Operationalisation of Research Questions of the Humanities within the CLARIN Infrastructure – An Ernst Jünger Use Case

Dirk Goldhahn
Natural Language Processing Group
University of Leipzig
Germany
dgoldhahn@informatik.uni-leipzig.de

Thomas Eckart
Natural Language Processing Group
University of Leipzig
Germany
teckart@informatik.uni-leipzig.de

Thomas Gloning
Department of German Philology
University of Gießen
Germany
thomas.gloning@germanistik.uni-giessen.de

Kevin Dreßler
Natural Language Processing Group
University of Leipzig
Germany
kvndrsslr@gmail.com

Gerhard Heyer
Natural Language Processing Group
University of Leipzig
Germany
heyer@informatik.uni-leipzig.de

Abstract

CLARIN offers access to digital language data for scholars in the humanities and social sciences. But how can the linguistic resources help to answer real research questions in the respective fields? By adressing research questions concerned with the work of German author Ernst Jünger an example of a successful operationalisation within the CLARIN infrastructure will be introduced. In addition a new versatile Web application for answering a wide range of research questions based on vocabulary use will be presented.

1 Introduction

CLARIN envisions itself as a research infrastructure for the humanities and social sciences. This makes the concerns of scholars of the respective fields a central aspect for CLARIN. A very important question which arises in this regard is: How can the linguistic resources offered by CLARIN be used to answer research questions in the humanities or social sciences?

In the following a use case of the humanities for the CLARIN infrastructure will be depicted. Research questions concerned with the work of German author Ernst Jünger will be adressed and operationalised. The CLARIN infrastructure will be used to search for necessary resources, to process the data and to analyse and visualise the results.

Part of this analysis is the Web application Corpus Diff for difference analysis currently developed in the context of CLARIN. It allows to answer a wide range of research questions which can be mapped to differences in vocabulary use. Typical workflows within this Web application will be presented utilizing the Jünger use case.

2 Research Questions

Ernst Jünger's political texts ("Politische Publizistik") from the years 1919 to 1933 are available in a philologically reviewed and well annotated edition (Berggötz, 2001), which has been digitalized for the purpose of internal investigation. The explosiveness of these texts lies in a wide range of topics regarding the development of Germany in the 1920s. This contains dealing with front experiences in World War I, consequences of the lost war, issues of national reorientation, and superordinated aspects of time interpretation. Jünger's texts change considerably in their thematic priorities and in their linguistic form in the 15 years of their creation.

Key issues that arise from a linguistic and discourse historical perspective on this corpus, include:

1. How does language use, in particular the use of words, correlate with specific topics and "perspectives", which are expressed in the texts?
2. How can the lexical profile of Jünger's political texts and its development be characterized in the temporal dimension?
3. How can the lexical profile of the political texts be characterized in comparison with contemporary material such as newspaper texts of the 1920s or the literary works of authors from the same period?

3 Operationalisation

In order to answer these research questions in a systematic matter, they need to be operationalised. Important aspects of this process are:

- data: collections of texts matching the research question (including reference corpora)
- algorithms: methods to carry out the desired analysis and their combination to more complex applications or workflows
- results and visualisation: structure, size, presentation and how to browse or search the data that lead to the results

Focus of the operationalisation will be on using the CLARIN infrastructure for searching for data and algorithms and performing the analysis by combining them to workflows.

First of all, texts matching the research question are required. As mentioned before, a digitized version of Ernst Jünger's political texts from 1919 to 1933 was already available. This corpus includes publishing dates for all included texts and will be the starting point for further analyses.

Next, a method needs to be chosen to discover differences in the use of vocabulary. One method that allows for such insights is difference analysis (Kilgarriff, 2001). Using this analysis we can investigate differences between different years of Jünger's work or between Ernst Jünger's texts and some reference corpora.

This will then allow to:

- quantify corpus similarity,
- discover differences in vocabulary and
- analyse prominent results (vocabulary) further.

3.1 Reference Data – DWDS

Another requirement for a difference analysis is the availability of reference data. A central entry point for scholars searching language resources in CLARIN is the Virtual Language Observatory (VLO). Using the VLO's search features such as facettes, it is easy to navigate and narrow down resources and identify those of interest for the respective research questions of the social sciences or humanities.

As our Jünger texts are in German and from the years 1919 to 1933, the same needs to hold for our reference corpora. When restricting the facets to "corpus" in "German" and adding the search term "20th century" one of the most prominent results is the DWDS Kernkorpus[1].

