Umanistica Digitale – ISSN: 2532-8816 – n. 11, 2021
D. Del Fante, G. M. Di Nunzio – *OCR Correction for Corpus-assisted Discourse Studies*
DOI: http://doi.org/10.6092/issn.2532-8816/13689

OCR Correction for Corpus-assisted Discourse Studies: A Case Study of Old Newspapers

[1]Dario Del Fante, [2]Giorgio Maria Di Nunzio

[1]Istituto di Linguistica Computazionale "A.Zampolli" - Consiglio Nazionale delle Ricerche, Italy
[2]Università di Padova, Padua, Italy
[1]dario.delfante@ilc.cnr.it
[2]giorgiomaria.dinunzio@unipd.it

Abstract

The use of OCR software to convert printed characters to digital text is a fundamental tool within diachronic approaches to Corpus-assisted discourse Studies. However, OCR software is not totally accurate, and the resulting error rate may compromise the qualitative analysis of the studies. This paper proposes a mixed qualitative-quantitative approach to OCR error detection and correction in order to develop a methodology for enhancing the quality of historical corpora. We applied the developed methodology to two projects on newspapers of the beginning of the 20th century for the linguistic analysis of the metaphors representing migration and pandemics. The outcome of this project consists in a set of rules which are, eventually, valid for different contexts and applicable to different corpora and which can be reproduced and reused. The proposed procedure, in terms of computational readability, is aimed at making more readable and searchable the vast array of historical text corpora which are, at the moment, only partially usable given the high error rate introduced by an OCR software.

L'uso di software di riconoscimento OCR per convertire i caratteri stampati in testo digitale è uno strumento fondamentale per quanto riguarda l'ambito di studio degli approcci diacronici all'analisi del discorso politico attraverso i corpora (CADS studies). Tuttavia, i software OCR non sono totalmente affidabili, e il loro tasso di fallibilità può compromettere l'analisi. Questo articolo propone un approccio qualitativo-quantitativo al rilevamento e alla correzione degli errori post scansione OCR al fine di sviluppare una metodologia per migliorare la qualità dei corpora all'interno degli studi storici. Abbiamo applicato la metodologia sviluppata a due casi di studio su giornali dell'inizio del XX secolo per l'analisi linguistica delle rappresentazioni metaforiche delle migrazioni e delle pandemie. Il risultato di questo progetto consiste in un insieme di regole che sono valide per diversi contesti e applicabili a diversi corpora e che possono essere riutilizzate. La procedura proposta, in termini di leggibilità computazionale, ha lo scopo di rendere più leggibile e ricercabile la vasta gamma di corpora di testi storici che sono, al

momento, solo parzialmente utilizzabili dato l'alto tasso di errore derivante da un software di riconoscimento OCR.

1. Introduction

The method we present in this work has been developed and tailored to a study of figurative language used in newspapers between the beginning 1900s and 2000s. The approach followed in this research study lies within the Diachronic Corpus-assisted Discourse Studies (henceforth D-CADS) framework ([28];[43]). D-CADS can be considered as a set of studies into the form and/or function of language as *communicative discourse* in history, which incorporates the use of computerised corpora in their analyses ([33]: 10). In D-CADS, the qualitative approach of Critical Discourse Analysis is combined with the quantitative techniques of Corpus Linguistics. Corpora provide empirical evidence to the analyst in support of any statement made about language. Therefore, the processes of corpus-design and corpus-compilation have a marked impact on the research structure ([33]). D-CADS research is interested in the study of language change over time ([5]), and the data analysed are usually extracted from old paper documents.[1] Since these documents are rarely available into computer-readable formats, a digital conversion is often required. Given the amount of data, a manual transcription of the documents would not be feasible. For this reason, the employment of an Optical Character Recognition (OCR) software represents an optimal solution to automatise this step ([22]) and to increase the number of usable documents for the analysis of the language. OCR is a technology which has been developed "to transform paper-based documents into digital documents" ([31]). OCR plays a fundamental role in Digital Humanities, giving access to many documents which were originally handwritten or printed but not made digitally available. Although the reliability of commercial OCR software has significantly improved well beyond 80% in terms of accuracy,[2] its robustness largely depends on the quality of the original documents.

For this study, we compiled two historical newspaper corpora in order to analyse the limitations in terms of OCR readability: the effects of time and the low quality of print had an impact on the quality of the OCR scan. The resulting corpora contain many errors and the lexical analysis software[3] cannot retrieve the searched information easily. Therefore, considering the importance of the corpus for CADS studies and the several issues related to the OCR scan with historical documents, a methodology for post-processing OCR error correction has been primarily

1 Historical speech and spoken data are really difficult to collect.

2 https://www.digitisation.eu/fileadmin/Tool_Training_Materials/Abbyy/PSNC_Tesseract-FineReader-report.pdf

3 The lexical analysis software WordSmith Tools ([40]) has been employed in both the projects.

developed in order to reduce the number of errors within the corpora and thus to enhance the capacity of the lexical analysis tools for retrieving more information.

The present methodology has been developed within two larger projects on the study of figurative language used in north-american newspapers between the beginning of the XX century and the beginning of the XXI century.

a) *Metaphors of migration between XX and XXI century* (2017-2020)
This project ([12]) concerns the investigation of metaphors used to represent migration to/from the United States of America and Italy between 1900-14 and 2000-2014. The aim of the project was to determine which were the most prevalent linguistic metaphors and the relative conceptual frames used by the press to speak about migration by adopting a comparative diachronic and cross-cultural approach.

b) *Metaphors and pandemics: Spanish Flu and Coronavirus in US newspapers.* (2020-ongoing)
The second project ([13]) regards the historical investigation of metaphors used to represent Spanish Flu and Coronavirus between 1918-1920 and December 2019- April 2020. The aim of the project was to define differences or similarities regarding the metaphorical representation of two pandemics in two different historical contexts by adopting a diachronic approach.

The definition of the OCR correction rules were based on the "Migration" project while the evaluation of the same rules were applied to the Spanish Flu corpus.

The remainder of the paper is organized as follows: Section 2 will be devoted to the description of how the OCR workflow is organized and a review of the state of art on OCR post-processing will be provided. In Section 3, the compilation process of the corpora used will be shown. In Section 4, the methodology will be described in detail. Section 6 will be devoted to the experimental study and analysis of the post-processing methodology. In the last section conclusions will be drawn and future developments will be explained.

