Cristina Bosco, Sara Tonelli and Fabio Massimo Zanzotto (dir.)

Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015
3-4 December 2015, Trento

Accademia University Press

The OPATCH corpus platform – facing heterogeneous groups of texts and users

Verena Lyding, Michel Généreux, Katalin Szabò and Johannes Andresen

DOI: 10.4000/books.aaccademia.1502
Publisher: Accademia University Press
Place of publication: Torino
Year of publication: 2015
Published on OpenEdition Books: 11 November 2016
Serie: Collana dell'Associazione Italiana di Linguistica Computazionale
Electronic ISBN: 9788899200008

The OPATCH corpus platform – facing heterogeneous groups of texts and users

Verena Lyding[1], Michel Généreux[1], Katalin Szabò[2], Johannes Andresen[2]
[1]EURAC research, viale Druso 1, 39100 Bolzano, Italy
[2]Dr. Friedrich Teßmann library, Via A.-Diaz 8, 39100 Bolzano, Italy
firstname.lastname@eurac.edu,firstname.lastname@tessmann.it

Abstract

English. This paper presents the design and development of the OPATCH[1] corpus platform for the processing and delivery of heterogeneous text collections for different usage scenarios. Requirements and technical solutions for creating a multipurpose corpus infrastructure are detailed by describing its development.

Italian. *L'articolo presenta il design e lo sviluppo della piattaforma OPATCH[1] per l'elaborazione e la distribuzione di testi eterogenei in differenti contesti d'uso. La procedura evidenzia le esigenze e le soluzioni tecnologiche nella creazione di una ampia infrastruttura per i corpora.*

1 Introduction

Nowadays, electronic text collections have become an important information source in different usage contexts, including linguistic research and research in the humanities. With regard to application contexts and user groups, requirements related to tools for analyzing the text collections can differ. This mainly concerns the level of search interfaces, while the underlying data processing and annotation procedures typically build on standard NLP technologies.

In this paper, we are presenting the OPATCH project[1], which aims at creating a multipurpose corpus platform, by combining a uniform system back-end with varied front ends for different usage scenarios. Its flexible use is illustrated through two sample portals: one for content research and one for a linguistic search scenario.

2 Related work

Recent works related to the use of NLP technologies for enhancing humanities research environments include ALCIDE (Moretti et al., 2014), a platform for the automatic textual analysis of historical text documents, and GutenTag (Brooke et al., 2015), a tool which serves as a reader, subcorpus builder and tagging tool for literary texts.

3 Overview of the OPATCH platform

The OPATCH platform aims at providing a comprehensive infrastructure for the processing and delivery of digital text documents. The platform is designed to serve multiple purposes with regard to both the textual resources and the usage contexts it can accommodate. The platform consists in a system back-end, a unique component for text processing, and an extendable number of front-end portals, the user interfaces.

The *system back-end* combines a variety of standard text processing and annotation tools into ready-to-use tool chains. These tool chains are designed in such a way that intermediate text and corpus formats are in accordance with established standards for corpus data exchange, and output data is made compliant with standard formats of corpus and text search environments. Furthermore, the platform back-end strictly relies on open source tools.

The *system front-end* consists of an extendable series of portals for specific usage scenarios. Within the project scope, its applicability for different use cases is demonstrated by two portals:

(1) a *historical newspaper portal* for research and investigations on cultural content
(2) a *linguistic corpus portal* for language research purposes

[1] OPATCH - *Open Platform for Access to and Analysis of Textual Documents of Cultural Heritage*, financed by the 'Provincia Autonoma di Bolzano-Alto Adige, Diritto allo studio, università e ricerca scientifica, Legge provinciale 13 dic. 2006, n. 14'; project duration: Jan 2014 – Mar 2016; http://commul.eurac.edu/opatch/website/index.html

4 Data

The two use cases deliver two different text collections: The content search on local history is based on a collection of historical newspapers and the linguistic search is based on a collection of documents of current South Tyrolean German.

4.1 Historical newspaper data

The newspaper corpus contains 100.000 pages of German newspapers from (South) Tyrol for the years 1910 to 1920. They are part of the historical newspaper archive from the Alpine region held at the Dr. Friedrich Teßmann library and were selected with regard to maximizing OCR (Optical Character Recognition) quality and to including full (not partial) news issues. Table 1 shows the division by newspaper title and years.

