CHAPTER 32

@PhilosTEI: Building Corpora for Philosophers

Arianna Betti[a], Martin Reynaert[c,d] and Hein van den Berg[a,b]

[a]Axiom Group/University of Amsterdam, [b]Vrije Universiteit Amsterdam, [c]TiCC/Tilburg University, [d]CLST/Radboud University Nijmegen, The Netherlands

ABSTRACT

For philosophers to be able to take a computational turn in their field, especially if that field relies heavily on historical material, it is crucial to be able to build high-quality, easily and freely accessible corpora in a sustainable format composed from multi-language, multi-script books from different historical periods. At the moment, corpora matching these needs are virtually non-existent. Within the CLARIN-NL project @PhilosTEI, we have addressed the problem of building this kind of corpora by developing an open-source, web-based, user-friendly workflow from textual images to TEI, based on state-of-the-art open-source OCR software Tesseract, and a multi-language version of TICCL, a powerful OCR post-correction tool. We have demonstrated the utility of the @PhilosTEI tool by applying it to a multilingual, multi-script corpus of important 18th to 20th century European philosophical texts.

32.1 Introduction

The main objective of the CLARIN-NL project @PhilosTEI was to develop a web-based, user-friendly workflow from scanned images of text to TEI (Text Encoding Initiative) (Betti and van den Berg, 2014b).[1] The workflow in question integrates state-of-the-art open-source OCR (Optical Character Recogniton) software Tesseract and a multi-language version of TICCL, a powerful OCR post-correction tool developed by Martin Reynaert at Tilburg University (Reynaert, 2010). Through building @PhilosTEI, we address a major challenge faced by researchers in philosophy and the digital humanities today: the lack of existing high-quality, easily accessible corpora.

---

[1] In this chapter, we sometimes lift text from the (unpublished) (Betti and van den Berg, 2014b).

How to cite this book chapter:

Betti, A, Reynaert, M and van den Berg, H. 2017. @PhilosTEI: Building Corpora for Philosophers. In: Odijk, J and van Hessen, A. (eds.) *CLARIN in the Low Countries*, Pp. 379–392. London: Ubiquity Press. DOI: https://doi.org/10.5334/bbi.32. License: CC-BY 4.0

@PhilosTEI is meant to offer a user-friendly way to transform images of texts into a machine-readable format, and thus provides researchers with an easy way to build exactly the kind of corpora they need. In particular, the machine-readable format delivered by @PhilosTEI, TEI XML[2] (Burnard and Bauman, 2007), is today's standard for digital editions, and is also a most suitable format to make texts ready for further digital exploration.

In Section 32.2 we describe related work and discuss the problems researchers in philosophy and history of ideas face when building philosophical corpora. In Section 32.3 we describe currently available solutions for building the corpora in question, while also highlighting some of the shortcomings of these solutions. Section 32.4 describes @PhilosTEI, the tool we have developed in order to help philosophers to build the corpora they need. An evaluation of @PhilosTEI is given in Section 32.5, and we discuss future work and challenges in Section 32.6.

32.2 The Problem: Building Philosophical Corpora

Computational tools and methods have significantly impacted philosophical research (van den Berg et al., 2014; Ess, 2004).[3] Many different computing technologies have been applied within the field of logic (Barwise and Etchemendy, 1998); philosophers have been involved in the field of computer ethics (Bynum, 2001), and computing technologies have influenced and changed specific philosophical disciplines, such as epistemology, philosophy of science, and metaphysics (Bynum and Moor, 1998). Finally, the last few years have seen a discipline called 'philosophy of information' blossom - which comprises the study of the application of computing technologies to philosophy (Floridi, 2011; van den Berg et al., 2014; Ess, 2004).

A number of philosophers are involved in computational methods in a somewhat different way, that is, they apply computational tools to study (large amounts of) textual material. For example, Overton (2012) has applied text mining techniques to scientific articles in order to philosophically explore the phrase 'explain', while Herbelot et al. (2012) have analysed phrases such as 'man' and 'woman' from the perspective of gender theory by applying distributional techniques to Wikipedia. Formal ontologies aiding philosophical research have been constructed on the basis of the *Stanford Encyclopedia of Philosophy* and writings of Wittgenstein (Buckner et al., 2011; Pichler and Zöllner-Weber, 2013; Pasin et al., 2008).[4] There are also examples of applications of ontologies to the history of philosophy (e.g. to texts by Kierkegaard and Schelling (McKinnon, 1977; Ziche et al., 2014), the field on which we focus in the present chapter.

A crucial prerequisite to apply computational tools and methods to textual material in the way just mentioned is access to high-quality corpora (van den Berg et al., 2014). The work cited in the previous paragraph is produced by researchers who have access to suitable corpora, but the vast majority of researchers do not have such access. A particularly disadvantaged group are historians of philosophical ideas, i. e. researchers working with (massive amounts of) philosophical texts stretching across centuries, written by many different authors in multiple languages and printed in a variety of scripts. Philosophers working with a history-of-ideas approach face harder challenges than historians of philosophy concentrating on works by only one author, one language, one short period, or even one work.

---

[2] `http://www.tei-c.org/`

[3] We here partly follow (van den Berg et al., 2014): the reader can consult this article for more information on the methods and tools used by historians of ideas. On these topics, see also (Betti and van den Berg, 2014a).

[4] Another ontology aiding philosophical research is given by (Grenon and Smith, 2009).

