A Quantitative/Qualitative Approach to OCR Error Detection and Correction in Old Newspapers for Corpus-assisted Discourse Studies

Dario Del Fante[1][0000−0002−1650−273X] and Giorgio Maria Di Nunzio[2,3][0000−0001−9709−6392]

[1] Dept. of Linguistic and Literary Studies, University of Padua, Italy
dario.delfante@phd.unipd.it
[2] Dept. of Information Engineering, University of Padua, Italy
[3] Dept. of Mathematics, University of Padua, Italy
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 because allow researchers to expand their interest by making many texts available and analysable through a computer. However, OCR software are not totally accurate, and the resulting error rate compromises their effectiveness. This paper proposes a mixed qualitative-quantitative approach to OCR error detection and correction in order to develop a methodology for compiling historical corpora. The proposed approach consists of three main steps: corpus creation, OCR detection and correction, and application of the automatic rules. The rules are implemented in R using a "tidyverse" approach for a better reproducibility of the experiments.

Keywords: Corpus-assisted Discourse Studies · OCR detection · OCR correction.

1 Introduction

In Corpus-assisted Discourse Studies (CADS) [14], the processes of corpus design and corpus compilation have a marked impact on the entire research and, depending on it, the results may shift dramatically. Especially for diachronic studies, there is a scarcity of digitized version of paper documents; for this reason, it is often necessary to manually transcribe the texts under analysis or to use Optical Character Recognition (OCR) software which plays a fundamental role in the study of digitizes manuscripts [10]. However, OCR technologies do

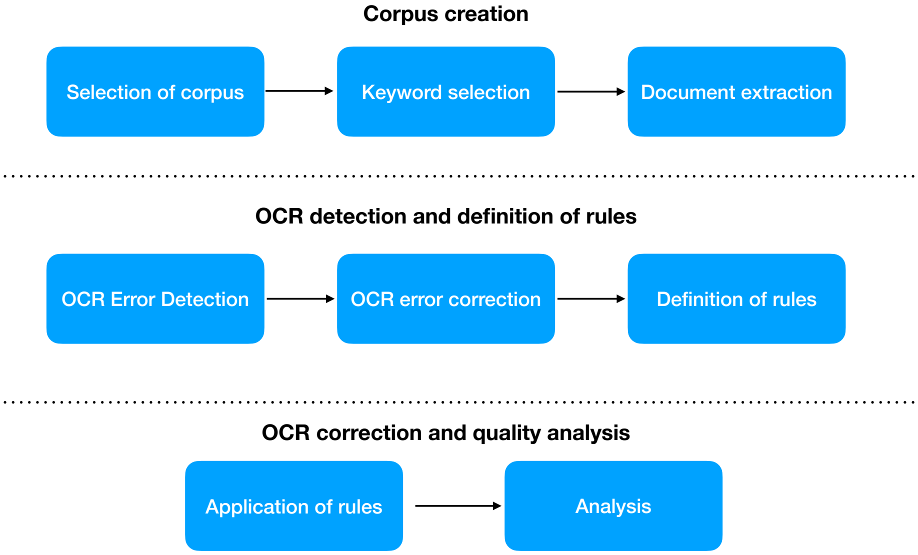

Fig. 1. The three steps of the proposed procedure for the qualitative and quantitative analysis of OCR Error and detection.

not always achieve satisfactory results because of several aspects that affects the quality of the original scan: the quality of the camera through which the image has been taken, the image compression algorithm, the quality of the paper especially when working with ancient or easily perishable texts such as old newspapers. These errors may affect in a crucial way the results of a search for documents which may compromise the compilation of a corpus in CADS [2].

In this paper, we propose a procedure for collecting and creating corpora for discourse analysis from old paper documents and we present a semi-automatic method for detection and correction of OCR errors. The outcome of this work consists in a set of rules which are, eventually, valid for different context and applicable to different datasets. 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.

In Figure 1, we show an overview of the proposed approach which consists of three main steps: corpus creation, OCR detection and correction, and application of the automatic rules. The details of each step are described in the following sections.

The remainder of the paper is organized as follows: in Section 2, we describe our case study and the choices of the corpora; in Section 3, we present a brief overview of the state-of-the art in OCR correction, we describe our proposal for the error detection and correction, and we analyze the preliminary results; in Section 4, we give our final remarks and discuss our future work.

