Korpus

Parzival

Der *Parzival* Wolframs von Eschenbach ist ein arthurischer Gralroman in mittelhochdeutscher Sprache, entstanden zwischen 1200 und 1210. Der *Parzival*, der seiner Überlieferung zufolge als der beliebteste höfische Roman des Mittelalters gilt, basiert auf einer französischen Vorlage, dem unvollendet gebliebenen *Conte du Graal* (1180–90) Chrétiens de Troyes. Thematisch knüpft Wolfram mit dem *Parzival* an den „klassischen" Artusroman in der Tradition Hartmanns von Aue (*Erec*, *Iwein*) an, löst sich aber gleichzeitig von diesem, indem er mit der Gralthematik eine neue religiöse Dimension einbringt.

Der in knapp 25.000 paarweise gereimten Versen und in 16 Bücher unterteilte Text gliedert sich in zwei Haupthandlungsstränge, die Parzival- und die Gawan-Handlung. Unsere Annotationen beschränken sich bisher auf die erste Parzival-Partie (Buch 3–6): Sie beginnt nach der Elternvorgeschichte (Buch 1–2) und endet vor der ersten Gawan-Partie (Buch 7–8). Somit wird eine relativ abgeschlossene Handlungseinheit erfasst.

Adorno

Adornos *Ästhetische Theorie* ist eine der wirkmächtigsten philosophischen Ästhetiken des 20. Jahrhunderts. Sie gilt als letzte große Ästhetik der Moderne. Auf Basis der Besonderheiten moderner Kunstwerke untersucht Adorno Status und Charakter der Kunst, ihr Verhältnis zur Gesellschaft sowie ihren utopischen Kern. Adorno hat die *Ästhetische Theorie* aufgrund seines unerwarteten frühen Todes nicht selbst abgeschlossen. Der Text liegt nur in einer 1970 von Adornos Frau Gretel und Ralf Tiedeman erstmals herausgegeben Leseausgabe vor.

Die bisherigen Annotationen beschränken sich auf das Kapitel *Zur Theorie des Kunstwerks*, das von Adorno als einer der zentralen Abschnitte des Buches erachtet wurde.

Werther

Bei Johann Wolfgang von Goethes *Die Leiden des jungen Werthers* von 1774 handelt es sich um einen komplex aufgebauten Briefroman. Eine Gattung, die sich orientierend an den englischen und französischen Vorbildern von Samuel Richardsons *Pamela* (1740) und *Julie ou La Nouvelle Heloise* (1761) von Jean-Jaques Rosseau in der Mitte des 18. Jahrhunderts auch in Deutschland etablierte. 1787 entstand eine von Goethe überarbeitete Fassung des *Werthers*. Die Änderungen manifestieren sich in einer stilistischen Aufwertung des Textes, der Tilgung übermäßiger Gefühlschilderungen des Protagonisten und der Hinzufügung einzelner

Briefe (Bauernburschenepisode). Unsere Annotationen umfassen die einleitenden Worte des fiktiven Herausgebers sowie die ersten Briefe von Werther an seinen Freund Wilhelm, dessen Antworten nur stellenweise und implizit in den Briefen enthalten sind. Darin schildert Werther ihm einen Ball-Abend, an dem es zur Bekanntschaft mit der verlobten Lotte kommt.

Bundestagsdebatten

Das Plenardebattenkorpus des deutschen Bundestages besteht aus den von Stenografinnen und Stenografen protokollierten Plenardebatten des Bundestages und umfasst 1.226 Sitzungen zwischen 1996 und 2015. Die Sitzungen stehen als sogenannter Stenografischer Bericht der Öffentlichkeit zur Verfügung und wurden im Rahmen des PolMine[1]-Projektes aufbereitet und verfügbar gemacht.

Unsere Annotationen beschränken sich auf Auszüge aus insgesamt vier Plenarprotokollen, wobei pro Protokoll jeweils die gesamte Rede eines Politikers oder einer Politikerin annotiert wurde. Die Auszüge behandeln inhaltlich allesamt Debatten über die Europäische Union (Beitrittsverhandlungen mit der Türkei, europäische Sozialpolitik, EU-Verfassungsdebatte).