The humanities researchers within @PhilosTEI (the 'Axiom group/Concepts in Motion' at the University of Amsterdam[5]) trace shifts of meaning of philosophical concepts such as *truth*, *explanation*, and *life* by studying texts in multiple languages published from the 17th up to the 20th century (Betti and van den Berg, 2014a,b). These researchers are keenly aware that their work would benefit significantly from the use of computational methods, especially natural language processing, text-mining, and machine learning. However, currently these methods can only be applied in an extremely limited way due to the lack of high-quality philosophical corpora in a sustainable and suitable open format.[6]

32.2.1 Availability of Historical Texts and Use Restrictions

In the last decades there have been several attempts to create high-quality digital editions of historical texts. For example, the Thesaurus Linguae Graecae (TLG)[7] is a digital library of Greek literature, providing access to all extant Greek texts from Homer to AD 600, thus including many ancient philosophers. Similarly, the Perseus Project[8] provides online access to works of ancient philosophers such as Plato and Aristotle. The famous Index Thomisticus provides a complete lemmatization of the works of Saint Thomas Aquinas.[9] The Bonner Kant-Korpus[10] provides an online and searchable digital edition of the complete works of Immanuel Kant, whereas the project Transcribe Bentham[11] provides digital images and transcriptions of the writings of Jeremy Bentham. Other historical works, including those of Nietzsche and Wittgenstein, have been made available within the PhiloSource federation.[12] There are, moreover, commercial companies that sell CD-ROMs or downloads of the works of famous philosophers such as Spinoza, Leibniz, Husserl, and others.[13]

Importantly, many among these editions are not in open access, so their use within digital philosophy projects is severely limited. This applies to e.g. commercial electronic editions, and to the content provided by the TLG. The TLG materials are copyrighted: users can browse and search the TLG but are not allowed to download material. Similarly, users can browse but not download the contents of the Bonner Kant-Korpus. Obtaining a licence for scholarly use of materials is sometimes possible, but one should keep in mind that the threshold for endeavouring to obtain appropriate licences might be too high for many users, and certain publishers will not give any licences.

32.2.2 Multilinguality, Minor Authors and Machine-Unreadability

Historians of ideas need to be able to build corpora of texts written in diverse languages such as Latin, German, French, Dutch, Polish, and English; they also need suitable editions of many works written by relatively unknown or 'minor' philosophers, whose works are not digitally available (Betti et al., 2014); finally, and most importantly, even many texts by known and important thinkers are not yet digitized in a way suitable for computational exploration. Our Axiom Group/Concepts

---

[5] `http://www.axiom.humanities.uva.nl/`

[6] For an example of what this group achieved by constructing and applying a simple text-mining tool on a single, albeit long (2,000 pages) text of reasonable quality, see (van Wierst et al., 2016).

[7] `http://stephanus.tlg.uci.edu/index.php`

[8] `http://www.perseus.tufts.edu/hopper/`

[9] `http://www.corpusthomisticum.org/`

[10] `https://korpora.zim.uni-duisburg-essen.de/kant/`

[11] `http://blogs.ucl.ac.uk/transcribe-bentham/`

[12] `http://www.discovery-project.eu/philosource.html`

[13] `https://www.infosoftware.de/index.htm`

in Motion philosophers study the works of e. g. Bernard Bolzano (1781–1848) and Gottlob Frege (1848–1925). Digitizations of a number of these philosophers' works are scattered across different repositories, such as Google Books, Hathi Trust, and Europeana, and researchers can often only download low-quality, scanned PDF images of original printings in Gothic typefaces (Betti and van den Berg, 2014b; van den Berg et al., 2014). Texts of this kind are unsuitable for minimally sophisticated computational exploration.

The challenges above are faced by historians of ideas and other researchers working in text-based digital humanities alike. For printed texts, Gothic typefaces – also known as Fraktur or blackletter – emerged in the 16th century and have been widely used up into the 20th century.[14] Hence, many historians work with material of this kind. Researchers working with texts published in different languages, typefaces, and formats would also profit from having a simple tool that allows them to create high-quality, easily and freely accessible corpora in a sustainable format. The purpose of @PhilosTEI was to develop such a tool.

32.3 Existing OCR Solutions to Building Corpora, and their Shortcomings

The main challenge faced by historians of philosophical ideas who wish to build corpora is how to transform scanned PDFs of texts printed in historical typefaces such as Gothic into a machine-readable format. In other words, the main challenge is how to perform automatic image-to-text conversion or OCR on images of texts written in a Gothic typeface (Furrer and Volk, 2011), such as depicted in Figure 32.1.

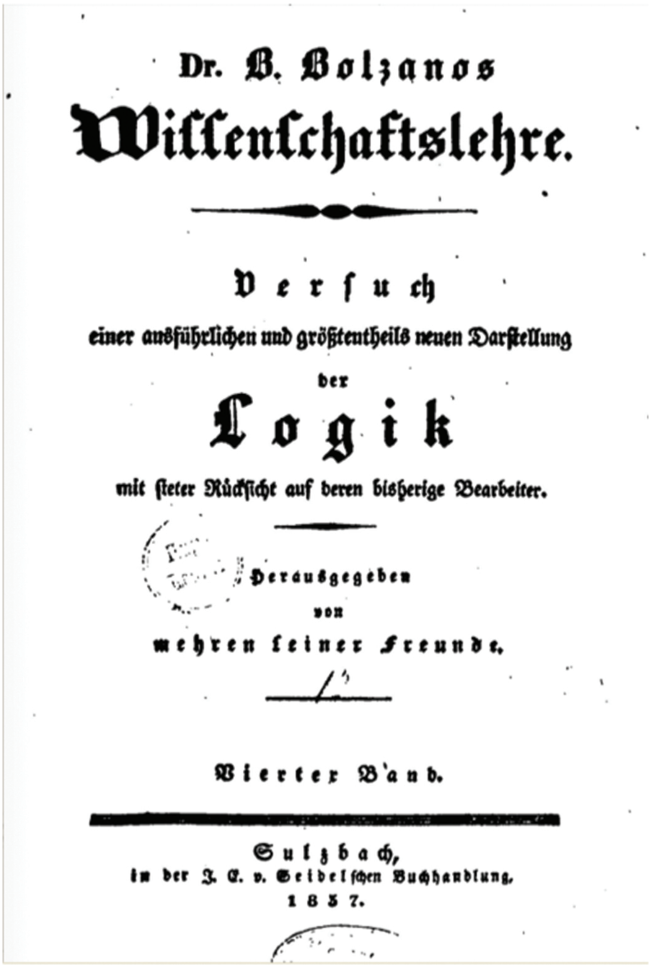
Historians who wish to OCR images of texts typeset in Gothic currently have a limited number of options. The first option is to use the OCR software developed by ABBYY (Fuchs, 2016). The downside to using ABBYY is that the software is not open-source and that users have to pay a specified amount of money for OCRing volumes of texts. At the time of writing of this chapter, the pricing for ABBYY Recognition Server 4 with Gothic/Fraktur was 999 euros for 50,000 pages, which equals approximately 160 digitised books, and might be considered a corpus of acceptable size for a single-researcher project in the history of ideas. Though not prohibitive, this is still a significant cost as it equals two thirds of the travel money for an entire year for a senior researcher in a Humanities Faculty in the Netherlands.

