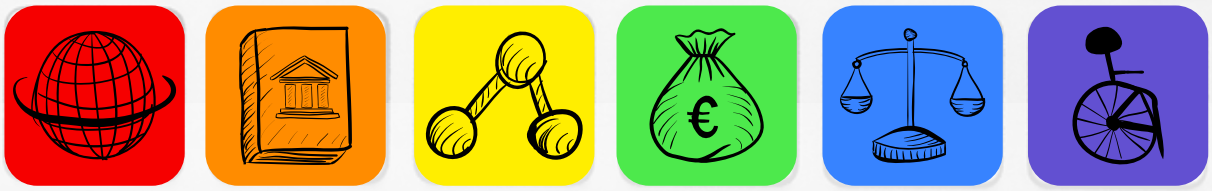
DIGITAL PUBLIC HUMANITIES

OPEN CULTURE

RETI SOCIALI

ASSOCIAZIONE per l'INFORMATICA UMANISTICA e la CULTURA DIGITALE



AIUCD2021

AIUCD 2021 - DH per la società: e-guaglianza, partecipazione, diritti e valori nell'era digitale.

Raccolta degli abstract estesi della 10° conferenza nazionale

AIUCD 2021 - DHs for society: e-quality, participation, rights and values in the Digital Age.

Book of extended abstracts of the 10th national conference

TECH ECONOMY E-PARTICIPATION TECNOLOGIE ASSISTIVE

ISBN: 9788894253559

ASSOCIAZIONE PER
L'INFORMATICA UMANISTICA
E LA CULTURA DIGITALE

Federico Boschetti, Angelo Mario Del Grosso, Enrica Salvatori (edd.), *AIUCD 2021 - DH per la società: e-guaglianza, partecipazione, diritti e valori nell'era digitale. Raccolta degli abstract estesi della 10ª conferenza nazionale*, Pisa, 2021.
Federico Boschetti, Angelo Mario Del Grosso, Enrica Salvatori (edd.), *AIUCD 2021 - DHs for society: e-quality, participation, rights and values in the Digital Age. Book of extended abstracts of the 10th national conference*, Pisa, 2021.

Immagine di copertina | cover image: Chiara Mannari, Università di Pisa.
Editing: Mario Valori
Gestione bibliografica | Bibliographic standardization: Gaia Sitri

Ogni link citato era attivo al 14 gennaio 2021, salvo ove diversamente indicato.
All links have been visited on 19th January 2021, unless otherwise indicated

Si prega di notificare all'editore ogni omissione o errore si riscontri, al fine di provvedere alla rettifica.
Please notify the publisher of any omissions or errors found, in order to rectify them.
aiucd.segreteria [at] aiucd.org

Gli abstract estesi pubblicati in questo volume hanno ottenuto il parere favorevole da parte di valutatori esperti della materia, attraverso un processo di revisione anonima mediante double-blind peer review sotto la responsabilità del Comitato Scientifico di AIUCD 2021.

All the extended abstracts published in this volume have received favourable reviews by experts in the field of DH, through an anonymous double-blind peer review process under the responsibility of the AIUCD 2021 Scientific Committee.

Il programma della conferenza AIUCD è disponibile online
The AIUCD 2021 conference program is available online
https://aiucd2021.labcd.unipi.it/

Enti organizzatori / Organizing institutions:

AIUCD, LabCD dell'Università di Pisa, CLARIN, ILC-CNR, ISTI-CNR

Lista dei revisori - List of the reviewers

Agnese Addone, Alessandra Donnini, Alessandro Lenci, Alessandro Perissinotto, Alina Sirbu, Ana Iglesias Maqueda, Andrea Bellandi, Angelica Lo Duca, Angelo Mario Del Grosso, Anna Galluzzi, Anna Maria Tammaro, Antonio Narzisi, Armando Stellato, Barbara Guidi, Barbara Leporini, Benedetta Iavarone, Caterina Senette, Chiara Mannari, Christian D'Agata, Claudio Forziati, Cristiano Amendola, Cristina Marras, Cristina Pattuelli, Daniela Fogli, Daniela Rotelli, Daria Spampinato, Deborah Paci, Dino Buzzetti, Dominique Brunato, Edmondo Grassi, Elisa D'Argenio, Emanuela Reale, Emanuele Luigi Colazzo, Emiliano Giovannetti, Emma Pietrafesa, Enrica Salvatori, Eva Sassolini, Fabiana Guernaccini, Fabio Ciotti, Fabio Ciracì, Fabio Pacini, Federico Boschetti, Federico Meschini, Francesca Di Donato, Francesca Frontini, Francesca Levi, Francesca Pratesi, Francesca Tomasi, Francesco Grisolia, Francesco Romano, Franz Fischer, Gino Roncaglia, Giorgio Guzzetta, Giorgio Maria Di Nunzio, Giovanni Michetti, Giovanni Scarafile, Giuliana Vitiello, Giulio Mori, Giulio Rossetti, Giuseppe Andrea L'Abbate, Hansmichael Hohenegger, Ioanna Miliou, Irene Sucameli, Javier Gomez, Jisu Kim, Jorge Morato, Juan Carlos Torrado Vidal, Laura Giarré, Laura Pollacci, Letizia Cortini, Lucia Ferlino, Luigi Bambaci, Luigi Catalani, Luigi Siciliano, Marco Manca, Maria Chiara Pievatolo, Maria Claudia Buzzi, Maria Simi, Mariasole Bondioli, Marilena Daquino, Marina Buzzi, Marina Pietrangelo, Marion Hersh, Massimiliano Gambardella, Massimiliano Grava, Massimo Magrini, Massimo Zancanaro, Matteo Sanfilippo, Maurella Della Seta, Maurizio Caminito, Maurizio Lana, Maurizio Vivarelli, Michela Natilli, Michele Coscia, Mirko Tavosanis, Nadia Sansone, Nicola Lettieri, Paola Galimberti, Paolo Bottoni, Paolo Mogorovich, Paolo Monella, Paolo Rossi, Pierluigi Feliciati, Rachele Sprugnoli, Raffaele Gareri, Riccardo Del Gratta, Roberto Delle Donne, Roberto Pellungrini, Roberto Rosselli Del Turco, Rolando Minuti, Ruggero G. Pensa, Salvatore Citraro, Sara Conti, Sebastiano Faro, Serge Noiret, Silvia Corbara, Silvia Mirri, Silvia Piccini, Simona Turbanti, Simone Rebora, Simonetta Montemagni, Stefano Chessa, Stefano Rollo, Susanna Pelagatti, Timothy Tambassi, Tito Orlandi, Tiziana Mancinelli, Tiziana Possemato, Tommaso Agnoloni, Vindice Deplano, Vittore Casarosa.

Indice - Table of Contents

Prefazione

Il decimo convegno annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale ha nell'edizione di quest'anno un titolo peculiare e importante: *DH per la società: e-guaglianza, partecipazione, diritti e valori nell'era digitale*.

Questo è in linea con lo spirito più profondo dell'Associazione, che fin dalle sue origini ha voluto essere un luogo dove filologi, bibliotecari, storici, giuristi, informatici e ingegneri informatici si incontrano e si confrontano (informatica umanistica) per incidere su aspetti rilevanti della società (culture digitali). Abbiamo deciso quindi di discutere su domande forti invitando ognuno a uscire dalla propria *comfort zone* per praticare l'impegno civile necessario a far parlare accademia e società.