2. Background

2.1 OCR Workflow

The OCR workflow can be divided into three fundamental phases ([34], Figure 1):

1. Pre-processing: the input images are prepared for processing by simplifying the original color image into binary conversion, by conducting a deskewing operation, or by additionally performing works such as cropping, denoising, or despeckling. Lastly, the text is segmented and text regions are separated from non-text areas. Within this step, individual text lines and words and single glyphs are identified.
2. Processing: the OCR software recognizes the segmented lines and words or glyphs by producing a digital textual representation of the printed input. There are two basic core

algorithms on which OCR recognition is based. The first one is the 'matrix matching' which works through the isolation and identification of the single character and the comparison of it with existing models or by training book-specific models with the support of a required Ground Truth (GT)[4]. The second algorithm is called 'feature extraction' and it is the extraction of features from a text - each text is decomposed into different vector-like representations of a character corresponding to a feature.

3. Post-processing: the raw textual output is improved by adopting different dictionaries, language models or GT versions of the processed document in order to correct the resulting texts.

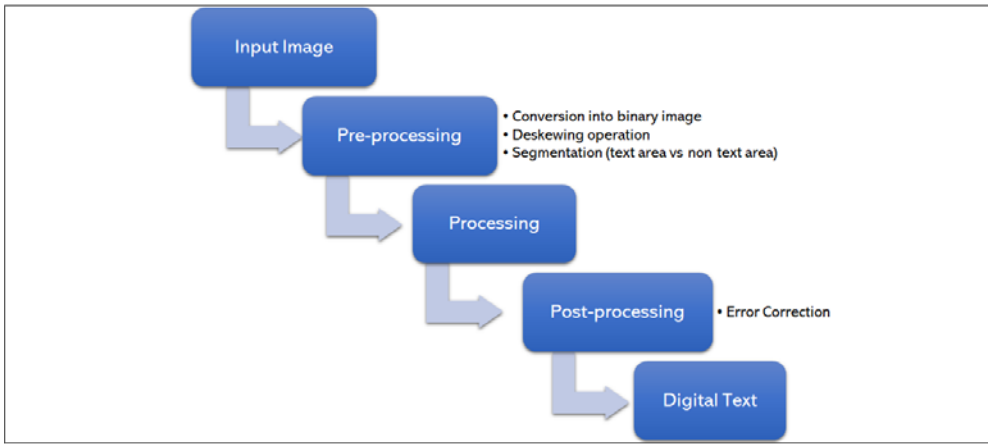

Figure 1: OCR workflow.

However, such technologies do not always achieve satisfactory results and many challenges are posed to OCR software developers and users. For example, OCR-related errors may significantly affect the compilation and the investigation of a corpus ([4]). Different factors in relation to the resulting digitized text have to be considered:

- The state of preservation of the original manuscript/document;[5]

4 The Ground Truth of an image corresponds to "the objective verification of the particular properties of a digital image, used to test the accuracy of automated image analysis processes. The ground truth of an image's text content, for instance, is the complete and accurate record of every character and word in the image. This can be compared to the output of an OCR engine and used to assess the engine's accuracy, and how important any deviation from ground truth is in that instance." (Glossary entry Ground Truth. <https://www.digitisation.eu/glossary/ground-truth/> (last visited, October 3rd, 2021)

5 The low contrast (faded) of the print itself makes it difficult to recognize the characters even with the naked eye.

- The poor quality of the camera through which the image has been taken;
- The quality of the photo taken from the original manuscript;
- The image compression algorithm through which the photograph of the text is produced (especially when working with ancient or easily perishable texts such as newspapers);
- Various layouts of the original texts (especially for newspapers);
- The presence of a GT version of the image under analysis.

There are two possible ways which can be adopted to improve the quality of the documents under analysis: on the one hand, a focus on all the aspects of the pre-processing - namely the optimization of documents quality and OCR systems; on the other hand, a focus on the aspects of the post-processing - namely the optimization and development of post-processing OCR error detection and correction methods. In this paper, we focus our attention on the latter.

2.2 OCR Post-Processing Approaches - Related Work[6]

OCR post-processing is essential for detecting and correcting errors and for improving the quality of OCR scanned documents. As argued by Nguyen *et al.* ([32]), there are two types of errors in an OCR-scanned document:

- *Non-word errors*: invalid lexicon entries which are not attested in any dictionary of the analysed language: e.g. 'tli' for 'the' or 'nigrants' for 'migrants'.
- *Real-word errors*: valid words which are in the wrong context.

The former are easier to identify but more difficult to correct. In contrast, the latter are easier to correct but more difficult to identify. Generally, both errors are analysed according to the number of edit operations used to transform an error to its corresponding GT. The number of edit operations corresponds to the number of changes that have to be made to correct a word. For example, the word 'tli' is a non-word error of distance 2, because it requires two substitutions to be corrected. Its GT is 'the' and the 'l' and the 'i' have to be substituted with 'h' and 'e'. The phrase "capture a near" is a real-word error of distance 1, because it requires only a substitution, 'b' instead of 'n', to correct in its GT "capture a bear".

Following Nguyen et al. ([32]), there are two main approaches to OCR post-processing error correction. Firstly, the *dictionary-based approach* aims at the correction of isolated errors. This lexically-driven method is characterized by the use of string distance metrics or word comparison between an erroneous word and a dictionary entry to suggest candidates for correcting OCRed errors. Traditionally, lexical resources or n-grams are applied in combination with statistical language modelling techniques in order to generate a list of error candidates ([44]). Kolak and

6 Nguyen et al. ([30]) produced an extensive and detailed survey on the state of art of Post-OCR Processing. Approaches which can be found at <https://dl.acm.org/doi/pdf/10.1145/3453476> (Last access on 3rd October 2021).

Resnik ([24]) adopted this approach by implementing a probabilistic scoring mechanism to calculate correct candidates. Bassil and Alwani ([3]), instead of using a lexicon, exploited Google's online spelling suggestion. This approach is easily applicable to any dataset and because it is not language-dependent considering that the error detection and correction tasks depend on the lexicon chosen. However, there may also be additional difficulties given by historical documents. In fact, these types of documents may have different spellings than contemporary documents which, in their turn, are generally used to compose lexicons for OCR candidates. Moreover, these approaches have the limit of being effective only on single words without considering the context. Secondly, the *context-based approach* takes into account the grammatical context of occurrence ([23]) and it is characterized by the use of different models to detect and automatically correct errors mostly from a quantitative statistically-driven perspective. In this case, the methodologies adopted generally overcome the word-level and both the single tokens and their surrounding contexts are considered. For example, Reynaert ([35];[36]) adopts an anagram hashing algorithm for identifying spelling variants of the OCRed errors focusing on both character and word levels. Despite the good results, as discussed and revised by Reynaert ([37]), this approach needs improvements in terms of accuracy and in terms of number of characters which can be simultaneously corrected. Volk *et al.* ([45]) use two different OCR softwares and then merge and compare the output of the systems to perform a disambiguation of the error. However, the context is not properly considered.