A second option for historians is to use the open source OCR engine Tesseract. The advantage of using Tesseract is that it is free. In addition, the quality of the OCR output seems to be comparable to that of the output of the ABBYY software. In a 2012 report conducted within the IMPACT project, the quality of the output of ABBYY Finereader and the quality of Tesseract were deemed to be relatively similar (Heliński et al., 2012). However, Tesseract also has its downsides. Setting up Tesseract properly is highly challenging. Tesseract would benefit from proper documentation as it comes with over 2,000 option settings that are impenetrable to researchers in the humanities who lack a strong technical background. Without appropriate setup and training, the quality of Tesseract OCR on diverse historical material (different languages, periods, and scripts) remains unsatisfactory (Betti and van den Berg, 2014b).

At the time of drafting our @PhilosTEI proposal (September 2012) an open-source alternative to Tesseract, namely OCRopus (Breuel, 2008) seemed set to help overcome Tesseract's drawbacks, while in fact it incorporated the system. In the following section we explain why in the end we did not follow this path.

---

[14] `https://en.wikipedia.org/wiki/Fraktur`



Dr. B. Bolzanos
Wiſſenſchaftslehre.

Verſuch
einer ausführlichen und größtentheils neuen Darſtellung
der
Logik
mit ſteter Rückſicht auf deren bisherige Bearbeiter.

Herausgegeben
von
mehren ſeiner Freunde.

Vierter Band.

Sulzbach,
in der J. E. v. Seidelſchen Buchhandlung.
1837.

Figure 32.1: Cover of philosopher Bolzano's work *Wisschenschaftslehre*, printed in Fraktur.

32.4 The Solution: @PhilosTEI

In order to provide philosophers with an easy way to build philosophical corpora, and thus to solve some of the problems mentioned in the previous section, we have developed the web-based demonstrator tool @PhilosTEI.[15] This tool provides researchers with a free, open-source workflow from (scanned) images of texts to TEI, which is today's standard for digital editions. The system is easy to use and was developed in such a way that users with little technical knowledge can use the workflow. The system performs automatic OCR error post-correction on output delivered by the OCR software built in the workflow, namely Tesseract, in order to improve the quality of the output. It comes with at least basic lexica and provisions for 18 languages and diachronic language varieties.

---

[15] In this section, we draw on (and occasionally lift sentences from) the description of the TICCLops system given in (Reynaert, 2014c) and on the description of the @PhilosTEI system in (Betti and van den Berg, 2014b).

The tool is currently hosted by the Institute for the Dutch Language (INT).[16] When visiting the online system, the user sees the main interface page (Figure 32.2). There, the user can upload input files for a transcription project. These can be e.g. scanned images in PDF, TIFF or DjVu (Betti and van den Berg, 2014b; Reynaert, 2014c). The user needs to specify the language of the original text (necessary for having Tesseract load the appropriate training files for the language) and needs to provide the project with a name (to help the user locate the output for their own project). Advanced users can select a number of further options, which allows them, for example, to choose to modernise spelling to contemporary or to the original diachronic spelling on the basis of the specific lexicon selected or to select how many ranked post-correction variants TICCL will return (Figure 32.3).

When the user presses the 'Process files' button, the system runs and eventually provides different kinds of output. These include, most importantly, the OCR output and the OCR output as corrected by the OCR post-correction system that is an integral part of @PhilosTEI (Betti and van den Berg, 2014b). Users are provided with a reader that allows them to visually compare the original image files with the fully-automatically corrected OCR output (Figure 32.4).

Of necessity, the @PhilosTEI is internally far more complex than will ever be apparent to its users. Tesseract, to start with, delivers its output in hOCR HTML format (Breuel, 2007). This is converted to FoLiA XML (van Gompel and Reynaert, 2013) before being delivered to the OCR post-correction system we describe next.

The CLARIN-NL Call 1 project TICCLops delivered the Text-Induced Corpus Clean-up system (or TICCL) as a web application and service thanks to the development of the Computational Linguistics Application Mediator (or CLAM) (van Gompel and Reynaert, 2014), a generic solution for turning linguistic applications with a command-line interface into web applications and RESTful services.