The DWDS corpus (Digitales Wörterbuch der deutschen Sprache) (Geyken, 2006) was constructed at the Berlin-Brandenburg Academy of Sciences between 2000 and 2003. The DWDS Kernkorpus contains approximately 100 million running words, balanced chronologically and by text genre.

[1] http://catalog.clarin.eu/vlo/search?q=20th+century&fq=languageCode:code:deu&fq=resourceClass:Corpus

The main purpose of the DWDS Kernkorpus is to serve as the empirical basis of a large monolingual dictionary of the 20th century. The Kernkorpus is roughly equally distributed over time and over genres: journalism, literary texts, scientific literature and other nonfiction.
Using the webservices of the DWDS we extracted texts for all genres. We collected corpora for each year and genre separately, allowing for analyses using both dimensions.

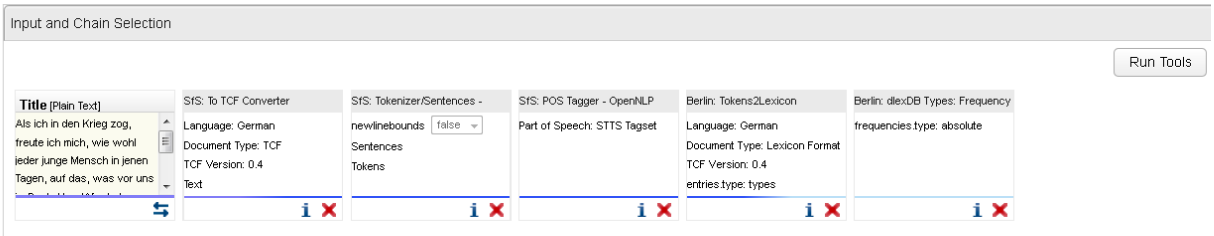
4 Combination to workflows

4.1 Preprocessing

Preprocessing of the raw material is the basis for conducting a difference analysis as word frequencies for the underlying texts need to be computed. Therefore, especially sentence segmentation and tokenization are relevant preliminary work. In addition, to allow for part of speech specific analyses, POS tagging needs to be performed.

For data processing WebLicht (Hinrichs, 2010), the Web-based linguistic processing and annotation tool, is an obvious choice. Within WebLicht one can easily create and execute tool chains for text processing without downloading or installing any software.

Figure 1 depicts a preprocessing chain created in WebLicht. It includes import of plain text files, conversion to an internal format (TCF), sentence segmentation, tokenization, part of speech tagging and counting of word frequencies. This processing can be run online and the results can then be downloaded. For an even more convenient data transfer an export of results to the personal workspace of CLARIN users will be available in the future.



Title [Plain Text]
Als ich in den Krieg zog, freute ich mich, wie wohl jeder junge Mensch in jenen Tagen, auf das, was vor uns

SfS: To TCF Converter
Language: German
Document Type: TCF
TCF Version: 0.4
Text

SfS: Tokenizer/Sentences -
newlinebounds: false
Sentences
Tokens

SfS: POS Tagger - OpenNLP
Part of Speech: STTS Tagset

Berlin: Tokens2Lexicon
Language: German
Document Type: Lexicon Format
TCF Version: 0.4
entries.type: types

Berlin: dlexDB Types: Frequency
frequencies.type: absolute

Figure 1: Preprocessing chain in WebLicht.

The tool chain was executed on the Jünger texts for all annual time slices from 1919 to 1933 separately. This resulted in information on word frequencies for 15 corpora, one for each year. Furthermore we used the preprocessing for the annual corpora of each of the four DWDS-genres, resulting in 60 subcorpora, four for each year.

