

Language Resources for Historical Newspapers: the *Impresso* Collection

Maud Ehrmann*, Matteo Romanello*, Simon Clematide†, Phillip Benjamin Ströbel†, Raphaël Barman*

*Digital Humanities Laboratory, EPFL

†Institute for Computational Linguistics, Zurich University

*{maud.ehrmann, matteo.romanello, raphael.barman}@epfl.ch.ch

†{siclemat, pstroebel}@cl.uzh.ch

Abstract

Following decades of massive digitization, an unprecedented amount of historical document facsimiles can now be retrieved and accessed via cultural heritage online portals. If this represents a huge step forward in terms of preservation and accessibility, the next fundamental challenge—and real promise of digitization—is to exploit the *contents* of these digital assets, and therefore to adapt and develop appropriate language technologies to search and retrieve information from this ‘Big Data of the Past’. Yet, the application of text processing tools on historical documents in general, and historical newspapers in particular, poses new challenges, and crucially requires appropriate language resources. In this context, this paper presents a collection of historical newspaper data sets composed of text and image resources, curated and published within the context of the ‘*impresso* - Media Monitoring of the Past’ project. With corpora, benchmarks, semantic annotations and language models in French, German and Luxembourgish covering ca. 200 years, the objective of the *impresso* resource collection is to contribute to historical language resources, and thereby strengthen the robustness of approaches to non-standard inputs and foster efficient processing of historical documents.

Keywords: historical and multilingual language resources, historical texts, multi-layered historical semantic annotations, OCR, named entity processing, topic modeling, text reuse, digital humanities

1. Introduction

Digitization efforts are slowly but steadily contributing an increasing amount of facsimiles of cultural heritage documents. As a result, it is nowadays commonplace for many memory institutions to create and maintain digital repositories which offer rapid, time- and location-independent access to documents (or surrogates thereof), allow to virtually bring together disperse collections, and ensure the preservation of fragile documents thanks to on-line consultation (Terras, 2011). Beyond this great achievement in terms of preservation and accessibility, the next fundamental challenge—and real promise of digitization—is to exploit the *contents* of these digital assets, and therefore to adapt and develop appropriate language technologies to search and retrieve information from this ‘Big Data of the Past’ (Kaplan and di Lenardo, 2017).

In this regard, and following decisive grassroots efforts led by libraries to improve OCR (Optical Character Recognition) technology and generalize full-text search over historical document collections (see, e.g., the *Impact*¹ and *Trove*² projects), the Digital Humanities (DH), Natural Language Processing (NLP) and Computer Vision (CV) communities are pooling forces and expertise to push forward the processing of facsimiles, as well as the extraction, linking and representation of the complex information enclosed in transcriptions of digitized collections. These interdisciplinary efforts were recently streamlined within the far-reaching Europe Time Machine project³ which ambitions, in general, the application of artificial intelligence technologies on cultural heritage data and, in particular, to achieve text understanding of historical material.

This momentum is particularly vivid in the domain of digitized newspaper archives, for which there has been a notable increase of research initiatives over the last years. Besides individual works dedicated to the development of tools (Yang et al., 2011b; Dinarelli and Rosset, 2012; Moreux, 2016; Wevers, 2019), or to the usage of those tools (Kestemont et al., 2014; Lansdall-Welfare et al., 2017), events such as evaluation campaigns (Rigaud et al., 2019; Clausner et al., 2019) or hackathons⁴ based on digitized newspaper data sets have multiplied. Additionally, several large consortia projects proposing to apply computational methods to historical newspapers at scale have recently emerged, including *ViralTexts*⁵, *Oceanic Exchanges*⁶, *impresso*⁷, *NewsEye*⁸, and *Living with Machines*⁹ (Ridge et al., 2019). These efforts are contributing a pioneering set of text and image analysis tools, system architectures, and graphical user interfaces covering several aspects of historical newspaper processing and exploitation.

Yet, the application of text processing tools on historical documents in general, and historical newspapers in partic-

⁴See the 2017 edition of the Coding Da Vinci cultural hackathon, <https://www.deutsche-digitale-bibliothek.de/content/journal/aktuell/kicking-coding-da-vinci-berlin?lang=en>

⁵A project aiming at mapping networks of reprinting in 19th-century newspapers and magazines (US, 2012-2016): <https://viraltexts.org>

⁶A project tracing global information networks in historical newspaper repositories from 1840 to 1914 (US/EU, 2017-2019): <https://oceanicexchanges.org>

⁷<https://impresso-project.ch>

⁸A digital investigator for historical newspapers (EU, 2018-2021): <https://www.newseye.eu>

⁹A project which aims at harnessing digitised newspaper archives (UK, 2018-2023): <https://www.turing.ac.uk/research/research-projects/living-machines>

¹<http://www.impact-project.eu>

²<https://trove.nla.gov.au>

³<https://www.timemachine.eu>

ular, poses new challenges (Sporleder, 2010; Piotrowski, 2012). First, the language under study is mostly of earlier stage(s) and usually features significant orthographic variation (Bollmann, 2019). Second, due to the acquisition process and/or document conservation state, inputs can be extremely noisy, with errors which do not resemble tweet misspellings or speech transcription hesitations for which adapted approaches have already been devised (Linhares Pontes et al., 2019a; Chiron et al., 2017; Smith and Cordell, 2018). Further, and due to the diversity of the material in terms of genre, domain and time period, language resources such as corpora, benchmarks and knowledge bases that can be used for lexical and semantic processing of historical texts are rather sparse and heterogeneous. Finally, archives and texts from the past are not as anglophone as in today’s information society, making multilingual resources and processing capacities even more essential (Neudecker and Antonacopoulos, 2016).