Il direttivo dell'associazione e gli organizzatori si sono di fatto impegnati in questa occasione per fornire un aiuto concreto al cambiamento epocale in atto. Il mondo dei negozi sotto casa, delle lezioni tenute solo in presenza, della tv, dell'ufficio postale, del telefono ha lasciato spazio ad aziende di commercio elettronico su scala mondiale, alla didattica a distanza (DAD), alle *newsletter*, ai *social network*, alle *chat* sugli *smartphone*. È una rivoluzione in atto già da tempo e, oggi, abbiamo urgente e profondo bisogno di strumenti concettuali e di ricerca per gestirla. Perché le rivoluzioni non portano solo benefici, ma anche disagi e diseguaglianze. Le *Digital Humanities* (DH) sono il campo di studi in cui si elaborano e si applicano gli strumenti per gestire questo cambiamento. La conferenza vuole quindi rappresentare proprio un momento di approfondimento e di riflessione dell'informatica umanistica come luogo privilegiato di incontro tra i diversi bisogni della società contemporanea, perché siamo convinti che si debba restituire all'umanista il ruolo di chi interpreta e accompagna il cambiamento, di chi favorisce la cultura aperta, la partecipazione, i diritti, i nuovi e antichi valori.

In sostanza c'è la necessità di passare da una fase in cui le *Digital Humanities* sono prevalentemente concentrate sui metodi e sui mezzi, ad una fase in cui esse? tornino a ragionare sui fini, vale a dire sugli obiettivi di ricerca e sui benefici prodotti dal raggiungimento di tali obiettivi per la società nel suo complesso. Nel decimo libro della *Repubblica*, Platone fa notare che «per ogni oggetto esistono tre arti: quella che ne farà uso, quella che lo realizzerà e quella che lo imiterà ..» «Ma la virtù, la bellezza, la perfezione di ogni singolo oggetto riguardano soltanto l'uso per il quale ciascuno di essi è fabbricato», e allora «chi adopera ogni singolo oggetto deve per forza averne la maggiore esperienza e riferire al fabbricante i pregi e i difetti che si rivelano all'uso; ad esempio un flautista dà spiegazioni al costruttore di flauti sugli strumenti che gli servono nel suo mestiere e gli ordinerà come fabbricarli» (601d-e, trad. Caccia). L'umanista digitale non costruisce né flauti né pifferi, anche se talvolta viene scambiato ingiustamente per un pifferaio magico, ma collabora alla progettazione di strumenti che servono ad abbattere le barriere fra la conoscenza o l'esperienza estetica e la nuova società inclusiva. Non si deve tuttavia cadere nel tranello secondo cui il digitale per sé, grazie ad un indefinito potere taumaturgico, abbatte le barriere. Al contrario, uno strumento digitale può innalzare barriere nuove, se non è stato progettato bene, se non è stato pensato per essere così duttile da adattarsi alle esigenze specifiche di chi lo usa.

Ecco perché ad alcuni di noi stanno particolarmente a cuore, fra le molteplici tracce del convegno, il tema della cultura aperta – o cultura dell'apertura – e il tema dell'accessibilità. La cultura dell'apertura ci spinge ad essere esigenti, a voler essere sempre più liberi di usare gli strumenti che ci vengono consegnati in modi inattesi rispetto alle previsioni dei costruttori stessi: noi i libri elettronici non li vogliamo solo leggere (per quello basta il libro a stampa!) o leggere ovunque (anche in questo caso, un libro tascabile può bastare); noi vogliamo costruire indici, estrarre e mettere in relazione dati, visualizzare nuvole di parole, usufruire dei risultati di raffinate analisi semantiche..

Infine, ragionare tutti insieme sul tema dell'accessibilità alle risorse digitali ci fa vedere come negli ultimi decenni, silenziosamente e quasi invisibilmente, la società sia diventata di fatto più inclusiva, dove le persone con disabilità sensoriali, motorie o cognitive, o i discenti nelle varie fasi dell'età evolutiva, non sono affatto beneficiari passivi di risorse e strumenti creati per loro, ma sono protagonisti attivi nella progettazione di risorse, strumenti, percorsi ideati e, grazie a questo, migliori per tutti.

