DHQ: Digital Humanities Quarterly
2023
Volume 17 Number 3

Category Development at the Interface of Interpretive Pragmalinguistic Annotation and Machine Learning: Annotation, detection and classification of linguistic routines of discourse referencing in political debates

Michael Bender <mbender_at_linglit_dot_tu-darmstadt_dot_de>, Technical University of Darmstadt https://orcid.org/0000-0001-6498-7236

Maria Becker <maria_dot_becker_at_gs_dot_uni-heidelberg_dot_de>, University of Heidelberg https://orcid.org/0000-0003-2596-9152

Carina Kiemes <carina_dot_kiemes_at_tu-darmstadt_dot_de>, Technical University of Darmstadt

Marcus Müller <marcus_dot_mueller_at_tu-darmstadt_dot_de>, Technical University of Darmstadt https://orcid.org/0000-0003-4921-4512

Abstract

In this paper, we present a case study on quality criteria for the robustness of categories in pragmalinguistic tagset development. We model a number of classification tasks for linguistic routines of discourse referencing in the plenary minutes of the German Bundestag. In the process, we focus and reflect on three fundamental quality criteria: 1. segmentation, i.e. size of the annotated segments (e.g. words, phrases or sentences), 2. granularity, i.e. degrees of content differentiation and 3. interpretation depth, i.e. the degree of inclusion of linguistic knowledge, co-textual knowledge and extra-linguistic, context-sensitive knowledge. With the machine learnability of categories in mind, our focus is on principles and conditions of category development in collaborative annotation. Our experiments and tests on pilot corpora aim to investigate to which extent statistical measures indicate whether interpretative classifications are machine-reproducible and reliable. To this end, we compare gold-standard datasets annotated with different segment sizes (phrases, sentences) and categories with different granularity, respectively. We conduct experiments with different machine learning frameworks to automatically predict labels from our tagset. We apply BERT ([Devlin et al. 2019]), a pre-trained neural transformer language model which we finetune and constrain for our labelling and classification tasks, and compare it against Naive Bayes as a probabilistic knowledge-agnostic baseline model. The results from these experiments contribute to the development and reflection of our category systems.

1. Introduction

1 This study investigates discourse referencing practices in parliamentary debates from a linguistic perspective. *Discourse referencing* is present in sentences in which a speaker makes references to preceding utterances within the discourse. We therefore study intertextual references to oral utterances and written texts. The visualization and automated recognition specifically of such practices opens up relevant new perspectives of insight. Firstly, they serve as a starting point to uncover and analyze intertextual reference structures in more detail in subsequent applications. These could be analyses according to subject areas or discourses (the parliamentary-procedural, the economic, the academic), or according to types of reference objects (written text types, oral utterances), or to the relation of references to party affiliation, for example. Secondly, the study of communicative practices in parliaments is fundamentally relevant for understanding the mechanisms of Western parliamentary democracies. The analytical annotation and automated recognition of different types of such practices are important prerequisites for the further investigation of their mutual interaction in different contexts. And thirdly, categorizing practices of discourse referencing is methodologically interesting for digital pragmalinguistics, because it addresses a fundamental challenge in the field: On the one hand, pragmalinguistic phenomena can be indicated on the linguistic surface and thus be recognized, also in an automated way, but on the other hand, the capture of implicit and inferred aspects as well as the inclusion of contextual knowledge are of central importance. This makes interpretative analysis indispensable and requires the training of algorithms through manual annotation. [Archer et al. 2008, 615] have

pointed out this particular difficulty for annotation studies with an automation perspective.

In linguistic heuristics, discourse referencing belongs to pragmatics because it involves linguistic practices whose function can only be inferred based on contextual knowledge. This may not seem to be evident at first glance. Some forms of discourse referencing are easily detectable on the linguistic surface. Consider, for example, explicitly marked quotations or communication verbs (such as *say* or *promise*). However, discourse referencing can also be indicated implicitly. Formulations, such as *"With your behavior […] you have placed yourself in an improper proximity […]"* (1) in the following example, require interpretation based on contextual knowledge to be identified as practices of discourse referencing. Here, an interpretative effort would lead to understanding "behavior" as a linguistic action rather than, say, physically violent behavior: 2

> **1.** *With your behavior [...] you have placed yourself in improper proximity to your neighbors here further to the right.* (all examples are translated by the authors)
> [Sie haben sich mit Ihrem Verhalten […] in eine ungute Nähe zu Ihren Nachbarinnen und Nachbarn hier weiter rechts begeben.]

Such contextual and interpretative phenomena cannot be simply captured by corpus linguistic or algorithmic access to the linguistic surface, which makes them difficult to analyze in an automated way: While linguistic surface patterns (e.g., word order, collocations, word frequencies or the distributions of words or larger linguistic constructions) can be detected easily, their exact meaning and pragmatic function may not be fully captured on this level by machines due to the missing context knowledge. One approach to solving this problem is to combine interpretive-categorizing annotation and machine learning. The application of this methodological approach to the subject of discourse referencing has so far been a research desideratum, much more so with a focus on category development with automatability in mind. 3

While discourse referencing as our linguistic research object is important in its own right for understanding the mechanism of parliamentary discourse, here we focus on the methodological aspect of category development concerning the automated detection of such references in large datasets. For this purpose, we conduct a collaborative annotation study and run experiments with probabilistic classifiers such as Naive Bayes [Jurafsky and Martin 2022] and transformer language models such as BERT [Devlin et al. 2019]. As part of this study, we methodologically describe and discuss the development of an annotation category system on the object of discourse referencing with automation possibilities. 4

We obtain the dataset for our case study from the linguistically preprocessed corpus of the plenary minutes of the German Bundestag ([Müller and Stegmeier 2021], cf. [Müller 2022b]). 5

The category system combines deductive and inductive categorizations. In a first step, we form categories for discourse referencing that stem from linguistic theories. In a second step, we have to adapt these or create new categories for forms and cases that we only recognize in the course of data exploration – especially for cases of implicit discourse referencing, such as *"you have placed yourself […],"* and others. The central challenge with this approach is to capture the phenomena under investigation as precisely as possible and at the same time to maintain a certain balance of granularity and variance in the category contents. 6

In the following, we first provide an overview of preliminary work on category design in pragmalinguistics and linguistic discourse research. We focus on already-known success factors in the formal and contextual tailoring of categories. Next, we introduce the pragmatic phenomenon of discourse referencing and describe the properties that are relevant to our heuristic model building. Subsequently, we describe and discuss our dataset and the collaborative annotation of discourse referencing practices in terms of assumptions, process and results. The annotation process consists of two phases: 1. We test the aspect of categorization granularity by modelling a binary classification task (discourse referencing present or not). 2. We tag our data in a more fine-grained way, focusing on the actors (authors/speakers) of referenced utterances (actors mentioned or not), and additionally extracting phrases that have been identified to indicate discourse referencing. In addition to this, we run linguistic experiments using probabilistic and neural classifiers to detect discourse referencing. In this set of experiments, we test the influence of different input data in terms of taxonomies (number of categories) and segment sizes (phrase input vs. sentence input). By doing so, we also investigate the interplay between form and meaning. We analyze its impact on both algorithmic models and collaborative manual annotation: Does annotating smaller segment sizes, which are more specific to the phenomenon under investigation, or entire sentences containing the phenomenon, align better with the content-conceptual granularity of the category in question? Finally, we discuss our results on the question of category design 7 and conclude with a summary.