Newspaper	Years	Pages
Bozner Nachrichten	1910-20	24786
Der Tiroler	1910-20	19933
Meraner Zeitung	1910-20	19286
Bote für Tirol	1910-19	9497
Volksblatt	1910-20	9275
Lienzer Zeitung	1910-15, 19	8479
Tiroler Volksbote	1910-19	6746
Bozner Zeitung	1911, 13, 15	1100
Pustertaler Bote	1917-20	898
Total		**100000**

Table 1. Subdivision of newspaper corpus

4.2 'Korpus Südtirol' core corpus

The other text collection consists in the core part of the 'Korpus Südtirol' (KST) initiative (Abel et al., 2009) enhanced with further newspapers. The core corpus consists of balanced texts of four genres: fiction, informative, functional (e.g. user manuals) and journalistic texts. It has a size of 3.5 Mio tokens and spans the entire 20th century.

5 Platform design

The platform design has been informed by user requirements and specified with reference to format standards and available open source tools.

5.1 User requirements

Both portals deliver documents of South Tyrolean cultural heritage and aim at fostering research on local history and language. The newspaper portal serves humanities related research with a focus on *textual content*, while the linguistic portal targets studies on *linguistic characteristics* of the South Tyrolean texts. Accordingly, the user studies addressed different target groups: historians, town/family chroniclers, teachers and students for the newspaper portal, and linguists and language planners/testers for the linguistic portal.

For the newspaper portal, two user studies were performed. In study (1), interviews with 13 library users yielded the following requirements:

- Research topics: local history, family history, world war, media history, literary study
- Objectives: research work, thesis preparation (students), preparation of teaching material
- Modes of access: by date, by topic, full text search with focus on names and events
- Use of results: saving results, references and query history, export and printing, notes
- Additional features: highlighting of search terms, overview of data base, user space

Study (2) evaluated an early interface prototype via an online questionnaire compiled by 55 respondents. It yielded the following results:

- Navigation: clear interface structure (80%)
- Modes of access:[2] text based search (80%), by title (35%), by date (25%)
- Required search facilities: [3] multiword searches, Boolean, Regular Expressions, searches, combination of text-based and filters, search by page number, fuzzy search
- Results display: standard view, pdf and download are most used (>60% often/ always; < 18% rarely/never); animated and tiles view are hardly used (< 10% often/always; > 65% rarely/never)
- Additional features:[3] persistent links to results, more (Italian) content, query storage, download of entire articles, ordering of results, adaptation to mobile devices, API

For the linguistic search, OPATCH relied on user studies from previous corpus projects (cf. Wisniewski et al. (to appear), Lyding et al. (2013)). Accordingly, primary requirements are:

- Powerful query facilities
- Search on linguistic annotations / metadata
- Focus on frequencies
- Visualization of frequencies in concordances

5.2 Format and annotation requirements

The design of the OPATCH platform is oriented on established standards for the description of language resources. With reference to the European infrastructure initiative CLARIN [4] ,

[2] respondents who "often/always" use this search
[3] as listed in free text field by several respondents
[4] http://clarin.eu

OPATCH aims at compatibility with CMDI (Component MetaData Infrastructure) for the exchange of metadata, and at providing an FCS (Federated Content Search) endpoint for the final linguistic corpus portal. Furthermore, different standard formats for encoding texts throughout processing are employed: METS/ALTO[5] format for OCRed newspaper issues files, ALTO format for single pages of text with linguistic annotations, Lucene/SOLR indexes for newspaper portal back-end, IMS Open Corpus Workbench (openCWB)[6] vertical format for linguistic portal back-end, and custom text format for the *Double Tree* (Culy/Lyding 2010) visualization front-end.

Regarding linguistic annotations and mark-up, all text documents are required to be tokenized and split into sentences, and to be annotated with metadata, lemma, part-of-speech and NE (Named Entity) information.

5.3 Portal specifications

The configuration of each portal has been specified in relation to its general purpose and in response to results of the user studies.

The **newspaper portal** aims at serving research on cultural topics and thus targets the retrieval of textual *content*. The design foresees:
(1) different search modes
(2) full access to the source data

Search options are designed to combine the access via metadata filtering (e.g. by year, title, etc.) with full text search and search by linguistic annotations (e.g. NE). This way, the portal offers text browsing and targeted searches.

The presentation of search results is designed to allow for a comprehensive view on the data, by providing the digitized text as well as the original image files. Furthermore, the possibility to highlight search terms, and download or print search results and related documents is foreseen.

The **linguistic corpus portal** aims at serving research on structural language characteristics. Accordingly, it aims at providing:
(1) powerful query facilities
(2) access to contextualized text and statistics

The query facilities are designed to support searches on text and annotation layers, including Regular Expression searches and the use of Boolean operators.

The presentation of search results, next to a standard KWIC display, foresees a *Double Tree* view, which highlights frequent word combinations and allows for the interactive exploration of results with regard to sequential and frequency characteristics of the data.

