Corpus

Parzival

Parzival, written between 1200 and 1210 by Wolfram von Eschenbach, is an Arthurian grail novel in Middle High German. *Parzival*, by transmission regarded as the most popular courtly novel of the Middle Ages, is based on *Conte du Graal* (1180-90) by Chrétien de Troyes, the never finished French reference. In *Parzival*, Wolfram thematically builds on the "classic" Artus novel in the tradition of Hartmann von Aue (*Erec*, *Iwein*), but at the same time breaks away from it by introducing the grail theme as an additional religious dimension.

The text comprises approximatively 25,000 pairwise rhymed lines and is divided into 16 books (in modern editions). It is organised into two main plots: the Parzival and the Gawan storyline. Our annotations currently are limited to the first Parzival part (books 3-6): It begins after the parents' prologue (books 1 and 2) and ends just before the first Gawan part (books 7 and 8). As such, it contains a complete plot unit.

Adorno

The *Aesthetic Theory* by Adorno is one of the most impactful philosophical aesthetics of the 20th century. It is regarded as the last great aesthetics of the modern era. On the basis of peculiarities of modern artworks, Adorno investigates the status and nature of art, its relation to society, and its utopic core. Because of his sudden early death, Adorno did not finish the *Aesthetic Theory* himself. The text only exists as a reading edition, initially published in 1970 by his wife Gretel and Ralf Tiedemann.

Currently, annotations are limited to the chapter *On theory of artwork* which is one of the central parts of the book as regarded by Adorno.

Werther

The *Sorrows of Young Werther*, written 1774 by Johann Wolfgang von Goethe is a sophisticatedly structured epistolary novel. This literary form was established during the middle of the 18th centruy in Germany, following the English and French models of *Pamela* (1740) by Samuel Richardson and *Julie ou La Nouvelle Heloise* (1761) by Jean-Jaques Rosseau. 1787 Goethe wrote a revised edition of *Werther*. The modifications comprised a stilistic revaluation of the text, a deletion of excessive sentiment narratives by the protagonist, and the addition of individual letters (swains episode). Our annotations are included in the introductory words of the fictional publisher and the first few letters of Werther directed to his friend Wilhelm, whose responses are contained only fragmentarily

and implicitly in these letters. Here, Werther describes a ball event where he gets to know Lotte being engaged.

Plenary Debates

The corpus consists of 1.226 protocols of the plenary debates of the German Bundestag from 1996 to 2015. Every session of plenary debate is publicly accessible online as so-called stenographic report. In CRETA, we use a corpus from the PolMine[1]-research project, headed by Andreas Blätte, in which all individual reports were collected, subjected to various preprocessing steps and integrated into an overall corpus. So far, our annotations are confined to extracts of four plenary minutes. For each of these protocols we annotated one entire speech of a given politician. All statements deal with issues connected to the European Union (accession negotiations with Turkey, European social policy and debates on a European Constitution).

Entity annotations

Our corpus contains annotated *typed entity references*. We annotate text strings as entity references when they refer to clearly distinguishable individual instances in the real or fictional world. This makes entity references highly related to *mentions* from coreference resolution, although there are some conceptual differences. We further annotated the type of the referred entity as part of the entity reference. In determining the type of the referred entity, we handle fictional worlds in the same way as the real world. In the absence of other indications, we assume that fictional worlds follow the same rules as the real world.

The types we distinguish are

- Person: Reference to humans or humanly acting animals or things
- Organisation: Reference to an organisation
- Location: Reference to a place or location
- Work: Reference to a cultural artefact (publication, law)
- Event: Reference to an event
- Concept: Reference to domain-specific theoretical concept

Please find more detailed information as well as examples and discussion of special cases in the annotation guidelines (only in German, separate PDF file).

Named entities are a subset of the potentially referring expressions.

[1] http://polmine.sowi.uni-due.de/polmine/

Corpus size

Subkorpus	Tokens	Entities	Version
Werther (1787)	41.505	331	1
Adorno	13.233	929	1
Parzival	30.491	2.001	1
Bundestagsdebatten	6.371	488	1

For legal reasons, the data for Adorno can only be made available after the requesters or participants have given evidence that they have a license for that literary work.

Data formats

CoNLL (TSV)

A token per line. For sentence boundaries an empty line. Annotations are tab-separated (TSV data format)[2]. Multiple annotations are put into different columns. `B-PER` denotes the first token of a person annotation, `I-PER` denotes all following tokens of that person annotation. `O` (the letter) denotes 'no annotations'.

Note that the training data for Werther with the ID `3_34_12` contains only part of the original text. Due to a technical problem earlier versions alos contained unchecked annotations.

Example

```
Die B-PER
geringen I-PER
Leute I-PER
des I-PER B-LOC
Orts I-PER I-LOC
kennen O
mich O
schon O
, O
und O
lieben O
mich O
, O
```

[2] https://de.wikipedia.org/wiki/CSV_(Dateiformat)

```
besonders O
die B-PER
Kinder I-PER
. O
```

Apache UIMA XMI

XML-based, for processing with Apache UIMA[3]. The relevant types[4] are subtypes of `de.unistuttgart.ims.creta.api.Entity`. The annotation category is found in the subtype denotation (e.g. `de.unistuttgart.ims.creta.api.EntityPER`) or in the value of the attribute `category`.

Markdown

The pandoc[5] Markdown is used, above all for manually reading the annotations. Annotations are marked by square brackets, followed by the category in subscript. This format should not be used for automatic processing.

Example

```
[Die geringen Leute [des Orts ]~LOC~]~PER~kennen mich schon, und
lieben mich, besonders [die Kinder]~PER~.
```

[3] https://uima.apache.org/

[4] The type system can be downloaded from https://www.creta.uni-stuttgart.de/wp-content/uploads/2016/09/typesystem.xml.

[5] http://pandoc.org/MANUAL.html#pandocs-markdown