Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 803–812
Marseille, 11–16 May 2020

Corpus REDEWIEDERGABE

Annelen Brunner[1], Stefan Engelberg[1], Fotis Jannidis[2], Ngoc Duyen Tanja Tu[1], Lukas Weimer[2]
[1]Leibniz-Institut für Deutsche Sprache Mannheim; [2]Julius-Maximilians-Universität Würzburg
[1]R5 6-13, 68161 Mannheim, Germany; [2]Am Hubland, 97074 Würzburg, Germany
{brunner, engelberg, tu}@ids-mannheim.de; {fotis.jannidis, lukas.weimer}@uni-wuerzburg.de

Abstract
This article presents the corpus REDEWIEDERGABE, a German-language historical corpus with detailed annotations for speech, thought and writing representation (ST&WR). With approximately 490,000 tokens, it is the largest resource of its kind. It can be used to answer literary and linguistic research questions and serve as training material for machine learning. This paper describes the composition of the corpus and the annotation structure, discusses some methodological decisions and gives basic statistics about the forms of ST&WR found in this corpus.

Keywords: corpus, annotation, speech thought writing representation, machine learning

1. Introduction and Motivation

The corpus REDEWIEDERGABE is a historical corpus of fictional and non-fictional German texts from 1840 to 1919 annotated with various forms of speech, thought and writing representation (ST&WR). It was created by the project 'Redewiedergabe' (www.redewiedergabe.de, funded by the Deutsche Forschungsgemeinschaft). The corpus can be used for quantitative literary and linguistic studies (e.g. about the development of ST&WR forms, differences between fiction and non-fiction) and also serve as training material for the development of automatic recognizers for ST&WR. The use of the corpus as training material for machine learning was a main focus of the project 'Redewiedergabe', which influenced some aspects of its design and structure. This paper describes the composition of the corpus and the annotation structure, discusses some methodological decisions and gives basic statistics about the forms of ST&WR found in this corpus.

2. Related Work

ST&WR has been extensively researched in German linguistics and literary studies. In linguistics, different types of representation are described structurally (e.g. Fabricius-Hansen, 2002; Weinrich, 2007) and specific linguistic features have been studied, especially the use of subjunctive verb mode in the context of representation (e.g. Zifonun, Hoffmann and Strecker, 2011; Fabricius-Hansen, Solfjeld and Pitz, 2018) but also other aspects like verb selection in framing phrases (Hauser, 2008; Tu, Engelberg and Weimer, 2020). In literary studies, different modes of representation for a fictional character's voice are a central part of many narrative theories (e.g. Stanzel, 2008; Genette, 2010; Leech and Short, 2013; Martinez and Scheffel, 2016). Some studies focus on a particular aspect such as the representation of thoughts (e.g. Cohn, 1978; Palmer, 2004) or free indirect ST&WR in particular (e.g. Banfield, 1982; Fludernik, 1993) which received much attention in literary studies.
Our corpus strives to capture the phenomenon ST&WR in a structured way by systematic annotation. A comparable effort is the corpus by Semino and Short (2004) who annotated a corpus of modern English texts according to the ST&WR schema defined by Leech and Short (1981). A direct predecessor of our corpus is Brunner (2015), a corpus of 13 German narratives from the 18th and early 19th century (approx. 57,000 tokens). Our corpus uses an annotation schema very similar to Brunner (2015), but is considerably larger and more diverse, contains non-fictional material and implements a more complex annotation process yielding more reliable results.
More corpora annotated with ST&WR can be found, but they only deal with one type of representation and mostly with representation in other languages, e.g. corpora with direct speech: Krug et al. (2018b) (German novels); Elson and McKeown (2010) (English literature); Haan-Vis and Spooren (2016) (Dutch newspapers); Lee and Yeung (2016) (English biblical texts); Weiser and Watrin (2012) (French newspapers); corpora with non-direct speech: Krestel, Bergler and Witte (2008) (English newspapers).

3. Composition of the Corpus

3.1 Premises

The following aims guided the composition of our corpus: 1) diversity. The corpus strives for a general understanding of ST&WR and its textual material should be as diverse as possible. Therefore we opted to use shorter excerpts from multiple texts rather than longer, complete texts and also tried to represent many different authors, newspapers and magazines. 2) balanced representation of fictional and non-fictional material 3) balanced representation of each decade in our time period (1840-1919) to allow diachronic studies.

3.2 Sources

The texts come from three sources: The 'Digitale Bibliothek' collected by the project TextGrid[1], the 'Mannheimer Korpus Historischer Zeitschriften und Zeitungen' (MKHZ)[2] and the journal 'Die Grenzboten'[3].
The Digitale Bibliothek is a collection of German nonfiction and narrative texts that has been converted to XML/TEI and made publicly available in the TextGrid Repository. For our corpus only short and medium-length

[1] https://textgrid.de/digitale-bibliothek.
[2] https://repos.ids-mannheim.de/mkhz-beschreibung.html.
[3] http://www.deutschestextarchiv.de/doku/textquellen#grenzboten.

narrative texts were selected, resulting in 258 narrative texts from 79 different authors being used for the corpus.
The MKHZ is a collection of 26 German newspapers and magazines from the 18th and 19th century. It was digitized by the Leibniz Institute for the German Language and converted by the Deutsches Text Archiv (DTA) into the DTA basic XML format. 19 of these newspapers and journals fall into the time period of our corpus and were integrated.
The journal 'Die Grenzboten' was published regularly from 1841 to 1922 in 81 volumes with diverse content. It was digitized and made available by the Bremen State and University Library and converted into TEI format by the DTA. We used texts from 70 volumes that fit into the time period of our corpus.