5.4 Back-end specification

The specification of the system back-end is based on the functional demands that have been derived from the user studies for both portals and the technical requirements for text processing and annotation related to them. In order to serve the two portals, the OPATCH system has to accommodate a series of tools into a flexible tool chain. The processing measures specified for the OPATCH system are presented in Table 2, in order of their application. The Table also indicates the tools used and the applicability of measures to the data of each portal.

Processing measure	**Tool**	**News**	**KST**
OCR recognition	ABBYY's fine reader	yes	partly
OCR post-processing	Custom model	yes	-
layout recognition	Formal description	-	-
metadata collection	manually	yes	yes
tokenization	Treetagger	yes	yes
lemmatization	Treetagger	yes	yes
part-of-speech tagging	Treetagger	yes	yes
NE recognition	Stanford NER and lists	yes	yes
transformation to format of retrieval tools	Lucene/SOLR and open CWB	yes	yes

Table 2. Text processing and annotation

6 Processing and annotation chain

The following subsections describe in detail the processing procedures listed in Table 2 and discuss particular challenges related to the two types of text collections and portals

6.1 OCR and post-processing

Processing of the newspaper data started from OCRed text files (METS/ALTO format), which had been produced using ABBYY's Fine Reader.[7] Due to the printing in 'Fraktur' font and a partly deteriorated paper quality, the data

[5] flexible XML schema for describing complex digital objects, maintained by the Library of Congress
[6] http://sourceforge.net/projects/cwb/

[7] OCR recognition was done within the *Europeana Newspaper Project (Europeana)*, see: http://www.europeana-newspapers.eu/

showed a very low quality. An evaluation of 10 pages of the OCRed collections gives an average bag-of-word (BoW) index success rate of 67.5%. BoW evaluations apply well to texts with complex layout structures (newspapers), cf. Pletschacher et al. (2014), while more refined evaluations that go beyond Levenshtein or edit distance may be better suited for more uniform layout such as books, cf. Reynaert (2014).

The post-correction of the OCRed texts was approached by applying a multi-step transformation model of edit operations on single or multiple letters, trained on manually corrected data. In an experiment, we could show significant reductions in error rates for words no further than two edit-operations from their true value. The task of correcting OCRed texts of newspapers is made more difficult by complex layouts, dislocated or merged words and incomplete dictionaries (Généreux et al., 2014).

At project start, KST texts have been available in digital format. OCR post-correction has been no issue, as texts were either genuine electronic text or high quality prints in modern fonts. [8]

6.2 Layout recognition

The feasibility of automatic layout recognition for historical newspapers has been investigated, related to experimentations of the project partner library Teßmann.[9] A manual analysis of section headings and layouts of ten newspapers showed:

- Text appears in three columns (rarely two)
- Vertical and horizontal separation of articles by lines or little star signs
- No headlines, but titles in 2-3 font sizes
- Very compact printing, little free space
- Advertisements with varied layout/fonts
- Content-related subdivisions are recurrent

Within OPATCH, the automatic layout division has been limited to separation of pages and distinction of header data and core text.

6.3 Metadata collection

For the newspaper corpus, metadata was collected for entire issues and includes 'title/publisher', 'publication date', 'no. of printed issues', 'no. of pages' and 'font type'. The metadata was recorded during the OCR step or added after.

For the KST corpus, great effort has been dedicated to the systematic collection of detailed metadata sets (Anstein et al., 2011). In particular, literary, functional and informative texts are associated with detailed descriptions of the author, publisher or text content, which for newspaper data would need to be assigned on article level.

The shared set of metadata among both collections covers: title, publisher, publication date.

6.4 Linguistic annotations: tokenization, lemmatization, part-of-speech and NER

For tokenization, lemmatization and part-of-speech tagging the IMS Treetagger (Schmid, 1994) trained for German has been used. Regarding Named Entity Recognition (NER), for both, the newspaper collection and the KST corpus, two approaches were combined: the Stanford NER tool re-trained for South Tyrolean German and the exact matching of texts with detailed lists of South Tyrolean names (place names[10], person names[11], addresses and organization names[12]).

The corrected OCRed output complies with the latest ALTO-XSD specification (v2.1, Feb. 20, 2014), which enforces a consistent enumeration of all entities, including multi-word entities.

6.5 Transformation for retrieval engines

The newspaper portal and the linguistic portal are based on different retrieval engines, in order to respond to the relative requirements. The newspaper portal relies on Lucene/SOLR which allows for the efficient retrieval of plain text and facetted searches based on metadata.