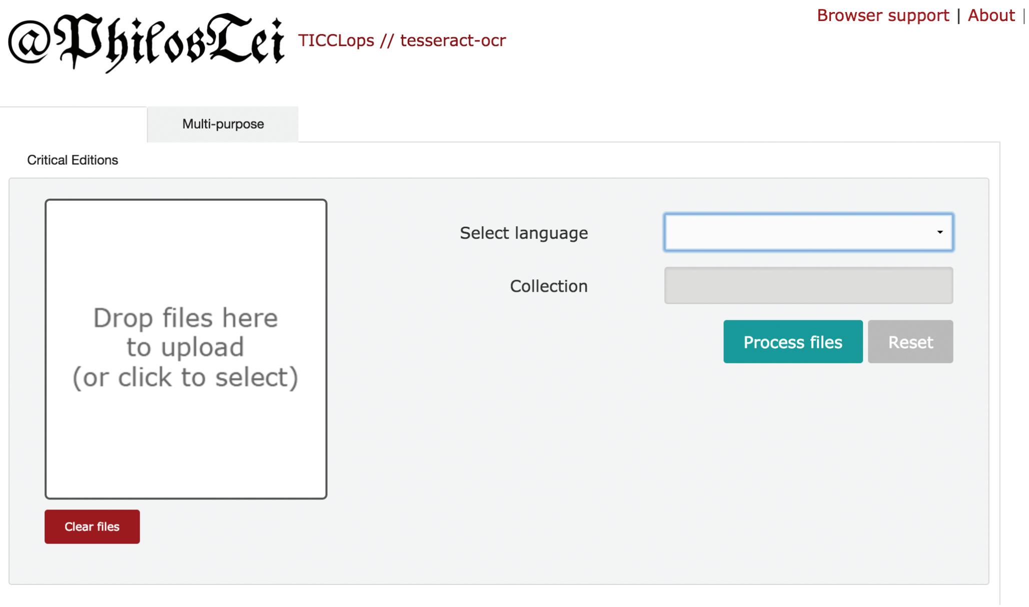
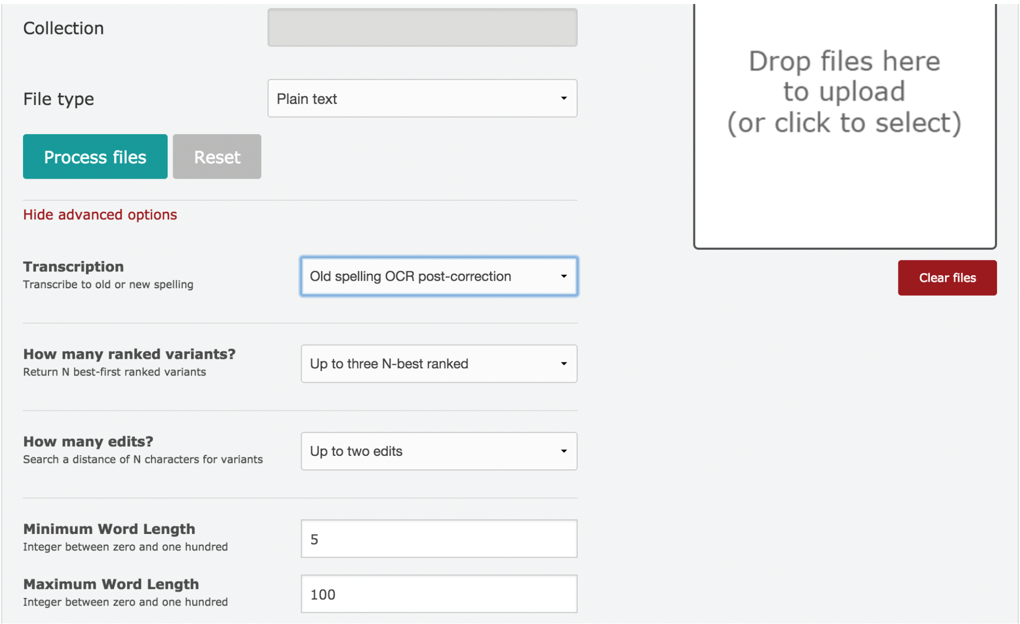

@PhilosTei
TICCLops // tesseract-ocr
Browser support | About |
Multi-purpose
Critical Editions
Drop files here to upload (or click to select)
Clear files
Select language
Collection
Process files
Reset

Figure 32.2: Main interface page.

---

[16] `http://ticclops.clarin.inl.nl/philostei/`



Collection

File type: Plain text

Transcription
Transcribe to old or new spelling: Old spelling OCR post-correction

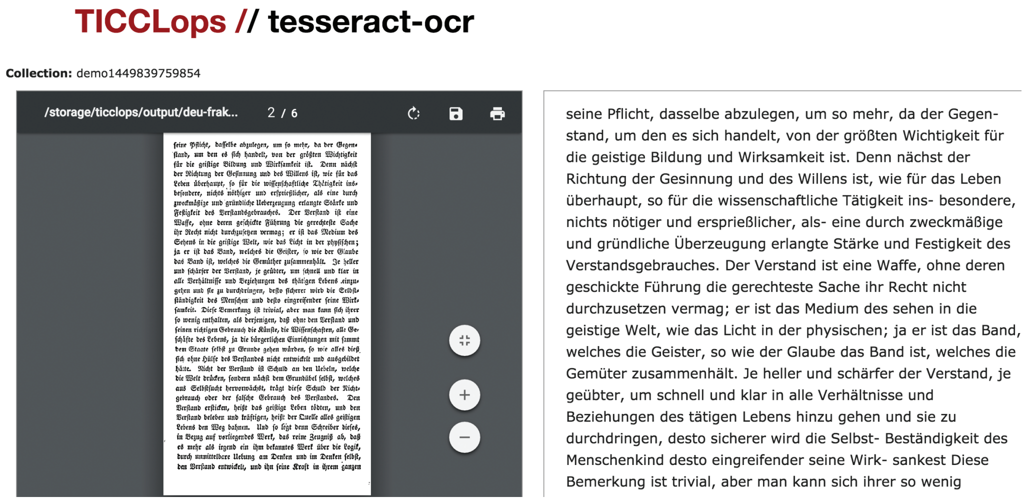
How many ranked variants?
Return N best-first ranked variants: Up to three N-best ranked

How many edits?
Search a distance of N characters for variants: Up to two edits

Minimum Word Length
Integer between zero and one hundred: 5

Maximum Word Length
Integer between zero and one hundred: 100

Figure 32.3: Interface page for advanced users.