Many studies tackle the OCR post-correction by approaching it as a translation problem. Afli et al. ([1]) adopt a Statistical Machine Translation system (SMT) trained on OCR output texts which have been post-edited and manually corrected. Mokhtar et al. ([29]) and Hämäläinen and Hengchen ([19]) base their approach on using sequence-to-sequence models in neural machine translation (NMT). Amrhein and Clemantide ([2]) use borth NMT and SMT and show how these two approaches could benefit each other: SMT systems produce better performance than NMT systems in error correction, while NMT systems obtain better results in error detection. Schaefer and Neudecker ([39]) use a bidirectional LSTM to predict if a character is incorrect or correct. Within the ICDAR2019 competition[7], a post-OCR sequence-to-sequence correction model ([38]) adjusted to BERT ([14]) won the competition. On the same line, Nguyen et al. ([31]) extended this model and proposed an approach developed from BERT and a character-level NMT. The *dictionary-based approach* is not able to capture all the real-word errors. For example, the English expression 'a flood of irritants' is not recognized as an error because all the words are part of the dictionary. However, analyzing the context, it should be corrected in 'a flood of immigrants'. The *context-based approach* intends to overcome the problems of the

7 This competition is organized in two tasks, error detection and error correction for two datasets, an English one and a French one. At the time of writing, two editions took place: in 2017 and 2019.

dictionary-based approach, however it requires more computational effort and it shows a more complex implementation[8].

2.3 Project objectives

The development of the methodology presented in this paper was subject to different restrictions and forced decisions due to a series of issues relating to the data under analysis. Firstly, considering the problem of finding and studying metaphors (which is the objective of the projects related to this study), the accuracy of the OCR version of the documents must be higher than the "average" 80% accuracy. This is due to the fact that a small spelling error may compromise the efficacy of the linguistic analysis. In this sense, a manual and qualitative approach is more appropriate. However, considering the size of the dataset (see Section 3), the use of OCR is useful as a starting point for the manual post-editing. Secondly, we have no access to the original documents of the 20th century newspapers. It is still possible to get the data as OCRed text documents but the quality of the image could only be slightly enhanced by using image enhancer. Lastly, the GT version of the documents was not available; as a consequence it was not possible to use automatic post-OCR error correction models based on GT ([39]).

Therefore, having considered that the margins for improvement by only adopting a single approach have revealed small without a proper GT and a model trained on similar language, we decided to adopt a semi-automatic mixed approach to error detection and correction by bringing together three already existing approaches:

- The *dictionary-based approach* - error detection and compilation of an error candidate list by comparing the OCRed corpus with the error-free corpus;
- The *context-based approach* - error categorisation and compilation of error list by statistically calculating the collocates of each candidate;
- A *manual approach* - manual formulation of the correction rule for each error in the list; further confirmation by in its context of occurrence.

Therefore, we have three main objectives in this paper: present the qualitative-quantitative mixed approach to post-OCR correction; test the efficacy of the method by the application to an alternative comparable dataset; analyze the quality of the application of the rules to a different corpus.

8 A valuable way of resolving the limitation of the dictionary-based approach and the complexity of the context-based approach would be to manually correct the errors ([7];[21]). This approach shows great accuracy, however it heavily relies on human work and on the use of crowd-sources. It also requires more effort in terms of time and energy invested and is characterized by a lower level of efficiency in terms of automation.

3. Corpus Compilation

Considering the diachronic perspective of our research, there were a set of issues we considered during the compilation of the dataset.

The first aspect regards the size of the corpus and its "diachronic representativeness" ([5]). A corpus is a specific form of linguistic data, a collection of written texts searchable by a computer and it represents a sample of language, namely a (small) subset of the language production of interest. In the case of diachronic studies, a corpus should systematically reflect the population (language use) under analysis at different points in time. In this sense, a corpus should contain a variety of texts which cover different types of styles and genres to be diachronically representative in a broad sense. However, not all the corpora have this quality[9] and, in this sense, the corpora compiled for these projects (see Section 1) cannot be defined as diachronically representative of American English of the past. In fact, they represent a topic-specific sample of a newspaper language relative to newspapers and have been collected by means of a set of search terms. However, the corpora might be defined as representative of two special languages in two time periods: the newspaper lexicon of migration and the newspaper lexicon of disease.

The second aspect regards the resolution of the analysis when looking over stretches of time, as in this research study. This depends on how the corpus is designed and on how many sampling points in time the corpus comprehends. Considering a time span of ten years, a corpus of newspapers could cover the whole period by containing all the articles published daily in the time selected. Otherwise, a researcher could decide to select a number of sampling points, so a representative sample containing only a part of the whole number of articles published. The higher the resolution, the higher the number of sampling points in time and, consequently, the higher the accuracy of our analysis. Moreover, as noted in McEnery and Baker [28], when looking at frequent grammatical features, the number of sampling points in time plays a minor role in contributing to the accuracy of the analysis, because the probability of finding such a phenomenon is high. Contrarily, when looking at non-grammatical features of language or discursive representation of a specific group like migration, the number of sampling points plays a fundamental role because "it is possible that, in the gaps between the sampling points, change is happening that we simply cannot measure because we have not sampled in that area" ([28]). For this reason, to produce an appropriate corpus, a researcher needs a priori reflection on how to determine data segmentation and to identify time units:

9 The Time Magazine Corpus, for example, consists of 100 million words from TIME magazine from 1923 and 2006 and can be considered diachronically representative of newspaper language because it exclusively pertains to journalistic language, therefore the results based on this corpus cannot be generalized to the English language. The Brown family corpus is a good example of a diachronic representativeness in a broad sense.

- Text-lifecycle segmentation: based on the periodicity characterising the text type (e.g. a daily edition for a newspaper);
- Top-down segmentation: based on standardised spans (typically a calendar year), or on contextual historical knowledge (e.g. a timeline of events), which is used to inform the segmentation;
- Bottom-up segmentation: based on internal variability, i.e. using distributional information to identify milestones in time, which can in turn be used to divide the corpus in subsets of data to compare against one another ([27]:180).

We adopted a top-down segmentation because we selected periods of time according to our research aims and to contextual historical knowledge. Therefore we defined the following time periods.

- As for "Metaphor of Migration" project:
  - 1900-1914: intense migration movements to USA particularly from Europe ([8]);
  - 2000-2014: the highest decade of immigration in USA.
- As for "Metaphor of Pandemics" project:
  - 1918-20: official duration of Spanish Flu ([47]).