2 Case Study: Searching for Metaphors to Represent Immigrants

Our case study focuses on the analysis of the metaphors used in the newspapers to represent migration to/from the United States of America and Italy from a diachronic perspective between the beginning of the XX century and the beginning of the XXI century. Given the vast amount of documents available, we needed to define a criterion in order to select a representative sample of documents that allows the comparison for the type of discourse analysis which is the object of our work.

2.1 Choice of the Corpus

In order to reduce the amount of data to collect, we selected two moments in history which represent two sampling points in time which have a significant value in relation to migratory movement: 1900-14 and 2000-2014. The decision to focus on the aforementioned time periods lies in the fact that these represent important moments for migratory movements, for both USA and Italy:

- As for USA:
    - 1900-1914: intense migration movements to USA particularly from Europe [5];
    - 2000-2010: the highest decade of immigration in USA.
- As for Italy:
    - 1900-1914: intense period of emigration and internal migration [9, 6];
    - 2000-2014: a dramatic increase of the immigration phenomenon which lead to the "2015 migration crisis".[7]

The availability of data, the newspaper political leaning, and the registration fees were additional constraints that narrowed the range of options down to three newspapers.

For USA, we selected the New York Herald[4], for the time period 1900-1914, and the New York Times[5], for the time period 2000-2014, because they are both examples of quality press published in New York.[6] Even though the analysis of

[4] `http://chroniclingamerica.loc.gov`

[5] `http://www.lexisnexis.com`

[6] By quality press we mean the more accurate newspapers which give detailed accounts of events, as well as reports on business, culture, and society, which contrasts with tabloid newspapers which are more devoted in giving sensational news.

the same newspaper was preferable for a matter of homogeneity and integrity, we could not find an American newspaper available for both time period.

Regarding Italy, we selected La Stampa,[7] a newspaper belonging to the category of quality press and which is published in Turin, a crossroads for many migration routes, both internal and from foreign countries. Fortunately, La Stampa provides an archive concerning all of its daily editions in digital format from 1867 to nowadays.

2.2 Selecting Search Terms

Having chosen the newspapers and the historical period, we needed to select the keywords to filter the articles useful for our study. We decided to use search-terms to sort and collect the newspaper articles. On the one hand, not using search terms would have provided for more versatile corpora, that could be used for other research purposes. On the other hand, a corpus collected by narrowing down the amount of texts retrieved using search terms is more manageable. In addition, as shown by [8], sometimes the idiosyncrasies of the online database from which texts are retrieved pose limitations.

When compiling a specialised corpus using keywords, there is a trade-off between precision and recall. That is, there is a tension between, on the one hand, creating a corpus in which all the texts are relevant, but which does not contain all relevant texts available in the database, and, on the other, creating a corpus which does contain all available relevant texts, albeit at the expense of irrelevant texts also being included. Seen from a different perspective, the trade-off is between a corpus that can be deemed incomplete, and one which contains noise (i.e. irrelevant texts). Therefore, considering our research purposes which essentially consisted in identifying metaphors of migration, we decided to define a set of search terms to retrieve texts, in order to create a specialised corpus.

In particular, given the task of identifying metaphor of migration, we needed to identify a set of search terms which would fit into both time periods and which would be comparable and should denote the same meaning [16].

English keywords As for English, the starting point was the set of words identified by [8] named under the acronym RASIM: *refugee**,[8] *asylum seeker**, *immigrant**, and *migrant**. We added a fifth word to this list: *emigrant**. These set of words, in fact, has received great attention within corpora and discourse studies and is generally recognized as fully accounting for migration. In order to study the use of these words, we consulted two diachronic corpora: the Contemporary Corpus of Historical American English (COHA)[9] corpus and the US Supreme Court Opinions.[10] The former consists of a collection of 400 million

[7] http://www.archiviolastampa.it/

[8] We use the symbol '*' to indicate the possibility of plural, or feminine/masculine for the Italian words.

[9] https://www.english-corpora.org/coha/

[10] https://www.english-corpora.org/scotus/

words from a balance set of sources. The latter contains approximately 130 million words in 32,000 Supreme Court decisions from the 1790s to the current time. We also used the Corpus of Contemporary American English (COCA)[11], which contains more than one billion words of text (25 million words each year 1990-2019) of different genres, and the Sibol Corpus,[12] which contains newspapers data from 1993 to 2013. The comparison of the relative frequency of the selected terms (the full table is not displayed for space reason) shows that in particular the two terms *emigrant* and *immigrant* changed the relative frequency across time: in the past the term *immigrant* was less frequent than the present, while *emigrant* was more frequent in the past. Asylum seeker and refugee are two relatively recent terms (at least their use).

