and informative of the nature of these document types. We provide our dataset annotated with discourse relations (Rhetorical Structure Theory), German discourse connectives, and discourse particles.

1 Introduction

The comparison of spoken and written language has a long tradition. Distinctions between different types of language use are usually based on both language-internal (e.g., lexical) and -external (e.g., situational) features. Since the description of conceptional orality as a continuum by Koch and Oesterreicher (1985) and others, studies on variation between spoken and written, personal and interpersonal communication have been conducted by, e.g., Biber et al. (1999) and Biber and Conrad (2019). Describing different *registers* in the continuum of language, the authors identify various features typical for specific types and settings of communication.

But not only do texts differ in how closely they match typical patterns of written or spoken language, the introduction of digital communication channels has also added to the variety of the continuum. On the one hand, we can find characteristics of spoken language in written computer-mediated communication (CMC) (Scheffler, 2017), and on the other hand, spoken computer-mediated interaction differs from the speech style before the introduction of digital communication channels, as the interlocutors are aware of a broader range of communicative practices and make use of them (Soffer, 2010; Heyd, 2021).

However, even if there have been comprehensive corpus-based studies on variation in spoken and written language that cover many different features, most of these features are lexical or syntactic in nature. Related work on the orality of texts shows that classification between spoken and written modes based on various feature levels works quite well. Though language and language variation is thought to be influenced by non-linguistic factors such as a speaker's communicative goal and the context of language production (Biber and Conrad, 2019), only few of these studies take explicitly discourse-related phenomena into account.

We study whether structural differences in a text, such as its discourse structure, together with discourse-related lexical features, such as discourse connectives and discourse particles, can be leveraged for a classification of language mode in CMC. As features like the average word length of a text and type-token ratio have been shown to systematically vary between spoken and written language (Kunz et al., 2018), we compare our results to a classification based on these two features. To account for different communicative practices, we compare spoken and written communication in digital communication channels: blog posts and podcasts. The dataset used in our study contains excerpts from blog posts and podcasts aligned for topic, to further reduce factors that might influence a text outside of its mode.

We provide our dataset of German blog posts and podcast transcripts, annotated with RST-style discourse relations as well as discourse connectives and particles, enabling further studies on both discourse structure in two modes and implicit and explicit discourse relations. So far, there are only very few discourse-annotated German corpora

(Stede, 2016a; Hewett, 2023), and thus our work contributes both additional training data (for example for RST parsing) and material for qualitative studies on the interactions of discourse phenomena.

2 Related Work

There are three types of related studies: Studies that identify relevant dimensions for distinguishing between different modes of language, studies that work on the automatic classification of language modes based on various linguistic features, and work that identifies discourse-related features that vary between language modes.

Kunz et al. (2018) compare various spoken and written registers (in English and German) based on "shallow" lexical features of lexical cohesion (Kunz et al., 2018, p. 176), such as type-token ratio, word class profiles, and frequently used words. They find less lexical variability in spoken than in written texts,[1] and that this difference is even more prominent in English than in German.

Ortmann and Dipper (2019) study a German corpus of five registers on different positions on the 'conceptual orality' scale (Koch and Oesterreicher, 1985) – from written newspaper texts to informal spoken conversations – showing that a decision tree classifier can reliably distinguish between them based on various linguistic features. The features used in the study are lexical (e.g., use of particles and interjections), syntactic (e.g., use of active vs. passive voice), measures of complexity (sentence and word length), co-reference and deixis (e.g., pronouns), and punctuation.

Lapshinova-Koltunski and Zampieri (2018) use part-of-speech n-grams to distinguish between German and English translations of fictional texts and political speeches, finding that the n-grams differ systematically enough between the two modes for a reliable classification (81.25% accuracy) based on a Bayesian classifier with Laplace smoothing.

The above-mentioned approaches rely primarily on lexical differences between speech and writing, or on a mix between all feature types. In contrast, some recent work more directly addresses the question of whether the discourse structure must be adapted when information is presented in spoken or written mode. Existing work focuses on co-reference phenomena as well as the marking of coherence relations.

Aktaş and Stede (2020) quantitatively study the features of co-reference in spoken and written English language. Comparing language corpora of both modes, they show that the distance between anaphoric pronouns and their antecedents is longer in spoken texts than in written texts. Co-reference thus seems to be a discourse feature that can be used reliably to distinguish spoken and written texts.