L'incarico di organizzare il decimo convegno dell'Associazione AIUCD fu dato a Pisa un anno fa e, in particolare, annunciato durante la sessione di chiusura della nona conferenza annuale dell'associazione, svoltasi a Milano dal 15 al 17 gennaio 2020, presso l'Università Cattolica del Sacro Cuore. Per gli amanti della cabala, erano le 17 circa del 17 gennaio 2020, ed era un venerdì. Già allora era manifesto il presagio che il nostro sarebbe stato un lavoro complicato: ma all'epoca la straordinaria emergenza sanitaria dovuta alla diffusione del virus SARS-COV-2 era ancora soltanto una epidemia locale di una regione della Cina, e la sola sfida che ci sembrava di dover affrontare era relativa alla proposta del nuovo sguardo che le DH avrebbero avuto verso la società.

Da quel giorno fino ad oggi, molte istituzioni del variegato mondo della ricerca pisana hanno lavorato in grande sinergia per preparare al meglio l'evento associativo più importante per la comunità italiana di informatica umanistica.

Nonostante le notevoli difficoltà incontrate a causa del COVID-19 e alle conseguenti misure di sicurezza via via deliberate, possiamo dire che i "numeri" del convegno danno la misura di un serio interesse verso i temi proposti e di un conseguente indice di successo.

Le date e i dati principali della decima conferenza si possono riassumere in 118 contributi inviati, dalla data di pubblicazione della call for paper – il 19 giugno 2020 – fino al termine stabilito del 2 ottobre 2020. Dei contributi inviati, 96 lavori sono stati presentati come comunicazione orale e i restanti 22 presentati come poster. Il processo per l'invio delle proposte ha previsto la scelta di uno tra i sei possibili temi della conferenza: *Digital public humanities, Open culture, Reti sociali, Tech-economy, e-Participation* e infine *Tecnologie assistive per l'inclusione.*

La distribuzione dei contributi ha visto 35 proposte inviate per il tema *Digital public humanities*; 43 per *Open Culture*; 14 *Reti sociali;* 3 *Tech-economy*; 6 *e-Partecipation*; e 17 *Tecnologie assistive per l'inclusione.*

Promuovendo una politica di inclusione intesa a concedere ampio spazio di discussione e promozione per il maggior numero di attività di ricerca possibile, sono state accettate 100 proposte (85% circa), revisionate da un totale di 125 colleghi per un totale di 318 revisioni (per una media di 2.7 revisioni per contributo).

Delle 100 proposte, 60 sono state accettate come presentazioni orali (63% circa delle proposte inviate per una presentazione orale) e 40 come poster, di cui 18 proposte erano originariamente state inviate come poster (82% circa sul totale delle proposte inviate come poster) e 22 erano state pensate dagli autori come talk (23% circa del totale delle proposte originariamente inviate come presentazioni orali).

Il numero totale degli autori relativi ai contributi inviati è stato pari a 295, di cui 268 sono autori di proposte accettate.

Nella giornata di pre-convegno del 19 gennaio sono stati inoltre organizzati quattro workshop:

1. *Introduzione alle edizioni digitali: preparazione con codifica XML TEI e visualizzazione con il software EVT*, a cura di Roberto Rosselli Del Turco;
2. *Narrativa e divulgazione scientifica delle DH: l'esperienza dei QUARANTIP*, a cura di Krino;
3. *Qui CLARIN-IT: posso aiutarti? / This is CLARIN, how can we help you?* a cura di Monica Monachini e Francesca Frontini;
4. *ALDiNa, Archivi Letterari Digitali Nativi*, a cura di Tiziana Mancinelli, Emmanuela Carbè e Federico Boschetti

Infine l'ultimo numero, quello dei circa 650 iscritti al convegno: frutto non solo dell'apertura delle tematiche, ma, indubbiamente, anche della nuova modalità di partecipazione a distanza. Questo credo ci dovrà far molto riflettere su cosa fare in futuro, a pandemia speriamo archiviata, quando dovremo ragionare su come conciliare il piacere e l'utilità dell'incontro in presenza con il dovere morale di favorire una partecipazione democratica.