Overall, and as demonstrated by Vilain et al. (2007), the transfer of NLP approaches from one domain or time period to another is not straightforward, and performances of tools initially developed for homogeneous texts of the immediate past are affected when applied on historical material (Ehrmann et al., 2016). This echoes the statement of Plank (2016), according to whom what is considered as standard or canonical data in NLP (i.e. contemporary news genre) is more a historical coincidence than an objective evidence or reality: non-canonical, heterogeneous, biased and noisy data is more prevalent than is commonly believed, and historical texts are no exception. In this respect, and in light of the above, it can therefore be considered that historical language(s) belong to the family of less-resourced languages for which further efforts are still needed.

To help alleviate this deficiency, this paper presents a ‘full-stack’ historical newspaper data set collection composed of text and image resources produced, curated and published within the context of the ‘*impresso* - Media Monitoring of the Past’ project¹⁰. These resources relate to historical newspaper material in French, German and Luxembourgish and include: OCRed texts together with their related facsimiles and language models, benchmarks for article segmentation, OCR black letter and named entity processing, and multi-layer semantic annotations (named entities, topic modeling and text reuse). The objective of the *impresso* resource collection is to contribute to historical language resources, and thereby strengthen the robustness of approaches to non-standard inputs and foster efficient processing of historical documents. More precisely, these resources can support:

- (a) NLP research and applications dealing with historical language, with a set of ‘ready-to-parse’ historical texts covering 150 years in French and German, and a set of language models;
- (b) Model training and performance assessment for three tasks, namely article segmentation, OCR transcription and named entity processing (for the first time on such material for the latter), with manually transcribed and annotated corpora;

- (c) Historical corpus exploration and digital history research, with various stand-off semantic annotations.

To the best of our knowledge, the *impresso* resource collection represents the most complete historical newspapers data set series to date. In the following, we introduce the *impresso* project (Section 2), present the *impresso* resource collection (Sections 3, 4 and 5), account for major existing historical language resources (Section 6), and conclude (Section 7).

2. Mining 200 years of historical newspapers: the *impresso* project

impresso - Media Monitoring of the Past’ is an interdisciplinary research project in which a team of computational linguists, designers and historians collaborate on the semantic indexing of a multilingual corpus of digitized historical newspapers¹¹. The primary goals of the project are to apply text mining techniques to transform noisy and unstructured textual content into semantically indexed, structured, and linked data; to develop innovative visualization interfaces to enable the seamless exploration of complex and vast amounts of historical data¹²; to identify needs on the side of historians which may also translate into new text mining applications and new ways to study history; and to reflect on the usage of digital tools in historical sciences from a practical, methodological, and epistemological point of view.

In doing so, *impresso* addresses the challenges posed by large-scale collections of digitized newspapers, namely: (1) newspaper silos: due to legal restrictions and digitisation policy constraints, data providers (libraries, archives and publishers) are bound to provide incomplete, non-representative collections which have been subjected to digitization and OCR processing of varying quality; (2) big, messy data: newspaper digital collections are characterised by incompleteness, duplicates, and abundant inconsistencies; (3) noisy, historical text: imperfect OCR, faulty article segmentation and lack of appropriate linguistic resources greatly affect image and text mining algorithms’ robustness; (4) large and heterogeneous corpora: processing and exploitation requires a solid system architecture and infrastructure, and interface design should favor efficient search and discovery of relevant content; and (5) transparency: critical assessment of inherent biases in exploratory tools, digitized sources and annotations extracted from them is paramount for an informed usage of data in digital scholarship context.

With respect to source processing, *impresso* applies and improve a series of state-of-the-art natural language and image processing components which produce, *in fine*, a large-scale, multilingual, semantically indexed historical newspaper collection. The various lexical and semantic annotations generated thereof are combined and delivered to

¹¹The project is funded by the Swiss National Science Foundation for a period of three years (2017-2020) and involves three main applicants: DHLAB from the Ecole polytechnique fédérale de Lausanne (EPFL), ICL from the University of Zurich, and C²DH from the University of Luxembourg.

¹²<https://impresso-project.ch/app/#>

¹⁰<https://impresso-project.ch>

digital scholars via a co-designed, innovative and powerful graphical user interface. Furthermore, and this is the focus of the present paper, those sources and annotations are also published apart from the interface for further usage by cultural heritage partners, and DH and/or NLP communities. Finally, some of the text and image mining components are subject to systematic evaluation, for which ground truth data are produced.

All publicly released *impresso* resources, i.e. corpora, benchmarks and annotations, are published on the project's website¹³ and on *impresso* zenodo community¹⁴ with detailed documentation. Table 2 summarizes the links and DOIs of the datasets.

3. *Impresso* Corpora

The first resource is a set of normalized, 'ready-to-process' newspaper textual corpora which, for copyrights reasons, do not correspond to the full *impresso* newspaper collection accessible through the interface.

3.1. Original Sources

impresso gathers a consortium of Swiss and Luxembourgish research and cultural heritage institutions and focuses primarily on sources of these countries in French, German, and Luxembourgish. Provided by its partners,¹⁵ *impresso* original sources correspond as of November 2019 to 76 newspapers. Concretely speaking, sources consist of either both OCR output and images, or only OCR. Regarding images, they are thus either served online via the IIF Image API¹⁶ of the *impresso* infrastructure, or accessed directly via the data provider's IIF endpoint. Text and layout acquisition outputs (i.e. OCR and OLR) come, for their part, in a variety of METS/ALTO format flavors, sometimes complemented by proprietary formats of private service providers. Overall, the current collection amounts to ca. 77TB, text and image combined. More newspaper titles in French and English will be acquired and ingested during the last year of the project.