3.3 Sampling and Preprocessing

As the corpus should be as diverse as possible, text excerpts were sampled from the selected narrative texts, newspaper or magazine articles. These samples have a minimum size of 500 tokens for texts from the 'Digitale Bibliothek' and 200 tokens for newspaper and magazine texts. The latter limit is lower to allow for complete short articles which are typical for newspapers and magazines. The samples were drawn randomly, but with some restrictions: Firstly, the decades should be represented in a balanced way. Secondly, each author available in a decade should be represented as evenly as possible. Therefore, an author of the 'Digitale Bibliothek' could only be selected again after all other authors available in this decade had already been drawn. An analogous process was established for the newspaper and magazines of MKHZ. Note that only 'Die Grenzboten' was available for all decades and is represented evenly. For MKHZ, the availability of different newspapers/magazines over the time period of the corpus varied: On average we had three to four different sources per decade, but the extremes are only one source (in 1860) and six sources (in 1850).[4] In summary, we prevented overrepresentation of any author, newspaper or magazine to the best of our abilities, considering the available texts and technical challenges. The resulting corpus contains at least 79 clearly distinguishable authors[5] and 20 different newspapers and magazines.
Apart from that, only small changes were made to the texts: The journal 'Die Grenzboten' had been digitized automatically and the samples thus contained some OCR errors, which were corrected manually. In addition, a few frequent obsolete characters were replaced by their modern equivalents. Remaining idiosyncrasies such as old spellings have been left untouched.
We deliberately kept samples with unusual content such as dialect text and newspaper excerpts containing lists or tables. There are also some samples which do not contain any instances of ST&WR. This leads to a realistic representation of the distribution of ST&WR and the diverse textual material available during the time period of our corpus.

3.4 Metadata

Table 1 lists the metadata for each sample. The metadata were assigned partly automatically, partly in single annotation and were checked in several random checks.

metadata	**value(s)**	**description**
year	[Integer]	year of first publication
decade	[Integer]	decade of first publication
source	*digBib*, *grenz*, *mkhz* (*mkhz* has subtypes for periodicals)	text source
filename	[String]	name of the source file the sample was pulled from
title	[String]	title, if available
author	[String]	author, if available
fictional	*yes*, *no*, *unsure*	information on fictionality
narrative	*yes*, *no*, *unsure*	information on narrativity
text_type	*Anzeige* (advert), *Biographie* (biography), *Erzähltext* (narrative), *Kommentar* (commentary), *Nachrichten* (news), *Reisebericht/Brief* (travelogue/letter), *Reportage* (report), *Rezension* (review), *Unsure*	predominant text type
dialect	*yes*, *yes_DS* (dialect in direct speech), *no*	information on dialect
perspective	*first*, *first_plural* ('we'), *third*, *unsure*	predominant perspective
quotes	*german*, *chevron*, *chevron_single*, *ascii*, *dash*, *none*, *other*, *undef*	predominantly used quotation marks

Table 1: Metadata assigned to samples.

The metadata 'fictional' and 'narrative' are based on established definitions from literary theory ('fictionality':

[4] In addition to that, we received some of the MKHZ texts late, which disrupted the sampling process and led to an overrepresentation of one of the five available newspapers in 1900. The exact distributions can be studied in our metadata.

[5] Most likely over 150 different authors are represented in the corpus, but we only have reliable author information for the texts from 'Digitale Bibliothek'.

Gabriel, 2007[6]; 'narrativity': Nünning, 2013[7]). Both refer to the sample specifically and not to the text from which it originates. 13.2% of our fictional samples originate from newspapers and magazines where fiction was part of the feuilleton.

The distinction between text types for newspaper and magazine samples was added to reflect the diversity within these texts. Figure 1 shows the distribution of the text types within the samples drawn from MKHZ and 'Die Grenzboten'. Note the high number of narratives found in these sources.

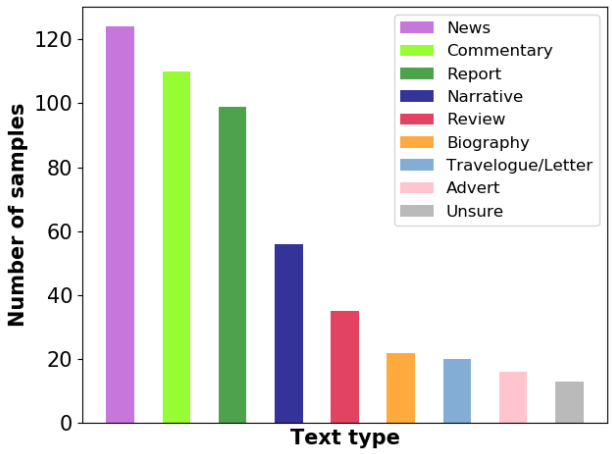

Figure 1: Distribution of text types in newspapers and magazines.

The rich metadata makes it possible to filter the samples in multiple ways and thus answer specific research questions or recognize and, if necessary, eliminate possible problematic factors when using the corpus for machine learning.

4. Annotation

The annotation system used for our corpus is based on Brunner (2015) with some adaptations. It has many similarities to the systems of categories defined by narratologists Genette (2010) and Leech and Short (2013). While based in narratology, it still relies on surface and linguistic indicators for category distinctions as much as possible. We will first outline the system and annotation process and then address some methodological decisions and difficulties in section 4.3. The complete annotation guidelines with many examples are available at Brunner et al. (2019a).