2. Capturing discourse referencing by annotation

2.1 Criteria for the development of machine-learnable categories in a pragmalinguistic annotation approach

Numerous issues, aspects and criteria for the development of category systems have been discussed in the literature on pragmalinguistic annotation. [Archer et al. 2008, 615] differentiate five levels of pragmatic information relevant to category development: the formal, the illocutionary, the implied/inferred, the interactional and the contextual level. The consistent consideration of these level differences is seen as an important criterion for the design of annotation schemes. In particular, [Archer et al. 2008, 633] highlight segmentation: “Segmentation requires us not only to state what unit we will be analysing, but also to define it in a way that will enable us to measure one unit against another, and, by so doing, ensure a level of consistency.” Segmentation thus refers to the size of annotated units on the linguistic surface (e.g., phrases or sentences) chosen according to the conception of the category system. This aspect has been described as an important quality criterion in other works in the field as well, e.g., in the annotation of speech acts (c.f. [Leech and Weisser 2003]). Teufel also addresses the segmentation problem – from a more computational linguistic point of view – she reflects on the difficulty of assigning abstract categories to linguistic units. She also addresses the problem that categories can overlap but is critical of multiple annotations with regard to evaluability (cf. [Teufel 1999, 108]). Instead, she opts for selective annotation with exclusive categories and consistent segmentation ([Teufel 1999, 111]; cf. [Weisser 2018, 213–277]). 8

These aspects – consistent segmentation and a distinctive category system – have likewise proven crucial in our previous studies on pragmalinguistic annotation, also concerning the combination of pragmatic annotation and machine learning. In addition to these two aspects, we have worked out the factors of granularity of categories and context sensitivity/depth of interpretation in prior studies ([Becker et al. 2020]; [Bender 2023]). To give an example of different category granularities in a system: In [Becker et al. 2020], we treated discourse referencing as a subcategory of relevance marking and again distinguished more fine-grained between directed and undirected discourse referencing, thus had three levels of granularity in one category. We developed a complex annotation scheme with pragmalinguistic categories at different levels of granularity to study academic text routines (e.g., relevance marking, definition, argumentation). We used this scheme to manually annotate sentences in a corpus of texts from different academic disciplines and then to train a recurrent neural network for classifying text routines. The experiments showed that the annotation categories are robust enough to be recognized by the model, which learns similarities between sentence surfaces represented as vectors. Nevertheless, the accuracy of the model depended strongly on the granularity of the category level [Becker et al. 2020, 450–455]. 9

In general, pragmalinguistic questions raise the challenge of operationalizing and segmenting phenomena that are context-dependent rather than bound to a formal segment. In a great number of cases, discourse referencing acts can be delimited to certain phrases. However, there are cases – e.g., certain anaphoric references – where the indicators of discourse referencing can only be fully captured in the extended cotext, i.e., the surrounding sentences/utterances at a definable distance from the focused utterance – as opposed to context as extra-linguistic, e.g., social and situational conditions and knowledge backgrounds. Thus, in addition to the aspect of segmentation consistency, the granularity of segmentation and the size of the cotext window are also important. 10

Both granularity and distinctiveness are relevant factors for the segmentation and also for the robustness of the category system as a whole. Granularity determines the semantic and pragmatic content of the categories in annotation schemes. The granularity of the tagset influences the accuracy of the algorithm (cf. [Becker et al. 2020, 455]). This does not mean that schemes with few categories or tags are always better. Rather, it is important to capture a certain phenomenon as well as possible through the operationalization in the scheme and to make it analyzable at first. Secondly, insufficiently differentiated tagsets lead to overly heterogeneous categories, which in turn limits machine learnability. 11

The annotation guidelines need to consider this. For instance, they need to specify exactly how much communicative and contextual knowledge may be included and how deeply it is to be interpreted to determine whether an utterance is a reference to a communicative act – even in cases where this is not made explicit through according lexis (see example in the introduction). 12

To achieve agreement in the annotation process, the team of annotators must reach explicit common ground on the depth of interpretation when assigning segments to categories. The more cotext/context is available to annotators, the more they will interpretatively work out what was "actually" meant by a sentence, and the higher the risk that annotators will disagree. Therefore, it may be useful to deliberately limit the co-textual information and thus limit the depth of interpretation. Categories designed to be distinctive (allowing no overlap of categories) and exhaustive (covering the whole variety of phenomena in the data) have proven to optimize machine learning [Becker et al. 2020, 430]. This robustness can be evaluated by calculating the inter-annotator agreement [Artstein and Poesio 2008, 555–596]. The above-mentioned factors also represent quality criteria for the explicitness and intersubjective comprehensibility of interpretative categorizations in annotation studies, i.e., they determine whether categorizations are compatible with machine learning, for one, and comprehensible for human addressees, such as other annotators or recipients of the respective study, for another. Besides this, the accuracy values of the different algorithmic models we will test represent verification results. 13

In summary, our category development considers the factors of segmentation, granularity, distinctiveness and context sensitivity/depth of interpretation on different levels as well as in their mutual interaction with the machine learnability of the category system in experiments. In this study, we draw on these findings and test the effects of changes in these factors as well as their impact in various experiments (on the Inter-annotator agreement and the learning success of different algorithmic models). Furthermore, we test whether the trained algorithmic models cope better with sentence segmentation or with phrase-level segmentation. 14

2.2 Linguistic routines of discourse referencing

By discourse referencing, we mean referring to a preceding communicative act within discourse ([Müller 2007, 261]; [Feilke 2012, 19]). We are thus dealing with particular cases of intertextuality [Allen 2000]. These are characterized by concrete and explicit references to other communicates, which can be called "texts" in a broad sense. They include not only pre-texts such as laws, templates, drafts, and policy papers but also oral utterances. In all cases, the referenced act is in the past from the speaker's point of view. The reference can be uttered as a complete proposition, as a verbal phrase (VP), or as a noun phrase (NP) (see examples in section 3.2), with the subject of the utterance fully named, metonymically named, or without naming the subject of the utterance. Discourse referents in this sense are constitutive of many genres, e.g., academic or legal discourse. 15

Communicative practices in parliaments are fundamentally relevant for understanding the mechanisms of Western parliamentary democracies. But discourse references in parliamentary discourse also have functions that are interesting in terms of linguistic systematics: First, they serve to orient and co-orient political statements in different discourses (citation 2; e.g., the parliamentary-procedural, the economic, the academic); second, they are used to index institutional and situational coalitions or oppositions (3); and third, they are used to invoke the legal basis of parliamentary action (4; laws, directives, regulations). In the sense of this last point, discourse references serve to recall the distinguished function of the parliamentary arena as a laboratory in which the legal framework of our social life is forged. 16

> **2.** *Those who say this are subject to an essential misjudgment because they do not know or misjudge what great added value it means in terms of acceptance and industrial peace when important decisions are discussed beforehand in the works council and then implemented together in the company.*
> [Die, die das äußern, unterliegen einer wesentlichen Fehleinschätzung; denn sie wissen nicht oder schätzen falsch ein, welchen großen Mehrwert es im Hinblick auf Akzeptanz und Betriebsfrieden bedeutet, wenn wichtige Entscheidungen zuvor im Betriebsrat besprochen und dann gemeinsam im Betrieb umgesetzt werden.]

> **3.** *The suitable and also still possible minimal invasive solution in the remaining weeks is an opening of the contribution guarantee, which leads also according to the opinion of science to more net yield and more security.*
> [Die passende und auch noch mögliche minimalinvasive Lösung in den verbleibenden Wochen ist eine Öffnung der Beitragsgarantie, die auch nach Meinung der Wissenschaft zu mehr Rendite und mehr Sicherheit führt.]