3.2. Legal Framework

Original sources are subject to copyright law and *impresso* has received permission from its partners to use them, provided that legal terms of use are respected upon online access and/or download. More specifically, digital documents are subject to two different right statements: (1) public domain, or unrestricted: documents are no longer in copyright and may be used without restriction for all purposes, including commercial; (2) academic use, or restricted: documents are still under copyright and their use is restricted to personal and/or academic purposes, with the possibility

to download the text or not. The present *impresso* corpus release includes unrestricted documents and a part of restricted ones (for personal and academic usage). Depending on negotiations with data providers and on the inclusion of new collections, the situation is very likely to evolve in the future and *impresso* original source release will be complemented.

3.3. Source Processing

The original files provided by our partners encode the structure and the text of digital objects according to METS/ALTO XML library standards. METS (Metadata Encoding and Transmission Standard¹⁷) encodes various metadata as well as information on the physical and logical structure of the object, while ALTO (Analyzed Layout and Text Object¹⁸) represents information of OCR recognized texts, i.e. describes text and layout information (coordinates of columns, lines and words on a page). While very precise and complete, these XML files contain more information than necessary in a text mining context, and are cumbersome to process. Moreover, METS and ALTO schemas are flexible and libraries usually adapt them according to their text acquisition capacities, resulting in a variety of input variants. Combined with the existence of different file hierarchies, source identifiers and image mappings, as well as other OCR/OLR proprietary formats, these inputs require, to say the least, a great deal of processing before they can finally be parsed.

To this end, each library input is converted into 'canonical' files where information is encoded according to *impresso* JSON schemas,¹⁹ from which 'ready-to-process' files can easily be derived. Defined iteratively and shared with other newspaper projects, these JSON schemas act as a central, common format which a) allows the seamless processing of various data sources; b) preserves the information necessary for NLP processing and interface rendering only; and c) drastically reduces file sizes, thereby allowing easier processing in distributed environments.

Schemas and converters are published and documented online and are not described further here. An important point to mention, though, is that we mint and assign unique, canonical identifiers to newspaper issues, pages as well as content items (i.e. newspaper contents below the page level such as articles, advertisements, images, tables, weather forecasts, obituaries, etc.)

3.4. Release

The *impresso* corpora are released in two versions, both distributed as compressed archives (bzip2) of data in newline-delimited JSON format: 1) the 'canonical' version, with a fine-grained logical and physical representation of newspaper contents, including image coordinates and 2) the 'ready-to-process' version, which offer 'reconstructed' content item full texts, that is to say continuous strings non divided by OCR token units. This reconstruction significantly reduces the overhead when parsing the entire dataset,

¹³<https://impresso-project.ch/project/datasets>

¹⁴<https://zenodo.org/communities/impresso>

¹⁵Namely: the Swiss National Library, the National Library of Luxembourg, the Media Center and State Archives of Valais, the Swiss Economic Archives, the journal *Le Temps* (Ringier group), the journal *Neue Zürcher Zeitung*, and other local and international data providers.

¹⁶Defined by the International Image Interoperability Framework, an interoperable technology and community framework for image delivery: <https://iiif.io>

¹⁷<http://www.loc.gov/standards/mets>

¹⁸<https://www.loc.gov/standards/alto>

¹⁹<https://github.com/impresso/impresso-schemas>

Number of items	Unrestricted	Restricted (with download)	Restricted (w/o download)	Total
# issues	79,746	337,163	187,860	604,769
# pages	399,363	4,132,821	399,363	4,931,547
# tokens	572,030,104	9,374,592,395	2,641,896,310	12,588,518,809
# content items	1,461,700	38,948,561	4,269,189	44,679,450
# images	32,964	3,030,126	417,732	3,480,822

Table 1: Global statistics on the *impresso* corpora.

which amount to 145GB compressed (restricted and unrestricted).

The *impresso* corpus currently contains 76 newspapers: 50 from Switzerland and 26 from Luxembourg. AS mentioned previously, contents are subject to different license regimens, depending on the permissions given by cultural heritage institutions and rights holders. In Table 1 we provide some basic statistics about our corpora, divided by license type. The release will contain all contents in the public domain (unrestricted), as well as those available for academic use and for which the text can be downloaded (restricted with download, negotiations ongoing).

The released corpora amount to almost 10 billion tokens of textual contents, covering a time span of more than 200 years (see Fig. 1), and contain roughly 3 million images.

3.5. Metadata

Contextual information about digital collections is essential and we attempt to provide as much information as possible, even though this is neither the core expertise nor part of the main objectives of the project. *Impresso* newspaper metadata corresponds to descriptive (e.g. title, dates, place of publication), structural (issue, page, content items), and administrative metadata (file timestamps, file creator, preservation metadata). These metadata were given by cultural institutions and, most of the time, completed by the *impresso* team (either technical or descriptive metadata). Since this metadata set does not intend to replace library professional information but is rather meant for statistical ‘data science’ purposes, each record contains links to authority information such as the original bibliographic notice and the library portal. *Impresso* newspaper metadata is encoded in JSON format, covers all newspapers and is published under a CC-BY 4.0 license.²⁰

4. *Impresso* Benchmarks

In order to support the training and evaluation of some processing components, several benchmarks were produced. They include material from both restricted and unrestricted collections, for which right clearance has been achieved. All are released under open licenses.

4.1. Article Segmentation Ground Truth

Exploration and automatic processing of digitized newspaper sources is greatly hindered by the sometimes low qual-

ity of legacy OCR and OLR (when present) processes: content items are incorrectly transcribed and incorrectly segmented. In an effort to address these shortcomings, *impresso* developed an approach for content item recognition and classification exploiting both textual and visual features (Barman et al., 2020). The objectives were, on the one hand, to filter out noisy or unwanted material before the application of subsequent NLP processes (e.g. removing all meteo tables and title banners before running topic modeling or text re-use) and, on the other hand, to allow faceted search on content item types (e.g. search “xyz” in type of items ‘editorials’).