Italian keywords As for Italian, we needed to select a set of comparable search terms between English and Italian [16]. We initially checked different sources (newspaper articles, glossaries, books on migration) in order to identify a preliminary list of plausible candidate query terms. We identified the following words: *migrante/i*, *immigrato/i/a/e*, *immigrante/i*, *emigrante/i*, *emigrato/i/a/e* as translations of *migrant/s*, *immigrant/s*, and *emigrant/s*, *rifugiato/i/a/e*, *profugo/i/a/e*, *clandestino/i/a/e* and richiedente/i asilo as translations of *refugee/s* and *asylum seeker/s*. We excluded[13] straniero/i/a/e (foreigner) because, as argued by [15], it is used more in its adjectival function than as a noun and this would be problematic since it would introduce data which are not relevant for my research purpose in the corpus. We consulted four different corpora: the diachronic Diacoris Corpus,[14] a 15 million words collection of written Italian texts produced between 1861 and 1945; the Pais,[15] a 250 million words corpus of Italian web texts produced in 2010; ItTenTen16,[16] a 5 million words collection of Italian web texts produced in 2016; La Repubblica,[17] a 380 million words corpus of Italian newspaper texts published between 1985 and 2000. These corpora can be regard as representative dataset of the Italian language, including both the 20th and 21st century, because it spans more than 150 years, from 1861 to 2016. Focusing on the aforementioned terms, we looked at the most frequent words over time to in order to define a representative set of search terms for both the past and the present. This way, we discarded terms which did not have a significant relative frequency. The comparison of the relative frequency of the selected terms (the full table is not displayed for space reason) in the aforementioned corpora shows that the best candidate translation for migrant, immigrant and emigrant were *migrant**, *immigrat**, *immigrant**, *emigrant**, *emigrat**; for

[11] https://www.english-corpora.org/coca/

[12] https://www.sketchengine.eu/sibol-corpus/

[13] Our main interest is on the people who actually migrate and not on the issue of migration in general.

[14] http://corpora.dslo.unibo.it/DiaCORIS/

[15] https://www.corpusitaliano.it/

[16] https://www.sketchengine.eu/ittenten-italian-corpus/

[17] https://corpora.dipintra.it/public/run.cgi/first_form

Table 1. Statistics about each corpus with Type/Token ratio (TTR)

Corpus	Years	Documents	Tokens	Types	TTR
New York Herald	1900-1914	9,119	64,061,101	3,085,080	4.82%
La Stampa	1900-1914	3,092	19,396,796	899,688	4.64%
New York Times	2000-2014	125	58,915,060	308,251	0.52%
La Stampa	2000-2014	62	15,324,728	282,318	1.84%

refugee and asylum seeker the candidate Italian terms were *rifugiat**, *profug**, *clandestin** and *richiedent* asil**.

2.3 Corpora Statistics

After the identification of the two sets of query terms, we compiled four datasets. In Table 1, we show a summary of the statistics for each corpus. The tokens and types values represent the total number of occurrences versus the number of unique words, respectively. We report the type/token ratio (TTR) which serves as indicator of lexical diversity [3]. The differences between the older and the newer datasets were unexpectedly high and it is very unlikely due to chance. As shown in Table 1, the older datasets relative to the period 1900-14 show a dramatically higher number of TTR.

A careful analysis of a sample of texts showed that in both the old corpora there were a lot of misspellings or non-meaningful words caused by the OCR software which produced those documents. For example, there are many occurrences of the sequence *tbe* instead of *the* in the English corpus, as well as many occurrences of *cho* in the Italian corpus instead of *che* (that). In the following section, we describe the semi-automatic procedure for the detection and correction of these OCR errors.

3 OCR Error Detection and Correction

3.1 Background

As argued by [12], 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;
- Real-word errors: valid words which are in the wrong context.

The former are easier to identify but more difficult to correct. Contrarily, the latter are easier to correct but more difficult to identify.

The main approaches to OCR post-processing error correction are

1. a dictionary-based approach which aims at the correction of isolated errors without considering the context [1];

2. a context-based approach which takes into account the grammatical context of occurrence [11].

The former 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 latter approach intends to overcome the problems of the dictionary-based, however it requires more effort in terms of time and energy invested and is characterized by a lower level of efficiency in terms of automation. Moreover, the procedures which are generally adopted to overcome OCR errors [17, 1] do not work properly in these particular cases because these method make use of the linguistic context which is, in turn, compromised and non-corrected. For this reason, it is necessary to develop a semi-automatic approach which mix quantitative and qualitative methodologies.