Lastly, the availability of data also had to be considered. Whilst 21st century texts are relatively easier to collect, texts published during the 20th century are not always available in digital format. Some of them have not been digitized, some others are not easily retrievable or accessible. For this study, we identified two databases which fitted our needs. *Lexis Nexis*[10] archive for the XXI century data and the *Chronicling America*[11] archive for the XX century data.

Given the vast amount of documents available in both archives, we defined a set of search terms in order to retrieve only the text which supposedly contained the information needed and to select a representative sample of documents that allows the comparison for the type of discourse analysis which is the object of our work.

With respect to the "Metaphor of Migration" project, we produced a set of search terms which would account how *migration* is lexically shaped in North-America in two different moments in history. Regarding the 2000-14 period, we referred as our baseline to Gabrielatos ([18]) who

10 LexisNexis is an on-line archive which has a large electronic database for legal and public-records–related information. There is a fee to access <http://www.lexisnexis.com> (Last access 10th October 2021), which has been available by The University of Padova library.

11 Chronicling America is a free-access repository which provides access to information about historic newspapers and select digitized newspaper pages produced by The National Digital Newspaper Program (NDNP). It represents a long-term effort to develop an Internet-based, searchable database of U.S. newspapers with descriptive information and select digitization of historic pages. Available at <https://chroniclingamerica.loc.gov> (Last access 10th October 2021)

coined the RASIM acronym (refugee*, asylum seeker*, immigrant* and migrant*) as designating the group of lexical items which represent the group of people who migrate, both legally and illegally. We then checked the presence of these terms within four corpora: the *COHA corpus*[12] ([11]) and the *US Supreme Court Opinions*[13] [20] for the period 1900-1914 and the *COCA – Corpus of Contemporary American English* [10] and the Sibol corpus[14] for the period 2000-2014. The data suggest that there are only minor changes between the past and the present and we confirmed the RASIM terms in addition to 'emigrant/s' which was used more often than 'immigrants' and 'migrants' in the past.

With respect to the "Metaphor of Pandemics" project, two sets of search terms have been selected on the basis of existing literature on the history of pandemics ([9];[41]) and on a careful analysis of the main contemporary newspapers.

Therefore, we produced the following set of search terms:

- 'Emigrant/s', 'Immigrant/s', 'Migrant/s', 'Refugee/s' and 'Asylum seeker/s' for *Migration*
- 'Spanish Flu', 'Influenza', 'Flu Pandemic/s', 'Pandemic', 'Epidemic' for *Spanish Flu*

Chronicling America provides access to these digitized historic materials through a Web interface enhanced with dynamic HTML interactivity for magnification and navigation. Searches are available for both full-text newspaper pages and bibliographic newspaper records (the Newspaper Directory). Pages are displayed in JPEG format and can be downloaded in three different formats:

- JPEG2000, Part 1; 8-bit component; 6 decomposition layers; 25 quality layers; 8:1 compression; with XML Box with specified RDF metadata;
- Single page PDF with hidden text; downsampled to 150 dpi, using JPEG compression; with XMP containing specified RDF metadata;
- Single page machine-readable text encoded in unicode UFT-16; in column-reading order, created with Optical Character Recognition.

12 *The Contemporary Corpus of Historical American English* consists of 400 million words from a balanced set of sources from the 19th century onwards.

13 The *US Supreme Court Opinions* contains approximately 130 million words in 32,000 Supreme Court decisions from the 1790s to the present.

14 The SiBol corpus is available through the *Sketch Engine interface* at https://www.sketchengine.eu/sibol-corpus/#toggle-id-3 (Last access on 28th September 2021). More information is available from http://www.lilec.it/clb/?page_id=8 (Last access on 28th Sepember 2021).

Considering that the JPEG2000 is a high quality format, we originally downloaded a data sample in order to test if different OCR software would have worked better than the OCR provided by Chronicling America. We adopted the open-source OCR4all software[15] ([34]). However, the resulting OCR version did not encourage us to carry on the time-consuming operation of OCR scanning all the documents retrieved. Therefore, we decided to adopt the OCRed text provided by the Chronicling America itself.[16]

3.1 The dataset - an overview

Three corpora were compiled using the aforementioned keywords and time spans. The details are shown in the following Table 1.

	Documents	Tokens	Types	TTR
Flu Corpus (henceforth Flu1920)	484	1.850.886	140.160	7,57%
New York Herald 1900-1914 (henceforth NYHC1900)	9119	64.061.101	3.085.080	4,82%
New York Times 2000-2014 (henceforth NYTC2000)	125	58.915.060	308.251	0,52%

Table 1: Details of the compiled corpora.

NYTC2000 was compiled by downloading raw text files in .txt format from *LexisNexis.* NYTC2000 is yearly clustered in 15 folders - from 2000 to 2014. NYHC1900 is yearly clustered in 15 folders - from 1900 to 1914. NYHC1900 and Flu1920 have been compiled by downloading OCRed raw text files in .txt format from *Chronicling America.* Flu1920 is yearly clustered in three folders - from 1918 to 1920. Each folder contains all the articles retrieved for the selected year by using the aforementioned set of keywords. NYTC2000 documents do not

15 Available at <http://www.ocr4all.org/en/software_download.php>

16 More detailed informations on OCR guidelines can be found at <https://www.loc.gov/ndnp/guidelines> (Last access on 7th October 2021).

result from OCR scans. NYHC1900 and Flu1920 are the OCRed documents which will be the object of the methodology we are going to present.

The corpora have different sizes in terms of number of tokens, but this variable is strongly dependent on external factors such as the number of words actually published in both of the time periods and it is not simple to be predicted or controlled. However, some atypical data regarding the number of types and the ratio between *types* and *token* can be identified. The term *type* refers to the number of distinct words in a text or corpus, whilst the term *token* refers to the total number of words in a text or corpus. So, looking at the sentence "A good car is a car that goes fast", it contains 9 tokens, but only 7 types, because the type 'a' and 'car' count as two tokens each. TTR[17], namely type-token ratio, is a statistic which measures lexical diversity/lexical richness of a corpus: it expresses the proportion of types relative to the number of tokens. It is calculated by dividing the number of type (*t)* by the number of token (*n*) as the following formula shows:

$$TTR = \frac{t}{n}$$

The larger the TTR results, the less repetitive the vocabulary usage. On the contrary, the lesser the TTR results, the more repetitive the vocabulary used. An optimal TTR score would be '1'.

