Overview of the GermEval 2023 Shared Task on Speaker Attribution in Newswire and Parliamentary Debates

Ines Rehbein
DWS
Mannheim U
rehbein@uni-mannheim.de

Fynn Petersen-Frey
HCDS
Hamburg U
fynn.petersen-frey@uni-hamburg.de

Annelen Brunner
Lexik
Leibniz-IDS
brunner@ids-mannheim.de

Josef Ruppenhofer
CATALPA
Hagen U
josef.ruppenhofer@fernuni-hagen.de

Chris Biemann
LT
Hamburg U
chris.biemann@uni-hamburg.de

Simone Paolo Ponzetto
DWS
Mannheim U
ponzetto@uni-mannheim.de

Abstract

This paper gives an overview of the GermEval 2023 Shared task on Speaker Attribution in Newswire and Parliamentary Debates (SpkAtt2023) and describes the data, annotation guidelines and results of the evaluation campaign. The task targets the identification of speech events in text and their attribution of the respective speakers, including the detection of other roles that might be expressed, such as the addressee or the topic of the speech event. The shared task includes two subtasks, (i) the identification of speech, thought and writing in parliamentary debates and (ii) in newswire text. Being able to identify *who* says *what* to *whom* is crucial for in-depth analyses and enables researchers to extract more meaningful information from unstructured text.

1 Introduction

Identifying who says what to whom is an essential prerequisite for analysing human communication. The complexity of the task, however, is often underestimated by assuming that the words produced by the speaker only reflect his or her own point of view. Figure 1 shows an excerpt from a parliamentary debate of the German Bundestag, illustrating how speakers frequently switch perspectives, at times presenting their own views and sometimes reporting and citing the views of others. Thus, it is crucial to identify the correct source for each speech event when analysing text. Furthermore, studying how speakers construct their own arguments relative to the views of other speakers, either to back up their own claim or to attack the others' perspective, is an intruiging research question in itself.

In order to investigate these questions, we need annotated resources that allow us to train models that learn to predict speech events in unstructured text, together with their respective speakers, messages and addressees. This overview paper presents two new resources for speaker attribution in German text, based on parliamentary debates from the German Bundestag and on newswire text. We first review previous work on quote detection and speaker attribution before we describe our data and the annotation process. Then we provide a description of the shared task settings and report baseline results for each of the two new resources. Finally, we present the results of the shared task, with an evaluation of the system output for the participating systems.

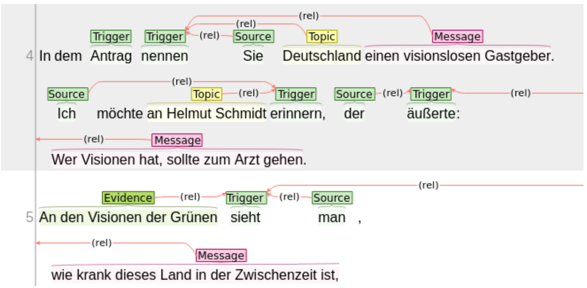
4 In dem Antrag nennen Sie Deutschland einen visionslosen Gastgeber.
Ich möchte an Helmut Schmidt erinnern, der äußerte:
Wer Visionen hat, sollte zum Arzt gehen.
5 An den Visionen der Grünen sieht man ,
wie krank dieses Land in der Zwischenzeit ist,

Figure 1: Example for speaker attribution in parliamentary debates (Task 1).

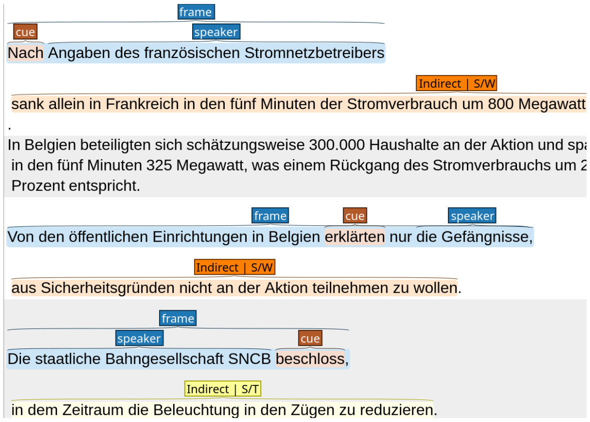
Nach Angaben des französischen Stromnetzbetreibers
sank allein in Frankreich in den fünf Minuten der Stromverbrauch um 800 Megawatt
.
In Belgien beteiligten sich schätzungsweise 300.000 Haushalte an der Aktion und spa
in den fünf Minuten 325 Megawatt, was einem Rückgang des Stromverbrauchs um 2
Prozent entspricht.
Von den öffentlichen Einrichtungen in Belgien erklärten nur die Gefängnisse,
aus Sicherheitsgründen nicht an der Aktion teilnehmen zu wollen.
Die staatliche Bahngesellschaft SNCB beschloss,
in dem Zeitraum die Beleuchtung in den Zügen zu reduzieren.

Figure 2: Example for speaker attribution in news articles (Task 2).