3.2 Our Proposal

In this paper, we propose a semi-automatic mixed approach to OCR detection which brings together the dictionary-based and the context-based approaches. The first problem in our case study concerns the fact that we did not have the corresponding ground truth version of the corpora. Therefore, we decided to use the contemporary corpora where the text was digital since the beginning. The error detection correction task consisted in a three-step procedure:

1. Definition of a list of plausible error candidates by comparing the list of words of the old corpus with the new corpus. The words that do not appear in the latter, or that have a statistically significant difference in frequency, compose a list of plausible error candidates. For example, the previously mentioned expressions *tbe* or *te*, in the English dataset, and *cho* and *olla* in the Italian dataset. Successively, each error candidate has been qualitatively analysed by being manually observed through concordance lines within its context of occurrence in order to verify if it was an error or not. Lastly, a list of detected errors has been produced.
2. Analysis and categorization of the error in the list of candidates. Each error is categorised according to three categories: i) Standard Mapping: the error contains the same number of characters than the respective correct form. For example: 'hear' (correct) vs 'jear' (error); ii) Non-standard Mapping: the error contains a higher or a smaller number of characters than the correct form. For example: 'main' (correct) vs 'rnain' (error); iii) Split errors: the word is interpreted by the OCR as two distinct words. This is a very common error when digitalizing newspapers because of the shape of the column in which articles are written. For example: 'department' vs 'depart' and 'ment'.
3. Define 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)[18].

[18] Ambiguous and dubious cases where two or more plausible corrections were available, were not inserted in the list to avoid compromising the validity of the corpora.

Table 2. Statistics about errors before and after OCR corrections.

	Before OCR correction		After OCR correction		Difference	
Corpus	Tokens	Types	Tokens	Types	Δ Tokens	Δ Types
NY Herald 1900-1914	**64,061,101**	**3,085,080**	**64,246,208**	**3,082,880**	**+0.29%**	**-0.04%**
La Stampa 1900-1914	**19,396,796**	**899,688**	**19,396,558**	**899,676**	∼**-0.0%**	∼**-0.0%**

Table 3. Examples of standard and non-standard errors and corrections.

Type	Error	Correction
Standard	olla	alla
	alia	alla
	cho	che
	cne	che
	ohe	che
	die	che
Non-standard	clic	che
	clie	che
	Clie	che
	colleglli	colleghi
	eia	da
	eli	di

3.3 A 'Tidy' Implementation

The implementation of these procedure follows the principles described by [18] where the idea is to efficiently and effectively mine textual information from large text collections by means of pipelines in order to 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' [19], that implements this idea of pipelined in a clear way. For space reasons, we are not going to describe in detail the code and we will make the source code used in our experiments available online.[20]

3.4 Post-hoc analysis

A total of 476 errors for English and 80 errors for Italian have been manually individuated and, respectively, as many correcting rules have been written for each language.[21] The automatic application of the rules produced 722,371 substitutions for English and 99,255 substitutions for Italian. As the Table 3 shows, for both the American and the Italian corpora, the number of Tokens and Types have been moderately changed. As a general comment, it is not easy to predict in what way OCR correction will work. On one side, an increase in the number of tokens might happen because many errors, such as *p/r* or *th/* were not

[19] `https://www.tidyverse.org`

[20] `https://github.com/gmdn`

[21] In this analysis, we excluded the split errors since this type of error require a longer evaluation procedure given the amount of false positives errors.

previously recognized as valid tokens. On the other side, 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 four words are counted as four different types. Then, by correcting substituting all these words with *the*, the number of types is reduced of three units. Similarly, the Italian female article *la* has been misspelled as *jd*, *ja*, *ln*. These three words are counted as three different types. Then, by correcting all these words with *the*, the number of types is reduced of three units. The correction task has been repeated four times for English corpus and two times for Italian. Any ambiguous and dubious case, such as *ii* and *ih* in Italian which could be corrected were not corrected to not compromise the validity of the corpora.

4 Final Remarks and Future Work

In this paper, we described a procedure for collecting and creating corpora for discourse analysis from old paper documents and we presented a semi-automatic method for detection and correction of OCR errors. The semi-automatic approach for correcting OCR errors developed for this project has proved to be effective. Despite the fact that the rules produced for the corrections may be less useful with other corpora, the methodology itself is applicable to different contexts.