Using different corpora of spoken Italian and Irish English, Tonelli et al. (2010) and Rehbein et al. (2016) report on the distribution of discourse relations and their explicit marking through connectives. Tonelli et al. (2010) observe a smaller variety of connectives being used for multiple relations in telephone speech, whereas in the written newspaper texts of the Penn Discourse Treebank (PDTB, Prasad et al., 2008, 2019), more specific connectives are used. They also observe that compared to the PDTB, explicit relations are twice as frequent as implicit ones in speech, and TEMPORAL as well as CAUSAL relations are more frequent than other relations. These results are confirmed by Rehbein et al. (2016), who report twice as many explicit than implicit relations in speech. Both studies concede that the differences in relation frequency could arise based on the domain differences between the two corpora, so Rehbein et al. (2016) additionally compare spoken broadcasts to telephone conversations. They find that in conversations, CAUSAL relations are by far more frequent than in the broadcasts, thus the distribution of relation types seems to depend on a text's register rather than its mode.

In the following, we analyze a corpus that is matched for topic and audience in order to investigate whether discourse level differences between the spoken and written mode can be shown there as well.

3 Data & Preprocessing

PARADISE (PARAllel DIScoursE)[2] is a corpus of German computer-mediated communication in different media – written blog posts, and the transcripts of spoken podcast episodes. It is in parts annotated for discourse structure, and comprises 69 blog posts and the corresponding podcast transcripts from two different topic domains: science/culture and business. Each blog post is written

[1] We use 'text' to refer to any coherent linguistic entity, independent of its mode. In this sense, a podcast is a spoken text.

[2] The corpus is freely available at: https://osf.io/59acq/

by the podcast host to describe the content of the podcast, so the same topics are covered in both modes of communication.

Episodes from the science/culture domain are taken from Metaebene[3] and cover topics like scientific achievements, astronomy, German media, and politics. Episodes from the business domain are produced by companies or registered associations like Deutsche Telekom or Verbraucherzentrale, topics covered are digitalization, start-ups, health and food as well as the food industry. The design of the podcasts differs between the two domains. Podcasts in the science/culture domain are either 1:1 interviews with one host and one guest who is an expert on the episode's topic, or group conversations with varying members from a fixed pool of participants. In the business podcasts, one host or a team of hosts invites a number of guests to advertise a product or concept, which resembles a scripted interview more than a naturally flowing conversation about a fixed topic.

If no transcript of the podcast was provided by the host, the audio files were transcribed automatically. For both blog posts and podcasts, we used SoMaJo[4] for tokenization and sentence splitting (Proisl and Uhrig, 2016). We manually checked for and corrected errors in the word level transcription of the audios as well as errors in sentence splitting.

To allow for an exact analysis of how similar content is talked about in the two modes 'spoken' and 'written', we identified text chunks with parallel content: For each sentence/text chunk in a blog post covering a specific topic, the corresponding parallel segment in the podcast is identified (if there is one). To do so, one annotator was instructed to annotate all parts in the transcript that are parallel to each sentence in the blog post by marking keywords in the blog post and searching for them in the podcast. The resulting parallel segments may cover turns of more than one speaker if the utterances are related to the topic in question. In an additional step, two annotators rated how parallel each previously annotated segment is. The degree of parallelism is indicated by the following scale:

A: Perfectly parallel segments, where the blog and the transcript address the same content and even use the same wording/share the same expressions.

B: Good parallelism, the blog and the transcript address the same content but one goes into more detail than the other.

C: Medium parallelism, the blog and the transcript have the same topic and share some part of the content, but they also contain 'non-parallel' information, i.e. some content that the other one is lacking.

D: Low parallelism, the blog and the transcript address the same topic but do not have the same content.

E: Non parallel segments, where the blog and the transcript do not even share the same topic.

For example, the parallelism between the segments below – part of the blog post in (1) and part of the transcript in (2) – is rated as belonging to category B: The podcast elaborates on the topic mentioned in the blog.

(1) And: What goals is the Stifterverband pursuing with its new podcast project?[5]

(FG000B)

(2) And that was probably kind of the idea behind this new project, which we have christened Forschergeist. Which may initially give the impression that we only want to talk about research, and that is what we want. But we actually want to talk a bit more about the spirit of research. In other words, what drives people, what drives foundations, what, let's say, certain foundations actually aim to achieve with their objectives. But also to present the work of science.[6]

(FG000T)

We used weighted Krippendorff's α to calculate inter-annotator agreement (α = 0.53). Out of 406

[3] https://metaebene.me/

[4] https://github.com/tsproisl/SoMaJo

[5] German original: *Und: Welche Ziele verfolgt der Stifterverband mit seinem neuen Podcast-Projekt?*

