Text Technology Project Topic Description

Sebastian Sammet, Thomas Bott

January 9, 2021

Relationships between Characters in German Fairy Tales

We plan to work with a corpus of 86 German fairy tales by the brothers Grimm available in TCF (XML) format. We intend to extract and then annotate nouns which would be typical for fairy tales (e.g. king, princess, witch, magic, ...), starting with characters. Our goal is then to research the relationships between these nouns (e.g. if there is a relationship and maybe even what kind of relationship this is). We still have to determine on how we want to annotate this (e.g. occurrence in the same sentence → in relationship). Finally we want to visualize the annotated data in the NoSQL graph database Neo4j and therefore import XML to Neo4j, which would be our extension.

The collect, prepare and access steps could roughly be described as follows:

- collect
    - from DTA: http://www.deutschestextarchiv.de/book/show/grimm_maerchen01_1857
    - corpus containing 86 German fairy tales by the brothers Grimm
    - available in TCF (XML) format with some annotation layers (e.g. token, lemma, pos)

- prepare
    - add more annotation layers
    - e.g. document/sentence boundaries, characters/nouns to investigate, relations between these characters/nouns, etc.
    - we still have to determine what we really need and how to annotate it

- access
    - import to Neo4j (NoSQL, extension)
    - investigate research questions with regard to the relationships between characters/nouns
    - visualize the data