> **4.** *Please read the act first, before you argue in a populist way here.*

> [Lesen Sie doch bitte erst das Gesetz, bevor Sie hier populistisch argumentieren.]

One can see from these first examples that the focus and concreteness of the intertextual reference varies considerably. (2) contains a reference to a concrete and theoretically precisely determinable group of speakers antecedent in the discourse, but introduced into the discourse only unspecifically (*those who say this*). In (3), there is a similarly unspecific reference that is introduced with a metonymic shift (*according to the opinion of science* instead of *“according to the opinion of some academic scholars who are concerned with this issue”*). In (4), a legal statute is referred to as the manifest result of a communicative act, without addressing the actors involved in the writing of the statute at all. Such a reference to texts as instances independent of the author, as it were autonomously effective, is a common rhetorical procedure in parliamentary debates. 17

In other cases, of course, utterances refer to concrete empirical persons. These can be groups (see example 5), or individuals (6). Besides this, there are (albeit rare) cases in which reference is made to a preceding text in the discourse, such that the text itself takes the place of the actor in a communicative action (7). These metonymic shifts are interesting because they give a different hue to the action structure of the discourse that is being produced using discourse referencing: the cognitive focus, the claim of validity, and also the authority are shifted from the author to the text in such cases. Methodologically, what is interesting here is the extent to which such metonymic constructions can be found automatically, especially since they are rare. 18

> **5.** *After all, the concern of the democratic opposition groups is a correct one.*
> [Denn das Anliegen der demokratischen Oppositionsfraktionen ist ja ein richtiges.]
>
> **6.** *Ladies and gentlemen, Kohl, a historian by training, once said: “Those who do not know the past cannot understand the present and cannot shape the future.”*
> [Meine Damen und Herren, der gelernte Historiker Kohl hat einmal gesagt: “Wer die Vergangenheit nicht kennt, kann die Gegenwart nicht verstehen und die Zukunft nicht gestalten.”]
>
> **7.** *The report confirms: Inner cities are losing their individuality and thus their attractiveness.*
> [Der Bericht bestätigt: Die Innenstädte verlieren ihre Individualität und damit Attraktivität.]

We exemplify our methodological considerations and experiments on category design with the following research questions: 1. Which types of discourse referents occur in our data set and in which distribution? 2. What role do actors play in discourse referencing? That is, when are the speakers and writers of utterances explicitly named, and when, instead, in a metonymic thrust, does the text itself move into the position of the actor (as in evidence 7)? 19

3. Dataset and annotation workflow

3.1 Dataset

To investigate discourse referencing in parliamentary discourse, we draw on the plenary minutes of the German Bundestag [Müller and Stegmeier 2021]. Discourse Lab [Müller 2022a] hosts a linguistically processed and metadata-enriched corpus of the plenary minutes that currently covers the period from 1949 to May 2021, i.e., all completed election periods from 1 to 19. The corpus contains about 810,000 texts (debate contributions) and about 260 million tokens. It is expanded at regular intervals with current data [Müller 2022b] which is provided by the German Bundestag (https://www.bundestag.de/services/opendata). Pre-processing includes tokenization, sentence segmentation, lemmatization, part-of-speech tagging, marking of speakers’ party affiliation, and separate marking of heckling. This way, speeches with and without heckling or even heckling separately can be searched. The basic unit (`<text>`) of the corpus is the parliamentary speech. It is subclassified by speakers’ texts `<sp>` and heckling `<z>`. Text attributes are speaker’s fraction, year, month, speaker, session, legislative period, text ID and day of the week. The corpus is managed via the IMS Corpus Workbench [Evert and Hardie 2011]. For our categorization experiment, we draw a random sample of 6,000 sentences from the May 5–7, 2021 plenary transcripts. We exclude hecklings in the process. The sample is homogeneous across time and actors: Since our study is about methodological experiments on category formation, the variation of parameters should be controlled. With the sample design, we exclude diachronic variation and variation caused by changing groups of actors. We include various types of discourse referencing in that our dataset covers functional, thematic, and interpersonal variation. 20

3.2 Collaborative annotation

The first part of our experimental annotation study on discourse referencing focuses on collaborative manual annotation. We consider collaborative annotation to mean not only that several annotators assign categories, but also that categories and guidelines are developed in a team (cf. [Bender and Müller 2020]; [Bender 2020]). The understanding of discourse referencing described in Chapter 3 requires linguistic expertise – at least in less explicit cases. Thus, we cannot simply assume everyday linguistic intuition to be sufficient but must develop criteria and guidelines and make them available to annotators, or at least train them to some extent in the application of the guidelines. Of course, it is best to involve all annotators in the development of the categories as well, if possible. We have been able to do this, at least in part, in the study described here. For this purpose, we discussed the theoretical concept of categorization in the team and, on this basis, first established criteria for assigning categories to segments. 21

The basic unit of annotation was set to be sentences. The reason for this is that linguistic actions are typically represented in sentences. Co-textual information was intentionally narrowed down in this study by extracting individual sentences and making them available to annotators in random order. Within this cotext window, not all discourse referencing can be fully resolved even in terms of unambiguous attribution to prior utterances, but the indicators of discourse referencing can be detected at the individual sentence level by context knowledge/language knowledge (without further cotext). In this respect, the unit sentence, which can also be delimited and quantified for algorithmic models, was given preference here over, for example, freely selectable text sections as larger cotext windows. No overlap of categories was allowed in the annotation. The next smaller unit in the linguistic system is phrases, which were used in this case for the extraction of classification-relevant indicators. Evident indicators of discourse referencing are phrases with communication verbs and noun phrases that introduce sources of referenced utterances (i.e., authors, speakers). Other – context-sensitive – indicators were identified in the collaborative data analysis in the course of pilot annotations. For example, discourse references in parliamentary discourse are also made with action verbs in conjunction with nominal mentions of texts or utterances (e.g., “with the draft we initiated the debate”). 22

After determining relevant categories deductively, trial annotations were carried out. The category system was revised inductively in a data-driven manner and team members discussed cases of doubt. The abductive differentiation or reconfiguration of the scheme is necessary when the assignment of text segments ([Pierce 1903] calls it “percept”) to categories (“percipuum,” [Pierce 1903]) by qualitative induction fails in the course of annotation. In our annotation process, however, we understand this new construction or configuration not as a result of purely individual insights, but as a collaborative-discursive process of negotiating categories that are plausible for all annotators. 23

An additional goal that made this collaborative discursive negotiation process even more complex was to combine a linguistic analysis perspective with computational linguistic expertise to better anticipate what different machine learning algorithms can capture. For example, we decided against annotating verbatim quotations and indirect speech because we wanted to train the algorithmic models primarily on indicators which show that referencing is taking place, instead of focusing on what is being referenced. After all, the formation of linguistic routines occurs at the level of referencing, while what is referenced can vary indefinitely. Since we aim to discuss the question of category design at the intersection of disciplinary heuristics and machine learning, we developed an annotation workflow that allows us to conduct machine learning experiments on categories of varying complexity in terms of form and content. 24

We decided on different levels of annotation complexity for which we developed the appropriate categories: 25

Annotation step	Complexity level	Category	Segment	Classification decision	Possible numbers of segments per instance
1	1	discourse referencing	sentence	yes/no	1
2a	2	mention of the source (author/speaker) of the referenced utterance	sentence	explicit/metonymic/none	1
2b	3	discourse referencing	phrase	yes/no	n

Table 1. Manual annotation – workflow.