TICCLops // tesseract-ocr
/storage/ticclops/output/deu-frak...
2 / 6

Collection: demo1449839759854

seine Pflicht, dasselbe abzulegen, um so mehr, da der Gegenstand, um den es sich handelt, von der größten Wichtigkeit für die geistige Bildung und Wirksamkeit ist. Denn nächst der Richtung der Gesinnung und des Willens ist, wie für das Leben überhaupt, so für die wissenschaftliche Tätigkeit ins- besondere, nichts nötiger und ersprießlicher, als- eine durch zweckmäßige und gründliche Überzeugung erlangte Stärke und Festigkeit des Verstandsgebrauches. Der Verstand ist eine Waffe, ohne deren geschickte Führung die gerechteste Sache ihr Recht nicht durchzusetzen vermag; er ist das Medium des sehen in die geistige Welt, wie das Licht in der physischen; ja er ist das Band, welches die Geister, so wie der Glaube das Band ist, welches die Gemüter zusammenhält. Je heller und schärfer der Verstand, je geübter, um schnell und klar in alle Verhältnisse und Beziehungen des tätigen Lebens hinzu gehen und sie zu durchdringen, desto sicherer wird die Selbst- Beständigkeit des Menschenkind desto eingreifender seine Wirk- sankest Diese Bemerkung ist trivial, aber man kann sich ihrer so wenig

Figure 32.4: Original page image and corrected text output.

One of the great advantages of using a system such as CLAM is that developers get the opportunity to select the linguistic application's features and parameters they choose to confront the web application's users with or to shield users from.

TICCL's availability as a web application and service made it the natural OCR post-correction system of choice. In the @PhilosTEI project we set out to expand the existing TICCLops system (Reynaert, 2014c) with OCR facilities. Our first ideas for the OCR-engine went to OCRopus (Breuel, 2008), but preliminary tests did not deliver the necessary results. In fact, the latest version of OCRopus no longer incorporated Tesseract, and very few trained models for its new OCR-engine were as yet available. This meant that our workflow building project might get side-tracked into an OCR-engine training project, and this we could not afford. We soon found out that the Andrew W. Mellon Foundation-funded Early Modern OCR Project[17] had project plans very similar to our own and much the same experience with the system. In consultation with eMOPS at the Dutch National Library KB, we too redirected our attention to the open source OCR solution Tesseract and co-opted it for our own pipeline. Tesseract's broad range of pre-trained languages made it perfectly suited for our purposes.

In @PhilosTEI we incorporated a totally new implementation of TICCL. This now consists of a series of C++ modules which are wrapped in a Perl script which takes care of a large part of the file handling peculiarities. TICCL emphatically does not handle a single page of text of a single document at a time; rather it derives the unigram frequency list of the full batch of documents to be corrected, uses a range of derived statistics to rank its list of correction candidates for the focus words and delivers a list of focus words paired with their ranked correction candidates.

TICCL was further made multilingual in the sense that it was equipped with available open source spelling dictionaries for 18 languages. From these, per language, a language-specific alphabet is first derived on the basis of the dictionary's character frequency list. In order to restrain the spelling variation search space, the characters below an ad-hoc frequency threshold are in essence disregarded. The alphabet is next used to precalculate the anagram hash values for the character confusions possible up to a particular Levenshtein distance (LD) (Levenshtein, 1966), in practice LD 2, given the characters available. Given the 25 characters in the smallest alphabet in our language selection, that for Latin, this gives 72,009 character confusion values. For the largest alphabet, i.e. Classical Greek with 100 characters, this amounts to 13,802,616 values. It should be noted that the actual alphabet is supplemented with two extra wildcard values, one encoding for any punctuation mark, another for all the characters not deemed to belong to the particular language. This has the handy result that if - say - a Latin text has Greek words in it, the Greek words will get an anagram value which is equal to the wildcard value times its length in characters. This in effect puts foreign words in another script automatically and neatly out of reach of TICCL's lexical variant retrieval mechanism for the given language.

On the basis of the character confusion anagram value list, TICCL performs an efficient exhaustive search for all the word string pairs in the corpus at hand that differ by no more than the LD threshold set.

The two main components of @PhilosTEI are integrated into CLAM in an extended Perl wrapper which encompasses both Tesseract and TICCL and the further assistive components such as convertors for the diverse page image formats that are supported, such as TIFF, JPG and PDF. The wrapper allows for flexible handling of numbers of input/output files, taking e.g. x PDF input files for separate book chapters apart into y (where $y \geq x$) image files, one per page, to be sent to the OCR engine Tesseract, then presenting the y OCRed files as a single batch to TICCL, which eventually corrects the y FoLiA XML page files to be collated into a single output FoLiA XML book and also, as the philosopher-user desires, a TEI XML output e-book, using TEI Lite.

32.5 Evaluation

Within @PhilosTEI, we have conducted two types of evaluation: (a) a quantitative evaluation of the OCR post-correction tool TICCL, and (b) a qualitative user evaluation of the web-based workflow from textual images to TEI. Below we present the main results of these evaluations.