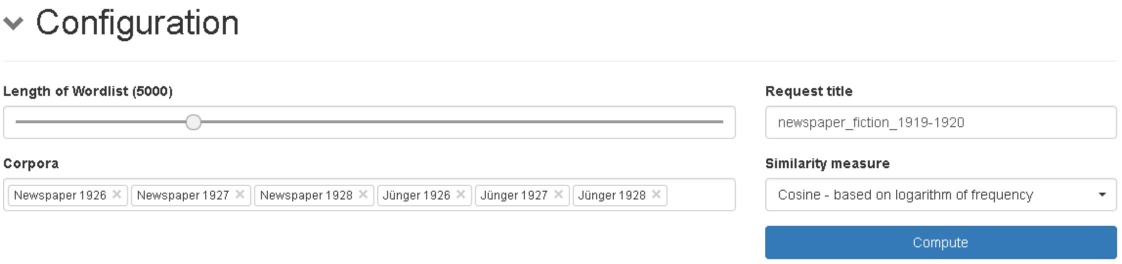
4.2 Data Analysis

The actual analysis can be carried out via the dedicated Web application Corpus Diff[2]. An easy to use JavaScript-based interface allows creating several analysis processes in parallel. The generation of corpus similarity is solely based on lists of word frequencies or word rank representing their respective corpus. The user interface provides several similarity measures that are using cosine similarity on these word vectors. Cosine similarity is just one possible option for the proposed comparison but has the advantage of flexibility concerning features and their weighting (e.g. logarithmic scaling). The result is a normalised value for every pairwise comparison between 0 (no resemblance of the word lists) and 1 (identically distributed vocabulary). The application is completely based on a RESTful webservice that delivers all information necessary: an overview of all provided corpus representations and the complete word lists for every corpus.
Using word frequency lists for corpus comparison has several advantages: these lists are dense representations of the content of a corpus, but due to their small size easy to transfer and to process. Furthermore there are hardly any copyright restrictions as no access to full texts or text snippets is necessary. This means that even for textual resources with very restrictive licences an exchange of this data is in most cases feasible.

By using the Web interface a user can select a set of corpora, the used similary measure and how many of the most frequent words of a word list should be taken into account for the analysis (figure 2). As a

[2] http://corpusdiff.informatik.uni-leipzig.de

result the user is provided with a matrix visualising pairwise similarity of corpora using different colour schemes. These colour schemes also emphasize clusters of similar corpora. In addition, a dendogram shows a representation of a single-linkage clustering for all word lists used in the analysis. Both, matrix and dendogram, are a means of identifying interesting corpora with an especially high or low similarity of their vocabulary. This can be used to perform a diachronic comparison to identify changes over time, but also for comparing corpora of different genre or origin with each other.



Configuration

Length of Wordlist (5000)

Request title: newspaper_fiction_1919-1920

Corpora: Newspaper 1926 × Newspaper 1927 × Newspaper 1928 × Jünger 1926 × Jünger 1927 × Jünger 1928 ×

Similarity measure: Cosine - based on logarithm of frequency

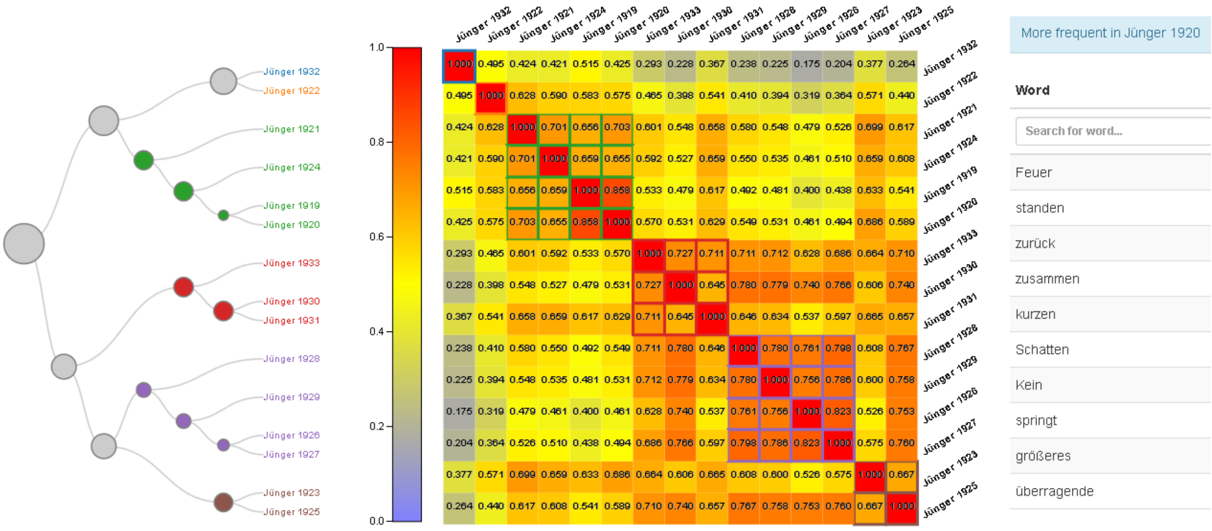
Figure 2: Configuration of corpus comparison tasks.

By selecting two corpora more detailed information about differences in their vocabularies is shown. This especially includes lists of words that are more frequent or that exclusively occur in one of the corpora. Both are valuable tools to identify terms that are specific or more significant for the respective resource. Moreover the results are a starting point for further analyses with hermeneutic approaches by experts of the respective fields.