Entitätsannotationen

Unser Korpus enthält annotierte *typisierte Entitätenreferenzen*. Wir annotieren Zeichenfolgen als Entitätenreferenzen, wenn sie auf eindeutig unterscheidbare individuelle Instanzen in der realen oder fiktionalen Welt referieren. Damit sind unsere Entitäten ähnlich zu *mentions* (Erwähnungen) aus dem Bereich der Koreferenzresolution, obwohl auch konzeptuelle Unterschiede bestehen. Zusätzlich markieren wir bei jeder Entitätsreferenz, zu welcher Kategorie die referierte Entität gehört. Dabei gehen wir mit fiktionalen Welten genauso um wie mit der realen Welt. Wenn keine anderen Umstände dem widersprechen, nehmen wir an, dass fiktionale Welten denselben Regeln folgen wie die reale Welt.

Die unterschiedenen Typen sind

- Person: Referenz auf Menschen und menschenähnlich agierende Tiere oder Dinge
- Organisation: Referenz auf eine Organisation
- Location: Referenz auf einen Ort oder Schauplatz
- Work: Referenz auf ein kulturelles Artefakt (Veröffentlichung, Gesetz)
- Event: Referenz auf ein Ereignis
- Concept: Referenz auf ein theoretisches Konzept

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

Eine genauere Beschreibung finden Sie in den Annotationsrichtlinien (in einer separaten PDF-Datei), in denen auch Beispiele und die Diskussion einiger Spezialfälle enthalten sind.

Eigennamen (*named entities*) sind eine Teilmenge der möglichen referierenden Ausdrücke.

Korpusgröße

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

Aus lizenzrechtlichen Gründen können die Daten zu Adorno nur herausgegeben werden, wenn die Anfordernden bzw. Teilnehmenden eine Lizenz zu diesem Werk nachweisen können.

Datenformate

CoNLL (TSV)

Je Zeile ein Token. An Satzgrenzen eine Leerzeile. Annotationen sind tab-separiert (CSV-Dateiformat)[2]. Mehrere Annotationen können in verschiedenen Spalten notiert werden. `B-PER` bezeichnet das erste Token einer Personen-Annotationen, `I-PER` bezeichnet folgende Tokens einer Personen-Annotation. `O` (der Buchstabe) bezeichnet keine Annotationen.

Update, 29.11.: Da sich die Trainingsdaten für Werther nur auf einen Teil des Textes beziehen, enthält die Datei mit der ID `3_34_12` nur noch einen Teil des Textes. Frühere Versionen der Datei enthielten aufgrund eines technischen Fehlers leider auch nicht-geprüfte Annotationen.

Beispiel

```
Die B-PER
geringen I-PER
Leute I-PER
```

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

```
des I-PER B-LOC
Orts I-PER I-LOC
kennen O
mich O
schon O
, O
und O
lieben O
mich O
, O
besonders O
die B-PER
Kinder I-PER
. O
```

Apache UIMA XMI

XML-basiert, für die Verarbeitung mit Apache UIMA[3]. Die relevanten Typen[4] sind Untertypen von `de.unistuttgart.ims.creta.api.Entity`. Die Annotationskategorie ergibt sich zum einen aus dem Untertyp (z.B. `de.unistuttgart.ims.creta.api.EntityPER`) und zum anderen aus dem Wert von Attribut `category`.

Markdown

Benutzt das pandoc[5] Markdown, vor allem zum manuellen Lesen der Annotationen. Annotationen sind mit eckigen Klammern gekennzeichnet, tiefgestellt folgt dann die Kategorie. Das Format sollte nicht zur automatischen Verarbeitung verwendet werden.

Beispiel

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

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

[4] Das Typsystem kann von https://www.creta.uni-stuttgart.de/wp-content/uploads/2016/09/typesystem.xml heruntergeladen werden.

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