TEI Annotation and Network Analysis of Diplomatic History Documents

Florentina Armaselu
CVCE Luxembourg
florentina.armaselu@cvce.eu

Marten Düring
CVCE Luxembourg
marten.during@cvce.eu

Veronica Martins
CVCE Luxembourg
veronica.martins@cvce.eu

In this paper we explore the potential of network analysis for the exploration of TEI [1] annotated documents.

Network visualizations provide insight in highly complex relations between any type of entities such as individuals or locations. Visualizations are used to reveal patterns in data which are otherwise impossible or very hard to detect. We will discuss three use cases: 1) networks as means to create an interlinked taxonomy in which hierarchies are expressed through centrality scores. 2) Centrality algorithms describe structural properties of nodes and networks mathematically. We identify possible synonyms based on the axiom of homophily which is also entailed in the proverb that “birds of a feather flock together”. 3) We discuss the added analytical value of network visualizations for a domain expert in diplomatic history with in-depth knowledge of the underlying primary sources.

The studied corpus is part of a larger research project on the diplomacy within Western European Union (W.E.U.) and contains documents (French) on the production, standardisation and control of armaments (1954 to 1982) from the W.E.U. archives which are based at the Archives Nationales de Luxembourg. It includes different types of materials encoded in XML-TEI P5, e.g. notes from the Secretary-General or Secretariat-General, minutes of meetings, memoranda and studies. Three categories of encoding are provided: metadata (title, author, availability date, origin place, confidentiality status, etc.), structural markup (headers, footers, sections, paragraphs, line breaks), content-related annotations (discourse of country/institutional representatives, named entities). The Named Entity Recognition (NER) task involved a semi-automatic approach using GATE [2], i.e. the French NE system, Gazetteer and Gazetteer List Collector plugins. Seven classes of entities were identified and annotated in the texts: persons, places, organisations, events, dates, products and functions (official positions).

Through network analysis via Gephi [3], we will address questions related to topics like: use of variants for the same entity, the so called “synonyms” in the context (Union de l'Europe occidentale / Union de l'Europe Occidentale / UNION DE L'EUROPE OCCIDENTALE / U.E.O. / U. E. O.); hierarchical relations between entities (Conseil / Conseil ministériel / Conseil ministériel de l'U.E.O.); interpretation from a historical perspective of the different types of networks built with the annotated data (persons, organisations, etc.).

References

GATE (General Architecture for Text Engineering) [WWW Document], n.d. URL https://gate.ac.uk/ (accessed 3.30.15).

Gephi (The Open Graph Viz Platform) [WWW Document], n.d. URL http://gephi.github.io/ (accessed 3.30.15).

TEI (Text Encoding Initiative) [WWW Document], n.d. URL http://www.tei-c.org/index.xml (accessed 3.30.15).