4.1 Annotation System

The annotation system has two main axes: 1) *What* is represented? Here we distinguish between the three media *speech*, *thought* and *writing*. Multiple media are allowed for ambiguous cases. 2) *How* is the content represented? We distinguish between four main types of representation: *direct*, *indirect*, *free indirect* and *reported*. A fifth type, coded as *indirect/free indirect*, was added to account for the category known in German linguistics as 'Berichtete Rede' (e.g. Fabricius-Hansen, Solfjeld and Pitz, 2018). Each annotation can be further specified by optional attributes that mark special or borderline cases. We will describe the types using examples from the corpus. The parts of the text covered by the annotations are marked by underlining.

Direct ST&WR is a literal quote of a character's speech, thought or writing. It can be introduced by a framing clause and is often embraced by quotation marks.

1. *Und Gurow, der heftiges Herzklopfen hatte, dachte: »Mein Gott! Wozu diese Menschen, wozu dieses Orchester...«*
   (And Gurow, whose heart was beating rapidly, thought: "My God! Why these people, why this orchestra...")

Free indirect ST&WR – also known as "erlebte Rede" in German or "style indirect libre" in French – is defined as a blending of the character's and the narrator's voice. It is mainly used for thought representation, typically as a literary device in fictional texts. Indicators are elements of the narrator's voice, such as third person and past tense, in combination with elements of the character's voice, such as questions, exclamations and informal language.

2. *Dreimal hatte sie den Brief gelesen – war es wirklich erst gestern gewesen? – ohne ihn zu verstehen.*
   (She had read the letter three times – had it really just been yesterday? – without understanding it.)

Indirect ST&WR is a paraphrase of a character's speech, thought or writing by the narrator. In our annotation system it is distinguished from its neighboring category, *reported* ST&WR, mainly by its specific form that makes it appear like a straightforward transformation of *direct* ST&WR[8]: It is composed of a framing phrase with a dependent subordinate clause, which often uses subjunctive mode.

3. *Lilli hoffte, er werde dasselbe thun…*
   (Lilli hoped he would do the same…)

Constructions with a framing phrase and a dependent infinitive phrase are also considered indirect ST&WR (cf. ex. 5).

Reported ST&WR is the mention of the act of speaking, thinking or writing. The topic and content may be specified, but the actual wording never is. It is thus on average the most summarizing type of ST&WR and farthest removed from a direct quotation. The most important indicator of *reported* ST&WR is the use of words referring to speech, thought or writing.

4. *Vor der Behandlung der Tagesordnung sprach BG. Franz Witzmann dem Wasserleitungskomittee für die bisher geleisteten großen Arbeiten den Dank aus.*

[6] Gabriel defines 'fictionality' as an invented fact or a combination of such facts into a made-up story (cf. Gabriel 2007, S 594).

[7] Nünning defines 'narrativity' als a temporally organized sequence of actions in which an event leads to a change of situation (cf. Nünning 2013, S. 555).

[8] Note however that *indirect* ST&WR can sometimes be highly summarizing as well (cf. section 4.3).

(Before discussing the agenda, BG. Franz Witzmann thanked the Water Pipeline Committee for the great work done so far.)

A special case are independent sentences in subjunctive mode that are used for representation (called 'Berichtete Rede' in German linguistics). This phenomenon occurred often enough in our corpus that we decided not to subsume it under one of the other categories. It is marked as *indirect/free indirect* representation, as it shows characteristics of both these forms: On the one hand, these sentences have the independence of *free indirect* ST&WR and may also take the form of questions and exclamations, though they are generally used to represent speech rather than thought. On the other hand, they use subjunctive mode which is typical for *indirect* ST&WR. They can appear independently, but often follow directly after *indirect* ST&WR, like in ex. (5), where the preceding sentence contains indirect ST&WR with an infinitive phrase.

5. *Er überwand die kleine Enttäuschung und versprach, den Wunsch der Nichte zu erfüllen. Aber erst später, man brauche Zeit*
   (He overcame the small disappointment and promised to fulfill the niece's wish. But only later, time would be needed.)

In addition to these main ST&WR annotations, we also annotate framing phrases for direct and indirect ST&WR (*frame*) and link them to their corresponding ST&WR annotations. Within those phrases, the key word that indicates the speech, thought or writing act is marked separately (*intExpr*). Finally, we also annotate and link the speaker/source of the speech, thought or writing act (*speaker*) if it is available in the close context of the ST&WR. In ex. (1), this would be marked as follows:

- [frame:] *Und Gurow, der heftiges Herzklopfen hatte, dachte: (And Gurow, whose heart was beating rapidly, thought:)*
- [intExpr:] *dachte (thought)*
- [speaker:] *Gurow*

The attributes listed in table 2 can optionally be assigned to any ST&WR annotation to mark special cases.

4.2 Annotation Process and Tooling

Annotating ST&WR is not a trivial task and we put great effort in consistent and high-quality annotations. For this reason, each sample went through a multi-step process. First, it was independently annotated by two primary annotators. Then a third annotator compared the annotations, adjudicated discrepancies and created a consensus annotation. So, each sample was handled by three persons, which reduced bias and increased consistency. Our annotators were thoroughly trained on the annotation system, received regular feedback and had the opportunity to discuss difficulties in monthly team meetings.[9]

attribute	description
level	level of embedding (one instance of ST&WR containing another), counted in integers
nonfact	non-factual ST&WR, e.g. negations, hypotheticals, plans etc. (e.g. *She wanted to ask him about the restaurant.*)
border	borderline cases where the represented content does not conform to the prototypical definitions of speech, thought or writing (cf. section 4.3)
prag	using the patterns of ST&WR with a different pragmatic intent, e.g. emphasis, politeness, idioms (e.g. *I tell you this is wrong!*)
metaph	metaphorical use of ST&WR (e.g. *His heart told him to go.*)

Table 2: Optional attributes

The annotations were created using the annotation tool ATHEN (Krug et al., 2018a). We developed a custom annotation view for our annotation system to support the annotators with a clear and fast way to assign the complex annotation. ATHEN (including the ST&WR view) is freely available at http://ki.informatik.uni-wuerzburg.de/nappi/release.