[6] German original: *Und das war dann wahrscheinlich auch ein wenig der Gedanke hinter diesem neuen Projekt, was wir Forschergeist getauft haben. Was vielleicht so erst mal den Eindruck macht, als würde man jetzt nur über Forschung reden wollen, aber das wollen wir auch. Aber wir wollen eigentlich ein bisschen mehr über den Forschergeist eigentlich sprechen. Also das was so die Leute treibt, was Stiftungen treibt, was sagen wir mal vielleicht auch bestimmte Stiftungen überhaupt mit ihren Stiftungszielen bezwecken. Aber eben auch die Arbeit der Wissenschaft vorzustellen.*

parallel segments, the two annotators agreed in 223 (54.92%) of all cases. In 150 cases (36.94%), their ratings deviated from each other by one step, e.g., annotator 1 assigned label 'B' and annotator 2 assigned label 'C'. In 33 cases (8.12%), the ratings lay farther apart. A third annotator assigned a final label for each case where the two annotators did not initially agree. The distribution of the final labels is presented in Table 1.

A	B	C	D	E
6.87	58.47	27.51	6.38	0.73

Table 1: Distribution of parallelism labels, in %.

For the task presented here, we use chunks that are parallel but not identical (= category B). The size of the sub-corpus used is presented in Table 2.

	Blogs	Transcripts	Total
Science	2,411	30,416	32,827
Business	788	4,814	5,602
Total	3,199	35,230	38,838

Table 2: Token count in the sub-corpus used for this task, by medium and domain.

Table 3 shows the type-token ratio and the average word length in our dataset. In general, the type-token ratio and average word length are higher in blogs than in podcasts, which matches the characteristics of spoken and written language.[7] The mean length of blog chunks is 70 token (with a standard deviation of 44), compared to 756 (SD 649) in podcasts, reflecting that there is usually more elaboration on a certain point in the spoken conversation.

	Blogs	Transcripts
TTR	0.83	0.49
Avg. word length	5.52	4.51

Table 3: Type-Token Ratio and average word length, by medium.

[7]A reviewer pointed out that these measures are not normalized and do reflect differences in the document's length. However, normalization methods are influenced and limited by the length of the shortest document, which is rather short in our case. The resulting measure would not necessarily be more informative and therefore, we decided against normalizing type-token ratio and average word length.

4 Annotation of Discourse Features

We manually annotate discourse relations, discourse connectives and discourse particles in the dataset described above. This combination of structural and lexical discourse-level features allows us to account for variability between modes as well as the structural differences between modes that go beyond the lexical level.

4.1 Rhetorical Structure Theory

Rhetorical Structure Theory (RST, Mann and Thompson, 1988) is a model of discourse structure that captures the intentions of and rhetorics used by an author of a written text. In this model, discourse structure is represented as a hierarchical tree built from *Elementary Discourse Units* (EDUs, often clauses) and the discourse relations that span between them as well as hierarchically between complex units. RST was mainly designed for written monologues but has been applied to spoken data before (Stent, 2000; Shahmohammadi et al., 2023). Example (3) shows two EDUs connected by a CAUSE relation.

(3) [The cat was tired during the day] [because it was up running around the whole night.]

In our RST annotation, we follow the guidelines defined in Stede (2016b). To account for particularities of spoken language, we added a COMPLETION relation that holds between segments of interrupted utterances. Most of the RST annotation was conducted by one trained annotator and a second trained annotator annotated parts of the corpus. We used RST-Tace (Wan et al., 2019) to evaluate the agreement between the two annotators. With Fleiss $\kappa = 0.49$,[8] our inter-annotator agreement is comparable to similar complex annotation tasks. Except for Hewett (2023), who reports on averaged Fleiss $\kappa = 0.27$, a direct comparison to other RST annotations is not possible, either because no agreement is reported or a different evaluation method is used.

We grouped the annotated discourse relations by their semantic characteristics: All relations that express a causal relation between two segments are grouped as CAUSAL, etc. An overview of our groups is shown in Table 4. For the classification task, we use the proportion of each relation type

[8]The evaluation score we report is based on the annotation of 27 documents (= 470 EDUs) and averaged over nuclearity, spans, attachment points, and relation agreement.

Semantic Category	Discourse Relations
CAUSAL	Cause, Justify Evidence, Reason-N Reason, Result
CONTRAST	Antithesis, Concession Contrast
HYPOTHETICAL	Condition, Enablement Means, Motivation Otherwise, Purpose
JUDGEMENT	Evaluation-N, Evaluation-S, Interpretation, Solutionhood
TEMPORAL	Circumstance, Sequence
INFORMATION	Background E-Elaboration Elaboration, Preparation Restatement, Summary Question
ADDITIVE	Conjunction, Joint, List
STRUCTURAL	Attribution, Completion Sameunit

Table 4: Discourse relations annotated in our corpus, grouped by semantic category.

compared to the overall distribution of discourse relations in a document, though we exclude the group of STRUCTURAL relations from our analysis since we judge them as text-specific necessities rather than semantically motivated discourse relations.