The linguistic portal relies on the openCWB which provides support for linguistic annotations on token level and a powerful query processor which allows for Regex and Boolean searches. Transformations towards the required input data formats have been handled by custom scripts.

7 Conclusion

This article reported on the design and development of the OPATCH corpus platform. Based on two usage scenarios for different target groups, relevant considerations concerning requirements towards a comprehensive corpus infrastructure

[8] today, texts in standard fonts yield OCR accuracies of 90% (Kettunen et al., 2014; Kettunen, 2015)

[9] experimentations carried out within *Europeana*

[10] from database for South Tyrolean place names, http://www.uibk.ac.at/germanistik/fachbereiche/germanistische_linguistik/forschung_flurnamen.html

[11] list of South Tyrolean names, cf. Strickner (2011)

[12] taken from historical address books, 1911-1922

have been illustrated and the technical solutions chosen in OPATCH have been presented.

References

Andrea Abel, Stefanie Anstein, and Stefanos Petrakis. 2009. Die Initiative Korpus Südtirol. In *Linguistik Online*, vol. 38, no. 2, 2009.

Stefanie Anstein, Margit Oberhammer, and Stefanos Petrakis. 2011. Korpus Südtirol - Aufbau und Abfrage. In A. Abel & R. Zanin (eds.), *Korpora in Lehre und Forschung*. Bozen - Bolzano: University Press, 15-28.

Julian Brooke, Adam Hammond, and Graeme Hirst. 2015. GutenTag: an NLP-driven Tool for Digital Humanities Research in the Project Gutenberg Corpus. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, held at NAACL HLT 2015, 42-47.

Chris Culy and Verena Lyding. 2010. Double Tree: An Advanced KWIC Visualization for Expert Users. In *Information Visualization, Proceedings of IV 2010, 14th International Conference Information Visualization*, 98-103.

Michel Généreux, Egon Stemle, Lionel Nicolas, and Verena Lyding. 2014. Correcting OCR Errors for German in Fraktur Font. In *Proceedings of the First Italian Conference on Computational Linguistics (CLiC-It 2014)*, edited by R. Basili, A. Lenci & B. Magnini, Pisa, Italy.

Kimmo Kettunen, Timo Honkela, Krister Lindén, Pekka Kauppinen, Tuula Pääkkönen, and Jukka Kervinen. 2014. Analyzing and Improving the Quality of a Historical News Collection using Language Technology and Statistical Machine Learning Methods. In *IFLA World Library and Information Congress Proceedings: 80th IFLA General Conference and Assembly*.

Kimma Kettunen. 2015. Keep, Change or Delete? Setting up a Low Resource OCR Post-correction Framework for a Digitized Old Finnish Newspaper Collection, presented at the *11th Italian Research Conference on Digital Libraries - IRCDL 2015,* Bozen-Bolzano, Italy, 29-30 January, 2015, http://ircdl2015.unibz.it/papers/paper-01.pdf

Verena Lyding, Claudia Borghetti, Henrik Dittmann, Lionel Nicolas, and Egon Stemle. 2013. Open Corpus Interface for Italian Language Learning. In *Proceedings of ICT4LL 2013, 6th edition of the ICT for Language Learning Conference*. libreriauniversitaria.it

Giovanni Moretti, Sara Tonelli, Stefano Menini, and Rachele Sprugnoli. 2014. ALCIDE: An online platform for the Analysis of Language and Content In a Digital Environment. In *Proceedings of the First Italian Conference on Computational Linguistics (CLiC-It 2014)*, edited by R. Basili, A. Lenci & B. Magnini, Pisa, Italy.

Stefan Pletschacher, Christian Clausner, and Apostolos Antonacopoulos. 2014. *Performance Evaluation Report of Europeana Newspapers*, *A Gateway to European Newspapers Online*, D3.5, http://www.europeana-newspapers.eu/wp-content/uploads/2015/05/D3.5_Performance_Evaluation_Report_1.0.pdf

Martin Reynaert. 2014. On OCR ground truths and OCR post-correction gold standards, tools and formats. In *Proceedings of Digital Access to Textual Cultural Heritage, Datech 2014*. New York: ACM, 159-166.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Sieglinde Strickner. 2011. *Nachnamen in Südtirol 2010*. Autonome Provinz Bozen-Südtirol, Landesinstitut für Statistik – ASTAT. Bozen 2011.

Katrin Wisniewski, Andrea Abel, and Verena Lyding. 2015. The MERLIN platform: exploring CEFR-related learner texts. Software demo presented at the *Third Learner Corpus Research Conference*, Nijmegen, 11-13 Sept. 2015, In *LCR 2015 Book of Abstracts*, 172-174.