Cue/Role name	description	example
Cue	the cue that triggers the STW event	Merkel spoke to the people.
Source	Source of the STW event	Merkel spoke to the people.
Medium	Medium of the STW event	The Basic Law reads ...
Message	Message / content of the STW event	She said that she would resign.
Topic	Topic of the STW event	Merkel addressed the theme of taxation.
Evidence	Evidence for the message	The survey shows that ...
Addressee	Addressee of the STW event	Merkel spoke to the people.
Particle	Separated verb prefix or	Merkel schlug vor (proposed) ...
(PTC)	obligatory particle	Merkel Cue sich vor (imagines herself) ...

Table 1: Overview over our classification scheme for annotating events of **S**peech, **T**hought and **W**riting (STW).

2 Related Work

2.1 Work on speaker attribution

Much recent work has been devoted to quote detection, mostly with the goal of extracting information from newswire text (Pouliquen et al., 2007; Krestel et al., 2008; Pareti et al., 2013; Pareti, 2015; Scheible et al., 2016). Other related work comes from the field of opinion mining and has targeted the identification of opinion holders (speakers) and the targets of the opinions (Choi et al., 2005; Wiegand and Klakow, 2012; Johansson and Moschitti, 2013).

Many studies have addressed speaker attribution in novels and other literary works, in the context of computational literary studies. Elson and McKeown (2010) were among the first to propose a supervised machine learning model for quote attribution in literary text. He et al. (2013) extended their supervised approach by including contextual knowledge from unsupervised actor-topic models. Almeida et al. (2014) and Fertmann (2016) combined the task of speaker identification with coreference resolution. Grishina and Stede (2017) test the projection of coreference annotations, a task related to speaker attribution, using multiple source languages. Muzny et al. (2017) improved on previous work on quote and speaker attribution by providing a cleaned-up dataset, the QuoteLi3 corpus, which includes more annotations than the previous datasets. They also present a two-step deterministic sieve model for speaker attribution on the entity level and report a high precision for their approach.[1] Papay and Padó (2020) annotate direct and indirect quotations in 19th century English literature while Kim and Klinger (2018) extend the speaker attribution task to capture emotion trigger phrases and the experiencers, targets and causes of the emotion.

While many studies have addressed the task of quote detection or speaker attribution in English text from the literary domain or in news articles, less work has been done for other languages and genres. Brunner (2015), Krug et al. (2018), Brunner et al. (2019) and Brunner et al. (2020) have focused on German literary text and created several resources. The DROC corpus (Krug et al., 2018) includes around 2,000 manually annotated quotes and annotations for speakers and their mentions in 90 fragments from German literary prose and the RedeWiedergabe corpus substantially extends this work by presenting a German-language historical corpus with detailed annotations for speech, thought and writing (Brunner et al., 2020). Dönicke et al. (2022) address a task related to speaker attribution, i.e., identifying whether a certain text passage is written from the perspective of the narrator of the novel or from the author's point of view, or whether it reflects the view of a character in the novel. Interestingly, they show that including annotator bias in the model can improve results.

Less work has been done for other domains. A noteworthy exception is Ruppenhofer et al. (2010) who present preliminary work on speaker attribution in text from the political domain, using German cabinet protocols. As the focus of SpkAtt2023 Task 1 is also on analysing the language of political debates, we extended the work of Ruppenhofer et al. (2010) and created a new, manually annotated resource for speaker attribution with around 13,000 clauses and more than 200,000 tokens. Our second research focus is on analysing who says what to whom according to German news media (SpkAtt2023 Task 2). For this, we created a new, manually annotated dataset for speaker attribution in German news articles with almost 250,000 tokens.

Brunner et al. (2020) was an important basis for our annotation in both tasks in that we take into account not only speech events, but also thought

[1] When optimised for precision, the system obtains a score >95% on the development set from *Pride and Prejudice*.

Cue/Role	freq.	avg. len
Cue	7,706	1.1
Source	4,663	1.7
Message	4,578	9.7
Topic	1,188	5.4
Addressee	717	3.2
Particle	561	1.0
Medium	321	3.2
Evidence	151	4.3

Table 2: Statistics for the Task 1 dataset (GePaDe).

and writing. While the annotation scheme used in Task 2 can be seen as a direct adaptation, the annotation scheme for Task 1 shares several ideas with Brunner's work but has a somewhat different label inventory inspired by work in Automatic Semantic Role Labeling in the FrameNet mode (Baker et al., 1998). We describe the creation of these resources in the next section.

3 Data and Annotation

3.1 Task 1: Speaker attribution in German parliamentary debates

We present a new dataset for speaker attribution in data from the political domain, specifically, parliamentary debates from the German Bundestag. Our dataset includes manually annotated cues that trigger events of speech, writing and thought.[2] In addition, we annotate the arguments of the trigger, including the Source, Addressee, Message, Medium, Topic and Evidence for the speech event. Table 1 shows examples for the different categories in our schema.[3] We now describe our data, annotation setup and annotation procedure.