La conferenza, allargandosi a temi fino ad oggi non molto dibattuti entro l'associazione, ha comportato diverse aperture verso i ricercatori attivi negli ambiti della sanità, dell'economia, della giurisprudenza, delle reti sociali. Un'apertura che ci ha ovviamente arricchito, ma che ci ha anche posto nuove sfide. Le DH da anni sono una metadisciplina dai confini sfumati e in via di definizione, una galassia dalle numerose e diverse componenti: quale può essere la forza aggregante che le tiene insieme? Forse ce lo potranno dire i giovani partecipanti al convegno: numerosi studenti e neolaureati quest'anno hanno mandato le loro proposte che sono state valutate e accettate dai revisori. Un'ulteriore apertura di cui siamo particolarmente orgogliosi.

Ogni comunità di studiosi e studenti con cui AIUCD viene quotidianamente in contatto ha il proprio linguaggio specialistico e le proprie buone pratiche. Speriamo che la condivisione di tali linguaggi e di tali pratiche promuova la co-evoluzione delle diverse comunità che hanno sentito spontaneamente la necessità di accorciare le rispettive distanze. Paradossalmente – ma forse non tanto – la distanza fisica imposta dalle circostanze ha promosso, nelle estenuanti videoconferenze dell'ultimo anno, il contatto virtuale di persone appartenenti a mondi epistemologicamente paralleli.

Non ci bastava aver proposto un tema nuovo e nuove aperture. Ci ha messo lo zampino anche questa terribile pandemia, costringendoci a organizzare un evento completamente on line. Dal 12 marzo 2020 il nostro paese è entrato in lockdown e tutti gli eventi di socialità e collettivi sono stati di fatto fortemente limitati o comunque sono stati trasformati in eventi virtuali con partecipazione remota. La stessa sorte è toccata alla conferenza AIUCD2021.

Ci siamo trovati quindi subito a dover valutare le piattaforme più adatte e robuste per ospitare il convegno in modalità remota. Ovviamente non potevamo replicare in rete un convegno tradizionale: l'Associazione sa bene che il mezzo digitale può e deve mutare le forme di comunicazione. Si è trattato quindi di un'esperienza decisamente diversa, più interattiva, meno formale e con "ritualità" diverse rispetto al passato.

Chiusi nelle proprie case, isolati nelle proprie città, i nostri autori si sono trovati costretti a fare ciò che fino a poco tempo fa poteva essere, per l'umanista, solo un'opzione: molti si sono trovati a dover usare in modo massiccio strumenti per la scrittura collaborativa, a dover modificare quindi non solo le proprie abitudini ma anche i propri paradigmi cognitivi. In questo spirito di collaborazione sono state organizzate anche le sessioni del convegno, prendendo a modello l'ultima conferenza CLARIN.

Fortunatamente abbiamo avuto fin da subito un grande appoggio e una grande disponibilità dei mezzi e delle risorse proprio dall'infrastruttura CLARIN, che sia nella sua declinazione Italiana, che è partner della conferenza, sia nella sua declinazione di coordinamento europeo ci ha permesso di organizzare l'evento completamente online.

Per questo il comitato organizzatore e di programma della conferenza vuole rivolgere un sentito e accorato ringraziamento agli amici di CLARIN-IT, soprattutto nelle persone di Monica Monachini, coordinatrice del nodo Italiano di CLARIN, e di Francesca Frontini, da poco entrata nel board europeo di CLARIN-ERIC. In più, un doveroso ringraziamento anche per i tanti colleghi di CLARIN-ERIC, soprattutto nella persona di Franciska de Jong, direttrice esecutiva, e di Maria Eskevich (Central Office Coordinator), che ha tenuto i contatti con gli organizzatori della conferenza.