Table 1 presents the different annotation steps, which are designed according to increasing complexity: Step 1 is a binary classification task with two labels. In step 2, we ran two annotation tasks at the same time. First, different types of thematization of authors/speakers of the textual and oral utterances were classified – at the sentence level: explicit/metonymic/none. Second, within the sentences that were already classified as discourse referencing, those phrases that were relevant to the classification decision were identified (see Table 2). This step requires accurate annotation of phrases representing relevant actors, actions and products. Even though step 2b is a binary classification task, the decisions required for classification are even more complex because any number of segments can be annotated for each instance and the three-item classification from step 2a is presupposed. 26

The first annotation phase consisted of a binary classification task that required distinguishing between sentences with and without discourse referencing. According to this criterion, all 6,000 sentences of the corpus sample were double annotated (sentences as segments). Teams of two performed the annotation of 3,000 sentences each in Excel spreadsheets independently. The sentences were arranged in random order to avoid possible cotext/context effects. After double annotation, the inter-annotator agreement was calculated based on Cohen's kappa [Cohen 1960]. Agreement scores varied among groups in the first run. In group 1, 2,566 of 2,919 sentences were annotated in agreement (88%, Cohen's kappa: 72.87), in group 2, 2,408 of 2,883 (83.5%, Cohen's kappa: 57.44). The difference in kappa score between the groups is linked to the fact that in group 2 the rarer label ("+ discourse referencing") was assigned less frequently in agreement (in 487 cases), due to a misunderstanding that became apparent late in the annotation process. This had a disproportionately large impact on the calculation of agreement statistics using Cohen's kappa. That is because more infrequent labels are calculated to have a lower probability of overruling annotations by random chance than high-frequency ones. Cohen's kappa is designed to compute the randomly corrected matches of annotations from different annotators. This way, it expresses a ratio between the randomly expected agreement and the observed agreement, assuming that annotators can also assign the same label to an instance by random chance with a certain probability (cf. [Greve and Wentura 1997, 111]; [Zinsmeister et al. 2008, 765f]). 27

The average agreement score was nevertheless acceptable (Cohen's kappa: 65.02). Kappa scores are evaluated differently in the literature. [Greve and Wentura 1997] categorize kappa scores above 75 as excellent, and scores between 61 and 75 as good. In more recent NLP work, even lower values are accepted as good (e.g., [Ravenscroft et al. 2016]; cf. [Becker et al. 2020, 442]). Based on this assessment of the kappa value and the high degree of any other agreement between the annotations, the results of phase one were accepted as the basis for the second phase. That is, all cases in which different categories were assigned were filtered out. These cases were then decided by an independent annotator according to the criteria of the guidelines. 1,935 of 6,000 sentences (32.25%) were identified as discourse referencing, which indicates the importance of such practices in parliamentary discourse. 28

29 In the second annotation phase, these 1,935 sentences were annotated according to a more fine-grained scheme: The

classification task was to distinguish discourse references in which the actor (author/speaker) of the referenced utterance is explicitly named from those in which the text becomes the actor in a metonymic shift and those in which no actors are named (see Table 2).

Tag	Description	Example
1	Actor explicitly mentioned.	*Twelve years ago, the Chancellor, together with the prime ministers of the time, proclaimed the "7 per cent" goal.* [Die Kanzlerin hat gemeinsam mit den damaligen Ministerpräsidentinnen und Ministerpräsidenten vor zwölf Jahren das Ziel "7 Prozent" ausgerufen.]
2	Metonymic mention of the actor.	*Our Basic Constitutional Law obligates us to create equal living conditions in Germany.* [Unser Grundgesetz verpflichtet uns zur Schaffung gleichwertiger Lebensbedingungen in Deutschland.]
3	No actor mentioned.	*The recommended resolution is adopted.* [Die Beschlussempfehlung ist angenommen.]

Table 2. Tagset of the second annotation round.

As a result, we measured a very good agreement (Cohen's kappa: 84.35). After curating the annotations and producing the gold standard, 721 sentences (37.26%) were assigned to category 3, 1,155 (59.69%) to category 1, and 59 (3.05%) to category 2. 30

In the same step, we extracted the phrases that had been rated by the annotators as crucial for the categorization as discourse referencing. These included, for example, noun phrases (NP) representing communicative acts or texts or discourse actors (without heads of embedding phrases such as prepositions in a prepositional phrase) or relevant verb phrases (VP) (including verbs that express communicative action, as shown in the examples) without complements and adverbials. Table 3 gives an example of phrase extraction. 31

Categorized sentence	Extracted phrases critical to categorization	Phrase type	Referenced
For me, there are three good reasons to reject this proposal of the AfD today: The first is the sheer thin scope already mentioned by colleague Movassat; I do not need to say much more about it. [Für mich gibt es drei gute Gründe, diesen Antrag der AfD heute abzulehnen: Der erste ist der schon vom Kollegen Movassat erwähnte schiere dünne Umfang; dazu brauche ich nicht mehr viel zu sagen.]	*this proposal of the AfD* [diesen Antrag der AfD]	NP	text
	colleague Movassat [Kollege Movassat]	NP	actor
	mentioned [erwähnte]	VP	utterance

Table 3. Manual phrase extraction from sentences categorized as "discourse referencing."

The phrases *"to reject"* [abzulehnen] and *"I do not need to say"* [brauche ich nicht … zu sagen] were not extracted because they represent possible future utterance acts, not preceding ones. 32

This extraction was intended to work out what annotators are looking at when they detect discourse referencing. In machine learning, an "attention mechanism" is used to try to mimic human cognitive attention. The extraction of relevant phrases will be used to test whether this principle can be supported in this way. Which effects can be observed will be reflected in the next chapter. 33

4. Automatic classification/machine learning

In this section, we describe how we build and apply different machine learning algorithms to detect and classify discourse references in political debates. The goal of this research is to assess the ability of computational models such as traditional classification algorithms as well as Deep Learning techniques to detect discourse references in texts and to classify them. 34

4.1 Task Description

As mentioned before, we developed our category scheme with regard to the machine learnability of the different labels and paid particular attention to the factors segmentation, granularity, distinctiveness and context sensitivity. In line with the two phases of annotations, as described above, we designed two tasks for probing the ability of computational models to learn our category system: 35

> **Task 1: Detecting discourse references.** In the first annotation phase, our annotators had to distinguish sentences with discourse referencing from sentences without discourse referencing. For computational modelling, this can be framed as a binary classification task; the task of detecting discourse references in texts on the sentence level: Given a sentence, the task is to predict if this sentence contains a discourse reference or not. We use each of the given 6,000 sentences as input, and let the model predict for each of them one of the two labels *discourse referencing* (1) and *no discourse referencing* (0).
>
> **Task 2: Classifying types of discourse references.** The second task is to classify the discourse references into three categories: *Actor explicitly mentioned, Metonymic mention of the actor* and *No actor mentioned* (see Table 4). We use all instances that have been annotated as discourse references in the gold version of our annotations for training and testing our models (n=1,935). We experiment with three different input formats: providing the model (a) with the full sentence as input, (b) only with the phrase marked as relevant for discourse referencing as input, and (c) with both, the full sentence and the marked phrase, by concatenating the sentence and the phrase, separated by a separator token.