Looking at Table 1, the difference between the older corpora and the more recent one is unexpectedly high in terms of TTR. The types-token ratio[18] relative to the corpora from the 1900s are unusually higher than the two corpora from the 2000s. Flu1920 and NYCH1900 register respectively a TTR of 7.57% and 4.82 % with respect to the 0,52% for NYTC2000. This is because there are a lot of misspellings or non-real words in them which were due to the OCR processing. For example, we found many occurrences of 'tho' or 'whd' in both of the 1900s corpora.

Before presenting the methodology, we want to conclude this section with some remarks about the difficulty that these OCRed texts present in terms of automatic error recognition. For example, the following lines are a small portion of the raw text of the very first document (dated 18th february 1900) of the NYHC1900 corpus:

17 However, "the type/token ratio is very sensitive to the length of the text; it decreases as the text becomes longer and more words get used again (recycled)." ([5]: 58). In this sense, the STTR has been developed in order to normalize this value and to make it comparable across texts of different lengths. STTR has been proposed by Scott ([40]) and consists in simply dividing texts into standard-size segments (e.g. 1000 words), calculate the TRR for each segment and then take the mean value of the TTRs ([5]). We applied this formula to our dataset. However, even though the STTR registers results in line with expectations, the difference between the older corpora and the more recent ones is unexpectedly high.

18 In this case, we have converted the TTR value into a percent multiple in order to make it comparable.

Mr. Benson's
best work as King Henry is in the coun
cil chamber scene and in his bantering talk
with FluollPii and the soldier Williams, and
his stirring np^ches t« the army are w»ll
delivered. Mr. Alfred Brydone, as Charles VI.
1? excellent, and the minor parts are played
with a high order »f intelligence.

We can immediately see that the contextual approaches will fail in these cases given the amount of noise. See for example the sequence "**FluollPii** and the soldier Williams, and his stirring **np^ches t«** the army are **w»ll**" where in a window of less than 15 words there are 4 unknown sequence of characters.

4. Methodology

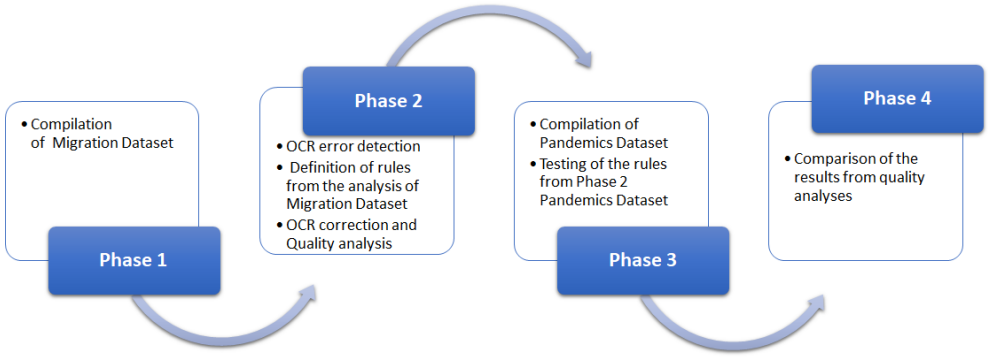

Figure 2: Workflow representation relative to this project.

The workflow of this paper can be divided in four different phases as displayed in Figure 2. For Phase 2, we decided to use the contemporary corpora as the GT version of our corpora. We compared the old OCRed corpora with the contemporary ones since the latter were free from OCR errors. We first produced four wordlists - one for each corpus processed by using the software *WordSmith tools 8* ([40]). A wordlist is defined as the list of all word types present in the corpus ordered by frequency of occurrence. Then, we proceeded as follows. The error detection and correction procedure consisted in a three-step procedure.

4.1 Detection of errors

Detection of errors is the first step. We use the *KeyWords* function in *WordSmith Tools*. The keywords are words which occur unusually frequently in a corpus in comparison with a reference corpus. We compared the wordlists of the corpora resulting from OCR with the contemporary corpora, then we listed the words according to the *LogRatio*[19]. This function provides a useful way to characterise a text or a genre. We found out that the words that appear more frequently in the OCRed corpus or do not appear in the error free corpus were plausible error candidates. The following Table 2 is an example of the keyword analysis. It shows the first 20 most frequent words of NYHC1900 compared to NYTC2000;

Key word	**Freq.**	**Log_R**
LR	5.224	147,69
PANY	4.656	147,48
TLI	4.189	147,42
TORK	5.738	147,42
IHE	6.439	147,37
JL	3.758	147,31
TII	4.123	147,28
IY	3.113	147,24
TBE	4.839	147,21
LF	2.969	147,13

19 https://resources.wolframcloud.com/FormulaRepository/resources/Log-Odds-Ratio

EF	2.687	147,07
XEW	2.814	147,07
LH	3.050	147,06
TERDAY	2.729	147,04
TLON	2.933	147,03
THF	2.514	146,98
SN	2.590	146,97
STH	4.404	146,96
TIIE	3.071	146,95
VL	2.240	146,85

Table 2: Keyword list - comparison between NYHC1900 and NYTC2000.

4.2 Analysis and categorization of error candidates

The second step concerns the *analysis and categorization of the error candidates* in the list. We analysed the collocates of each error candidate setting a left/right word span of 10 and we listed the results according to the *LogRatio*. By looking at the context of occurrence, each candidate was assigned to the error list or discarded. Each error in the list was categorised according to three categories:

- *Standard Mapping*: the error contains the same number of characters as the respective correct form – 1:1. For example: 'hear' (correct) vs 'jear' (error);
- *Non-standard Mapping*: the error contains higher or a smaller number of characters than the correct form – n:1 or 1:n. For example: 'main' (correct) vs 'rnain' (error);
- *Split errors:* the word is interpreted by the OCR as two distinct words. This is a very common error when digitizing newspapers because of the shape of the column in which articles are written. For example: 'department' vs 'depart' and 'ment'.

4.3 Selection of candidates and rule production

The third step regards the analysis of each error by focusing on the context of occurrence in order to determine the corresponding correct word. Then, we wrote the correction rules in the form of a tab separated values file where each entry represents the misspelled word (left) and the corresponding corrected word (right)

'WRONG WORD' \t 'CORRECT WORD'

We defined the error correction rule as a regular expression to match the pattern of the error (i.e. jear) and substitute it with the (supposedly) correct form (i.e. hear).

The implementation in R of this part of the code is a very simple regular expression substitution using the str_replace_all function of the stringr package.[20] For example:

str_replace_all(string = text, rule_errors)

The line above automatically substitutes in the string "text" all the patterns defined in the rules present in the tabular separated value file.