The resulting distribution of discourse relations in the two modes is presented in Figure 1. The plot shows the density of each relation's proportions: For example, the third graph on the left side shows that in all of the spoken podcasts, the proportion of HYPOTHETICAL relations is low, with almost all of the documents having a proportion of this relation group between 0.0 and 0.13. In the blog posts on the other hand, we find proportions of HYPOTHETICAL relations up to 0.6 – though most blog posts also make use of rather low proportions of this relation group, as the peak in the density plot is around 0.0, too. Given these density distributions, we can conclude that some relation groups like CONTRAST, HYPOTHETICAL, and TEMPORAL are rather infrequent in our dataset in both modes. The relations in the groups ADDITIVE, CAUSAL, and INFORMATION are typically more frequent in blog posts, whereas JUDGMENT relations are almost exclusively found in the podcast transcripts. Overall, our spoken documents show a broader distribution of discourse relations and in our written documents, we find more relations from a single group in one document. However, this effect is not necessarily driven by the document's mode. Instead, the blog posts are generally shorter than the parallel podcast chunks, thus there are overall fewer discourse relations present.

4.2 Discourse Connectives

Discourse connectives (*because, however, while, ...*) are items that explicitly mark discourse relations. They can belong to various syntactic categories, such as conjunctions, adverbs, or prepositional phrases. In example (3), the CAUSE relation between the two segments is explicitly signaled by the connective *because*. In most cases, there is no 1:1 correspondence between connectives and discourse relations, and not every discourse relation is always signaled by a connective – corpus studies report on 15%–50% of all relations being explicitly signaled by connectives (Das, 2014; Crible, 2020). Nonetheless, connectives can be used for the identification of discourse relations, as there typically is a set of possible connectives associated with a discourse relation, and since connectives can be identified relatively easily in text. Which connectives can be used to signal which relations in German is described in the lexicon of discourse markers, DiMLex (Stede, 2002; Scheffler and Stede, 2016).

Based on the connectives listed in DiMLex, we annotated a total of 1,117 connective instances in our corpus, with *und* ('and', 394 instances), *aber* ('but', 130 instances) and *wenn* ('if', 79 instances) being the most frequent connectives. The average relative frequency of connectives per text is 2.73 (SD 2.11) in blog posts and 2.95 (SD 1.29) in podcasts – showing only small differences between the two media. Inter-annotator agreement between two annotators was moderate to substantial with Cohen's $\kappa = 0.73$.[9] We provide both annotations as well as the curated set of connectives in the corpus repository. For our classification task, we use the count of connectives in a given document relative to the number of tokens in this document. Given the short length of each of our documents, we do

[9] Some of the differences arise from the annotators varying in their assessment of multi-word connectives, e.g., whether *nicht nur ... sondern auch* ('not only ... but also') should be annotated as two multi-word connectives or four separate connectives.