We then train and evaluate the models in **three settings**. In the first setting **A**, all three categories are taken into account. In the second setting B, the least frequently assigned category metonymy is excluded. The idea behind that is that most machine learning approaches suffer from imbalanced datasets and in particular from minor classes which are represented by too few examples. With setting **B**, we therefore can test how much the small size of our minor class metonymy affects our results. In the third setting **C**, we finally combine categories 1 and 2, which are both actor-naming categories, and contrast them with category 3, in which no actors are mentioned. In this way, we can reveal if our models can distinguish between instances that focus on the actors, and instances that leave the actors implicit. 36

4.2 Description of Models

To investigate to which extent our category system as described above can be learned by machine learning techniques, we test the ability of two different supervised machine learning approaches: (I) Naive Bayes, a traditional classification algorithm, serves as our baseline model and is compared to (II) BERT, a State-of-the-Art Transformer Language Model that has shown great success in various NLP tasks. Both models are applied to detect (Task 1) and classify (Task 2) discourse references in texts. 37

Baseline Model – Naive Bayes. Naive Bayes is a probabilistic classifier that makes assumptions about the interaction of features [Jurafsky and Martin 2022, 59]. The text is treated “as if it were a bag-of-words, that is, an unordered set of words with their position ignored, keeping only their frequency in the document.” [Jurafsky and Martin 2022] This means that first, the occurrence of words in a category is counted (“bag-of-words”). Then, for each word, the probability that it occurs in each category can be calculated. For each new observation, a probability value is calculated based on each category. That means, it is assumed at first, that the sentence belongs to category 1. The overall probability of category 1 to be classified is then added to the probabilities of each word to occur in category 1. In the next step, the same calculation is performed, assuming the new observation belongs to category 2. After calculating these values for each category, the values are compared with each other. The category with the highest value is the prediction of the classifier. 38

For our approach, we use the Multinomial Naive Bayes model as implemented in the Python package scikit-learn [Pedregosa et al. 2011]. We use 90% of the data for training and keep 10% for testing. 39

Transformer Language Model – BERT. The application of pre-trained language models, such as BERT [Devlin et al. 2019], GPT [Radford et al. 2019] or XLNet [Yang et al. 2020], has recently shown great success and led to improvements for various downstream NLP tasks. Through pre-training on large textual corpora, these models store vast amounts of latent linguistic knowledge ([Peters et al. 2018]; [Orbach and Goldberg 2020]). After pre-training, the models can be fine-tuned on specific tasks with a small labelled dataset and a minimal set of new parameters to learn. 40

Language models have been successfully applied to various language classification tasks, such as emotion classification 41

[Schmidt et al. 2021], sentiment analysis [Yin and Chang 2020], and relation classifications [Becker et al. 2021]. Inspired by these insights, we make use of the latent knowledge embodied in large-scale pre-trained language models and explore how we can finetune them for our two classification tasks – the detection of sentences with discourse referencing and the classification of different types of discourse references.

Initial experiments with different models had shown that the transformer language model BERT [Devlin et al. 2019], which is pre-trained on the Google Books Corpus and Wikipedia (in the sum of 3.3 billion words), yields the best performances for our two tasks. For efficient computing and robustness, we use the distilled version of BERT, DistilBERT [Sanh et al. 2019], for our experiments. DistilBERT uses the so-called knowledge distillation technique which compresses a large model, called the teacher (here: BERT), into a smaller model, called the student (here: DistilBERT). The student is trained to reproduce the behavior of the teacher by matching the output distribution. As a result, DistilBERT is 60% faster than the original BERT and requires less computing capacities, while retaining almost its full performance. 42

DistilBERT – as well as its teacher BERT – makes use of Transformer, a multihead attention mechanism that learns relations between words in a text. In contrast to other language models that process a text sequence from left to right, DistilBERT applies bidirectional training, which means that during training, it reads the entire sequence of words at once. More specifically, during training the model is provided with sentences where some words are missing. The task for the model is then to predict the missing (masked) words based on their given context. By learning to predict missing words, the model learns about the structure and semantics of a language during the training phase, which leads to a deeper sense of language context. 43

For our experiments, we use the pre-trained DistilBERT model from HuggingFace Transformers [Wolf et al. 2020] and finetune the training modules on our labelled training data. We use 70% of the data for training and keep 15% for validation and testing, respectively. We optimize the model parameters and configurations on the validation set and report results for the test set. The optimal hyperparameters for our two classification tasks are displayed in Table 4. As our output layer we use softmax. This function enables us to interpret the output vectors of the last layer from the model as probabilities, by mapping them to values between 0 and 1 that all add up to 1. 44

	Task 1	**Task 2**
Number of training epochs	4	4
Batch size	16	4
Learning rate	5e-5	5e-5

Table 4. Hyperparameter setting for DistilBERT.

4.3 Results

For both tasks, when evaluating the two models, respectively, we compare the predicted labels to the gold version of our annotations. We report results on the test sets and use the evaluation metrics Precision, Recall and F1 (we report all scores as micro scores, which means they are weighted according to the label distribution). 45

	Input	**Prec**	**Rec**	**F1**
Naive Bayes	Sentence	80.98	79.84	80.30
DistilBERT	Sentence	93.17	93.15	93.16

Table 5. Results for Task 1: Binary classification between discourse referencing and no discourse referencing.

Table 5 displays the results for our first task – which was, given a sentence, predict if this sentence contains a discourse reference or not. We find that both models – Naive Bayes and DistilBERT – outperform the majority baseline (64.48% for Label 0, No Discourse referencing) significantly. DistilBERT outperforms our baseline model Naive Bayes by 13 percentage points (F1 score), which matches our expectations that the latent linguistic knowledge that DistilBERT stores through its pre-training on large corpora can successfully be utilized for the task of detecting discourse references in political debates. 46

	Input	Prec	Rec	F1
Naive Bayes	**Sentence**	80.55	80.86	78.81
	Phrase	82.04	82.34	80.34
	Sent + phrase	**83.06**	**83.03**	**81.45**
DistilBERT	**Sentence**	92.44	92.44	92.41
	Phrase	**97.08**	**96.79**	**96.90**
	Sent + phrase	96.13	95.88	95.98

Table 6. Results for Task 2, Setting A: Classifying types of discourse references, three classes: “Actor explicitly mentioned” vs. “Metonymic mention of the actor” vs. “No actor mentioned.”

Tables 6–8 display the results of our second task – which was to classify the discourse references into different categories. The results for Setting A in which we distinguish the three categories *Actor explicitly mentioned, Metonymic mention of the actor* and *No actor mentioned* are shown in Table 3. Both models outperform the majority baseline (59.69% for the label *Actor explicitly mentioned*) significantly. For both models, we find that providing the model with relevant phrases instead of or in addition to complete sentences improves the model’s performance. The best results for the Naive Bayes model are obtained by combining the sentence with the relevant phrase as input to the model, while DistilBERT learns best when provided only with the relevant phrase. This indicates that the models are not always fully able to detect which parts of the sentences are relevant for classifying types of discourse references and can benefit from that information when provided with it as input. 47

When comparing the scores of the best input formats for each model, again we find that DistilBERT outperforms Naive Bayes significantly (15.5 percentage points F1 score), again demonstrating the superiority of pre-trained language models as opposed to knowledge-agnostic classification models. 48

	Input	Prec	Rec	F1
Naive Bayes	**Sentence**	79.13	79.30	79.20
	Phrase	**87.30**	**87.19**	**87.23**
	Sent + phrase	84.63	83.95	84.14
DistilBERT	**Sentence**	92.80	92.82	92.79
	Phrase	**98.48**	**98.47**	**98.47**
	Sent + phrase	98.01	98.00	97.99

Table 7. Results for Task 2, Setting B: Classifying types of discourse references, two classes: “Actor explicitly mentioned” vs. “No actor mentioned.”