To this end, a set of newspaper images was manually annotated and several experiments were conducted (Barman, 2019). Although newspaper content items can be of many types,²¹ we choose to focus on four classes that were deemed suitable for developing a first prototype, as well as meaningful within the *impresso* context, as follows:

1. *Feuilleton*, i.e. an excerpt of a bigger work published over time in several issues of a newspaper, corresponding to the French *roman-feuilleton* or the English *serial*;
2. *Weather forecast*, i.e. a text or image with the prediction of weather, or even a report of past weather measurements;
3. *Obituary*, i.e. a small notice published by relatives of a deceased person;
4. *Stock exchange table*, i.e. a table reporting the values of different national stocks.

Three newspapers from the French speaking part of Switzerland covering a period of ca. 200 years (1798-2017) were considered for the annotation.²² To obtain a diachronic ground truth, three issues were sampled every three or five years for the whole duration of each newspaper. The sampled images were annotated using the *VGG Image Annotator* v.2.0.8 (VIA), a simple web interface for annotating images with annotation export in JSON format (Dutta and Zisserman, 2019). Concretely speaking, each annotated image is associated with the list of its regions (i.e. coordinates) and their corresponding labels. Overall,

²¹There is little to no agreement among historians and/or librarians about a ‘base’ newspaper content items taxonomy.

²²The *Gazette de Lausanne*, the *Impartial* and the *Journal de Genève*.

²⁰<https://creativecommons.org/licenses/by/4.0/>

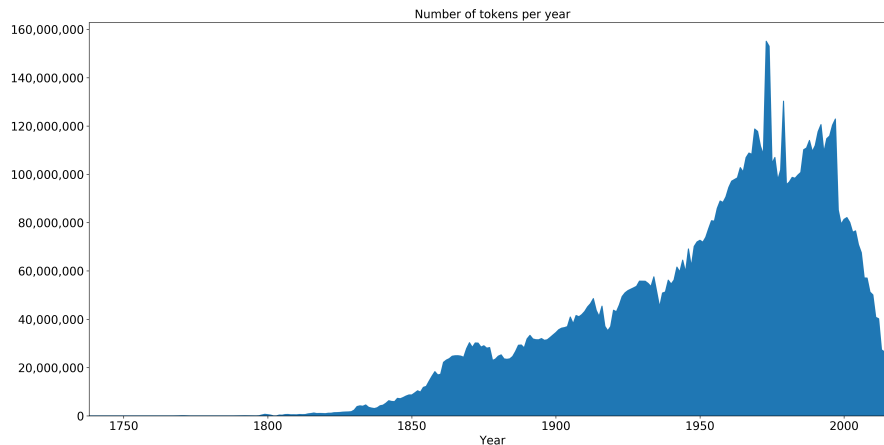


Figure 1: Distribution of tokens over years (whitespace tokenization was applied).

4624 page scans were annotated – among which 1208 with at least one annotation –, amounting to 2773 annotated regions.

Work is ongoing and once models would have reached a satisfying level of precision, they will be applied on the whole collection to filter out elements before text processing and enable faceted search over content item types.

This article segmentation data set (annotations and images) is published under a CC-BY-SA 4.0 license, using VIA as well as the standard object annotation COCO²³(Lin et al., 2014) formats.

4.2. Black Letter OCR Ground Truth

We created a publicly available ground truth (i.e., a manually corrected version of text) for black letter newspaper print for the assessment of the OCR quality of the German-language *Neue Zürcher Zeitung* (NZZ) (Ströbel, Phillip Benjamin and Clematide, Simon, 2019). We sampled one front page per year for the long period the NZZ has been published in black letter (1780 - 1947), resulting in a diachronic ground truth of 167 pages. We used the *Transkribus*²⁴ tool to complete the annotations. We published the ground truth as tiff images and corresponding XML files²⁵. First experiments on improving the OCR for this kind of data showed that elaborated deep learning models (Weidemann et al., 2018) reach character accuracies of 99.52% and that they are transferable to other newspaper data and to better images than present in the ground truth (Ströbel and Clematide, 2019).

4.3. Named Entity Processing Ground Truth

After image segmentation and transcription, the last *impresso* benchmark relates to an information extraction task, named entity (NE) processing. NE processing tools are increasingly being used in the context of historical documents and research activities in this domain target texts of

different nature (e.g. museum records, state-related documents, genealogical data, historical newspapers) and different tasks (NE recognition and classification, entity linking, or both). Experiments involve different time periods, focus on different domains, and use different typologies. This great diversity demonstrates how many and varied the needs –and the challenges– are, but also makes performance comparison difficult, if not impossible.

In this context, the *impresso* project organises a CLEF 2020 Evaluation Lab, named ‘HIPE’ (Identifying Historical People, Places and other Entities) (Ehrmann et al., 2020).²⁶ The HIPE shared task puts forward two NE processing tasks, namely: (1) the named entity recognition and classification (NERC) task, with two sub-tasks of increasing level of difficulty with high-level *vs.* finer-grained entity types, and (2) the named entity linking task.

The HIPE corpus is composed of content items from the *impresso* Swiss and Luxembourgish newspapers, as well as from American newspapers, on a diachronic basis.²⁷ For each language, articles of four different newspapers were sampled on a decade time-bucket basis, according to the time span of the newspaper (longest duration spans ca. 200 years). More precisely, articles were first randomly sampled from each year of the considered decades, with the constraints of having a title and more than 100 characters. Subsequently to this sampling, a manual triage was applied in order to keep journalistic content only and to remove undesirable items such as feuilleton, cross-words, weather tables, time-schedules, obituaries, and what a human could not even read because of OCR noise.