During annotation, we continuously calculated Kappa scores to monitor the annotation process. To give an impression of the difficulty of the annotation, table 3 lists the scores for the comparison between the two primary annotators. Seven different persons created these annotations over a period of three years and the annotation guidelines went through some minor adjustments during this period. Note that these primary annotations themselves are *not* part of corpus REDEWIEDERGABE. The corpus only includes the consensus annotation based on these competing annotations.[10] The consensus annotation also went through a final check before corpus release to eliminate inconsistencies.

The numbers in table 3 are values of Fleiss' Kappa calculated over 834 samples[11] on token basis, i.e. the annotation of each single token was compared. In case of overlapping annotations, partial matches were scored, e.g. if annotator 1 assigned the annotation *[direct]* and annotator 2 assigned *[direct, indirect]* (meaning *indirect* ST&WR embedded into *direct* ST&WR), the score for this token would be 0.5. Table 3 shows the comparison of only the type assignments as well as the type and medium assignments. The values of the optional attributes were not considered.

[9] We take the opportunity here to thank our diligent student annotators: Sarah Gorke, Anna Hartmann, Janne Lorenzen, Christoph Peterek, Laura Schäfer, Lisa Sergel and Theresa Valta.

[10] However, we are planning to release the primary annotations in their original form as part of our additional material.

[11] Four samples are excluded from this evaluation, because one of their primary annotations was from a very early training phase and has uncharacteristically low quality.

annotations	Fleiss' Kappa	
	type	type & medium
all types	0,73	0.72
only *direct*	0,92	0.89
only *indirect*	0,73	0.68
only *reported*	0,49	0.47

Table 3: Fleiss' kappa scores (token-based) over 834 samples for the primary annotations. We do not provide scores for 'only *free indirect*' and for 'only *indirect/free indirect*' because these types are so infrequent in our corpus that the Kappa values are not representative.

The scores were quite different between the ST&WR types, with *reported* being the most problematic type. The reasons for that will become clear in section 4.3. We also observed that the scores vary strongly between individual samples and that non-fictional samples scored on average lower than fictional samples. The latter is partly because of the higher percentage of direct – the 'easiest' type – in fictional texts, but even when looking at the three ST&WR types separately, the non-fictional samples scored lower on average for each type, i.e. posed more uncertainty for annotation. One reason for this is that the newspapers and magazines that were used for our corpus tend to contain quite complex texts (political commentary and reports in historical German). However, we also observed some systematic difficulties to apply our annotation system that is rooted in narrative theory to journalistic writing, e.g. citations integrated into the sentence structure in book reviews, highly summarizing forms of ST&WR and underspecified information about the medium.

4.3 Methodological Decisions and Challenges

When a complex literary or linguistic phenomenon shall be captured in annotations, one faces a multitude of difficulties. On the one hand, annotation guidelines have to be as clear and succinct as possible to ensure a fast and reliable annotation process; on the other hand, consideration must be given to literary and linguistic relevance and correctness in order to mark phenomena in such a way that they can later be distinguished and used in a meaningful way (cf. Hovy and Lavid, 2010; Ide and Pustejovsky, 2017; Gius and Jacke, 2017). In our project, this is complemented by a third consideration: We plan to use the corpus as material to train automatic recognizers for ST&WR. Our annotation system is thus the result of compromises and concessions. In this section, we will address some of the major challenges we faced when annotating ST&WR and describe how we handled them.

The categories used in our system, especially direct, indirect and free indirect (or 'erlebte Rede') are established distinctions in works dealing with ST&WR (cf. McHale, 2011). As mentioned above, our annotation system shows similarities to the system defined in the influential narratological theory of Genette (2010), and also to that defined by Leech and Short (2013), both fairly formal systems that incorporate linguistic features in their definitions. Thus, they were particularly suited to be adapted for annotation guidelines and also well suited to our other task of developing automatic ST&WR recognizers. For both goals, it is very helpful to have surface indicators to distinguish between categories and to have structural similarities reflected in similar categories. However, the decision for such a structured system has the consequence that there are aspects of narratological theory which are not clearly reflected in our annotation. In particular, we decided to handle thought representation parallel to speech representation, treating thought essentially as 'silent speech'. It is debatable whether this adequately reflects the reality of 'mind representation' (cf. Cohn, 1978; Fludernik, 1993; Palmer, 2004; McHale, 2011). One obvious consequence is that the well-known literary categories 'interior monologue' and 'stream of consciousness' are not present in our annotation. One can argue that *direct thought* is very close to what is defined as 'quoted interior monologue' by Cohn (1978: 15), however as McHale points out "stream of consciousness is best thought of not as a form but as a particular *content* of consciousness" (McHale, 2014: sec. 8). It would be orthogonal to our categories and is thus not included. In addition, many aspects of mind representation that are more removed from the idea of thought as speech and were pointed out by literary scholars (e.g. Palmer, 2004) are excluded in our annotation. Trying to incorporate these aspects would have added much additional complexity and also expanded the number of phenomena that had to be marked considerably. Though we cannot cover all nuances of literary analysis, we believe that we still provide an annotation that has internal consistency and a strong basis in literary theory as well as in linguistics so that it can be useful for studies in both fields.