Data The data for Task 1 includes debates from the German Bundestag, retrieved from Deutscher Bundestag – Open Data.[4] The data set includes 265 speeches from the German Bundestag, mostly from the 19th legislative term (2017-2021), given by 195 different speakers from 6 parties (CDU/CSU: 76, SPD: 57, AfD: 39, FDP: 33, Linke: 29, Grüne: 26, non-attached: 4). The total size of the data is >200,000 tokens. For more detailed information on the data, sampling and annotation process, please refer to the datasheet.[5]

Annotation process The data was annotated by four student assistants from different fields in the humanities. The annotators received extensive training. During the annotation phase, weekly meetings were held where we discussed open questions and problematic cases.

To ease the detection of speech events, we started with a list of cue words extracted from the RedeWiedergabe Corpus (Brunner et al., 2020). We marked all lemma forms from the list in our data and instructed the annotators a) to verify whether this instance is a Speech, Thought and Writing (henceforth: STW) event and, b) if true, to identify all of its arguments realised in the utterance. To increase recall, we asked the annotators to add new cue words to the list that were then included in the annotation. Table 2 shows the number of annotated cues and their roles in our corpus. Overall, we annotated more than 7,700 events of speech, thought or writing in the data.

Inter-annotator agreement We split the data into four samples that reflect the order of annotation. Table 3 shows the average percentage agreement of two coders for cue words and roles as the proportional token *overlap* between the annotated cues or roles. To augment this view, we also report a more lenient *binary* score which considers an annotation as correct if at least one token in the annotations overlaps and has been assigned the same label.[6] We can clearly see that inter-annotator agreement constantly improves with more training even after the third round of annotation.

Disagreements between the annotators Most questions during annotation concerned the class of Thought events. Our guidelines follow Brunner et al. (2020) and define Thought as "silent or inner speech which can be reproduced in the same way as verbalized speech". Brunner et al. (2020) conceptualise Thought as "a conscious, analytical, cognitive process" and exclude descriptions of emotional and mood states or passages that are told from a strongly personal perspective. This definition, however, is hard to operationalise and there

[2] In the reminder of the paper, we use the term "speech event" to refer to events of speech, thought or writing.

[3] The annotation guidelines are available at https://github.com/umanlp/SpkAtt-2023.

[4] https://www.bundestag.de/services/opendata.

[5] The datasheet is available from our github repository: https://github.com/umanlp/SpkAtt-2023/blob/master/doc/SpkAtt-Debates-Datasheet.pdf.

[6] For more details on the scoring method, see (Marasovic and Frank, 2018).

	overlap		*binary*
Sample	**Cue**	**Roles**	**Roles**
Sample 1	69.07	64.53	67.88
Sample 2	81.19	67.04	72.60
Sample 3	81.95	72.11	76.90
Sample 4	82.84	73.81	77.63

Table 3: Pair-wise percentage agreement between the annotators on the four samples from GePaDe (Task 1) (*overlap:* proportional token overlap between A1 and A2; *binary:* at least one token in the cue/role span has been identified and assigned the same label).

were many borderline cases that required discussion. We used our weekly meetings to decide which new cue words we would like to include. For more details, please refer to the annotation guidelines.

At the beginning of the annotation process, some annotators were eager to identify new cue words for thought events while others had a more conservative approach, considering only cues from our list. This is reflected in the high disagreement for sample 1. Sometimes new cues were included after one coder had already completed a document, ignoring those cues, while the second coder included the new cues in the annotations. The confusion matrix (Appendix, Table 11) shows that this is in fact the major source of disagreements: instances that were annotated by one annotator but not by the second coder (label NONE).

Other disagreements concern the distinction between MESSAGE and TOPIC (Example 3.1) and between MEDIUM and EVIDENCE (Example 3.2).

When distinguishing between TOPIC and MESSAGE, the annotators sometimes struggled to decide whether the speaker simply mentioned a certain topic or whether she also tried to convey a message. For instance, Example 3.1 may either be taken to mean that the addressee (“Sie” , 2Sg.formal) spoke about a democratic imposition (TOPIC) or that they said that something constituted a democratic imposition (MESSAGE).[7] Similarly, the distinction between MEDIUM and EVIDENCE was another case that was difficult for the annotators. Consider Example 3.2 where it is not clear whether the bold-faced text should be considered as the medium that transported the message or whether it should be interpretated as Evidence.

	freq.	**avg. len**
sentence	13,186	18.84
MESSAGE	4,182	16.69
CUE	2,929	1.57
ADDRESSEE	337	2.72
FRAME	3,038	8.95
SOURCE	3,908	3.53

Table 4: Statistics for the Task 2 dataset (news).

More details on the distinction between those labels can be found in the annotation guidelines.

Ex. 3.1 (Topic vs. Message)

Sie haben von einer „demokratischen Zumutung“ gesprochen.
You have spoken **of a "democratic imposition"**.

Ex. 3.2 (Medium vs. Evidence)

[...] die weltweite Stimmung mahnt uns, Erkämpftes zu erhalten [...]
[...] the global mood *urges us to preserve what we have fought for [...]*

3.2 Task 2: Speaker Attribution in German news articles

We present a new creative-commons-licensed dataset for speaker attribution in German news articles. The dataset consists of manually annotated articles from the German WIKINEWS website.[8] In total, these annotated articles contain almost 250,000 tokens. We manually annotated and curated MESSAGES in different forms of speech such as DIRECT, INDIRECT, FREE INDIRECT, INDIRECT/FREE INDIRECT, REPORTED together with the corresponding FRAME, SOURCE, CUE and ADDRESSEE. Table 4 reports the number and the average length of MESSAGE and the four roles used in Task 2. Examples for these roles can be found in Table 1. Table 5 shows the number and average length of the SPEECH/THOUGHT/WRITING representation (STWR) and the form of speech for our MESSAGE annotations. In the following subsections, we describe the raw source data, its pre-processing, the annotation process, the inter-annotator agreement and the handling of disagreements between annotators.