If the user is interested in a specific word, a frequency timeline is provided via a line chart. This will usually be relevant for important key words of the texts in question or words extracted in the previous analysis steps. Here the diachronic development of the word's usage can be easily analysed and put into relation to other words or comparisons can be made of its frequency in different genres over time.

5 Examples

Figure 3 (left) shows the similarity matrix and the dendogram for the Jünger texts between 1919 and 1933. One interesting pair of corpora are, among others, those from 1920 and 1927 since a low similarity holds for them. When looking at salient vocabulary for this comparison (figure 3, right), words like “Feuer” (“fire”) are much more prominent in the texts from 1920.



	Jünger 1932	Jünger 1922	Jünger 1921	Jünger 1924	Jünger 1919	Jünger 1920	Jünger 1933	Jünger 1930	Jünger 1931	Jünger 1928	Jünger 1929	Jünger 1926	Jünger 1927	Jünger 1923	Jünger 1925
Jünger 1932	1.000	0.495	0.424	0.421	0.515	0.425	0.293	0.228	0.367	0.238	0.225	0.175	0.204	0.377	0.264
Jünger 1922	0.495	1.000	0.628	0.590	0.583	0.575	0.465	0.398	0.541	0.410	0.394	0.319	0.364	0.571	0.440
Jünger 1921	0.424	0.628	1.000	0.701	0.656	0.703	0.601	0.548	0.658	0.580	0.548	0.479	0.526	0.699	0.617
Jünger 1924	0.421	0.590	0.701	1.000	0.659	0.655	0.592	0.527	0.659	0.550	0.535	0.461	0.510	0.659	0.608
Jünger 1919	0.515	0.583	0.656	0.659	1.000	0.858	0.533	0.479	0.617	0.492	0.481	0.400	0.438	0.633	0.541
Jünger 1920	0.425	0.575	0.703	0.655	0.858	1.000	0.570	0.531	0.629	0.549	0.531	0.461	0.494	0.686	0.589
Jünger 1933	0.293	0.465	0.601	0.592	0.533	0.570	1.000	0.727	0.711	0.711	0.712	0.628	0.686	0.664	0.710
Jünger 1930	0.228	0.398	0.548	0.527	0.479	0.531	0.727	1.000	0.645	0.780	0.779	0.740	0.766	0.606	0.740
Jünger 1931	0.367	0.541	0.658	0.659	0.617	0.629	0.711	0.645	1.000	0.646	0.634	0.537	0.597	0.665	0.657
Jünger 1928	0.238	0.410	0.580	0.550	0.492	0.549	0.711	0.780	0.646	1.000	0.780	0.761	0.798	0.608	0.767
Jünger 1929	0.225	0.394	0.548	0.535	0.481	0.531	0.712	0.779	0.634	0.780	1.000	0.756	0.786	0.600	0.758
Jünger 1926	0.175	0.319	0.479	0.461	0.400	0.461	0.628	0.740	0.537	0.761	0.756	1.000	0.823	0.526	0.753
Jünger 1927	0.204	0.364	0.526	0.510	0.438	0.494	0.686	0.766	0.597	0.798	0.786	0.823	1.000	0.575	0.760
Jünger 1923	0.377	0.571	0.699	0.659	0.633	0.686	0.664	0.606	0.665	0.608	0.600	0.526	0.575	1.000	0.667
Jünger 1925	0.264	0.440	0.617	0.608	0.541	0.589	0.710	0.740	0.657	0.767	0.758	0.753	0.760	0.667	1.000

More frequent in Jünger 1920

Word
Feuer
standen
zurück
zusammen
kurzen
Schatten
Kein
springt
größeres
überragende

Figure 3: Similarity matrix and dendogram for the Jünger texts 1919-1933 (left), List of words with higher relative frequency in 1920 when compared to 1927 (right).

The example of “Feuer” (in a military sense) shows the fruitfulness of the tool's visualizations. Both in respect of its usage from 1919 to 1933 in the edition of the “Politische Publizistik” and in comparison with newspaper texts from the same period, differences in word usage can be identified (figure 4), making it an ideal starting point for further analyses by experts of the respective fields. Since the focus of this paper is on the operationalization we will not follow up on these subject specific tasks.