While the basic structure of the annotation system is the result of a theoretical decision we made in advance, another difficulty only became clear when working with actual corpus data: It can be surprisingly hard to distinguish the representation of a speech, thought or writing act from 'pure' narration. We advised our annotators to always keep in mind the prototypical case of ST&WR, which we defined as the representation of a speech, thought or writing act performed by a character A by a character B or the narrator. Respecting this definition is particularly difficult for *reported* representation, which can be so close to pure narration that the boundary becomes blurred. For this reason, we added stricter criteria in our annotation guidelines: *Reported* representation must either contain explicit lexical reference to an act of speaking, thinking or writing or clearly communicate the content of such an act; ideally both.

We also gave definitions of what constitutes a prototypical act of speech, thought or writing. Thought proved especially difficult: As thoughts – other than speech and writing – do not manifest themselves in the (narrated) world, it is hard to decide what constitutes a thought at all. Our prototypical definition of thought is intentionally narrow: "a conscious, analytical, cognitive process; 'silent

speech'".[12] This made it possible to work with a scale similar to that of speech representation and excluded the narration of emotions and moods to a large extent. While *direct* thoughts are often marked by quotation marks and *indirect* thoughts are syntactically easy to identify, *reported thought* proved the most difficult form of representation. The following example may illustrate this:

6. *So ganz auf sich selbst gestellt, aufs höchste überzeugt von der Ueberlegenheit seines Geistes und der unwiderstehlichen Macht seines Willens, ohne eine Partei im Lande für sich zu haben, ja auch ohne die Nothwendigkeit einer solchen zu begreifen, stand Struensee, der Fremde, der Arzt, am Ruder des dänischen Staates.*
   (So completely on his own, convinced of the superiority of his spirit and the irresistible power of his will, without having a party in the country for himself, and without even understanding the necessity of such a party, Struensee, the foreigner, the doctor, stood at the helm of the Danish state.)

In this example a decision had to be made whether *überzeugt sein* (*to be convinced*) and *begreifen* (to *understand*) should be counted as thoughts. We opted to annotate these instances, but they are both marked with attributes, particularly the first one is considered a borderline case.[13] As such cases are quite common, the guidelines contain numerous examples for special and borderline cases as well as a list of problematic verbs and whether they usually indicate ST&WR or not. While such a list proved very helpful to increase the consistency of the annotation, annotators were advised to always consider context and textual meaning, which can override the recommendation. The decision to annotate or not cannot solely depend on an isolated verb. This is especially true for indirect and direct ST&WR where very unusual verbs can be used to introduce the representation (in the following examples, not the ST&WR passage, but rather other relevant features are underlined).

7. *Der Literat drohte ihm mit dem Finger: "So – so – gekauft?… Ei, Sie stiller Sünder!…"*
   (The writer threatened him with his finger: "So – so – bought?... Oy, you silent sinner!...")

In cases like these, content and structure drive the decision to annotate rather than the lexical material.

The attribute *border* can be used to mark cases that deviate from the prototypical definitions given in the annotation guidelines, but are still close enough to ST&WR to be marked. Consider these three examples from the corpus:

8. *Wir hören eben, daß der Stadtrath selbst, der bekanntlich wenig wühlerischer Natur ist, dennoch eben eine Adresse wegen Entfernung des 27. Regiments aus Köln beräth.*
   (We just hear that the city council itself, which is not known for its volatile nature, is nevertheless discussing a petition for the removal of the 27th regiment from Cologne.)
9. *Neulich las ich: sie konnten ihre Erbitterung nur unschwer unterdrücken,...*
   (The other day I read: they could hardly suppress their bitterness...)
10. *Es waren, Gott sei Preis und Dank, die Vorsichtigen und Sparsamen, die sich die Sache berechnet und anderswo für noch weniger Geld gesättigt hatten,...*
    (It were, thank God, the cautious and thrifty people who had calculated the matter and sated themselves elsewhere for even less money…)

In each example there must have been an act of speech (ex. 8), thought (ex. 10) or writing (ex. 9), but it is not addressed directly. Nevertheless, these examples are considered representations and given the attribute *border*.

Apart from the decision whether ST&WR is present at all, another difficulty we discovered during annotation concerns the distinction between indirect and reported ST&WR. In the narratological categorical systems that inspired our annotation system, the categories are arranged on a scale according to their effect in narration. Though Genette (2010) and Leech and Short (2013) differ in what their scale represents (for Genette it is dramatic vs. narrative mode, for Leech and Short it is a claim of 'faithfulness'), they both rank indirect closer to direct representation than reported. We generally follow this idea: reported is more summarizing and less precise, while indirect ST&WR can usually be read as a transformation of direct ST&WR that allows us to reconstruct the 'original' quote in more detail. However, there are sentences that follow the typical structure of indirect ST&WR – a framing clause and a dependent subordinate clause containing the content - but do not allow such a reconstruction:

11. *Was Ihr auf diesem Wege über die Beziehungen der Gräfin zu vornehmen Venezianern erfahrt, berichtet Ihr an diesem Ort.*
    (What you learn in this way about the Venetian countess's relations with the noble Venetians, you report at this place.)

In this example, the dependent clause does not specify the content of the report at all, but rather its broad theme. After much consideration, we decided to give the structural indicators precedence and stick to the category *indirect* even for extreme examples such as (11). This was because we found that once we break up the structural boundary between *indirect* and *reported* and instead take the level of detail or the 'closeness' to a hypothetical quotation as our deciding criterion, the lines become very blurred. In literature (and, in fact, in language in general) form does not necessarily force function (cf. Sternberg 1982: 112; also consider the approach of Schmid (2005), who describes ST&WR as an interference between narrator and character text where characteristics of the character's voice may seep through in even very 'distant' forms of representation).[14] However, sticking to formal criteria as much as possible is very beneficial when faced with the

[12] The other definitions are: speech: a verbal, coherent utterance with the intention of communication; writing: the process of writing or a written text with the intention of communication.