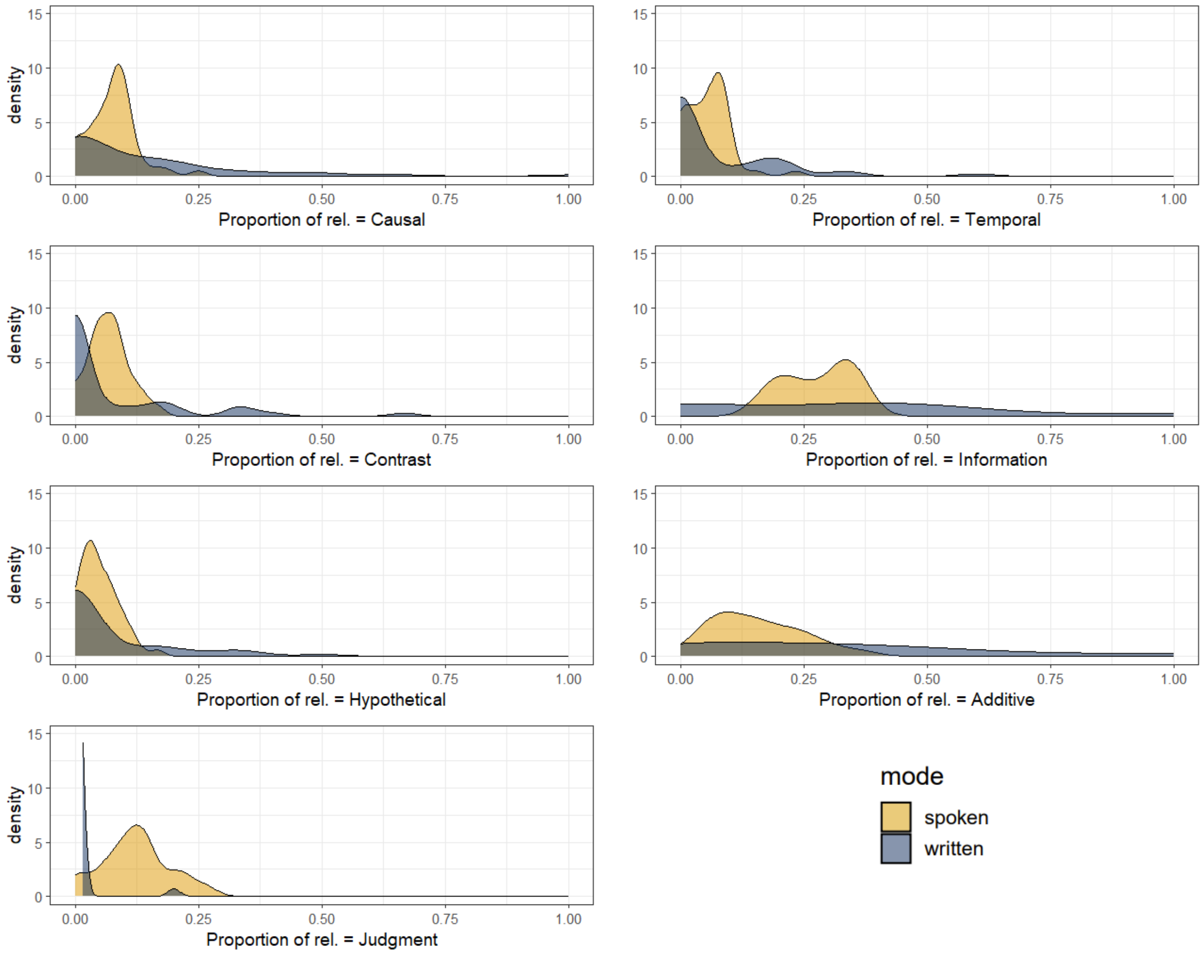

Figure 1: Density curves of discourse relation proportions, split by mode.

not distinguish between which discourse connectives are employed in a segment, but only record the overall frequency. (As this gives an indication of the text complexity, we regard this as a discourse-level feature rather than a solely lexical one.)

4.3 Discourse Particles

Discourse particles in a language like German are non-inflected sentence modifiers that convey discourse-related information (Zimmermann, 2011). They are employed to signal the epistemic state of the speaker and for common ground management between the interlocutors. Examples of two discourse particles are given in (4) and (5).

(4) Ich sollte nun wirklich gehen, es ist **ja** schon ziemlich spät.
‘I really should go now, it is (unarguably and obviously) quite late.’

(5) Meine Kollegin ist spät dran, sie hat **wohl** mal wieder den Bus verpasst.
‘My colleague is late, she (presumably) missed the bus again.’

In these examples, *ja* signals that the time being late is either known to all interlocutors or salient in the conversation’s context (possibly due to a setting sun or a clock nearby), and *wohl* indicates the speaker’s uncertainty about the expressed proposition. Discourse particles have been argued to be indicators of discourse relations, and corpus studies show that there is an interaction between discourse structure and discourse particles (Karagjosova, 2004; Döring, 2016), though the specific discourse function of these particles is not known. Because of their discourse-managing functions and often colloquial nature, the particles are used more often in informal spoken than in formal written language.

There are rarely two identical lists of elements considered to be discourse particles in German, but there is a ‘core class’ of discourse particles that is accepted by most researchers. A set of annotation guidelines for German particles compiling different lists from the literature (Kern and Scheffler,

2021) list 39 items that can be used as discourse particles, both elements from the core class and more rare discourse particles. Of those, we find 24 different items (518 instances in total) in our corpus: *aber, allerdings, auch, denn, doch, eben, eh, eigentlich, einfach, gleich, halt, irgendwie, ja, jetzt, leider, mal, schon, selbst, sogar, tatsächlich, vielleicht, wahrscheinlich, wirklich, wohl*, with *ja, eben* and *halt* being the three most frequent particles.[10] One annotator annotated all discourse particles in all parallel documents, and a second annotator annotated a subset of these documents. As the inter-annotator agreement is strong (Cohen's $\kappa = 0.81$), we consider one full annotation to be sufficient. The average relative frequency of discourse particles in blog posts is 0.30 (SD 0.83) and 1.46 (SD 0.84) in podcasts, showing a tendency for discourse particles to be more frequent in spoken podcasts than in written blog posts. We use the relative frequency of discourse particles as features for our classification task.