Table 7 displays the results for Task 2, Setting B in which we exclude the least frequently assigned category *metonymy* and only distinguish between the instances of the two classes *Actor explicitly mentioned* and *No actor mentioned.* We find that the Naive Bayes model significantly improves when excluding the small class *metonymy* (6 percentage points F1 score when provided with the marked phrase), while DistilBERT improves only by 1.5 percentage points compared to setting A (F1 score when provided with the marked phrase). Again, we find that providing both models with relevant phrases instead of complete sentences improves the model’s performance – which especially applies to Naive Bayes. 49

	Input	Prec	Rec	F1
Naive Bayes	**Sentence**	80.73	80.73	80.73
	Phrase	80.93	81.71	81.14
	sent + phrase	**82.84**	**81.88**	**82.25**
DistilBERT	**Sentence**	92.56	92.55	92.50
	Phrase	**97.83**	**97.82**	**97.82**
	sent + phrase	97.37	97.37	97.36

Table 8. Results for Task 2, Setting C: Classifying types of discourse references, two classes: "Actor explicitly mentioned+Metonymic mention of the actor" vs. "No actor mentioned."

In Table 8 we finally display the results for Task 2, Setting C, in which we subsume the category *Actor explicitly mentioned* with the category *Metonymic mention of the actor* under the main category *actor-naming references* and binarily distinguish between the categories *actor-naming references* and *No actor mentioned.* While the results for DistilBERT stay almost the same as in Setting B, we find that the performance of Naive Bayes drops drastically (-5 percentage points, F1 score when provided with a phrase as input). This indicates that the model struggles with the category *Metonymic mention of the actor* – even when this category is subsumed under one label with another category. 50

To summarize, our results show that both models are able to learn to detect and classify discourse references in political debates. The trained knowledge-rich model DistilBERT outperforms the knowledge-agnostic model Naive Bayes significantly on all tasks and settings. We furthermore find that providing the models with relevant phrases instead of or in addition to complete sentences improves the model's performance, which indicates that the models can benefit from being explicitly hinted at the parts of the sentences that are relevant for classifying different types of discourse references. It furthermore shows that those parts of the sentences which are not relevant for distinguishing between different types of discourse markers are not only useless for the classification, but even lower the model's performance. 51

5. Analysis of results

In this section, we present a deeper analysis of the predictions, performance, and errors of our best-performing model DistilBERT. 52

Figure 1 displays the error matrix for Task 1 where DistilBERT achieves a performance of 93.16 F1 score (cf. Table 5). We find that the cases where the model predicts a discourse reference but according to the gold data the respective instance contains no discourse referencing (false positives, n=37) and vice versa (false negatives, n=29) are almost balanced. 53

While the manual analysis of the 29 false negatives did not lead to any observation of linguistic patterns which might lead the model to wrong predictions, the analysis of the 37 false positives showed that in many cases, DistilBERT predicts a discourse reference for those instances that mention an actor, but not in a discourse referencing function such as in example 8 and 9: 54

> **8.** *When it comes to religious constitutional law and legal history at this late hour, I can understand that Mr. von Notz is not the only one who cannot wait to enter this debate.*
> [Wenn es zu vorgerückter Stunde um Religionsverfassungsrecht und Rechtsgeschichte geht kann ich verstehen dass Herr von Notz nicht der Einzige ist der es gar nicht abwarten kann in diese Debatte einzutreten.]
>
> **9.** *The Highway GmbH of the federal state examines the facts of the case.*
> [Die Autobahn GmbH des Bundes prüft den Sachverhalt.]

In both examples, actors are named (*Herr von Notz*; *Die Autobahn GmbH des Bundes*), which leads to the assumption that the model interprets explicit mentions of actors as indicators for discourse referencing. 55

Figure 1. Confusion matrix for DistilBERT on Task 1.

56 Figures 2–4 display the error matrices for the different settings of Task 2. Since the performance of DistilBERT on Task 2 is very high in all three settings, we find only very few errors. A systematic manual analysis of the misclassified instances revealed three main sources of errors:

Figure 2. Confusion matrices for DistilBERT on Task 2, Setting A.

Figure 3. Confusion matrices for DistilBERT on Task 2, Setting B.

Figure 4. Confusion matrices for DistilBERT on Task 2, Setting C.

Error type 1: The model confuses the labels actor and metonymy (Setting A)

57 One common error in setting A is that the category *Actor explicitly mentioned* and the category *Metonymic mention of the actor* are confused by the model. (10) displays an example that mentions an actor only m *etonymically* according to the annotation guidelines but is misclassified by DistilBERT (with all three input options) as an instance that *explicitly mentions the actor.*

> **10.** *Our Basic Law also protects freedom of occupation in Article 12.*
> [Unser Grundgesetz schützt in Artikel 12 auch die Berufsfreiheit.]

58 A reason for this type of error may be the small size of the class *Metonymic mention of the actor*, as it accounts for only 3.05% of the annotations in the gold standard. In the following discussion, we will also reflect on the distinctiveness of these two classes. This error type confirms our choice of Setting B and C, where metonymy is either excluded (B) or subsumed together with the frequent category *Actor explicitly mentioned* under the main category *actor-naming references* (C).

Error type 2: An actor was predicted when there were none

59 Similar to the first type of error, the model misclassifies instances as belonging to the class of actors being explicitly mentioned, whereas according to the gold standard, no actor is mentioned. An explanation may be the mentioning of actors that are not part of the discourse reference made (e.g., *"The Bundestag"* and *"US President Donald Trump"* in (11)), or the use of pronouns (*we* in (12)). This assumption is reinforced by the fact that this error mostly occurs when sentences build the input. When providing the model with a phrase (underlined in the examples), which usually does not contain a named entity/pronoun, the model makes correct predictions.

> **11.** *The Bundestag would be well advised to take this admonition to heart, also in order not to run the risk of being identified with a policy of racist claims of superiority against China, such as that put forward by former US President Donald Trump.*
> [Der Bundestag wäre gut beraten, sich diese Mahnung zu Herzen zu nehmen, auch, um nicht Gefahr zu laufen, mit einer Politik des rassistischen Überlegenheitsanspruchs gegenüber China, wie sie der vormalige US - Präsident Donald Trump nach vorne stellte, identifiziert zu werden.]

> **12.** *Unfortunately, we are increasingly seeing negative aspects of our digital world with disinformation and hate speech.*

[Leider sehen wir mit Desinformation und Hassrede vermehrt auch negative Aspekte unserer digitalen Welt.]

Error type 3: Metonymy is only recognized when the model is provided with a phrase

Lastly, we also find several cases where the model only predicts the category *Metonymic mention of the actor* correctly when provided with a phrase instead of the complete sentences, an example is given in (13). This error again emphasizes the importance of hinting the model to specific phrases for the detection and classification of discourse references, by providing it only with the phrase that has been marked manually as relevant for discourse referencing as input, as described above. 60

13. *On average, women do 1.5 hours more work a day in the household and raising children than their partners - that's what previous surveys tell us - and in return, they can work fewer hours.*
[Frauen leisten im Schnitt täglich 1,5 Stunden mehr Arbeit im Haushalt und bei der Kindererziehung als ihre Partner – das sagen uns die bisherigen Erhebungen – und im Gegenzug können sie weniger arbeiten gehen.]