This material was manually annotated according to HIPE annotation guidelines, derived from the *Quaero* annotation guide.²⁸ Originally designed for the annotation of ‘ex-

²³<http://cocodataset.org/#format-data>

²⁴<https://transkribus.eu/Transkribus>

²⁵<https://github.com/impresso/NZZ-black-letter-ground-truth>

²⁶See CLEF 2020: <https://clef2020.clef-initiative.eu> and HIPE: <https://impresso.github.io/CLEF-HIPE-2020>

²⁷From the Swiss National Library, the Luxembourgish National Library, and the Library of Congress, respectively.

²⁸See the original *Quaero* guidelines: <http://www.quaero.org/media/files/bibliographic/quaero-guide-annotation-2011.pdf>

tended' named entities (i.e. more than the 3 or 4 traditional entity classes) in French speech transcriptions, Quaero guidelines have furthermore been used on historic press corpora (Rosset et al., 2012). HIPE slightly recast and simplifies them, considering only a subset of entity types and components, as well as of linguistic units eligible as named entities²⁹.

The annotation campaign was carried out by the task organizers with the support of trilingual collaborators. We used INCEpTION as an annotation tool (Klie et al., 2018), with the visualisation of image segments alongside OCR transcriptions. For each language, a sub-sample of the corpus was annotated by two annotators and inter-annotator agreement is computed, before and after an adjudication. As of March 2020, 21000 top-level entity mentions were annotated and linked to Wikidata.

For each task and language the corpus is divided into training, dev and test data sets, with the only exception of English for which only dev and test are produced. These manually annotated materials are released in IOB format with hierarchical information.

Even though many evaluation campaigns on NE were organized over the last decades,³⁰ only one considered French historical texts (Galibert et al., 2012) and, to the best of our knowledge, this is the first multilingual, diachronic named entity-annotated historical corpus.

5. *Impresso* Lexical and Semantic Annotations

Finally, a wealth of annotations as well as language models are automatically computed over the *whole impresso* collection. They include: at lexical level, linguistic preprocessing (lemmatisation and historical spelling normalization), word embeddings, OCR quality assessment and n-grams; at referential level, NE mentions and linked entities; at conceptual level, topics, topic models, and topic-annotated content items; at collection level, text reuse clusters and passages; and, finally, visual signatures of photographs and pictures contained in newspapers. These enrichments of our content items are represented as stand-off annotations and are released under CC-BY or CC-BY-SA 4.0 license. However, not all annotation data sets are fully ready at the moment; the following sections present those which are part of the current release.

5.1. OCR quality assessment

In order to automatically assess the loss of information due to OCR noise, we compute a simple OCR quality measure inspired by spell-checker approach of Alex and Burns (2014). In our case, it basically corresponds to the proportion of words of an historical newspaper article that can be found in the Wikipedia corpus of the corresponding language. Given the multilingual nature of our texts and the large number of names in newspapers, this offers a practical approach, especially for German where normal nouns and

proper nouns are capitalized. Before actually comparing the words, we normalise diacritical marks the same way as our text retrieval system Solr does before indexing the content. Therefore, for instance, we consider the frequently occurring OCR errors *Bäle* or *Bàle* as equivalent to the correct spelling of the town *Bàle*, because they are all normalized to the same string *bale*. The reason for this normalisation approach in OCR assessment is that we want to inform our *impresso* users about the real loss of recall they should expect when actually running standard keyword queries over our text collection (*Bàle* will be found even is the user search for *Bàle*, but *Bâte* would not return any result, and this is the loss we want to account for).

The OCR quality assessment is a number between 0 and 1 that is distributed along with our data as stand-off annotation for each content item. *Impresso* interface users will probably quickly grasp the meaning of the numbers by just being exposed to texts and their corresponding OCR quality assessment, and learn to interpret them with respect to the type of article, e.g. stock market prices with many abbreviations that will lower the score. As our approach is unsupervised, we need to formally evaluate it similar to Alex and Burns (2014) by testing whether there is a reasonable correlation between the automatically computed quality and some ground truth character error rate.

5.2. Word Embeddings

As mentioned earlier, the full *impresso* collection cannot be distributed due to copyright restrictions. Having the material at hand, however, allows us to compute historical newspapers genre-specific lexical resources such as word embeddings that can be distributed to the public. Specifically, we build classical type-level word embeddings with fasttext³¹. This choice is motivated by fasttext's support for subword modeling (Bojanowski et al., 2016), which is a useful feature in the presence of OCR errors. There has been recent work on top of fasttext for bringing the embeddings of misspelled words even closer to the correct versions via supervised training material (Piktus et al., 2019). Well-known drawbacks of type-level word embeddings are that (a) they enforce their users to adhere to the same tokenisation rules that their producers applied and, more severely, (b) they cannot differentiate the meanings of ambiguous words, or words that change their meaning in certain constructions. The simple character-based approach proposed by Akbik et al. (2018) ("contextualized string embeddings"³²) has successfully tackled these two problems and led to excellent results for NER. Our own experiments with NER on noisy French historical newspapers additionally proved the resilience of these embeddings trained on in-domain material to OCR errors (Bircher, 2019).

Within the *impresso* interface, word embeddings are mainly used for suggesting similar words in the keyword search (including cross-lingual), thereby supporting query expansion by semantic or OCR noise variants. Query expansion is also offered for the lexical n-gram viewers.

Two types of word embeddings derived from the *impresso* text material are published: Character-based contextualized

²⁹HIPE guidelines are available at: <https://doi.org/10.5281/zenodo.3677171>

³⁰E.g. MUC, IREX, ACE, CoNLL, KBP, ESTER, HAREM, QUAERO, GERMEVAL, etc.

³¹<https://fasttext.cc>

³²<https://github.com/zalandoresearch/flair>

string embeddings and classical type-level word embeddings with subword information.

5.3. Topic Models

The *impresso* web application supports faceted search with respect to language-specific topics (French, German, Luxembourgish). We use the well-known MALLET³³ toolkit, which allows the training and inference of topic models with Latent Dirichlet Allocation (Blei et al., 2003).