[7] Based on the quotation signs used we think the latter interpretation is more likely to be correct but it’s a subtle judgment.

[8] URL: https://de.wikinews.org

	freq.	avg. len
Direct	873	17.54
Indirect	2250	14.71
Free Indirect	171	20.43
Indirect/Free Indirect	434	22.33
Reported	454	18.01
Speech	1906	16.75
Writing	572	19.13
Thought	2	10.5
Speech/Thought (ST)	322	14.95
Speech/Writing (SW)	1362	16.0
Writing/Thought (WT)	0	-

Table 5: Message statistics for the Task 2 dataset.

Sample	Form	STWR	Roles
Sample 1	0.56	0.37	0.61
Sample 2	0.76	0.51	0.75
Sample 3	0.77	0.40	0.76
Sample 4	0.77	0.68	0.76
Sample 5	0.86	0.51	0.83
Sample 6	0.78	0.61	0.78

Table 6: Krippendorff's Alpha agreement between the annotators on the six samples from Task 2

3.2.1 Source data

The data originates from news articles published on the German WIKINEWS website. We used the XML dump[9] available through the Wikimedia foundation. Our dataset is based on the dump from April 2022 that consists of 13,001 published articles. From these published articles, we sampled 1000 articles to annotate. These articles range from December 2004 to March 2022.

3.2.2 Data pre-processing

Since the articles are stored in MediaWiki markup with custom macros for the German WIKINEWS, we wrote a program to automatically convert this markup into plain text. The conversion is a recursive procedure in order to support the nested macros present in the markup. Using this approach, we stripped all markup like formatting (e.g. bold, italic), semantic information (e.g. links to entities on Wikipedia) and non-textual content (e.g. pictures, tables) from the documents. Further, we removed any text not belonging to the main text body such as publication metadata, comments, links to related articles or sources. The resulting plain text was tokenized and split into sentences using spaCy (Honnibal et al., 2020). Finally, the tokenized text was exported in a format compatible with our annotation software.

3.2.3 Annotation process

The annotation was carried out by three annotators with a background in German studies or Linguistics and an additional supervisor. The annotators were selected after performing a trial annotation on a handful of articles. The annotation team received training during a preliminary annotation before the actual annotation begun. Further, we held weekly meetings during the main annotation to discuss open questions and uncertain cases, thereby providing ongoing training to all annotators.

As outlined in Section 2, the annotation scheme is based on the Redewiedergabe project (Brunner et al., 2020). In an initial preliminary annotation, we tested the suitability of the annotation scheme in the news domain. We iteratively tested which attributes of the schema are necessary and which additional options we needed. Finally, we settled on the medium (referred to as STWR in the dataset) and type attribute for a Message and Frame, Cue, Source and Addressee as the other annotation parts (roles). STWR can either be Speech, Thought, Writing or one of the combinations Speech/Thought, Speech/Writing, Writing/Thought for cases where it is not possible to confidently decide on a single value from the text. The types of speech are taken from the Redewiedergabe project: Direct, Indirect, Free Indirect, Reported and Indirect/Free Indirect. For more details refer to the annotation guidelines (see supplementary materials).

For the annotation, we used the annotation software INCEpTION (Klie et al., 2018). The different parts are modeled as span annotations with relations between them to indicate e.g. which Source belongs to which Message.

3.2.4 Inter-annotator agreement

We used Krippendorff's Alpha to compute the agreement between two annotators per sample. The measure includes both the quality of the span annotation offsets (overlap) as well as their labels, but does not include the relations between the span annotations. However, the relations were typically made identically given the same annotation

[9] URL: https://dumps.wikimedia.org/dewikinews/

spans and labels. Moreover, for different annotation spans, there is no sensible way to compute an inter-annotator agreement on the relations.

Table 6 shows the inter-annotator agreement values for the six samples into which we divided the 1000 annotated documents. The inter-annotator agreement values increased strongly after the first sample, slightly increasing with additional experience and training over the course of the remaining samples. As such, the first sample required significant curation effort and discussion that ultimately led to improved skills of our annotators.

3.2.5 Disagreements between annotators

During the annotation phase we held weekly meetings to discuss general questions concerning how we would best annotate specific phenomena within our annotation scheme. After two annotators had finished annotating the documents, we employed curation by a third person to resolve differences between the annotations. In situations where the curator was not certain who (or if any) of the two annotators had annotated the sentences in question correctly, we discussed the issue in detail to resolve the disagreement, thereby potentially defining our annotation guidelines more precisely.