[13] *überzeugt von...* refers not to a thought process but a mind state, and is marked as "border:state", *ja ohne… zu begreifen* is negated and thus marked as "nonfact".

[14] From a linguistic perspective, the embedded clause in (11) is a free relative clause. It is assumed that such a clause can only

challenge of producing a consistently annotated corpus and also facilitates the task for a machine learner trained on the corpus data. We also believe that the *indirect* form is still a meaningful category due to its very specific way of presenting speech, thought and writing that sets it apart from *reported.* A more detailed study of its nuances may be one of the applications of our corpus.

A third unexpected difficulty concerns defining annotation boundaries. While these are mostly intuitive for *direct, free indirect* and *indirect* ST&WR, *reported,* being so closely integrated into the surrounding narrative, needed more formal rules: The framing word (referring to a speech, thought or writing act) had to be included into the annotation, as well as the content of the representation if it is specified. The *speaker*, on the other hand, was only part of the annotation if it was not too far from the rest of the material. A similar rule was implemented for *frame,* the framing clause for *(in)direct* ST&WR. In both cases this rather arbitrary rule was necessary as a lot of textual material (e.g. relative clauses, attributive modifiers, subclauses) can occur between the speaker and the rest of the *reported* ST&WR or the *frame*. We wanted to avoid bloating our annotations with material that does not relate to ST&WR and would be a distraction for machine learning as well as for most other types of studies. The speaker in these cases was still annotated and linked to the corresponding annotation, if it could be found in its close vicinity.

We could only address some of the most consequential and maybe controversial decisions here. More examples for difficulties and borderline cases and how they were dealt with can be found in the annotation guidelines and also in Tu, Engelberg and Weimer (2020).

5. Corpus Statistics

The corpus contains 838 samples with a total of 489,459 tokens. As described above, it is both balanced with regard to fictional and non-fictional material as well as material per decade (cf. table 4).

Figure 2 shows the token percentages for the ST&WR types. *Direct* and *free indirect* ST&WR are clearly more common in fictional texts. *Free indirect* even occurs almost exclusively there, but is very infrequent in general, due to the historical nature of our corpus. *Indirect* and *reported* ST&WR on the other hand are more frequent in non-fictional texts, though the difference is not as pronounced.

decade	tokens fictional	tokens non- fictional	total
1840	30,728	30,233	60,961
1850	30,258	30,426	60,684
1860	31,058	31,420	62,478
1870	30,436	30,568	61,004
1880	30,251	30,678	60,929
1890	30,963	30,273	61,236
1900	30,567	30,272	60,839
1910	30,430	30,898	61,328
total	244,691	244,768	489,459

Table 4: Corpus size

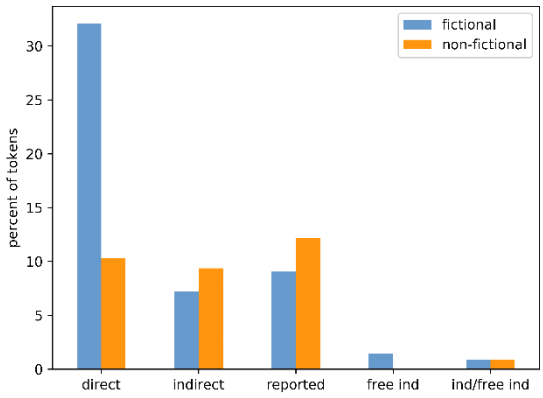

Figure 2: Proportion of all types of ST&WR in fictional vs. non-fictional texts

Figure 3 shows the percentages of the three main media and the most frequent ambiguous case, where it is unclear whether the represented content is speech or writing.[15] Speech representation is dominant in fictional texts and writing in non-fictional ones. The latter is due to book reviews and to written communication often being a topic in news stories. The high percentage of speech/writing also indicates that the medium tends to be underspecified and probably considered less important than the represented content in non-fiction.

induce indirect ST&WR when embedded under a verb of communication (e.g. *She asked what he did.*) (cf. Fabricius-Hansen, Solfjeld and Pitz, 2018). However, when dealing with corpus data we found that relying on verb semantics led to borderline cases as well (e.g. in the case of negated communicative verbs) and was not a satisfying solution.

[15] Other ambiguous media annotations were much more infrequent and are therefore not shown in figure 3.

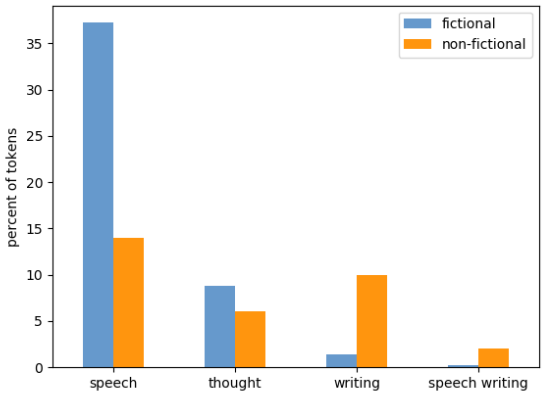

Figure 3: Proportion of the most frequent media of the represented content in fictional vs. non-fictional texts

Table 5 shows the distribution of ST&WR types with respect to instances. An instance is an unbroken stretch of annotation that may vary greatly in length between one token and several sentences. The average length varies between ST&WR types and between fictional vs. non-fictional texts. While *free indirect* and *indirect/free indirect* are too infrequent to draw robust conclusions, it is interesting to note that for the three remaining types the instances in non-fictional texts are longer on average.

	number of instances	**average length of instances**	**number of instances**	**average length of instances**
	fictional		**non-fictional**	
direct	3527	22.2	639	39.5
indirect	1424	12.4	1245	18.4
free ind	132	26.4	4	14.0
indirect/ free ind	65	32.8	66	31.7
reported	2778	8.0	2653	11.2

Table 5: Number and average token length of instances in fictional vs. non-fictional samples

6. Applications

With its rich annotation and metadata, the corpus allows for many quantitative evaluations and offers rich opportunities for linguistic and literary studies. It is also a rich resource for machine learning.