---

[17] `http://emop.tamu.edu/`

32.5.1 Quantitative Evaluation of TICCL

The @PhilosTEI project benefitted from a large-scale quantitative evaluation of TICCL undertaken in part in the framework of the NWO 'Groot' project Nederlab as described in (Reynaert, 2014b). As was described in its companion paper (Reynaert, 2014a), undertaking an evaluation of an OCR-post-correction system on the scale of even a single OCRed book is an extremely expensive and labour-intensive undertaking. The scale and scope of @PhilosTEI did not permit for this to be undertaken on a full philosophical work.

The evaluation on the Gold Standard book reported that TICCL improved the accuracy of the OCRed historical text by 5.5%, from 88.94% to 94.51%.

A new evaluation (Reynaert, 2016) on 1,000 randomly chosen word strings of the same post-correction of 10,333 Dutch mostly late-18th-century books has shown that the extremely high correction precision scores of over 99% reported on the single history book written for children are, as was to be expected, not obtained throughout the whole collection. However, at over 84% for fully automatic correction on the random sample of the whole collection, precision remains good. The score on recall, 35%, means one in three errors are fully-automatically corrected. Almost half of the errors are corrected when one takes into account the ten best-ranked correction candidates. This new evaluation points clearly towards necessary and possible future extensions of TICCL. These should finally allow for meaningful noisy-text improvement to be achieved fully automatically.

32.5.2 Qualitative User Evaluation of the @PhilosTEI Workflow

The qualitative user evaluation of the workflow from textual images focused on two main criteria: (i) user-friendliness, and (ii) quality of the output.[18]

Ad (i): Humanities users often have little experience with using computational tools for research purposes, and have little or no technical background knowledge. It is thus vital that the workflow be intuitive and very easy to use, and presupposes as little technical knowledge as possible.

During the development of the workflow, six users have provided evaluations by using a Google document template prepared by the user with the most extensive experience with this kind of testing (Hein van den Berg). The evaluations were collected at three different phases of development: after the development of the first user interface (July 2014, three researchers), after the development of the second user interface (October 2014, two researchers, one student), and after completion of the project (October 2015, students). The first batch of evaluations led to the development of a new interface. The first – classical CLAM – interface that was developed provided users with a lot of configuration options. Inexperienced users felt there were too many configuration options, which made the interface unintuitive to use. Based on this feedback, we built a new interface, presented in Section 32.4, enabling the user to select one out of two configuration methods. With this interface, novice users have no configuration options, and expert users have a fair range of configuration options.

The new interface has been evaluated very positively by the users within the project who had performed the first batch of evaluations. The experience of these users with many research tools produced within computational projects is that such tools are unappealing to use. By contrast, the new interface resembles the design of online tools for a large public that is typically found attractive and pleasant to interact with.

The evaluation of the two students who tested the new interface after completion of the project was less positive. Importantly, the students had no knowledge of the previous interface, and

---

[18] In this section, we draw on the description in (Betti and van den Berg, 2014b), which provides a more comprehensive qualitative evaluation than can be given in the current chapter.

had a different goal, namely identifying as quickly as possible the best tool available for building a highest-quality corpus for a text-mining project,[19] consisting, crucially, of modern English texts containing a huge amount of logico-mathematical formulas. With respect to other OCR tools available on the market, the main shortcoming was deemed to be usability (including e.g. the lack of an interactive correction panel). The choice of the students fell on ABBYY Finereader, a choice that posed additional institutional challenges with commercial licensing for the students involved, resulting in a great amount of institutional red tape.

Ad (ii): The two more experienced users from the project team have also evaluated the quality of the output. They converted scanned images of the texts mentioned in Table 32.1 below and evaluated both the OCR output and the TICCL output on how well the output matches the original text. What counts as 'good performance' has been evaluated on the basis of an intuitive measure of what our colleagues would consider useful results.

In general, the quality of the OCR and TICCL output was evaluated as having reasonably high quality, with ample opportunity for improvement. The OCR output for samples of the German texts (Bolzano, 1837; Frege, 1879) contains several spelling mistakes, and the OCR engine cannot handle end-of-line splits. TICCL also cannot yet handle end-of-line splits. In addition, TICCL currently does not correct words that contain multiple OCR-induced errors beyond the edit distance limit of 2 edits imposed and sometimes introduces incorrect changes by itself. The quality of the Polish OCR and TICCL output (Tarski, 1936) is comparable to that of the German. The quality of the Latin OCR output (Wolff, 1740) was not as good as that of the German or Polish. In general, each line of text contained multiple mistakes. TICCL corrects several of these mistakes, but, again, not all of them, and it also introduced errors of its own.

32.5.3 Discussion of the Evaluations

There is a major drawback of TICCL's incorporation in the @PhilosTEI web application that was insufficiently addressed during its development. This is that the online system on the whole is geared towards processing only a single book at a time. As TICCL's evaluations described in Reynaert (2014b) clearly show on the basis of comparison of the correction of a single book in isolation versus the correction of the same book within a batch of 10,333 books of the same era, the tool's performance on the isolated single book is far inferior to that on the full batch. This is due to the far poorer and sparser word string statistics obtainable from just a single book.

This is to be remedied in the current follow-up project PICCL (Reynaert et al., 2015) within the CLARIAH programme. We describe this future work in the next section.

The users in the project find it crucial to stress that improving TICCL's usability and performance would have important advantages over commercial solutions, namely cutting users' costs and institutional red tape to zero.

	Language	**Typeface**	**Period**	**Format**	**Location**
Wolff (1740)	Latin	Roman	middle 18th century	PDF	archive.org
Bolzano (1837)	German	Gothic	middle 19th century	PDF	dml.cz
Frege (1879)	German	Roman	last third 19th century	PDF	Gallica
Tarski (1936)	Polish	Roman	middle of 20th century	DjVu	TEL/Europeana

Table 32.1: Overview of philosophers' works, dates, print typefaces, languages, periods and source locations used in the @PhilosTEI system evaluations.