6. Discussion

First of all, it should be emphasized that the results can be considered very encouraging: The very high F1 values indicate the robustness of the category system and the high quality and homogeneity of the annotations. Not surprisingly, the results of the machine learning experiments show that the pre-trained BERT model outperforms the Naive Bayes model. This can be traced back to the fact that while traditional statistical models such as the Naive Bayes model are solely trained on the labelled training data, BERT is pre-trained on large amounts of data and then fine-tuned on the labelled training data, which makes it a knowledge-rich model. This aligns with the observation in various other NLP tasks such as sentiment analysis, text classification or summarization, where BERT (and other Large Language Models such as XLNet or GPT) usually outperform traditional statistical models (cf. [González-Carvajal and Garrido-Merchán 2020]). 61

In our experiments, especially in the rarer and more difficult category of metonymic actor mentions, the pre-trained model BERT performs well, while this more fine-grained distinction causes difficulties for the untrained Naive Bayes model. In addition to this content-categorical granularity, both models benefit from the higher granularity of segmentation. Phrase-accurate annotation produced better results than annotation with sentence-only segmentation. Thus, the attempt to introduce a kind of human “attention mechanism” into annotation has shown to be successful. 62

Concerning the category development, we observe how important it is to focus on the interplay between form and meaning – between segment size and conceptual granularity of categories: We showed that the annotation on smaller, customized segments that precisely indicate instances of categories improves the pre-trained BERT model's performance in detecting even fine-grained conceptual categories. In contrast to the larger and standardized segment “sentence,” the model could also learn differentiated category systems with high performance based on customized extracted phrases. The greater formal precision relieves the model of the multi-dimensional and highly complex inferential processes involved in human language understanding. In contrast, when categorizing based on sentences as input, the model must in principle mimic the full complexity of human language comprehension. 63

Against this background, we present a review of the course and results of the annotation workflow: The lowest inter-annotator agreement value was obtained for +/- discourse referencing, the least granular and, at first glance, the simplest distinction. 64

Since this presumably simple binary classification task was performed at the segmental level of sentences, the full range of linguistic, contextual, and also domain knowledge was required for classification. Even if the indicators were described as precisely as possible in the guidelines, the high variation of the form-function correlation still requires pragmatic consideration in most cases. This can only be done properly based on expert knowledge, which is acquired in the practice of everyday academic life. Accordingly, uncertainties and misunderstandings arose among student annotators, which could not be clarified by the guidelines alone, but by training and joint practice. Declarative factual knowledge is therefore not sufficient for such a classification task; procedural expert knowledge, as it were, is required. 65

7. Conclusion

With a focus on linguistic routines of discourse referencing, we conducted a collaborative annotation on a sub-corpus of the 66 plenary minutes of the German Bundestag in two steps: First, we performed a binary classification task (+/- discourse referencing). Second, we classified mentions of actors according to a three-item tagset (explicit/metonymic/none). Additionally, we extracted phrases that were identified to indicate discourse referencing. We then ran machine learning experiments with probabilistic and neural classifiers on our annotated dataset as training data. In these experiments, we tested the effect of different types of input data in terms of taxonomies (number of categories) and segment sizes (phrase input vs. sentence input). Our study has shown that the pre-trained neural transformer language model BERT achieves impressive learning results when provided with data annotated according to our category system.

It has been demonstrated that a more fine-grained segmentation on the linguistic surface (that means, the manual selection of relevant phrases) improves the model performance. This suggests that if fine-granular operationalization of pragmalinguistic phenomena in terms of indicators on the linguistic surface is possible, high machine learnability is achievable – probably even for more fine-grained as well as context- and background-knowledge-dependent categories. To summarize, our results show that the recognition and categorization of different types of discourse references can be modelled automatically with neural, knowledge-rich models. 67

In plenary debates, as our studies indicate, these practices of discourse referencing play an important role and are frequently applied. However, we believe that our methodological findings can be generalized to other text genres as well as to other complex linguistic categories. As a conclusion and reflection of our category system development process, it can be summarized that both the performance of the algorithmic models and the human inter-annotator agreement were positively affected by the refinement and specification of the segmentation. A prerequisite for this was the more precise operationalization of the phenomenon under investigation, i.e., the elaboration of more specific indicators on the linguistic surface that can be captured at the level of phrases. This was accompanied by an increase in the degree of granularity of the conceptual categories. Here it is necessary to find the right balance, depending on the object of investigation – also with regard to the machine and human learnability of categorization. An important part of the human learning process in the study took place in the course of the successively more precise operationalization, explication and description in guidelines as well as the accompanying meta-discussion among the annotators. Thus, the initially unclear scope of interpretation depth was gradually resolved by stronger operationalization and by explicit interpretation criteria. We consider this point to be the central success factor and key to collaborative category development and annotation with a view to automation. 68

Works Cited

Allen 2000 Allen, G. (2000) *Intertextuality*. Routledge. Available at: https://doi.org/10.4324/9780203131039.

Archer et al. 2008 Archer, D., J. Culpeper and M. Davies (2008) “Pragmatic Annotation,” in *Corpus Linguistics: An International Handbook*, pp. 613–641.

Artstein and Poesio 2008 Artstein, R. and M. Poesio (2008) “Inter-Coder Agreement for Computational Linguistics,” *Computational Linguistics*, 34(4), pp. 555–596. Available at: https://doi.org/10.1162/coli.07-034-R2.

Becker et al. 2020 Becker, M., M. Bender and M. Müller (2020) “Classifying heuristic textual practices in academic discourse: A deep learning approach to pragmatics,” *International Journal of Corpus Linguistics*, 25(4), pp. 426–460. Available at: https://doi.org/10.1075/ijcl.19097.bec.

Becker et al. 2021 Becker, M. et al. (2021) “CO-NNECT: A Framework for Revealing Commonsense Knowledge Paths as Explicitations of Implicit Knowledge in Texts,” in *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*. *IWCS 2021*, Groningen, The Netherlands (online): Association for Computational Linguistics, pp. 21–32. Available at: https://aclanthology.org/2021.iwcs-1.3 (Accessed: 24 July 2023).

Bender 2020 Bender, M. (2020) “Annotation als Methode der digitalen Diskurslinguistik,” *Diskurse digital. Theorien – Methoden – Fallstudien*. Band 2, Heft 1/2020: 1-35. DOI: https://doi.org/10.25521/diskurse-digital.2020.140.

Bender 2023 Bender, M. (2023) “Pragmalinguistische Annotation und maschinelles Lernen,” in S. Meier-Vieracker et al. (eds) *Digitale Pragmatik*. Berlin, Heidelberg: Springer (Digitale Linguistik), pp. 267–286. Available at: https://doi.org/10.1007/978-3-662-65373-9_12.

Bender and Müller 2020 Bender, M. and M. Müller (2020) “Heuristische Textpraktiken. Eine kollaborative Annotationsstudie zum akademischen Diskurs,” *Zeitschrift für Germanistische Linguistik* 48 (1)/2020: 1-46. DOI: https://doi.org/10.1515/zgl-2020-0001.

Cohen 1960 Cohen, J. (1960) “A coefficient of agreement for nominal scales,” *Educational and Psychological Measurement*, 20, S. 37–46. https://doi.org/10.1177/001316446002000104.

Devlin et al. 2019 Devlin, J. et al. (2019) “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. *NAACL-HLT 2019*, Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. Available at: https://doi.org/10.18653/v1/N19-1423.