5 Classification by Mode

As our dataset is quite small, we use a 5-fold cross-validation to evaluate all our classification results. For both the binary classification and the evaluation, we use scikit-learn Support Vector Classification with C = 2[11] and cross-validation (Pedregosa et al., 2011).

Table 5 presents the classification results, the confusion matrix is shown in Table 6. Overall, the accuracy is 0.90 with a SD of 0.10.

Out of eight misclassification cases, five blog posts are wrongly classified as podcasts, and three as blog posts. Except for one case, all of the misclassified blog posts have a high relative frequency of discourse particles (between $\sim$ 2–4.41). The exception is the blog post CRE222. Here, the relative frequency of discourse particles is low (0 compared to 0.30 on average for blog posts), but the relative frequency of connectives stands out (6.21 compared to an average of 2.73). In two blog posts classified as podcasts, the relation distributions stand out compared to the other blog posts, In the case of blog post FG089, all relation proportions except for INFORMATION are low and in FG082_Blog, the proportion of JUDGEMENT relations is unusually high compared to the other blog posts. In the other three cases, the distributions of discourse relations match the other blog posts.

Fold	Precision	Recall	F1
1	0.89	1.00	0.94
2	0.88	0.88	0.88
3	0.78	1.00	0.88
4	0.88	1.00	0.93
5	1.00	0.96	0.92
average	0.88	0.95	0.90

Table 5: Results of classifying *mode* using groups of discourse relations as well as the relative frequency of connectives and discourse particles in a 5-fold cross-validation.

Fold	TP	TN	FP	FN
1	9	9	1	0
2	10	9	0	0
3	8	6	3	1
4	8	9	0	1
5	8	8	1	1

Table 6: Confusion matrix for classifying *mode* using groups of discourse relations as well as the relative frequency of connectives and discourse particles in a 5-fold cross-validation.

A similar – though inverted – pattern can be detected for the podcast transcripts wrongly classified as blog posts. All of them have a low proportion of JUDGEMENT and instead high proportions of other relation groups (mostly INFORMATION and ADDITIVE), and a low relative frequency of discourse particles or discourse connectives.

To put the reported numbers into perspective, we compare our results to a classification based on lexical features that have been shown to reliably distinguish spoken and written language: type-token-ratio and average word length. As before, we use a 5-fold cross-validation to evaluate all of our results. Table 7 reports on the results of the classification, the confusion matrix is presented in Table 8. The overall accuracy is 0.96 with a SD of 0.02.[12]

[10]It should be noted that some of these items, like *irgendwie* or *sogar*, are by far more frequent in their adverb or focus particle meaning, but can be employed as discourse particle, too. We only annotated the discourse particle use. In addition, some of these discourse particles are also included in DiMLex, the basis for our discourse connective annotation. If a token was ambiguous between the two categories, it was only annotated as discourse particle.

[11]We tested C-values between 0.5 and 5 and report on the value that yielded the best classification results.

[12]We would like to point out again that these results are not based on normalized lexical measures and might be skewed

Fold	Precision	Recall	F1
1	0.88	1.00	0.94
2	1.00	1.00	1.00
3	0.88	1.00	0.93
4	1.00	0.86	0.92
5	1.00	1.00	1.00
average	0.95	0.97	0.96

Table 7: Results of classifying *mode* using average word length and TTR in a 5-fold cross-validation.

Fold	TP	TN	FP	FN
1	9	9	1	0
2	7	9	0	3
3	9	9	0	0
4	9	9	0	0
5	8	9	0	1

Table 8: Confusion matrix for classifying *mode* using average word length and TTR in a 5-fold cross-validation.

Four out of five cases of misclassification are podcast documents with relatively high TTR compared to the other podcast files (FG046: 0.72, FG050: 0.80, FG060: 0.71, TelekomS7E1: 0.67, average for podcasts: 0.48). In the case of FG060 and TelekomS7E1, the average word length (5.30 and 4.90) is also higher than the mean in this group (4.51). There are other outliers in each of the categories, though if one of the measures is higher, the other one is close to the group's mean, thus not leading to incorrect classification. The one blog post wrongly classified as podcast shows a lower average word length (4.8 vs. group mean of 5.52) and TTR (0.69 vs. group mean of 0.83) compared to the other blog posts.

6 Discussion

Our results indicate that – even though the classification of language mode based on discourse-related features only does not reach the results achieved based on shallow lexical features – a rather reliable classification of mode based on structural discourse features is both possible and informative about the nature of the studied documents. The presented results are higher than a classification based on n-grams (Kunz et al., 2018), indicating that discourse-level features can provide additional information on a document compared to certain other lexical frequency profiles. Our results further indicate that as for the discourse-level features, both an unusually high or low frequency of discourse particles and connectives as well as an unusual distribution of discourse relation may lead to a misclassification of mode.