---

[19] `https://quine1960.wordpress.com/the-quine-in-context-project/`

32.6 Future Work and Challenges

In 2016–2017 in the PICCL or 'Philosophical Integrator of Computational and Corpus Libraries' project we build further on the foundations laid in the @PhilosTEI project and expand the work flow into a full-fledged corpus building pipeline.

Next to the image input already catered for, we are to incorporate the necessary convertors for all manner of electronic text formats. We will revisit the current OCR and OCR-post-processing scene and see whether OCRopus or perhaps new approaches to OCR are now contenders for the Tesseract engine. In emulation of Volk et al. (2010), we will strive towards combining multiple OCR versions for the same works and let TICCL sort out statistically which of close but divergent renderings of the same word tokens – the assumption being that different OCR engines will produce differing results – best fit the actual text.

For TICCL, one pathway we will obviously follow is to allow the user to also provide the system with the frequency list obtained from contemporary and where possible comparable works, whether well-edited hand-keyed transcriptions or noisy OCRed versions. Users will likewise be able to furnish the system with domain-specific lexicons and name lists at their disposal. In the short run, TICCL is to be extended with word bigram information to allow for addressing split and run-on word errors and short word forms.

In order to allow the digital humanities scholar to obtain the best possible text result, we will provide solutions geared at manual and interactive text correction. This will be based on FLAT,[20] the 'FoLiA Linguistic Annotation Tool', which is a modern web application that offers an interface for the visualisation and editing of FoLiA documents. We will also see if some of the tools developed in the Impact[21] project – Aletheia[22] for text segmentation and PoCoTo[23] for OCR post-correction are interesting prospects, for example – may likely be enlisted.

The pipeline will next provide linguistic enrichments in the form of annotations for lemmata, part-of-speech and named entities towards more fine-grained exploration and analysis of the texts. To this end, the memory-based tool Frog[24] will be made part of the pipeline. In fact, we aim towards integrating all the available FoLiA XML tools[25] in PICCL.

Finally, indexing towards online availability in a corpus exploration and exploitation environment[26] will be provided.

In short, PICCL is about choosing the best possible tools currently available, wrapping them all in an environment that allows non-experts to nevertheless harness their contribution and making this environment freely and openly available to all.

32.7 Conclusion

The @PhilosTEI project has been successfully completed and has managed to bring its philosopher-users and its more technically directed developers closer together and more aware of each other's needs and limitations. While the system we built does not to-date allow for unfettered, fully-automatic corpus building, it does allow non-technical people to convert mere text images into electronic text presented in state-of-the-art corpus formats ready for further manual editing.

---

[20] `https://github.com/proycon/flat`

[21] `http://www.digitisation.eu`

[22] `http://www.primaresearch.org/tools`

[23] `https://github.com/cisocrgroup/PoCoTo`

[24] `https://languagemachines.github.io/frog`

[25] See Chapter 6 on FoLiA in this volume.

[26] See Chapter 19 on WhiteLab in this volume.

Some of the hurdles which have not been overcome yet are hoped to be overcome by follow-up project PICCL, which is currently underway in CLARIAH.

Acknowledgements

The authors acknowledge having been funded by CLARIN-NL project @PhilosTEI. Martin Reynaert further acknowledges being funded by the new Dutch national CLARIN project CLARIAH and by NWO in project Nederlab. Hein van den Berg further acknowledges being funded by Europeana eCloud (325091).

References

Jon Barwise and John Etchemendy. 1998. *Computers, Visualization, and the Nature of Reasoning*. Blackwell Publishers: Oxford.

Arianna Betti and Hein van den Berg. 2014a. Modelling the History of Ideas. *British Journal for the History of Philosophy*, 22(4): 812–835.

Arianna Betti and Hein van den Berg. 2014b. @PhilosTEI: Final user evaluation report. Technical report, Amsterdam, November.

Arianna Betti, Dirk Gerrits, Bettina Speckmann, and Hein van den Berg. 2014. Glammap: visualising library metadata. In *Proceedings of VALA* 2014 – Libraries, Technologies, and the Future: Melbourne, Australia.

Bernard Bolzano. 1837. *Wissenschaftslehre. Versuch einer ausführlichen und größtentheils neuen Darstellung der Logik mit steter Rücksicht auf deren bisherige Bearbeiter*. J.E. v Seidel, Sulzbach.

Thomas Breuel. 2007. The hOCR microformat for OCR workflow and results. In ICDAR '07 - *Proceedings of the Ninth International Conference on Document Analysis and Recognition* 2: 1063–1067. IEEE Computer Society: Washington, DC, USA.

Thomas Breuel. 2008. The OCRopus Open Source OCR System. In B.A. Yanikoglu and K. Berkner, editors, *Proceedings of SPIE 6815, Document Recognition and Retrieval XV, 68150F*. SPIE: San Jose, California, USA.

Cameron Buckner, Mathias Niepert, and Colin Allen. 2011. From encyclopedia to ontology: Toward dynamic representation of the discipline of philosophy. *Synthese*, 182(2): 205–233.

Lou Burnard and Syd Bauman, editors, 2007. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium.

Terrell Ward Bynum and James Moor, editors. 1998. *The Digital Phoenix: How Computers Are Changing Philosophy*. Blackwell Publishers: Oxford.

Terrell Ward Bynum. 2001. Computer ethics: Its birth and its future. *Ethics and Information Technology*, 3(2): 109–112.