One of the most frequent reasons of disagreement during the early phases of the annotation was the difficulty of choosing the correct STWR, usually the choice being between writing or speech. After many discussions, we concluded that it is sometimes impossible to decide from the text alone whether an utterance was produced in spoken or written form. As such, we modified our annotation scheme by adding three new labels to STWR (see Section 3.2.3).

4 Task Description

The SpkAtt2023 shared task included two tasks: (i) speaker attribution for parliamentary debates from the German Bundestag and (ii) speaker attribution in German newswire. The teams could participate either in both or just in one of the two tasks.

The terms of the shared task required that any data or models used outside of those that are provided should be publicly accessible or be made public by April 1, 2023 (release of the training data). Each team could submit multiple submissions, however, the last submission uploaded by the team was considered to be the official entry to the competition.

4.1 Task 1: Parliamentary debates

The goal of Task 1 was the identification of speakers in political debates and newswire, and the attribution of speech events to their respective speakers.

For this task, participants were asked to build a system that can identify all cue words that trigger a speech event and, for each speech event, all roles associated with this event (i.e., Source, Addressee, Message, Topic, Medium, Evidence). The task setup is thus similar to Semantic Role Labelling.

For Task 1, the participants could take part in the following subtasks:

- **Subtask 1 (full task):** Participants were asked to predict the cue words that trigger a speech event, together with the associated roles and their respective labels.

- **Subtask 2 (role labelling):** For this subtask, the gold cue words were given and the task consisted in identifying the spans for all associated roles expressed in the text, together with their respective labels.

A detailed description of the data format and the annotations can be found in the Task 1 GitHub repository: `https://github.com/umanlp/SpkAtt-2023` (see README and annotation guidelines). The trial and training data were made available from the same GitHub page.

4.2 Task 2: News articles

For this task, participants had to develop a system that identifies statements (MESSAGE), i.e. instances of speech (DIRECT, INDIRECT, FREE INDIRECT, INDIRECT/FREE INDIRECT, REPORTED) and the corresponding roles with it (FRAME, SOURCE, ADDRESSEE, CUE). Further, the system should identify the speech form and relevant medium (SPEECH, THOUGHT, WRITING) accordingly.

The participants could take part in the following task settings:

- **Subtask 1 (full task)**: Predict all parts of a statement, associate them, and label the form of speech and medium

- **Subtask 2 (simplified)**: Predict only the SOURCE (i.e. speaker) and MESSAGE (quotation) of top-level (i.e. not nested) annotations, then link both together. The annotation data contains a boolean flag to select only relevant annotations (`"IsNested": false`)

The technical data format description and some additional details are provided in the Task 2 GitHub repository at `https://github.com/uhh-lt/news-speaker-attribution-2023` (see the README file). This website is the place where the trial, training, development and blind test data were published.

5 Evaluation

We now present the experimental setup and report baseline results for both tasks.

5.1 Baseline system (Task 1 – GePaDe)

In order to automatically predict cue words for speech events and their roles, we split our data into training, dev and test sets with 9,298/927/3,067 sentences.[10] This amounts to 178/18/72 different speeches in each set, with 5,536 (train), 515 (dev) and 3,646 (test) annotated STW events.

For our baseline, we use two heuristic approaches. To predict the cue words, we extract all wordforms for cues from the training data. To reduce noise, we do not consider multiword triggers and also remove prepositions from the set of cue words. Then we search the test data for wordforms that match a cue word from our list and, if we find one, we insert a speech event for this cue.

To predict the roles, we use a dependency-based syntactic heuristic and assign all subjects of verbal cue words the label SOURCE and all direct objects of verbal cue words the label MESSAGE. For nominal cue words, we assign the label SOURCE to possessive pronouns (Ihren eigenen Antrag; *engl.: her own proposal*) and genitive NPs that bear the dependency label AG.

5.2 Evaluation metrics

The evaluation of system performance uses the familiar Precision, Recall and F1-metrics. Both cue and role labels can cover more than one token and therefore are represented as sets of (possibly discontinuous) tokens. The annotation scheme assumes that a given set of tokens can bear at most one cue annotation, that is, it can evoke at most one instance of speech, throught or writing. For roles this is not true: a set of tokens could bear multiple role labels, usually in relation to different cues.

According to our definition of the task, roles are dependent on cues and so system roles can match gold roles only if they are related to the same cue. In line with this, the evaluation first checks how system cues and gold cues align. In doing so the scorer matches at most one system cue to a at most one gold cue and the same in the other direction. System cues that cannot be aligned to gold cues produce false positives, including for their associated roles. In symmetric fashion, gold cues that cannot be aligned to a system cue result in false negatives.

For both cues and roles, alignment requires nonzero overlap with the tokens covered by a label of the same type on the other side. Each component token of aligned labels is counted as a true or false positive, or as a false negative. This means that longer spans contribute more to the overall score than shorter labels. In situations where a multi-token cue on one side overlaps with two or more separate cues on the other side, the scorer scores all possible alignments and chooses the one that maximizes the joint F1-score for cues and roles.