In our dataset, relations from the group of JUDGEMENT that evaluate content or express an opinion are found almost exclusively in the podcasts – it is only present in one out of all blog post. At the same time, the more argumentative relations from the group of CAUSAL relations are rarely found in podcasts. INFORMATION and ADDITIVE relations are found in both spoken podcasts and written blog posts, though more frequently so in the blog posts, whereas CONTRAST and HYPOTHETICAL relations are infrequent in both. Given the register of our data, this is unsurprising: The podcasts are conversations about a variety of topics, where interlocutors express their opinions or present information about the topic at hand. The blog posts on the other hand report on what is being talked about in the podcasts, informing the reader about content to be expected and possibly making an argument for why listening to the podcast is worthwhile. A register study on podcasts has shown that they are neither similar to other spoken registers nor other communication channels of CMC (Babyode et al., 2023). Analyses of blog posts find variation in the degree of formality and communicative purposes in this medium (Scheffler et al., 2022). This variation in register may also influence the representation of discourse coherence via RST-style discourse relations. Comparisons to other datasets will allow us to further delineate whether the results we find here are mainly driven by a difference in mode (spoken or written) or in register (e.g., conversation or presentation).

For the discourse particles, our results match the expectation that particles are more frequent in spoken language compared to written text. Given that there are only marginal differences between the frequency of connectives in blog posts and podcasts, no qualitative evaluation of the frequency of connectives in the two modes is possible. However, looking into the distribution of certain connectives in both blog posts and podcasts might yield different results.

We have shown that there are systematic dif-

to reflect the differences in document length as well as their difference in mode.

ferences in the distribution of discourse relations and discourse particles between our spoken podcasts and written blog posts. Therefore, our study contributes to the description of CMC practices as well as research on discourse structure, given that our dataset comprises documents coming from different registers than other corpora with discourse structure annotation. In addition, our dataset enables further studies on German discourse connectives as well as implicit and explicit discourse relations, as we provide both types of annotations.

Limitations

We have carried out all analyses according to our best abilities. Nevertheless, it should be noted that while the RST annotations were either done twice by different researchers or have been double-checked by at least one other expert for plausibility, in many cases there are alternative analyses of the texts which may also be applicable (as is usually the case for discourse structure). Since we do not have direct access to the discourse creators and their goals, this limitation is unavoidable in corpus studies.

In addition, annotation of discourse structure is quite costly and resource-demanding. It is therefore usual that datasets annotated for discourse structure are rather small, as is the case in our study. Further studies on different or combined datasets can help resolve this limitation.

Ethics Statement

The data reported on in this paper was collected within the research project named below. For the texts published, explicit consent was obtained from the creators to use and publish the data for annotation and scientific analysis. The data was automatically and manually preprocessed and reformatted, and manually annotated for discourse features. All annotations were carried out by researchers during their regular work time, either PhD students (authors of this paper) or student research assistants with regular work contracts and paid according to public pay scales. The data will be long-term archived and is available to other researchers according to the laws that apply. We see no other ethical issues with our data or research practices.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project ID 317633480 – SFB 1287.
We would like to thank Daniel Foppe, Freya Hewett, and Elisa Lübbers for their support with annotating and reviewing the data.

References

Berfin Aktaş and Manfred Stede. 2020. Variation in coreference strategies across genres and production media. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5774–5785, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Aminat Babyode, Laurens Bosman, Nicole Chan, Katharina Ehret, Ivan Fong, Noelle Harris, Alissa Hewton, Danica Reid, Maite Taboada, and Rebekah Wong. 2023. Structural linguistic characteristics of podcasts as an emerging register of computer-mediated communication. In *Proceedings of the 10th International Conference on CMC and Social Media Corpora for the Humanities*, pages 3–6, Mannheim, Germany.

Douglas Biber and Susan Conrad. 2019. *Register, Genre, and Style*, 2nd edition. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *The Longman Grammar of Spoken and Written English*. Longman, London.

Ludivine Crible. 2020. Weak and Strong Discourse Markers in Speech, Chat, and Writing: Do Signals Compensate for Ambiguity in Explicit Relations? *Discourse Processes*, 57(9):793–807.

Debopam Das. 2014. *Signalling of Coherence Relations in Discourse*. Dissertation, Simon Fraser University, Canada.

Sophia Döring. 2016. *Modal Particles, Discourse Structure and Common Ground Management*. Dissertation, Humboldt-Universität zu Berlin.

Freya Hewett. 2023. APA-RST: A text simplification corpus with RST annotations. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, pages 173–179, Toronto, Canada. Association for Computational Linguistics.