Type	Error	Correction
Standard	ail	all
	anl	and
	hnd	and
	rrivnte	private
	sofflcienl	sufficient
	thnt	that

20 https://cran.r-project.org/web/packages/stringr/index.html

Non-Standard	allesratlons	allegations
	ar\.d	and
	ar.nounced	announced
	<-hi!dren.	children
	/-ontlnue	continue
	proj>erty	property
	repre?ented	represented

Table 3: Example of Correction rules.

5. Experimental Study and Analysis

The implementation of the procedure follows the principles described by Wachsmuth et al. ([46]). They propose to mine textual information from large text collections by means of pipelines in order to efficiently and effectively allow for a sequential process of text analysis. For our experiments, we used the R programming language which has a set of packages, named *`tidyverse'*, that implements this idea of pipelines in a clear way.[21]

The following lines show the basic structure of the sequence of operations to find and substitute the selected errors:

```
data_corrected = data %>%
  mutate(text_corrected = str_replace_all(string = text, named_errors)) %>%
  mutate(text_corrected = ifelse(test = is.na(text_corrected),
                                 yes = text,
                                 no = text_corrected))
```

21 The source code of the experiments are available at https://github.com/gmdn

The symbol *%>%* is the operator that "pipelines" the functions and data transformation one after the other: starting from the "data" (the corpus), we first find and replace the errors ("named errors") with the str_replace_all function (section 4.3), then we adjust the results by checking the presence of empty values ("is.na") produced during the transformation/correction process. The result is a data structure with the original text paired with the corrected text ready to be qualitatively analysed by a linguist.

5.1 Migration Project

We manually individuated a total of 2313 errors (1837 split errors and 476 standard/non-standard errors respectively) and, respectively, as many correcting rules have been written. The automatic application of the rules produced 784,729 (62,358 split errors and 722,371 standard/non-standard substitutions). As Table 4 shows, the number of Tokens and Types have been moderately changed. The number of tokens has decreased by 0.09 %, and the number of types has decreased by 0.002%.

	Before OCR correction		After OCR correction		Difference	
Corpus	Tokens	Types	Tokens	Types	Δ Tokens	Δ Types
NYHC1900	64.061.101	3.085.080	64.001.438	3.085.024	- 0.09 %	- 0.002%

Table 4: OCR post-processing correction results for NYCH1900.

5.2 Pandemics Project

Successively, we applied the same correction rules to correct the FLU1920 corpus in order to test the validity and applicability of our rules to a different comparable dataset. What makes FLU1920 a comparable corpus with NYCH1900 is the fact that both of the corpora are composed of American English documents published in the same time period and also scanned by the same type of OCR software provided by the Chronicling America project. Therefore, on the basis of the list of errors we have previously manually identified for NYCH1900, we automatically identified a total of 728 errors (576 split errors and 152 standard/non-standard errors respectively). Then, the automatic application of the rules produced 12,555 substitutions. As Table 5 shows, this process gave similar results to NYCH1900: the number of tokens has decreased by 0.12 %, whilst the number has slightly decreased by 0.05%.

	Before OCR correction		After OCR correction		Difference	
Corpus	Tokens	Types	Tokens	Types	Δ Tokens	Δ Types
FLU1920	1.850.886	140.160	1.848.633	140.088	-0,12 %	-0,05 %

Table 5: OCR post-processing correction results for NYCH1900.

5.3 Post hoc Analysis

It is not easily predictable how OCR correction will work. In general, we would have expected a more significant decrease in the number of tokens and types, while the result of the thousands of corrections gave only a small reduction of these numbers. A possible explanation for this trend in the number of tokens might be that many errors, in particular non-standard errors such as 'ar\.d'(and), 'th/'(the) or 'ir\.'(in), were not previously recognized as valid tokens and were removed from the count. Whilst, the number of types are in general reduced for both corpora since different errors are mapped to the same type. For example, the English article `the' has been differently misspelled in many ways: 'tne', 'tha', 'tbe', 'tna'. These 4 words are counted as four different types. Then, by correcting substituting all these words with the, the number of types is reduced by three units. Similarly, the connector 'and' was misspelled as 'nnd', 'aad', 'snd', 'anl', 'hnd', 'nnd'. These 6 words are counted as 6 different types. Then, by correcting all these words with 'and', the number of types is reduced by 6 units.

In Table 6, we show the 20 most frequent substitutions made in the NYCH1900 dataset. The vast majority of substitutions are related to articles, prepositions, conjunctions. This is due to the "natural" distribution of words.[22]

word	substitution	count
th?	the	50,136
la	in	40,004
tho	the	28,835
th«	the	28,014

22 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4176592/

word	substitution	count
tha	the	17,257
th*	the	14,775
th	the	14,772
ln	in	14,624
cf	of	12,694
tbe	the	11,471
\\?\\?*e	the	11,178
o*	of	10,196
waa	was	9,908
hi*	his	9,157
nt	at	8,477
ot	of	8,188
tn	in	7,949
nnd	and	7,313
th*-	the	7,111
ihe	the	6,030

Table 6: OCR post-processing correction results for NYCH1900. Top 20 most frequent substitutions.

In fact, if we go down in the ranking of the list of frequency of substitutions of the same corpus, we can see examples of corrections that are related to nouns, adjectives, adverbs and so on (see Table 7).

word	substitution	count
newapapera	newspapers	16
tfiat	that	15
agaiust	against	12
Uovernor	Governor	10
wer»-	were	10
xew	new	9
rr\\.	m	8
o-jt	out	8
oontrol	control	7
pres-nt	present	7
?sp?cial	special	7
oj>en	open	7

Table 7: OCR post-processing correction results for NYCH1900. Examples of less frequent substitutions.

Lastly, the present results are significant in at least two major respects. Firstly, the correction results extracted from NYCH1900 and NYTC2000 proved to be operational in regards to our aims. The rules worked effectively and part of the corpus, even to a small extent, was corrected. Secondly, the test on FLU1920 was successful: the correction rules were able to correct a different corpus showing similar results than the previous application (on NYCH1900). Further research should be undertaken to investigate the number of errors which have not been individuated by the rules and which are context-dependent, namely peculiar errors to the FLU1920. Another possible line of future work should concern the extraction of correcting rules from FLU1920 following the proposed methodology and the application of the resulting rules to NYC1900. Furtherly, in future investigations, it might also be possible to collect a larger set of comparable American English historical corpora from the early 1900s and then compare the results after the post-OCR correction procedure.