5.3 Baseline system (Task 2)

We developed Quotes in Text (QUiTE) – a rule-based system to extract direct and indirect quotations with the speaker from text. The system follows ideas of an older system presented by Bögel and Gertz (2015). QUiTE uses rules and word lists on top of neural components for dependency parsing and named-entity recognition. DIRECT speech is identified by regular expressions looking for quotation marks. The SOURCE of the quotation (i.e. the speaker) is searched in the proximity, preferring candidates in the same sentence but outside of the quotation span. INDIRECT speech is identified through the grammatical structure of a sentence (using dependency parsing) and the main or auxiliary verb being a cue word that is looked up in a word list. The word list contains utterance verbs (verba dicendi) that can be used to indicate (in)direct speech. In addition, the system finds sentences in subjunctive mood that occur directly before or after a sentence containing quotation and source. These sentences are typically marked as INDIRECT/FREE INDIRECT in the dataset. Lastly, the system combines DIRECT and INDIRECT speech, enriching the information of identical quotations.

5.4 Evaluation metrics (Task 2)

Task 2 is evaluated similarly to Task 1 using the the usual Precision, Recall and F1-metrics on token overlap of possibly discontinuous spans (sets

[10] We use spaCy for sentence segmentation which results in segments on the clause level, with an average size of around 16 tokens/clause.

	Cues			**Roles**			**Joint**		
Team	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
baseline	57.34	82.96	67.81	67.02	32.00	43.32	64.33	37.73	47.56
aehrm2	89.70	88.87	89.28	77.64	87.06	82.08	78.85	87.26	82.84
nesasio	88.92	88.92	88.92	78.69	82.15	80.38	79.80	82.91	81.33
moiddes	67.48	66.08	66.77	56.67	84.30	67.78	57.51	82.25	67.69

Table 7: Evaluation results for Task1, subtask 1 (cues & roles).

Team	**Prec**	**Rec**	**F1**
baseline	89.08	33.66	48.86
aehrm2	91.12	90.23	90.67
nesasio	90.96	87.32	89.10
moiddes	53.49	85.35	65.76

Table 8: Evaluation results for Task1, subtask 2 (roles only).

of tokens). However, the roles are optional but always depend on a Message. Thus, predicted roles can only match gold roles if they belong to a matched Message. A span representing a role can be related to multiple Message spans, i.e. the same Source can utter multiple Messages. Roles or Message spans can be nested within another Message or Frame in the full task. To perform an evaluation, Messages from system and gold are assigned via linear sum assignment of the Message span's token overlap using form and STWR as tie-breakers. Each Message can only be matched to at most one other Message. The tie-breakers are needed to correctly assign Messages in rare cases as they can have the same offsets, yet use a different form or STWR. If a system predicts a Message that has no matching Message in the gold annotations, this increases the false positives for Message and each role system predicted as belonging to the unmatched Message. Vice versa, if a Message from the gold annotation has no match in the system prediction, the false negatives are increased. A correctly matched Message yields true positives for all correct roles according to the fraction of overlap and false negatives resp. false positives for tokens that were not identified resp. wrongly predicted by the system.

5.5 Baseline results

5.5.1 Task 1 – Parliamentary debates

Table 8 shows results for the baseline system (Task 1). The simple string match for the prediction of cues has a recall over 80% but precision is rather low with 57%. The heuristics-based role prediction thus suffers from error propagation (precision: 67%) and even more from the low coverage of our heuristic rules (recall: 32%). When applying the role prediction baseline to gold cues, we can see a substantial improvement for precision (89%) but not for recall.

A qualitative error analysis showed that, in addition to the low recall, many errors are due to incorrect syntactic parses. The dependency parser struggles with the long sentences and many parenthetical remarks included in the debates and, in addition, often fails to return the correct analysis for copula constructions.

5.5.2 Task 2 – News articles

Table 9 resp. Table 10 shows the results for the baseline system (Task 2) on the development resp. test set. The rule-based system is not tuned on the development set (and in fact not even trained on the training set). Consequently, there is almost no difference between the scores on the test and development set.

The results show that the system achieves decent precision while clearly suffering from low recall. The low recall mainly results from two causes. First, the system is not capable of predicting certain types of speech (Reported and Free Indirect) or roles (Addressee) that are present in the dataset. Second, the system was designed to prefer quality over quantity when automatically extracting quotations from large amounts of raw text. As such, the system has a preference for precision over recall even for types of speech that it can predict.

When comparing the results of the full task with

	Prec	Rec	F1
Subtask 1 (full task)			
Message	75.12	36.13	48.79
Roles	55.03	25.53	34.87
Joint	60.65	28.65	38.91
Form	57.78	29.56	39.11
STWR	56.59	28.94	38.30
Subtask 2 (simplified task)			
Message	71.29	36.46	48.25
Source	57.76	24.93	34.83
Joint	64.65	30.90	41.82

Table 9: Task 2 baseline results on the development set

	Prec	Rec	F1
Subtask 1 (full task)			
Message	70.75	36.22	47.91
Roles	55.60	26.05	35.48
Joint	59.86	28.99	39.06
Form	63.48	33.59	43.93
STWR	52.46	27.76	36.31
Subtask 2 (simplified task)			
Message	68.74	37.01	48.12
Source	53.98	22.47	31.73
Joint	61.56	30.02	40.36

Table 10: Task 2 baseline results on the test set

the simplified task, it can be seen that the system has worse Message precision but slightly better Message recall. This phenomenon can be attributed to the fact that the system produces the same output for both subtasks – it does not differentiate between the tasks. Since it predicts some cases of nested Messages (e.g. Direct speech within Indirect speech) the Message precision on the simplified task (that does not include nesting) is lower. As a side effect recall is slightly increased because in the reference data some instances of unsupported types of speech are excluded due to nesting. According to the joint score, the system performs better on the simplified task than the full task – while performing worse on Messages. The reason for this is the averaging over all correct resp. predicted spans: In the simplified task, there is only a single role (Source) and thus fewer role spans than in the full task. As the system is significantly better at predicting the Messages than the roles, the joint performance increases on the simplified task.