In the context of our project, the corpus was already used for linguistic studies on the lexical variance within framing clauses (Tu, Engelberg and Weimer 2020). In addition to that, the annotated material served as training material for automatic recognizers for ST&WR (cf. Brunner et al. 2019b), which were then used in a study comparing the use of STW&R in high and low brow literature (cf. Brunner et al., 2020). These automatic recognizers will be released on our Github page[16] in spring 2020.

7. Download

The corpus is available for download on https://github.com/redewiedergabe/corpus in three different formats: a column-based text format, a TEI compliant XML format, and an XMI format based on the UIMA framework (http://uima.apache.org). Full descriptions of these formats can be found on the Github page.

The column-based text format consists of UTF-8 encoded files with the extension .tsv (tab-separated value). Each of these files contains a sample in column format. Each row represents a token of the sample. In addition to several columns encoding the manual annotation, these files also contain automatically generated annotation, such as sentence boundaries, orthographic normalization, lemmatization, and part of speech tags, generated with the rfTagger (Schmid, 2008) and the CAB tool (Jurish, 2012). The metadata for all samples is listed in a separate .tsv table.

The XML version of the corpus consists of TEI compliant XML files. We provide a RELAX-NG syntax schema that adapts the TEI Module for Linguistic corpora to the annotation schema. Each file represents a sample with the manual annotations and contains the full metadata in an <fs> tag. We use the following XML tags to code the annotation: <said> (ST&WR annotation) and <seg> (annotation of *frame*, *speaker*, *intExpr*). Attributes are used to encode the specifics of the annotations and link them.

The XMI format is compatible with the ATHEN annotation tool and its ST&WR view. It contains the same additional automatically generated annotation as the column-based format.

In addition to the main corpus, we release additional annotated material. The annotation follows the same guidelines as for the main corpus, but is less reliable, as these texts were not processed by three people like the main corpus, but annotated by just one person. At the moment the additional material includes: a corpus of 256 fictional and non-fictional samples with 149,000 tokens, a corpus of 17 complete narratives (about 200,000 tokens) and a corpus of 12 complete newspaper and magazine articles (about 60,000 tokens). More additional material will be added in the future, such as a corpus containing only simplified annotation for *indirect* ST&WR (about 50,000 tokens) and a corpus of the primary annotations for the main corpus.

8. Licensing

[16] https://github.com/redewiedergabe

9. Bibliographical References

Banfield, A. (1982). Unspeakable Sentences. Narration and Representation in the Language of Fiction. Boston, MA: Routledge.

Brunner, A. (2015). Automatische Erkennung von Redewiedergabe. Ein Beitrag zur quantitativen Narratologie. Berlin, Germany: De Gruyter.

Brunner, A., Weimer, L., Engelberg, S., Jannidis, F. and Tu, N.D.T. (2019a). Annotationsrichtlinien des Projekts "Redewiedergabe. Eine literatur- und sprachwissenschaftliche Korpusanalyse", Zenodo, URL: http://doi.org/10.5281/zenodo.2634994.

Brunner, A., Tu, N.D.T., Weimer, L. and Jannidis, F. (2019b). Deep learning for Free Indirect Representation, *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, Erlangen, Germany, pp. 241-245.

Brunner, A., Jannidis, F., Tu, N.D.T. and Weimer, L. (2020). Redewiedergabe in Heftromanen und Hochliteratur, *DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts.* Paderborn, Germany, Zenodo, URL: http://doi.org/10.5281/zenodo.3666689, pp. 190-194.

Cohn, D. (1978). Transparent Minds. Narrative Modes for Presenting Consciousness in Fiction. Princeton, NJ: Princeton University Press.

Eisenberg, P. (2013). Grundriss der deutschen Grammatik, Bd. 2: Der Satz. Stuttgart and Weimar, Germany: J.B. Metzler, 4th edition.

Elson, D.K. and McKeown, K.R. (2010). Automatic Attribution of Quoted Speech in Literary Narrative, *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10)*, Atlanta, GA, pp. 1013-1019.

Engelberg, S. (2015). Quantitative Verteilungen im Wortschatz. Zu lexikologischen und lexikografischen Aspekten eines dynamischen Lexikons. In L.M. Eichinger (ed.), *Sprachwissenschaften im Fokus. Positionsbestimmungen und Perspektiven. Jahrbuch 2014 des IDS.* Tübingen, Germany: Narr, pp. 205-230.

Fabricius-Hansen, C. (2002). Nicht-direktes Referat im Deutschen – Typologie und Abgrenzungsprobleme. In C. Fabricius-Hansen, O. Leirbukt & O. Letnes (eds.), *Modus, Modalverben, Modalpartikeln*. Trier, Germany: Wissenschaftlicher Verlag, pp. 6-29.