First, linguistic preprocessing is applied to the data. For POS tagging, the spaCy³⁴ library is used because of its robustness in the presence of OCR noise. However, spaCy lemmatization is not always very satisfactory and further analyzers and sources are used to complement its results. For German, we rely mostly on the broad-coverage morphological analyser GERTWOL³⁵, and are currently working on the problem of lemmatization of words with historical spelling and/or OCR errors (see Jurish (2012) for earlier work based on finite-state approaches for German). For French, we use the full-form lexicon Morphalou³⁶ (ATILF, 2019) to complete lemma information not provided by spaCy. Dealing with the low-resourced Luxembourgish language is more difficult (although spaCy now has PoS tagging support for this language), mostly because of many spelling variants and reforms this language has seen over the last 150 years.

Then, under the assumption that topics are more interpretable if they consist of nouns and proper nouns only, we reduce the corpus size by excluding all other parts of speech based on the information obtained from spaCy. As an additional benefit, this filtering drastically reduces the number of tokens of the corpus that topic modeling has to deal with. Next, topics are computed on this reduced, preprocessed material. Although the German part of the collection is of reasonable size, the French material is however still too big for MALLET and sampling of articles containing at least 10 nouns and/or proper nouns is applied. In order to keep the facets for topic search manageable and interpretable, and at the same time account for the diversity of contents found in newspapers, we set the number of topics for German and French to 100. For the French topics, we directly fit topic distributions for about a third of our overall data. Topic inference with the model trained on the sample is used for the remaining articles. Topic inference also solves the problem that our collections is continuously growing, and recomputing topic models from scratch each time is not feasible. Additionally, historians prefer to have semantically stable topic models for their work. Therefore, we also apply topic inference on newly added German texts.

Topic models, as well as topics and content item topic assignments are released in JSON format.³⁷ Topics are also available within the *impresso* web interface, where they (a) serve as search facets, i.e., users can restrict their search results to articles containing only certain topics; or (b) the

users can select topics as entry points to explore the topic modeling based soft-clustering of articles over the entire corpus; or (c) they provide the basis for an article recommender system based on topic distribution similarity. Future work will focus on the evolution of topics over time and cross-lingual topic modeling.

5.4. Text Reuse

Text reuse can be defined as the meaningful reiteration of text beyond the simple repetition of common language. It is such a broad concept that it can be understood at different levels and studied in a large variety of contexts. In a publishing or teaching context, plagiarism can be seen as text re-use, should portions of someone else's text be repeated without appropriate attribution. In the context of literary studies, text re-use is often used as a synonym for literary phenomena like allusions, paraphrases and direct quotations.

Text reuse is a very common phenomenon in historical newspapers too. Nearly-identical articles may be repurposed in multiple newspapers as they stem from the very same press release. In newspapers from the period before the advent of press agencies, text reuse instances can be interesting to study the dynamics of information spreading, especially when newspapers in the same language but from different countries are considered. In more recent newspapers text reuse is very frequent due to cut-and-paste journalism being an increasingly common practice.

We used *passim*³⁸ (Smith et al., 2015) to perform the automatic detection of text reuse. *Passim* is an open source software that uses n-grams to effectively search for alignment candidates, the Smith-Waterman algorithm to perform the local alignment of candidate document pairs, and single-link clustering to group similar passages into text reuse clusters.

As a pre-processing step we used *passim* to identify boilerplate within our corpus. This step allows us to reduce the input size of approximately 10%, by removing mostly short passages that are repeated *within the same newspaper* within a time window of 30 days. We then run *passim* on the entire corpus after boilerplate passages have been removed: *passim* outputs all text passages that were identified as belonging to a text reuse cluster. As opposed to boilerplate detection, text reuse detection explicitly targets reuse instances across two or more sources (i.e. newspapers).

We post-process *passim*'s output to add the following information:

- size, i.e. the number of text passages in the cluster;
- lexical overlap, expressed as the proportion of unique tokens shared by all passages in a cluster;
- time delta: the overall time window covered by a given cluster (expressed in number of days);
- time gap: following Salmi et al. (2019), we compute the longest gap (expressed in number of days) between the publication of any two passages in a cluster.

³³<http://mallet.cs.umass.edu>

³⁴<https://spacy.io/>

³⁵<http://www2.lingsoft.fi/doc/gertwol>

³⁶<http://www.cnrtl.fr/lexiques/morphalou>

³⁷Also documented online at <https://github.com/impresso/impresso-schemas>

³⁸<https://github.com/dasmiq/passim>

Dataset	DOIs
Impresso Historical Newspaper Textual Material	10.5281/zenodo.3706823
Impresso Newspaper Metadata	10.5281/zenodo.3706833
Impresso OCR Quality Assessment	10.5281/zenodo.3709465
Impresso OCR ground truth	10.5281/zenodo.3333627
Impresso Article Segmentation Ground Truth	10.5281/zenodo.3706863
Impresso HIPE Shared Task Named Entity Gold Standard	10.5281/zenodo.3706857
Impresso Word Embeddings	10.5281/zenodo.3706808
Impresso Topic Modelling Data	10.5281/zenodo.3706840
Impresso Text Reuse Data	10.5281/zenodo.3706850

Table 2: *Impresso* datasets DOIs.

This information is added to each text reuse cluster with the goal of easing the retrieval as well as the analysis of detected text reuse. Since *passim* detects several million clusters in the entire *impresso* corpus, we need to further characterize each cluster if we want to enable historians to find instances of text reuse that are of interest to them.