Charles Ess. 2004. "Revolution? What Revolution?" Successes and Limits of Computing Technologies in Philosophy and Religion. In S. Schreibman, R. Siemens, and J. Unsworth, editors, *A companion to Digital Humanities*: 132–142. Blackwell Publishers: Oxford.

Luciano Floridi. 2011. *The philosophy of information*. Oxford University Press: Oxford.

Gottlob Frege. 1879. *Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens*. Louis Nebert: Halle.

Michael Fuchs. 2016. White paper. ABBYY historic OCR: the use of Gothic OCR in processing historical documents. http://www.frakturschrift.com/_media/en:white_paper_gothic-fraktur_ocr_e.pdf

Lenz Furrer and Martin Volk. 2011. Reducing OCR errors in Gothic-script documents. In Vertan et al., editor, *Proceedings of the RANLP 2011 workshop on Language Technologies for Digital Humanities and Cultural Heritage*: 97–103. Incoma, Shoumen, Bulgaria.

Pierre Grenon and Barry Smith. 2009. Foundations of an ontology of philosophy. *Synthese*, 182(2): 185–204.

Marcin Heliński, Miłosz Kmieciak, and Tomasz Parkoła. 2012. Report on the comparison of Tesseract and ABBYY Finereader OCR engines. PCSS, Poznań.

Aurélie Herbelot, Eva Von Redecker, and Johanna Müller. 2012. Distributional techniques for philosophical enquiry. In LaTeCH '12 - *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*: 45–54. Association for Computational Linguistics Stroudsburg, PA, USA.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, 10(8): 707–710.

Alastair McKinnon. 1977. From co-occurrences to concepts. *Computers and the Humanities*, 11(3): 147–156.

James A. Overton. 2012. "Explain" in scientific discourse. *Synthese*, 190(8): 1383–1405.

Michele Pasin, Milton Keynes, and Enrico Motta. 2008. PhiloSURFical: Browse Wittgenstein's World with the Semantic Web. In A. Pichler and H. Hrachovec, editors, *Wittgenstein and the Philosophy of Information: Proceedings of the 30th International Ludwig Wittgenstein-Symposium in Kirchberg, 2007*: 319–331. De Gruyter: Berlin.

Alois Pichler and Amélie Zöllner-Weber. 2013. Sharing and debating Wittgenstein by using an ontology. *Literary and Linguistic Computing*, 28(4): 700–707.

Martin Reynaert, Maarten van Gompel, Ko van der Sloot, and Antal van den Bosch. 2015. PICCL: Philosophical Integrator of Computational and Corpus Libraries. In *CLARIN Annual Conference 2015 – Book of Abstracts*: 75–79 CLARIN ERIC: Wrocław, Poland.

Martin Reynaert. 2010. Character confusion versus focus word-based correction of spelling and OCR variants in corpora. *International Journal on Document Analysis and Recognition*, 14: 173–187.

Martin Reynaert. 2014a. On OCR ground truths and OCR post-correction gold standards, tools and formats. In A. Antonacopoulos and K.U. Schulz, editors, DATeCH 2014 - *Proceedings of Digital Access to Textual Cultural Heritage*: 159–166. ACM: New York, NY, USA.

Martin Reynaert. 2014b. Synergy of Nederlab and @PhilosTEI: diachronic and multilingual Text-Induced Corpus Clean-up. In N. Calzolari et al. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. ELRA: Reykjavik, Iceland.

Martin Reynaert. 2014c. TICCLops: Text-Induced Corpus Clean-up as online processing system. In L. Tounsi and R. Rak, *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics System Demonstrations*: 52–56. Dublin City University and Association for Computational Linguistics: Dublin, Ireland.

Martin Reynaert. 2016. OCR post-correction evaluation of Early Dutch Books Online – revisited. In N. Calzolari et al., editor, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC-2016)*. ELRA: Portorož, Slovenia.

Alfred Tarski. 1936. O pojęciu wynikania logicznego. *Przegląd filozoficzny*, 39: 58–68.

Hein van den Berg, Gonzalo Parra, Anja Jentzsch, Andreas Drakos, and Erik Duval. 2014. Studying the History of Philosophical Ideas: Supporting Research Discovery, Navigation, and Awareness. In i-KNOW '14 - *Proceedings of the 14th International Conference on Knowledge Technologies and Data-driven Business*, 12:1–8. ACM: New York, NY, USA.

Maarten van Gompel and Martin Reynaert. 2013. FoLiA: A practical XML Format for Linguistic Annotation - a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, 3: 63–81.

Maarten van Gompel and Martin Reynaert. 2014. CLAM: Quickly deploy NLP command-line tools on the web. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*: 71–75. Dublin City University and Association for Computational Linguistics: Dublin, Ireland.

Pauline van Wierst, Sanne Vrijenhoek, Stefan Schlobach, and Arianna Betti. 2016. Phil@Scale: Computational Methods within Philosophy. In *Proceedings of the Third Conference on Digital Humanities in Luxembourg with a Special Focus on Reading Historical Sources in the Digital Age*, 1681. Aachen: CEUR-WS.org.

Martin Volk, Torsten Marek, and Rico Sennrich. 2010. Reducing OCR errors by combining two OCR systems. In C. Sporleder and K. Zervanou, editors, *Proceedings of the Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010)*: 61–65. Faculty of Science, University of Lisbon: Lisbon, Portugal.

Christian Wolff. 1740. *Philosophia rationalis sive logica*. Officina libraria Rengeriana: Frankfurt and Leipzig.

Paul Ziche, Dirk van Miert, Peter Sperber, Timmy de Goeij, Tom Giesbers, Daniel Meijer, et al. 2014. Mining for associated words in philosophical texts. *Schelling Studien: Internationale Zeitschrift zur Klassischen Deutschen Philosophie*, 2(1): 215–231.