OPTIMIZATION OF THE LINGUISTIC SEARCH FOR XML-ANNOTATED DISCOUNT OF LUDWIG WITTGENSTEIN

Anton Serjogin
Centre for Information and Speech Processing, LMU
anton.serjogin@gmail.com

22.05.2017

This work is a part of the Digital-Humanities-Project *"Wittgensten in Co-Text"*, in cooperation with the Wittgenstein Archives of the University of Bergen (WAB) in Norway. The documents consist of both typescripts and manuscripts. The goal of this work is to optimize the linguistic search for XML-annotated discount of Ludwig Wittgenstein through the optimal utilization of XML-annotation and the improvement of the XML-edition. The emphasis is on personal names and extension of the lexicon.

XSLT-data from Bergen for the conversion of the original XML-editions is presented in 3 different types:

- ORG (all options)
- NORM (the appropriate variant)
- DIPLO (what L.W. has written himself)

For this task a probabilistic POS-Tagger based on the Markov Moodel is used. The TreeTagger is a tool for annotating text with part-of-speech and lemma information. It was developed by Helmut Schmid in the TC project at the Institute for Computational Linguistics of the University of Stuttgart.

NORM-data → NORM-tagged.xml

- Step 1: localization of the personal names in NORM.xml - collect all the possible mistakes through tagging
- Step 2: improvement of the semantic search in WittFind: finding personal names

```
( ([ADJA] | [ADJD] ) | [NN]) & <+EN>
```

Suggestion: creating a new syntactic category “persname”:

- Adding a “persname” on tagged documents with each collected example.
- Adding a new category in WittFind

The results (from 20 documents):

- In 13 documents WittFind meets 168 results
- The recommended system searches and finds 833 results in all 20 documents

The poor results may also be affected by transcription errors that have been initially made at the University of Bergen. A transcription error is a specific type of data entry error that is commonly made by human operators and is usually a result of typographical mistakes affected by a cognitive bias. Improvements on the following aspects can be done in order to raise the number of successful results:

- Tokenization improvement
- Tagging improvement
- Personal name recognition improvement

The extension of the lexicon should be done by creating frequency distribution with the help of *etree* instead of *regex*, because it is less time-consuming and allows a more straight-forward approach.

To sum up the things mentioned, the work is split into 2 main parts:

- Improving the linguistic search in WittFind
- Recognition of personal names