Evert and Hardie 2011 Evert, S. and A. Hardie (2011) “Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium,” in *Proceedings of the Corpus Linguistics 2011 Conference*. *Corpus Linguistics 2011*, University of Birmingham, GBR. Available at: https://eprints.lancs.ac.uk/id/eprint/62721/ (Accessed: 24 July 2023).

Feilke 2012 Feilke, H. (2012) “Was sind Textroutinen? Zur Theorie und Methodik des Forschungsfeldes,” in H. Feilke and K. Lehnen (eds) *Schreib- und Textroutinen. Theorie, Erwerb und didaktisch-mediale Modellierung*. Frankfurt am Main u.a., pp. 1–31. Available at: https://www.academia.edu/77763867/Helmuth_Feilke_Was_sind_Textroutinen_Zur_Theorie_und_Methodik_des_Forschungsfeldes (Accessed: 24 July 2023).

González-Carvajal and Garrido-Merchán 2020 González-Carvajal, S. and E. C. Garrido-Merchán (2020) “Comparing BERT against traditional machine learning text classification.” Available at: https://doi.org/10.48550/ARXIV.2005.13012.

Greve and Wentura 1997 Greve, W. and D. Wentura (1997) *Wissenschaftliche Beobachtung: eine Einführung. [Scientific Observation: An Introduction]*. PVU/Beltz.

Hardie 2009 Hardie, A. (2009) “CQPweb - Combining power, flexibility and usability in a corpus analysis tool,” *International Journal of Corpus Linguistics*, 17. Available at: https://doi.org/10.1075/ijcl.17.3.04har.

Jurafsky and Martin 2022 Jurafsky, D. and J. H. Martin (2022) *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. III edn. draft. Stanford. Available at: https://web.stanford.edu/~jurafsky/slp3/ (Accessed: 24 July 2023).

Leech and Weisser 2003 Leech, G. and M. Weisser (2003) “Generic speech act annotation for task-oriented dialogues,” in T. McEnery et al. (eds) *Proceedings of the Corpus Linguistics 2003 Conference*. *Corpus Linguistics 2003*, Lancaster, UK (University Centre for Computer Corpus Research on Language Technical Papers 16(1)), pp. 441–446. Available at: https://www.semanticscholar.org/paper/Generic-speech-act-annotation-for-task-oriented-Leech-Weisser/5869397d550d8440fcd4724083a4b09375703e3b (Accessed: 24 July 2023).

Lüdeling and Kytö 2009 Lüdeling, A. and M. Kytö (eds) (2009) *Corpus Linguistics: An International Handbook*. Mouton de Gruyter. Available at: https://doi.org/10.1515/9783110213881.2.

Müller 2007 Müller, M. (2007) *Geschichte - Kunst - Nation: Die sprachliche Konstituierung einer ‘deutschen’ Kunstgeschichte aus diskursanalytischer Sicht*. Berlin, New York: De Gruyter. Available at: https://doi.org/10.1515/9783110969436.

Müller 2022a Müller, M. (2022a) “Die Plenarprotokolle des Deutschen Bundestags auf Discourse Lab,” *Korpora Deutsch als Fremdsprache*, 2(1), pp. 123–127. Available at: https://doi.org/10.48694/KORDAF-3492.

Müller 2022b Müller, M. (2022b) “Discourse Lab – eine Forschungsplattform für die digitale Diskursanalyse,” *Mitteilungen des Deutschen Germanistenverbandes*, 69, pp. 152–159. Available at: https://doi.org/10.14220/mdge.2022.69.2.152.

Müller and Stegmeier 2021 Müller, M. and J. Stegmeier (2021) “Korpus der Plenarprotokolle des deutschen Bundestags. Legislaturperiode 1–19. CQPWeb-Edition.” Darmstadt: Discourse Lab. Available at: https://discourselab.de/cqpweb/.

Orbach and Goldberg 2020 Orbach, E. and Goldberg, Y. (2020) “Facts2Story: Controlling Text Generation by Key Facts,” *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 2329–2345, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Pedregosa et al. 2011 Pedregosa, F. et al. (2011) “Scikit-learn: Machine Learning in Python,” *The Journal of Machine Learning Research*, 12(null), pp. 2825–2830.

Peters et al. 2018 Peters, M. et al. (2018) “Deep Contextualized Word Representations,” *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long Papers), pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Pierce 1903 Pierce, C. (1903) *CP 7.677*.

Radford et al. 2019 Radford, A. et al. (2019) *Language Models are Unsupervised Multitask Learners*. Online Resource: https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Ravenscroft et al. 2016 Ravenscroft, J. et al. (2016) “Multi-label Annotation in Scientific Articles - The Multi-label Cancer Risk Assessment Corpus,” in *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. *International Conference on Language Resources and Evaluation*, Association for Computational Linguistics, pp. 4115–4123. Available at: https://www.semanticscholar.org/paper/Multi-label-Annotation-in-Scientific-Articles-The-Ravenscroft-Oellrich/fe678e1311cc3ebf9b9a90428c3033269f19fe66 (Accessed: 24 July 2023).

Sanh et al. 2019 Sanh, V. et al. (2019) “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” in. *The 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing. Co-located with the 33rd Conference on Neural Information Processing Systems NeurIPS 2019*, arXiv, pp. 1–8. Available at: https://doi.org/10.48550/ARXIV.1910.01108.

Schmidt et al. 2021 Schmidt, T., K. Dennerlein and C. Wolff (2021) “Emotion Classification in German Plays with Transformer-based Language Models Pretrained on Historical and Contemporary Language,” in *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. *LaTeCHCLfL 2021*, Punta Cana, Dominican Republic (online): Association for Computational Linguistics, pp. 67–79. Available at: https://doi.org/10.18653/v1/2021.latechclfl-1.8.

Teufel 1999 Teufel, S. (1999) *Argumentative Zoning: Information Extraction from Scientific Text*. University of Edinburgh. Available at: https://www.cl.cam.ac.uk/~sht25/thesis/t1.pdf.

Weisser 2018 Weisser, M. (2018) *How to Do Corpus Pragmatics on Pragmatically Annotated Data*, *Studies in corpus linguistics* 84. Amsterdam, Philadelphia: John Benjamins Publishing Company. Available at: https://benjamins.com/catalog/scl.84 (Accessed: 24 July 2023).

Wolf et al. 2020 Wolf, T. et al. (2020) “Transformers: State-of-the-Art Natural Language Processing.” *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, p. 38–45, Online. Association for Computational Linguistics.

Yang et al. 2020 Yang, Z. et al. (2020) *XLNet: Generalized Autoregressive Pretraining for Language Understanding*. Online Resource: https://arxiv.org/abs/1906.08237.

Yin and Chang 2020 Yin, D., T. Meng and K.-W. Chang (2020) “SentiBERT: A Transferable Transformer-Based Architecture for Compositional Sentiment Semantics,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. *ACL 2020*, Online: Association for Computational Linguistics, pp. 3695–3706. Available at: https://doi.org/10.18653/v1/2020.acl-main.341.

Zinsmeister et al. 2008 Zinsmeister, H. et al. (2008) “Linguistically Annotated Corpora : Quality Assurance, Reusability and Sustainability,” in *Corpus Linguistics*. (HSK, 29). Available at: https://www.semanticscholar.org/paper/Linguistically-Annotated-Corpora-%3A-Quality-and-Zinsmeister-Witt/6d0570ac67ec0a231a249a81d96b4a50c045a0f4 (Accessed: 24 July 2023).