6. Final Remarks and Future Works

In this paper, we presented a semi-automatic method for detection and correction of OCR errors for the corpus-assisted discourse analysis of old newspaper documents. The outcome of this project consists in a set of rules which are, eventually, valid for different contexts and applicable to different corpora and which can be reproduced and reused. The methodology presented has revealed to be a useful strategy for those cases when the quality of the texts under analysis is low and, fundamentally, there is not a GT available but a comparable corpus without errors is present. Another peculiarity is represented by the fact that it is applicable to already OCRed data and it is not dependent on the type of software or on a specific language. We also consider that this methodology would be particularly profitable and useful for researchers from arts and humanities area (historians, linguists, literature scholar) where an expertise of computational techniques (e.g. AI, neural networks) is not a required skill and these might find some difficulties in accessing these approaches to OCR post-processing error corrections which have shown to be most successful.

The methodology we have proposed in this paper is general enough to be adapted in various situations; nevertheless, we believe that the interaction with the linguist has to be enhanced in order to improve the detection and classification of errors. Indeed, we think it is important during the validation process to show possible "false positives" substitutions: see for example in Table 5 the substitution "ot" into "of", are all the 8,188 substitutions correct? Is there any chance that one (or more) of the substitutions was meant to be "on" instead of "of"? In this sense, an interactive data visualization approach is needed to help the expert to support her/his decisions.

There are still open questions that we will investigate in this line of work: how many documents have we missed during the compilation of the corpus given that a search keyword may be subject to OCR correction as well? How can these types of keyword search error affect a CADS analysis? For this reason, we intend to use error models to predict the relative risk that queried terms mismatch targeted resources due to OCR errors, as suggested by [6]. We also want to compare our analysis with other approaches that make use of BERT pre-trained neural networks to post-hoc error correction [31], especially in those cases where the context is not clear given multiple OCR errors in the same paragraph, or that take advantage of multiple OCR engines by aligning and comparing their different outputs in order to reduce the error rate ([29]).

References

[1] Afli, Haithem, Zhengwei Qiu, Andy Way, and Páraic Sheridan. 2016. 'Using SMT for OCR Error Correction of Historical Texts'. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 962–66. Portorož, Slovenia: European Language Resources Association (ELRA). https://aclanthology.org/L16-1153.

[2] Amrhein, Chantal, and Simon Clematide. 2018. 'Supervised OCR Error Detection and Correction Using Statistical and Neural Machine Translation Methods'. https://doi.org/10.5167/UZH-162394.

[3] Bassil, Youssef, and Mohammad Alwani. 2012. 'OCR Post-Processing Error Correction Algorithm Using Google Online Spelling Suggestion'. *ArXiv:1204.0191 [Cs]*. http://arxiv.org/abs/1204.0191.

[4] Bazzo, Guilherme Torresan, Gustavo Acauan Lorentz, Danny Suarez Vargas, and Viviane P. Moreira. 2020. 'Assessing the Impact of OCR Errors in Information Retrieval'. In *Advances in Information Retrieval*, edited by Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, 12036:102–9. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-45442-5_13.

[5] Brezina, Vaclav. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. 1st ed. Cambridge University Press. https://doi.org/10.1017/9781316410899.

[6] Chiron, Guillaume, Antoine Doucet, Mickael Coustaty, Muriel Visani, and Jean-Philippe Moreux. 2017. 'Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information'. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 1–4. Toronto, ON, Canada: IEEE. https://doi.org/10.1109/JCDL.2017.7991582.

[7] Clematide, Simon, Lenz Furrer, and Martin Volk. 2016. 'Crowdsourcing an OCR Gold Standard for a German and French Heritage Corpus'.

[8] Cohen, Robin, ed. 1995. *The Cambridge Survey of World Migration*. 1st ed. Cambridge University Press. https://doi.org/10.1017/CBO9780511598289.

[9] Crosby, Alfred W. 2003. *America's Forgotten Pandemic: The Influenza of 1918*. 2nd ed. Cambridge ; New York: Cambridge University Press.

[10] Davies, Mark. 2008. *COCA – Corpus of Contemporary American English*.

[11] Davies, Mark. 2010. 'The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English'. *Literary and Linguistic Computing* 25, 4: 447–64. https://doi.org/10.1093/llc/fqq018.

[12] Del Fante, Dario. 2021. *The metaphorical representation of migration across time and cultures: the cases of USA and Italy.* Doctoral thesis, unpublished. Padua: University of Padua.

[13] Del Fante, Dario. 2021. 'Figurative Language and Pandemics: Spanish Flu and Covid-19 in US Newspapers. A Case-Study.', 95–96. TU Dortmund University. https://icame42.englisch.tu-dortmund.de/docs/Book_of_Abstracts.pdf.

[14] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 'BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding'. *ArXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805.

[15] Duguid, Alison. 2010. 'Investigating *Anti* and Some Reflections on Modern Diachronic Corpus-Assisted Discourse Studies (MD-CADS)'. *Corpora* 5, no. 2: 191–220. https://doi.org/10.3366/cor.2010.0105.

[16] Fairclough, Norman. 1995. *Critical Discourse Analysis: The Critical Study of Language*. Language in Social Life Series. London ; New York: Longman, 1995.

[17] Fairclough, Norman. 1989. *Language and Power*. First edition. London ; New York: Routledge, Taylor & Francis Group.

[18] Gabrielatos, Costas. 2007. 'Selecting Query Terms to Build a Specialised Corpus from a Restricted-Access Database.' *ICAME Journal* 31: 5–44.

[19] Hämäläinen, Mika, and Simon Hengchen. 'From the Paft to the Fiiture: A Fully Automatic NMT and Word Embeddings Method for OCR Post-Correction'. *ArXiv:1910.05535 [Cs]*, 12 October 2019. https://doi.org/10.26615/978-954-452-056-4_051.

[20] Harold J. Spaeth, Lee Epstein, Andrew D. Martin, Jeffrey A. Segal, Theodore J. Ruger, and Sara C. Benesh. 2020 Supreme Court Database, Version 2021 Release 01. URL: http://Supremecourtdatabase.org

[21] Holley, Rose. *Many Hands Make Light Work: Public Collaborative Text Correction in Australian Historic Newspapers.* Canberra: National Library of Australia, 2009. http://www.nla.gov.au/ndp/project_details/documents/ANDP_ManyHands.pdf.

[22] Kettunen, Kimmo, Eetu Mäkelä, Teemu Ruokolainen, Juha Kuokkala, and Laura Löfberg. 'Old Content and Modern Tools: Searching Named Entities in a Finnish OCRed Historical Newspaper Collection 1771–1910'. *Digital Humanities Quarterly* 11, no. 3 (November 2017).

[23] Kissos, Ido, and Nachum Dershowitz. 'OCR Error Correction Using Character Correction and Feature-Based Word Classification'. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, 198–203. Santorini, Greece: IEEE, 2016. https://doi.org/10.1109/DAS.2016.44.