Fabricius-Hansen, C., Solfjeld, K., and Pitz, A. (2018): Der Konjunktiv. Formen und Spielräume. Tübingen, Germany: Stauffenburg.

Fludernik, M. (1993). The Fictions of Language and the Language of Fiction. The Linguistic Representation of Speech and Consciousness. London and New York: Routledge.

Gabriel, G. (2007). Fiktion. In H. Fricke, K. Grubmüller & J.-D. Müller (eds.), *Reallexikon der deutschen Literaturwissenschaft*, Bd. 1: A-G. Berlin and New York: De Gruyter, 3rd edition, pp. 594-598.

Genette, G. (2010). Die Erzählung. Paderborn, Germany: Fink. 3rd edition.

Gius, E. and Jacke, J. (2017). The Hermeneutic Profit of Annotation: On Preventing and Fostering Disagreement in Literary Analysis. *International Journal of Humanities and Arts Computing 11(2)*: 233-254.

Haan-Vis, K. and Spooren, W. (2016). Informalization in Dutch journalistic subgenres over time. In N. Stukker, W. Spooren & G. Steen (eds.), *Genre in Language, Discourse and Cognition*. Berlin, Germany: De Gruyter, pp. 137-163.

Hauser, S. (2008). Beobachtungen zur Redewiedergabe in der Tagespresse. Eine kontrastive Analyse. In H.-H. Lüger & H.E.H. Lenk (eds.), *Kontrastive Medienlinguistik*. Landau, Germany: Verlag Empirische Pädagogik, pp. 271-286.

Hovy, E. and Lavid, J. (2010). Towards a 'Science' of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. *International Journal of Translation 22(1)*: 13-36.

Ide, N. and Pustejovsky, J. (eds.) (2017). Handbook of Linguistic Annotation. Dordrecht, Netherlands: Springer.

Jurish, B. (2012). Finite-state Canonicalization Techniques for Historical German. PhD thesis, University of Potsdam, Germany, URN: urn:nbn:de:kobv:517-opus-55789.

Krestel, R., Bergler, S. and Witte, René (2008). Minding the Source: Automatic Tagging of Reported Speech in Newspaper Articles, *Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC)*, Marrakech, Morocco, pp. 2823-2828.

Krug, M., Tu, N.D.T., Weimer, L., Reger, I., Konle, L., Jannidis, F. and Puppe, F. (2018a). Annotation and beyond – Using ATHEN, *Digital Humanities im deutschsprachigen Raum – Konferenzabstracts*, Cologne, Germany, pp. 19-21.

Krug, M., Weimer, L., Reger, I., Macharowsky, L., Feldhaus, S., Puppe, F. and Jannidis, F. (2018b). Description of a Corpus of Character References in German Novels - DROC [Deutsches ROman Corpus]. *DARIAH-DE Working Papers*, 27, Göttingen, Germany: DARIAH-DE, URN: urn:nbn:de:gbv:7-dariah-2018-2-9.

Lee, J. and Yeung, C.Y. (2016). An Annotated Corpus of Direct Speech, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia, pp. 1059-1063.

Leech, G. and Short, M. (1981). Style in fiction. A linguistic introduction to English fictional prose. London, UK: Longman.

Leech, G. and Short, M. (2013). Style in fiction. A linguistic introduction to English fictional prose. London and New York: Routledge, 2nd edition.

Martínez, M. and Scheffel, M. (2016). Einführung in die Erzähltheorie. Munich, Germany: C.H. Beck.

McHale, B. (2011). Speech Representation. In P. Hühn, J. Pier, W. Schmid & J. Schönert (eds.), *The Living Handbook of Narratology*. Hamburg, Germany: Hamburg University Press, URL: http://www.lhn.uni-hamburg.de/article/speech-representation.

Nünning, A. (2013). Narrativität. In A. Nünning (ed.), *Metzler Lexikon Literatur- und Kulturtheorie*. Stuttgart and Weimar., Germany: J.B. Metzler, pp. 555-556.

Palmer, A. (2004). Fictional minds. Lincoln, NE: University of Nebraska Press.

Schmid, H. & Laws, F. (2008). Estimation of Conditional Probabilities With Decision Trees and an Application to Fine-Grained POS Tagging, *The 22nd International Conference on Computational Linguistics (COLING)*, Manchester, UK, pp. 777-784.

Semino, E. and Short, M. (2004). Corpus stylistics. Speech, writing and thought presentation in a corpus of English writing. London and New York: Routledge.

Stanzel, F.K. (2008). Theorie des Erzählens. Göttingen, Germany: Vandenhoeck&Ruprecht.

Sternberg, M. (1982). Proteus in Quotation-Land: Mimesis and the Forms of Reported Discourse. *Poetics Today, 3(2)*, pp. 107-156.

Tu, N.D.T., Engelberg, S. and Weimer, L. (2020). Was für Enthüllungen! heulte die wohlgekleidete respektable Menge. – Eine korpus-linguistische Untersuchung zur lexikalischen Vielfalt von Redeeinleitern. In S. Engelberg, C. Fortmann and I. Rapp (eds.): Redewiedergabe. *Linguistische Berichte – Sonderhefte 27*: 13-53.

Weinrich, H. (2007). Textgrammatik der deutschen Sprache. Darmstadt, Germany: Wissenschaftliche Buchgesellschaft, 4th edition

Weiser, S. and Watrin, P. (2012). Extraction of Unmarked Quotations in Newspapers. A Study Based on Direct Speech Extraction Systems, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey, pp. 559-562.

Zifonun, G., Hoffmann, L. and Strecker, B. (2011): Grammatik der deutschen Sprache. Berlin, Germany: De Gruyter, reprint.