Each of these additional dimensions characterizes a certain aspects of reuse: *lexical overlap* allows for distinguishing almost exact copies of a piece of news from re-phrasings or paraphrases; *time delta* is an indicator of the longevity of a given piece of news; and, finally, *time gap* captures the viral nature of news spreading, especially its *pace* of publication. We release as a resource (in JSON format) the boilerplate and text reuse passages as detected by *passim*, as well as the additional information we compute at cluster-level. This data can be used to filter out duplicates from the input corpus, given the detrimental effects that such duplicates have on semantic models (e.g. topics, word embeddings) (Schofield et al., 2017).

Text reuse information is currently used in the *impresso* interface as an additional navigation aid, as it points users to existing reuses of the news article in focus. Future upgrades of the interface will include a dedicated text reuse explorer, which will allow users to search over and browse through all text reuse clusters, and to filter them based on several criteria (i.e. size, lexical overlap, time gap, time delta).

6. Related work

This section briefly summarizes previous efforts with respect to historical language resources. We focus here on historical newspapers and refer the reader to Sporleder (2010) and Piotrowski (2012) for further information on historical language in general.

Digitized newspaper corpora, understood here as consisting of both images and OCR, primarily exist thanks to the considerable efforts of national libraries, either as individual institutions, either as part of consortia, e.g. the Europeana Newspaper project (Neudecker and Antonacopoulos, 2016). Those institutions are the custodians of these digital assets which, after having been hidden behind digital library portals for long, are now increasingly making their way to the public via APIs and/or data dumps (e.g. the

French National Library APIs³⁹ and the National Library of Luxembourg open data portal⁴⁰). *Impresso* corpora are by no means meant to compete with these repositories, but rather to complement them, with derived, working ‘secondary’ versions of the material in a form that is suitable for NLP needs. To our knowledge, and since corpus preparation is often done by private companies mandated to develop digital portals, no ‘ready-to-process’ set of historical newspaper corpus such as the *impresso* one exists.

Several instances of OCR and article segmentation benchmarks exists thanks to, among others, the long-standing series of conference and shared tasks organized by the document analysis community⁴¹ *impresso* annotated data sets are, in this regard, not new but complementary: German Black Letter ground truth is not common and, given the variety of historical newspaper material, article segmentation over page scans of different sources is beneficial.

With respect to word embeddings, the companion website⁴² of Hamilton et al. (2016) provides word2vec embeddings for French and German derived from Google n-grams. More recently, Riedl (2019) released German word embedding data sets derived from historical newspapers.

In the last years, a few gold standards were publicly released for named entities: Galibert et al. (2012) shared a French named entity annotated corpus of historical newspapers from the end of the 19th century and Neudecker (2016) published four data sets of 100 pages each for Dutch, French, and German (including Austrian) as part of the Europeana Newspapers project. Besides, Linhares Pontes et al. (2019b) have recently published a data set for the evaluation of NE linking where various types of OCR noise were introduced. In comparison, the HIPE corpus has a broader temporal coverage and additionally covers English.

Regarding topic modeling, Yang et al. (2011a) gives an overview of earlier work on historical newspapers.

Finally, as far as text reuse is concerned, very few resources and/or benchmarks were published to date. Franzini et al. (2018) have published a ground truth dataset to benchmark

³⁹<http://api.bnf.fr>

⁴⁰<https://data.bnl.lu/data/historical-newspapers>

⁴¹In particular the ICDAR conferences, e.g. <http://icdar2019.org/>.

⁴²<https://nlp.stanford.edu/projects/histwords>

the detection of a specific type of text reuse (i.e. literary quotations). The Viral Texts project has published an online interface, the Viral Texts Explorer⁴³, which makes searchable and explorable text reuse clusters extracted from 19th century newspapers. A similar online interface was provided also by Salmi et al. (2019) for 13 million text reuse clusters extracted from Finnish press (1771–1920).

7. Conclusion and Perspectives

We have presented a series of historical newspaper datasets – the *impresso* resource collection – composed of corpora, benchmarks, semantic annotations and language models in French, German, Luxembourgish and English covering ca. 200 years. Produced in the context of a collaborative, interdisciplinary project which aims at enabling critical text mining of 200 years of newspapers, this collection includes different types of resources that could support the needs of several communities. The textual corpora we release are large-scale, diachronic, multilingual and with real-world OCR quality. Their availability will foster further research on NLP methods applied to historical texts (e.g. OCR post-correction, semantic drift, named entity processing). Similarly, our benchmarks will fill an important gap in the adaptation of existing approaches via e.g. transfer learning, as well as enable performance assessment and comparisons. Language models will naturally find their use in many applications, while lexical and semantic annotations will support historical corpus exploration and be suitable for use at public participatory events such as hackathons. As future work we attempt to integrate more textual material (French and English notably), to release additional annotations (image visual signatures, historical n-grams and named entities) and to serialize our data in more formats in addition to JSON.

8. Acknowledgements

We warmly thank the *impresso* team as well as student collaborators Camille Watter, Stefan Bircher, Julien Nguyen Dang for their annotation work. Authors also gratefully acknowledge the financial support of the Swiss National Science Foundation (SNSF) for the project *impresso* – Media Monitoring of the Past under grant number CRSII5_173719.

9. Bibliographical References

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Alex, B. and Burns, J. (2014). Estimating and rating the quality of optically character recognised text. pages 97–102. ACM Press.

ATILF. (2019). Morphalou. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

Barman, R., Ehrmann, M., Clematide, S., Oliveira, S. A., and Kaplan, F. (2020). Combining Visual and Textual Features for Semantic Segmentation of Historical Newspapers (submitted). *Journal of Data Mining and Digital Humanities*. <https://arxiv.org/abs/2002.06144>.

Barman, R. (2019). Historical newspaper semantic segmentation using visual and textual features. Master thesis, EPFL.