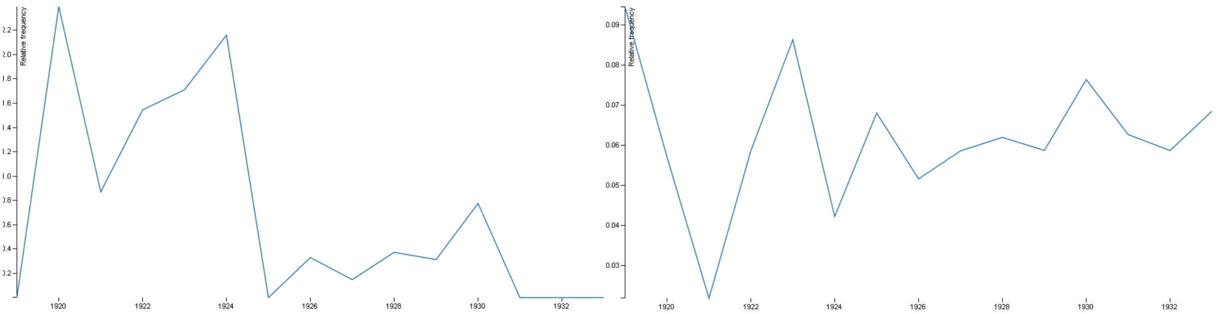


Figure 4: Frequency timeline for “Feuer” in texts of Ernst Jünger (left) and in newspaper texts (right).

A second example of the same kind are the dynamics of the word “Frontsoldat” which shows a comparable dynamical timeline. The word usage mirrors the status of the front soldier experience as a topic both in Jünger's texts and in the newspapers of the time.

6 Outlook

The current state is to be seen as a first working implementation to depict the workflow and its potential value for scholars from a broad range of scientific fields. For the near future it is planed to implement a seamless integration in the CLARIN infrastructure. This will especially include the support of TCF-based files and hence the usage of WebLicht as the first choice preprocessing chain.
It also planed to support the CLARIN personal workspaces as data storage to address copyright issues. As a consequence users will be enabled to store their data in a secure and private environment, and still make use of the full potential of a distributed infrastructure.

From the philological standpoint the following tasks are considered to be especially important for the further development:

- Methods to deal with polysemy and thematic specificity of terms (like “Feuer”/”fire”)
- Answers to the question what fields of “traditional” research (e.g. Gloning to appear) can be treated by means of Digital Humanities and in what fields this is not feasible yet. Reference point for such a comparison or evaluation of methods could be by contrasting the performance of DH tools with results of traditional approaches. Such a comparison could provide indications for the future development of DH tools for specific research questions. In the context of this usecase this may be especially relevant for the field of word formation (e.g. *Bierreden*, *Bierstimmung*, or *Biertisch* used for degradation).

In addition to the existing analysis and visualisation components it will be important to connect results with the original text material by providing an option to display and check the usage contexts. A Key Word in Context (KWIC) view would be one of these base functions where results generated by automatic analysis can be coordinated with the textual findings. As a typical user experience is that the corpus comparison tool provides surprising results that require further explanation, it would be desirable to provide easy to use methods for pursuing these traces in the original text. Besides a KWIC component (where available) an overview of typical contexts will be generated by using co-occurrences analysis (Büchler, 2006). This graph-based approach allows the visualisation of typical word contexts per time slice and especially following diachronic changes of these contexts.

7 Conclusion

In this paper we showed how a research question of the humanities and social sciences can be operationalised and answered using the CLARIN infrastructure. Different resources and services of CLARIN were utilized in this endeavour, such as the VLO or the WebLicht execution environment for automatic annotation of text corpora.

In addition we introduced a generic tool for difference analysis which can be used to answer a wide range of research questions based on vocabulary use. A workflow for utilizing this Web application and its different ways of visualisation to guide the exploration of the research results was presented. Finally, an outlook for a seamless integration into the CLARIN infrastructure was given.

References

Berggötz, S.O. (2001). Ernst Jünger. Politische Publizistik 1919 bis 1933. Klett-Cotta, 2001.

Büchler, M. (2006). Flexibles Berechnen von Kookkurrenzen auf strukturierten und unstrukturierten Daten. Diploma Thesis, University of Leipzig.

Geyken, A. (2006). A reference corpus for the German language of the 20th century. In: Fellbaum, C. (ed.): Collocations and Idioms: Linguistic, lexicographic, and computational aspects. London: Continuum Press, 23-40.

Gloning, Th. (to appear). Ernst Jünger Publizistik der 1920er Jahre. Befunde zum Wortgebrauchsprofil. To appear in: Benedetti, Andrea/Hagestedt, Lutz (eds.): Totalität als Faszination. Systematisierung des Heterogenen im Werk Ernst Jüngers. Berlin/Boston: de Gruyter (to appear january 2016).

Hinrichs, M., Zastrow, T., & Hinrichs, E. W. (2010). WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure. In Proceedings of LREC 2010, Malta.

Kilgarriff, A. (2001). Comparing corpora. International journal of corpus linguistics, 6(1), 97.