Ugualmente grati siamo ai giovani studenti di Krino per il supporto nelle giornate del convegno, all'Università di Pisa, all'ISTI-CNR e all'ILC-CNR, co-organizzatori assieme ad AIUCD, e infine allo staff di Media Events di UNIPI per il servizio di post-produzione video.

Federico Boschetti, Angelo Mario Del Grosso ed Enrica Salvatori

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

Dario Del Fante[1], Giorgio Maria Di Nunzio[2]
[1] Università di Padova – dario.delfante [at] phd.unipd.it
[2] Università di Padova – giorgiomaria.dinunzio [at] 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 compiling historical corpora. We present a case study on newspapers of the beginning of the 20th century for the linguistic analysis of the metaphors representing immigrants.

KEYWORDS

corpus-assisted discourse studies, OCR detection, OCR correction

TALK

1 INTRODUCTION

In Corpus-assisted Discourse Studies [9], 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, and it is often necessary to manually transcribe the texts under analysis or to use Optical Character Recognition (OCR) software which have a fundamental role in the study of digitizes manuscripts [6]. However, OCR errors may significantly affect the compilation of a corpus in CADS [2]. There are procedures which are adopted to correct OCR errors [1] may not work properly in those cases where the quality of the scan is poor. In this paper, we propose a replicable semi-automatic method for detection and correction of OCR errors. The outcome of this project consists of a set of rules which are, eventually, valid for a different context 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.

2 SEARCHING FOR METAPHORS TO REPRESENT IMMIGRANTS IN 1900

Our case study is 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 since the beginning of the XX century. Given the limited space available in this paper, we will refer to only one particular moment in history which have a significant value in relation to migratory movement: from 1900 to 1914, just before World War I, because this time period, as in [4], is a particularly significant moment for migratory movements, specifically from Europe to the U.S. The availability of data, the newspaper political leaning, and the registration fees were additional constraints that narrowed the selection of the newspapers to the New York Herald[1] for the U.S., and La Stampa[2] for Italy.

Afterwards, we needed to select the keywords to filter the articles useful for our study. The starting point for English was the set of words identified by [5] named under the acronym RASIM: *refugee**,[3] *asylum seeker**, *immigrant**, and *migrant**. We added a fifth word to this list: *emigrant**. As for Italian, we needed to select a set of comparable search terms between English and Italian. We consulted the diachronic Diacoris Corpus,[4] a 15 million words collection of Italian texts produced between 1861 and 1945. The best candidate translation for migrant, immigrant and emigrant were *migrant**, *immigrat**, *immi- grant**, *emigrant**, *emigrat**; for refugee and asylum seeker the candidate Italian terms were *rifu-giat**, *profug**, *clandestin** and *richiedent* asil**.

In the first two rows of Table 1, we show a summary of the statistics for each compiled corpus. The tokens and types of values represent the total number of occurrences versus the number of unique words, respectively. We report both the type/token ratio (TTR) and the standardized type/token ratio (STTR).

3 OCR ERROR DETECTION AND CORRECTION

In this section, we propose a semi-automatic mixed approach to OCR detection, which brings together the dictionary-based and the context-based approaches. A careful analysis of a sample of texts showed that there were a lot of misspellings or non-meaningful words in both corpora caused by the OCR software. 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 compile two contemporary newspaper corpora whose texts were digitalized since the beginning: The New York Times for the U.S., La Stampa for Italy (see last two rows of Table 1).

---

1 http://chroniclingamerica.loc.gov

2 http://www.archiviolastampa.it

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

4 http://corpora.dslo.unibo.it/DiaCORIS/

The error detection correction task consisted of a three-step procedure:

1. Detection of errors 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.
2. Analysis and categorization of the error in the list of candidates: i) an error containing the same number of characters than the respective correct form.; ii) an error containing a higher/ smaller number of characters than the correct form; iii) a word interpreted by the OCR as two distinct words (i.e., 'department' vs 'depart' and 'ment').