Theresa Heyd. 2021. Tertiary Orality? *Anglistik*, 32(2):131–147.

Elena Karagjosova. 2004. *The Meaning and Function of German Modal Particles*. Dissertation, Universiteit Utrecht.

Lesley-Ann Kern and Tatjana Scheffler. 2021. Annotationsrichtlinien für Modalpartikeln. Unpublished Manuscript.

Peter Koch and Wulf Oesterreicher. 1985. Sprache der Nähe - Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch*, 36:15–43.

Kerstin Kunz, Ekaterina Lapshinova-Koltunski, José Manuel Martínez Martínez, Katrin Menzel, and Erich Steiner. 2018. Shallow features as indicators of English-German contrasts in lexical cohesion. *Languages in Contrast: International Journal for Contrastive Linguistics*, 18(2):175–206.

Ekaterina Lapshinova-Koltunski and Marcos Zampieri. 2018. Linguistic features of genre and method variation in translation: a computational perspective. In Dominique Legallois, Thierry Charnois, and Meri Larjavaara, editors, *The Grammar of Genres and Styles. From Discrete to Non-Discrete Units*, pages 92–117. De Gruyter Mouton, Berlin, Boston.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Katrin Ortmann and Stefanie Dipper. 2019. Variation between different discourse types: Literate vs. oral. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 64–79, Ann Arbor, Michigan. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn discourse TreeBank 2.0. In *LREC 2008*.

Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. 2019. Penn Discourse Treebank Version 3.0.

Thomas Proisl and Peter Uhrig. 2016. SoMaJo: State-of-the-art tokenization for German web and social media texts. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 57–62, Berlin. Association for Computational Linguistics.

Ines Rehbein, Merel Scholman, and Vera Demberg. 2016. Annotating Discourse Relations in Spoken Language: A Comparison of the PDTB and CCR Frameworks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1039–1046, Portorož, Slovenia.

Tatjana Scheffler. 2017. Conversations on Twitter. In Darja Fišer and Michael Beißwenger, editors, *Investigating Computer-Mediated Communication: Corpus-based Approaches to Language in the Digital World*, pages 124–144. University Press, Ljubljana.

Tatjana Scheffler, Lesley-Ann Kern, and Hannah Seemann. 2022. The medium is not the message: Individual level register variation in blogs vs. tweets. *Register Studies*, 4(2):171–201.

Tatjana Scheffler and Manfred Stede. 2016. Adding semantic relations to a large-coverage connective lexicon of German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1008–1013, Portorož, Slovenia. European Language Resources Association (ELRA).

Sara Shahmohammadi, Hannah Seemann, Manfred Stede, and Tatjana Scheffler. 2023. Encoding Discourse Structure: Comparison of RST and QUD. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, pages 89–98.

Oren Soffer. 2010. "Silent Orality": Toward a Conceptualization of the Digital Oral Features in CMC and SMS Texts. *Communication Theory*, 20(4):387–404.

Manfred Stede. 2002. DiMLex: A Lexical Approach to Discourse Markers. In Alessandro Lenci and Vittorio Di Tomaso, editors, *Exploring the Lexicon - Theory and Computation*. Edizioni dell'Orso, Alessandria.

Manfred Stede. 2016a. Das Potsdamer Kommentarkorpus. In Hartmut Lenk, editor, *Persuasionsstile in Europa II. Kommentartexte in den Medienlandschaften europäischer Länder*, number 229-231 in Germanistische Linguistik, pages 177–202. Olms, Hildesheim.

Manfred Stede, editor. 2016b. *Handbuch Textannotation: Potsdamer Kommentarkorpus 2.0*. Number 8 in Potsdam cognitive science series. Universitätsverlag Potsdam, Potsdam.

Amanda Stent. 2000. Rhetorical structure in dialog. In *INLG'2000 proceedings of the first international conference on natural language generation*, pages 247–252.

Sara Tonelli, Giuseppe Riccardi, Rashmi Prasad, and Aravind Joshi. 2010. Annotation of discourse relations for conversational spoken dialogs. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Languages Resources Association (ELRA).

Shujun Wan, Tino Kutschbach, Anke Lüdeling, and Manfred Stede. 2019. RST-tace a tool for automatic comparison and evaluation of RST trees. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 88–96, Minneapolis, MN. Association for Computational Linguistics.

Malte Zimmermann. 2011. Discourse Particles. In P. Portner, C. Maienborn, and K. von Heusinger, editors, *Semantics*, number 33 in Handbücher zur Sprach- und Kommunikationswissenschaft HSK 2, pages 2011–2038. Mouton de Gruyter, Berlin.