Annotation Guidelines for CRETA Entity References

These guidelines were compiled and applied within the CRETA project in the course of 2016.

1 Definition: What is an entity?

We define entities as **specific objects that are distinguishable by naming in a real or fictional world**. These objects can be assigned to an entity category (e.g. persons or locations).
In texts, we annotate *references* to such entities, thus indicators which refer to a single entity or a group of entities. *Co-references* are not annotated.

2 General remarks

We use the term "entity references" for all text passages as described above. Entity references can be:

- **Proper names** (annotate): Werther, the Warsaw Pact, "The Trial" [by Kafka]
- **Generic Names** (sometimes annotate): We annotate generic names (appellativa, "the house", "an animal", "the peasant") if they refer to one or more *concrete instance(s)* of the class
  - [Three farmers] went to a birthday party (annotate)
  - Farmers usually live on farms (do not annotate)
  - The lion lives in a pack unlike other cats. (do not annotate)
- **Pronouns** (e.g. *I/she/we*) (do not annotate)

2.1 Annotation boundaries: Maximum nominal phrases

Entity references are (in these guidelines) *maximal noun phrases* (NPs), that is, nouns with preceding/subsequent text parts which further specify the noun.
Nominal phrases include, for instance: (for more examples, see Appendix)

- **Definite** and **indefinite articles**: [The fool] stumbled.
- **Complements** and **adjuncts**: [[Obelix] ' dog] is small and white.
- **Attributes**: [The magic wall] resisted the attempts to tear it down.
- **Relative clauses**: [The maid who had most responsibility] was Anna.
- **Appositions**: [My neighbor, a doctor named Doc Brown], recommended (…).
- **Nominalized adjectives**: [The beauty] threw her hair down.

2.2 Partial vs. full annotation

We annotate all entity references, irrespective of their status, importance or presumed interpretative value (except for events and complex concepts).

2.3 Nesting

Entity references can appear embedded, but are annotated separately only when they refer to different entities.

- [The EU summit in [Spain]] was a complete success.
- [Party colleagues [Merkel] and [Schröder]] came to the conclusion (…).
- [Chancellor Schröder] said 'Basta!'

3 Entity-Types Classification

If an entity reference is discovered in the text, it gets assigned one of the following semantic classes. References to entities outside outside these classes are left unannotated. The assignment of a reference to a class is not always clear. In cases of systematic ambiguity (e.g., states) we identify the class that predominates in the context. Cases that cannot be identified are marked as class ambiguous.

Persons (PER): We annotate references relating to persons, characters, or character-like entities (e.g., animals):

- [Mrs. Wedemeier]$_{PER}$ opened a new bank account today.

Places (LOC): Expressions referring to cities, countries or territories.

- [The Eiffel Tower]$_{LOC}$ is in [Paris]$_{LOC}$.

Organizations (ORG): Expressions that designate organizations.

- We will have to keep the demands of [the EU]$_{ORG}$ in mind.

Events (EVT): References to events.

- [September 11]$_{EVT}$ has changed everything.

Works (WRK): References to cultural artefacts in a relatively broad sense.

- [The Treaties of Rome]$_{WRK}$ were signed in Rome in 1957.

Abstract Concepts (CNC): References to other concepts that are important to the analysis. In most cases, those concepts (e.g. “identity”) could be operationalized in multiple ways, depending on the researcher’s theoretical framework.

- [Our common European values]$_{CNC}$ have to be defended.