Text Processing Interest Group

CLARIAH Tech Day

November 26th, 2020

Text Processing

Objective

The aims of the IG on Text are:

- Foster discussion and knowledge sharing regarding automatic text processing
- Enhance interoperability between various text processing solutions
- Develop and share best practices
- Inform development of CLARIAH text processing tools and services

Strong affiliation with WP3 and WP6.

Scope of the Interest Group

Our scope is **automatic** text processing, and roughly encompasses the following fields:

- Natural Language Processing
- Text Mining
- Text Search & Retrieval

This includes things like:

- automatic linguistic enrichment (PoS, lemma, parsing, etc..)
- sentiment analysis
- tokenisation
- OCR/HTR and normalisation
- language modelling & machine Translation

Scope of the Interest Group: Languages

Our language scope is not limited, but considering we are a Dutch project it is fair to say that extra attention goes to:

- Dutch, Flemish and dialects
- Frisian

Historical variants are also within our scope.

Scope of the Interest Group: Out of scope

Aspects that are outside the scope of this Interest Group (because they are covered by other IGs):

- *manual* text annotation (covered by the annotation group)
- annotation models and formats (covered by the annotation group)
- speech to text (speech recognition) and text to speech (covered by the AV group)

Tasks for the group

1. Provide an inventory of current text processing tools, services and models in CLARIAH, either developed in CLARIAH (WP3 or WP6), or third party projects that are adopted as solutions.
2. Identify connections that can be made between various tools (specific workflows/pipelines) to certain specific ends desired by the research community.
3. Specify what requirements we want text processing solutions to adhere to for CLARIAH, to facilitate interoperability between tools/services. Indicate to what extent the existing solutions adhere to these requirements.