[24] Kolak, Okan, and Philip Resnik. 'OCR Error Correction Using a Noisy Channel Model'. In *Proceedings of the Second International Conference on Human Language Technology Research*, 257–62. HLT '02. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002.

[25] Library of Congress., National Endowment for the Humanities., National Digital Newspaper Program (U.S.),. 'Chronicling America : Historic American Newspapers.', 2006. /z-wcorg/. http://bibpurl.oclc.org/web/19370.

[26] Lund, William B., and Eric K. Ringger. 'Improving Optical Character Recognition through Efficient Multiple System Alignment'. In *Proceedings of the 2009 Joint International Conference on Digital Libraries - JCDL '09*, 231. Austin, TX, USA: ACM Press, 2009. https://doi.org/10.1145/1555400.1555437.

[27] Marchi, Anna. 'Dividing up the Data: Epistemological, Methodological and Practical Impact of Diachronic Segmentation'. In *Corpus Approaches to Discourse: A Critical Review*, edited by Charlotte Taylor and Anna Marchi, 174–96. Milton Park, Abingdon, Oxon ; New York: Routledge, Taylor & Francis Group, 2018.

[28] McEnery, Tony, and Helen Baker. *Corpus Linguistics and 17th-Century Prostitution: Computational Linguistics and History*. Corpus and Discourse 3. London ; New York: Bloomsbury Academic, 2017.

[29] Mokhtar, Kareem, Syed Saqib Bukhari, and Andreas R. Dengel. 'OCR Error Correction: State-of-the-Art vs an NMT-Based Approach'. *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, 2018, 429–34.

[30] Nguyen, Thi Tuyet Hai, Adam Jatowt, Mickael Coustaty, and Antoine Doucet. 'Survey of Post-OCR Processing Approaches'. *ACM Comput. Surv.* 54, no. 6 (July 2021). https://doi.org/10.1145/3453476.

[31] Nguyen, Thi Tuyet Hai, Adam Jatowt, Nhu-Van Nguyen, Mickael Coustaty, and Antoine Doucet. 'Neural Machine Translation with BERT for Post-OCR Error Detection and Correction'. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, 333–36. JCDL '20. New York, NY, USA: Association for Computing Machinery, 2020. https://doi.org/10.1145/3383583.3398605.

[32] Nguyen, Thi-Tuyet-Hai, Adam Jatowt, Mickael Coustaty, Nhu-Van Nguyen, and Antoine Doucet. 'Deep Statistical Analysis of OCR Errors for Effective Post-OCR Processing'. *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2019, 29–38.

[33] Partington, Alan, Alison Duguid, and Charlotte Taylor. *Patterns and Meanings in Discourse: Theory and Practice in Corpus-Assisted Discourse Studies (CADS)*. Vol. 55. Studies in Corpus Linguistics. Amsterdam: John Benjamins Publishing Company, 2013. https://doi.org/10.1075/scl.55.

[34] Reul, Christian, Dennis Christ, Alexander Hartelt, Nico Balbach, Maximilian Wehner, Uwe Springmann, Christoph Wick, Christine Grundig, Andreas Büttner, and Frank Puppe. 'OCR4all—An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings'. *Applied Sciences* 9, no. 22 (13 November 2019): 4853. https://doi.org/10.3390/app9224853.

[35] Reynaert, Martin. 'Non-Interactive OCR Post-Correction for Giga-Scale Digitization Projects'. In *Computational Linguistics and Intelligent Text Processing*, edited by Alexander Gelbukh, 4919:617–30. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. https://doi.org/10.1007/978-3-540-78135-6_53.

[36] Reynaert, Martin .'On OCR Ground Truths and OCR Post-Correction Gold Standards, Tools and Formats'. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage - DATeCH '14*, 159–66. Madrid, Spain: ACM Press, 2014. https://doi.org/10.1145/2595188.2595216.

[37] Reynaert, Martin. 'OCR Post-Correction Evaluation of Early Dutch Books Online - Revisited'. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 967–74. Portorož, Slovenia: European Language Resources Association (ELRA), 2016. https://aclanthology.org/L16-1154

[38] Rigaud, Christophe, Antoine Doucet, Mickaël Coustaty, and Jean-Philippe Moreux. 'ICDAR 2019 Competition on Post-OCR Text Correction'. In *15th International Conference on Document Analysis and Recognition*, 1588–93. Sydney, Australia, 2019. https://hal.archives-ouvertes.fr/hal-02304334.

[39] Schaefer, Robin, and Clemens Neudecker. 'A Two-Step Approach for Automatic OCR Post-Correction'. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 52–57. Online: International Committee on Computational Linguistics, 2020. https://aclanthology.org/2020.latechclfl-1.6.

[40] Scott, Michael. *WordSmith Tools version 8*, Stroud: Lexical Analysis Software, 2020.

[41] Snowden, Frank M. *Epidemics and Society: From the Black Death to the Present*. Open Yale Courses Series. New Haven: Yale University Press, 2019.

[42] Sporleder, Caroline, Antal van den Bosch, and Kalliopi Zervanou. *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*. Theory and Applications of Natural Language Processing. Berlin Heidelberg: Springer-Verlag, 2011.

[43] Taylor, Charlotte, and Anna Marchi, eds. *Corpus Approaches to Discourse: A Critical Review*. Milton Park, Abingdon, Oxon ; New York: Routledge, 2018.

[44] Tong, Xiang, and David A. Evans. 'A Statistical Approach to Automatic OCR Error Correction in Context'. In *Fourth Workshop on Very Large Corpora*. Herstmonceux Castle, Sussex, UK: Association for Computational Linguistics, 1996. https://aclanthology.org/W96-0108.

[45] Volk, Martin, Lenz Furrer, and Rico Sennrich. 'Strategies for Reducing and Correcting OCR Errors'. In *Language Technology for Cultural Heritage*, edited by Caroline Sporleder, Antal van den Bosch, and Kalliopi Zervanou, 3–22. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. https://doi.org/10.1007/978-3-642-20227-8_1.

[46] Wachsmuth, Henning. *Text Analysis Pipelines*. Vol. 9383. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015. https://doi.org/10.1007/978-3-319-25741-9.

[47] Yang, Wan, Elisaveta Petkova, and Jeffrey Shaman. 'The 1918 Influenza Pandemic in New York City: Age-Specific Timing, Mortality, and Transmission Dynamics'. *Influenza and Other Respiratory Viruses* 8, no. 2 (March 2014): 177–88. https://doi.org/10.1111/irv.12217.