We are currently investigating how to evaluate the set of rules which have been developed for the actual project to other corpora in order to verify if it is successfully applicable to different contexts. Secondly, we would replicate the same methodology for other set of documents and produce a different set of rules which would be compared with the ones developed for the actual work. Our aim is to create rules which can be generally reused from everyone for correcting their own OCR processed documents. Given the number of substitutions (for English we were close to a million of substitutions), it is important to understand the number false positives introduced. In this sense, we will explore how to evaluate the rules in a semi-automatic way and produce a ground truth. Recent papers have explored advanced automatic corrections based on edit distances, n-grams and neural models [12]. They are successful indeed, but they all introduce some kind of error that may affect the qualitative analysis that CADS need.

There are still open questions that we will investigate in this line of work: for example, 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 these types of keyword search error can affect a CADS analysis? For this reason, following [4], we intend to use error models as a means to measure the relative risk of mismatch between search terms and the targeted resources posed by OCR errors. We also want to compare our analysis with recent approaches that make use of BERT pre-trained neural networks to post-hoc error correction [13], especially in those cases where the context is not clear given multiple OCR errors in the same paragraph.

References

1. Bassil, Y., Alwani, M.: OCR post-processing error correction algorithm using google online spelling suggestion. CoRR **abs/1204.0191** (2012), http://arxiv.org/abs/1204.0191
2. Bazzo, G.T., Lorentz, G.A., Suarez Vargas, D., Moreira, V.P.: Assessing the impact of ocr errors in information retrieval. In: Jose, J.M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M.J., Martins, F. (eds.) Advances in Information Retrieval. pp. 102–109. Springer International Publishing, Cham (2020)
3. Brezina, V.: Statistics in corpus linguistics: A practical guide. Cambridge University Press (2018)
4. Chiron, G., Doucet, A., Coustaty, M., Visani, M., Moreux, J.: 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). pp. 1–4 (2017)
5. Cohen, R.: The Cambridge Survey of World Migration. Cambridge University Press (1995). https://doi.org/10.1017/CBO9780511598289
6. Colucci, M.: Storia dell'immigrazione straniera in Italia. Carocci (2019)
7. Comte, E.: The history of the European migration regime: Germany's strategic hegemony. Routledge (2017)
8. Gabrielatos, C.: Selecting query terms to build a specialised corpus from a restricted-access database. ICAME Journal **31**, 5–44 (Apr 2007)
9. Gallo, S.: Senza attraversare le frontiere: le migrazioni interne dall'unità a oggi. Gius. Laterza & Figli Spa (2012)
10. Kettunen, K., Mäkelä, E., Ruokolainen, T., Kuokkala, J., Löfberg, L.: Old content and modern tools - searching named entities in a finnish ocred historical newspaper collection 1771-1910. Digit. Humanit. Q. **11**(3) (2017), http://www.digitalhumanities.org/dhq/vol/11/3/000333/000333.html
11. Kissos, I., Dershowitz, N.: OCR error correction using character correction and feature-based word classification. CoRR **abs/1604.06225** (2016), http://arxiv.org/abs/1604.06225
12. Nguyen, T., Jatowt, A., Coustaty, M., Nguyen, N., Doucet, A.: Deep statistical analysis of ocr errors for effective post-ocr processing. In: 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL). pp. 29–38 (2019)
13. Nguyen, T.T.H., Jatowt, A., Nguyen, N.V., Coustaty, M., Doucet, A.: 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. pp. 333–336. JCDL '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3383583.3398605, https://doi.org/10.1145/3383583.3398605
14. Partingron, A., Duguid, A., Taylor, C.: Patterns and meanings in discourse: Theory and practice in corpus-assisted discourse studies (CADS), Studies in Corpus Linguistics, vol. 55. John Benjamins (2013)
15. Taylor, C.: The representation of immigrants in the italian press. CIRCaP Occasional Papers **21**, 1–40 (2009)
16. Taylor, C.: Investigating the representation of migrants in the uk and italian press: A cross-linguistic corpus-assisted discourse analysis. International Journal of Corpus Linguistics **19**(3), 368–400 (2014). https://doi.org/10.1075/ijcl.19.3.03tay, http://sro.sussex.ac.uk/id/eprint/50044/
17. Tong, X., Evans, D.A.: A statistical approach to automatic OCR error correction in context. In: Fourth Workshop on Very Large Corpora (1996), https://www.aclweb.org/anthology/W96-0108

18. Wachsmuth, H.: Text Analysis Pipelines - Towards Ad-hoc Large-Scale Text Mining, Lecture Notes in Computer Science, vol. 9383. Springer (2015). https://doi.org/10.1007/978-3-319-25741-9, https://doi.org/10.1007/978-3-319-25741-9