5.6 Results of the SpkAtt2023 shared task

5.6.1 Task 1

The shared task had three participating teams that submitted their system results. Only two of the participating teams submitted a system description. Below we summarize the main features of each system. For details, see the system descriptions (Ehrmanntraut et al., 2023; Bornheim et al., 2023).

Speaker attribution with BERT The winning system is based on a large BERT model (deepset/gbert-large, Chan et al. (2020)) and divides the task into three subtasks. In the first step, the system tries to identify the cue words. Next, individual cue words are grouped into cue spans (i.e., multi-word cues) that trigger the same speech event. In the last step, given a group of cue words, the system predicts the associated roles for this cue as a multi-label classification task on the token level. To increase efficiency, the system does not fine-tune the full model parameters but inserts Low Rank Adapters (LoRA) (Hu et al., 2021) into the model that are then fine-tuned on the data, either in a token classification setup (cue word detection; role detection) or in a sequence classification task (detection of multi-word cues).

The participants also experimented with domain adaptation via continual pre-training on in-domain data but could not further improve their results.

Speaker attribution with Llama 2 The second-ranked system decided on a very different design for the speaker attribution task, using a prompt-based approach. The system is based on two fine-tuned Llama 2 models (Llama 2 70B) (Touvron et al., 2023), one for identifying the cues and one for role prediction. To reduce memory usage and make the system more efficient, QLoRA (Quantized Low-Rank Adaptation) (Dettmers et al., 2023) has been applied to quantize the model weights to four bits. Additionally, LoRA adapters are added to all linear transformer blocks of the model.

The prediction of cues and roles is done separately by means of two prompting mechanisms and postprocessing, in order to convert the system output into structured predictions for evaluation. More details on the implementation can be found in the system description (Bornheim et al., 2023).

Results A summary of the results can be seen in Table 8. All three systems beat the joint baseline for both, subtask 1 and 2. While the two best-ranked systems yield very similar results for cue prediction, the BERT-based system clearly outperforms the QLoRA-adapted Llama 2 model for role prediction with regard to recall (82% vs. 87%).

Interestingly, for role prediction on *automatically predicted* cues the QLoRA-adapted LLM seems to outperform the BERT-based system.[11] When predicting roles on gold cues, however, this advantage disappears and the BERT-based system beats the other systems in both, precision and recall.

5.6.2 Task 2

Since no team submitted an official run for Task 2, the only results on this task are the baseline results presented in Section 5.5.2. Thus, we are looking forward to task and dataset being used in future experiments and evaluations.

6 Conclusions

We presented an overview of the GermEval 2023 Shared Task on Speaker Attribution in Newswire and Parliamentary Debates. The shared task provided two new datasets, one including parliamentary debates from the German Bundestag (Task 1) and one from the news domain (Task 2). Each task consisted of two subtasks. All data is made available, either via a GitHub repository (train and dev sets) or in codalab (test sets for evaluation).

The outcome of the shared task showed results close to 90% F1 for the detection of cue words and well above 80% F1 for role prediction on automatically predicted cues (Task 1). When also providing the gold cues, we see a further increase in results for role prediction up to 90% F1. The high accuracy of the results should enable new applications in the computational social sciences and the release of the new datasets will provide the basis for further improvements for speaker attribution in German text.

Acknowledgements

We would like to thank the German Society for Computational Linguistics and Speech Technology (GSCL) for their generous financial support for the creation of the Task 1 dataset. We also thank our annotators for their dedicated work. Part of this project has been supported by a research grant from the Ministry of Science, Research and the Arts (MWK) Baden-Württemberg.

[11] But note that the results are based on one run only, with no information on the standard deviation of the systems. Thus, more evaluation is required before we can draw any final conclusions.

References

Mariana S. C. Almeida, Miguel B. Almeida, and André F. T. Martins. 2014. A joint model for quotation attribution and coreference resolution. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 39–48.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Thomas Bögel and Michael Gertz. 2015. Did I really say that? - Combining machine learning and dependency relations to extract statements from German news articles. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology, GSCL 2015, University of Duisburg-Essen, Germany, 30th September - 2nd October 2015*, pages 13–21. GSCL e.V.

Tobias Bornheim, Niklas Grieger, Patrick Gustav Blaneck, and Stephan Bialonski. 2023. Speaker Attribution in German Parliamentary Debates with QLoRA-adapted Large Language Models. In *The GermEval 2023 Shared Task on Speaker Attribution in Newswire and Parliamentary Debates, co-located with KONVENS 2023*, Ingolstadt, Germany.