Define the error correction rule as a regular expression.
The implementation of these procedure follows the principles described by [10] where the idea is to mine textual information from large text collections in an efficient and effective by means of pipelines allowing for a sequential process of text analysis. For our experiments, we used the R programming language, which has a set of packages, named 'tidyverse'[5], that implements this idea[6]. A total of 2,313 errors for English and 269 errors for Italian have been individuated and, respectively, as many correcting rules have been written for each language.

Corpus	Years	Documents	Tokens	Types	TTR	STTR
New York Herald	1900-1914	8,540	55,796,968	2,326,897	4.17%	50.24%
La Stampa	1900-1914	3,092	18,773,664	817,865	4.36%	56.63%
New York Times	2000-2014	125	58,915,060	308,251	0.52%	48.39%
La Stampa	2000-2014	62	15,332,063	275,103	1.79%	62.78%

Table 1: statistics

In general, it is not easy to predict in what way OCR correction will work (see Table 2). On one side, the Italian corpus dimensions have been increased in relation to the number of tokens. The increase of tokens might be because many errors were not previously recognized as valid tokens. On the other side, the English corpus dimension has been decreased in relation to the number of tokens. The decrease might be due to the correction of split errors, such as *depart* and *ment*, corrected in *department*.

Corpus	Tokens (B)	Types (B)	Tokens (A)	Types (A)	Δ Tokens	Δ Types
New York Herald 1900-14	55,796,968	2.326.897	55.555.708	2.323.790	-0.43 %	-0.13 %
La Stampa 1900-14	18.773.664	817.865	18.778.210	817.858	0.02%	-0.001%

Table 2: Statistics about errors before (B) and after (A) OCR corrections

5 https://www.tidyverse.org

6 https://github.com/gmdn

4 FINAL REMARKS AND FUTURE WORK

In this paper, we presented a semi-automatic method for detection and correction of OCR errors for the discourse analysis of old newspaper documents. The outcome of this project consists in a set of rules which are, eventually, valid for a different context and applicable to different corpora and which can be reproduced and reused. 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 these types of keyword search error can 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 [3]. We also want to compare our analysis with other approaches that make use of BERT pre-trained neural networks to post-hoc error correction [8], 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 [7].

REFERENCES

1. Bassil, Youssef, and Mohammad Alwani. 'Ocr Post-Processing Error Correction Algorithm Using Google Online Spelling Suggestion'. *ArXiv Preprint ArXiv:1204.0191*, 2012.
2. Bazzo, Guilherme Torresan, Gustavo Acauan Lorentz, Danny Suarez Vargas, and Viviane P. Moreira. 'Assessing the Impact of OCR Errors in Information Retrieval'. In *European Conference on Information Retrieval*, 102–109. Springer, 2020.
3. Chiron, Guillaume, Antoine Doucet, Mickaël Coustaty, Muriel Visani, and Jean-Philippe Moreux. '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. IEEE, 2017.
4. Cohen, Robin. *The Cambridge Survey of World Migration*. Cambridge University Press, 1995.
5. Gabrielatos, Costas. 'Selecting Query Terms to Build a Specialised Corpus from a Restricted-Access Database.' *ICAME Journal* 31 (2007): 5–44.
6. 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'. *ArXiv Preprint ArXiv:1611.02839*, 2016.
7. Lund, William B., and Eric K. Ringger. 'Improving Optical Character Recognition through Efficient Multiple System Alignment'. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, 231–240, 2009.
8. 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*, 333–336, 2020.
9. Partington, Alan, Alison Duguid, and Charlotte Taylor. *Patterns and Meanings in Discourse: Theory and Practice in Corpus-Assisted Discourse Studies (CADS)*. Vol. 55. John Benjamins Publishing, 2013.
10. Wachsmuth, Henning. *Text Analysis Pipelines: Towards Ad-Hoc Large-Scale Text Mining*. Vol. 9383. Springer, 2015.