Annelen Brunner. 2015. Automatic recognition of speech, thought, and writing representation in german narrative texts. *Literary and Linguistic Computing*, 28(4):563 – 575.

Annelen Brunner, Stefan Engelberg, Fotis Jannidis, Ngoc Duyen Tanja Tu, and Lukas Weimer. 2020. Corpus redewiedergabe. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC), May 11-16, 2020, Palais du Pharo, Marseille, France*, pages 803 – 812, Paris. European Language Resources Association.

Annelen Brunner, Ngoc Duyen Tanja Tu, Lukas Weimer, and Fotis Jannidis. 2019. Deep learning for free indirect representation. In *Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019), October 9 – 11, 2019 at Friedrich-Alexander-Universität Erlangen-Nürnberg*, pages 241 – 245, München [u.a.]. German Society for Computational Linguistics & Language Technology und Friedrich-Alexander-Universität Erlangen-Nürnberg.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Identifying sources of opinions with conditional random fields and extraction patterns. In *Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, HLT/EMNLP 2005, pages 355–362.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs.

Tillmann Dönicke, Hanna Varachkina, Anna Mareike Weimer, Luisa Gödeke, Florian Barth, Benjamin Gittel, Anke Holler, and Caroline Sporleder. 2022. Modelling Speaker Attribution in Narrative Texts With Biased and Bias-Adjustable Neural Networks. *Frontiers in Artificial Intelligence*, 4:725321.

Anton Ehrmanntraut, Leonard Konle, and Fotis Jannidis. 2023. Politics, BERTed: Automatic Attribution of Speech Events in German Parliamentary Debates. In *The GermEval 2023 Shared Task on Speaker Attribution in Newswire and Parliamentary Debates, colocated with KONVENS 2023*, Ingolstadt, Germany.

David K. Elson and Kathleen R. McKeown. 2010. Automatic attribution of quoted speech in literary narrative. In *The Twenty-Fourth AAAI Conference on Artificial Intelligence*, AAAI 2010.

Susanne Fertmann. 2016. Using speaker identification to improve coreference resolution in literary narratives. Master's thesis, Computational Linguistics.

Yulia Grishina and Manfred Stede. 2017. Multisource projection of coreference chains: assessing strategies and testing opportunities. In *The 2nd Coreference Resolution Beyond OntoNotes Workshop*, CORBON-2017.

Hua He, Denilson Barbosa, and Grzegorz Kondrak. 2013. Identification of speakers in novels. In *The 51st Annual Meeting of the Association for Computational Linguistics*, ACL 2013, pages 1312–1320.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, et al. 2020. spacy: Industrial-strength natural language processing in python.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Richard Johansson and Alessandro Moschitti. 2013. Relational features in fine-grained opinion analysis. *Computational Linguistics*, 39(3):473–509.

Evgeny Kim and Roman Klinger. 2018. Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

Ralf Krestel, Sabine Bergler, and René Witte. 2008. Minding the source: Automatic tagging of reported speech in newspaper articles. In *The International Conference on Language Resources and Evaluation*, LREC 2008.

Markus Krug, Frank Puppe, Isabella Reger, Lukas Weimer, Luisa Macharowsky, and Stephan Feldhaus. 2018. *Description of a Corpus of Character References in German Novels – DROC [Deutsches ROman Corpus]. DARIAH-DE Working Papers. Göttingen: DARIAH-DE.*

Ana Marasovic and Anette Frank. 2018. SRL4ORL: improving opinion role labeling using multi-task learning with semantic role labeling. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 583–594. Association for Computational Linguistics.

Grace Muzny, Angel X. Chang, Michael Fang, and Dan Jurafsky. 2017. A two-stage sieve approach for quote attribution. In *The 15th Conference of the European Chapter of the Association for Computational Linguistics*, EACL 2017, pages 460–470.

Sean Papay and Sebastian Padó. 2020. RiQuA: A corpus of rich quotation annotation for English literary text. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 835–841, Marseille, France. European Language Resources Association.

Silvia Pareti. 2015. *Attribution: a computational approach*. Ph.D. thesis, University of Edinburgh, UK.

Silvia Pareti, Timothy O'Keefe, Ioannis Konstas, James R. Curran, and Irena Koprinska. 2013. Automatically detecting and attributing indirect quotations. In *The 2013 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2013, pages 989–999.

Bruno Pouliquen, Ralf Steinberger, and Clive Best. 2007. Automatic detection of quotations in multilingual news. In *The International Conference on Recent Advances in Natural Language Processing*, RANLP 2007, pages 487–492.

Josef Ruppenhofer, Caroline Sporleder, and Fabian Shirokov. 2010. Speaker attribution in cabinet protocols. In *The Seventh conference on International Language Resources and Evaluation*, LREC 2010.

Christian Scheible, Roman Klinger, and Sebastian Padó. 2016. Model architectures for quotation detection. In *The 54th Annual Meeting of the Association for Computational Linguistics*, ACL 2016.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Michael Wiegand and Dietrich Klakow. 2012. Generalization methods for in-domain and cross-domain opinion holder extraction. In *The 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL 2012, pages 325–335.