Bircher, S. (2019). *Toulouse* and *Cahors* refer to locations, but *T<<i*louse* and *Caa.Qrs* as well. A Neural Approach for detecting Named Entities in Digitized Historical Newspapers. Master thesis, Zurich University.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *CoRR*, abs/1607.04606.

Bollmann, M. (2019). A large-scale comparison of historical text normalization systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3885–3898. Association for Computational Linguistics.

Chiron, G., Doucet, A., Coustaty, M., Visani, M., and Moreux, J.-P. (2017). Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information. In *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries, JCDL '17*, pages 249–252, Piscataway, NJ, USA. IEEE Press.

Clausner, C., Antonacopoulos, A., Pletschacher, S., Wilms, L., and Claeysens, S. (2019). PRImA, DMAS2019, Competition on Digitised Magazine Article Segmentation (ICDAR 2019).

Dinarelli, M. and Rosset, S. (2012). Tree-structured named entity recognition on OCR data: Analysis, processing and results. In *LREC*, pages 1266–1272.

Dutta, A. and Zisserman, A. (2019). The VIA annotation software for images, audio and video. In *Proceedings of the 27th ACM International Conference on Multimedia, MM '19*, New York, NY, USA. ACM.

Ehrmann, M., Colavizza, G., Rochat, Y., and Kaplan, F. (2016). Diachronic Evaluation of NER Systems on Old Newspapers. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 97–107. Bochumer Linguistische Arbeitsberichte.

Ehrmann, M., Romanello, M., Clematide, S., and Bircher, S. (2020). (submitted) Introducing the CLEF 2020 HIPE Shared Task: Named Entity Recognition and Linking on Historical Newspapers. In *European Conference on Information Retrieval*, Lisbon, Portugal, April.

Franzini, G., Moritz, M., Marco, B., Passarotti, M., and Cuore, S. (2018). Using and evaluating TRACER for an Index fontium computatus of the Summa contra Gentiles of Thomas Aquinas. In Elena Cabrio, et al., editors, *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*.

Galibert, O., Rosset, S., Grouin, C., Zweigenbaum, P., and

⁴³<https://viraltexts.northeastern.edu/clusters>

- Quintard, L. (2012). Extended named entities annotation on ocred documents: From corpus constitution to evaluation campaign. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).
- Hamilton, L. W., Leskovec, J., and Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501. Association for Computational Linguistics.
- Jurish, B. (2012). *Finite-state canonicalization techniques for historical German*. doctoral thesis, Universität Potsdam. <https://nbn-resolving.org/urn:nbn:de:kobv:517-opus-55789>.
- Kaplan, F. and di Lenardo, I. (2017). Big Data of the Past. *Frontiers in Digital Humanities*, 4.
- Kestemont, M., Karsdorp, F., and Düring, M. (2014). Mining the twentieth century's history from the time magazine corpus. *EACL 2014*, page 62.
- Klie, J.-C., Bugert, M., Boulosa, B., de Castilho, R. E., and Gurevych, I. (2018). The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9.
- Lansdall-Welfare, T., Sudhakar, S., Thompson, J., Lewis, J., , and Cristianini, N. (2017). Content analysis of 150 years of british periodicals. *Proceedings of the National Academy of Sciences*, 114(4):E457–E465.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.
- Linhares Pontes, E., Hamdi, A., Sidere, N., and Doucet, A. (2019a). Impact of ocr quality on named entity linking. In *Digital Libraries at the Crossroads of Digital Information for the Future*, pages 102–115. Springer LNCS, October.
- Linhares Pontes, E., Hamdi, A., Sidere, N., and Doucet, A. (2019b). Impact of ocr quality on named entity linking. In Adam Jatowt, et al., editors, *Digital Libraries at the Crossroads of Digital Information for the Future*, pages 102–115, Cham. Springer International Publishing.
- Moreux, J.-P. (2016). Innovative Approaches of Historical Newspapers: Data Mining, Data Visualization, Semantic Enrichment. In *Proceedings of IFLA WLIC 2016*, page 17, Columbus, OH.
- Neudecker, C. and Antonacopoulos, A. (2016). Making Europe's Historical Newspapers Searchable. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 405–410, Santorini, Greece, April. IEEE.
- Neudecker, C. (2016). An open corpus for named entity recognition in historic newspapers. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).
- Piktus, A., Edizel, N. B., Bojanowski, P., Grave, E., Ferreira, R., and Silvestri, F. (2019). Misspelling oblivious word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3226–3234, Minneapolis, Minnesota, June. Association for Computational Linguistics.
- Piotrowski, M. (2012). Natural language processing for historical texts. *Synthesis Lectures on Human Language Technologies*, 5(2):1–157.
- Plank, B. (2016). What to do about non-standard (or non-canonical) language in NLP. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*. Bochumer Linguistische Arbeitsberichte.
- Ridge, M., Colavizza, G., Brake, L., Ehrmann, M., Moreux, J.-P., and Prescott, A. (2019). The past, present and future of digital scholarship with newspaper collections. page 9. Multi-paper panel presented at the 2019 Digital Humanities Conference, Utrecht, July 2019.
- Riedl, M. (2019). German Word Embeddings for ShiCo based on historic newspapers, June.
- Rigaud, C., Doucet, A., Coustaty, M., and Moreux, J.-P. (2019). ICDAR 2019 Competition on Post-OCR Text Correction. In *15th International Conference on Document Analysis and Recognition*, Sydney, Australia, September.
- Rosset, S., Grouin, Cyril, Fort, Karen, Galibert, Olivier, Kahn, Juliette, and Zweigenbaum, Pierre. (2012). Structured Named Entities in two distinct press corpora: Contemporary Broadcast News and Old Newspapers. In *6th Linguistics Annotation Workshop (The LAW VI)*, pages 40–48, Jeju, South Korea, July.
- Salmi, H., Rantala, H., Vesanto, A., and Ginter, F. (2019). The long-term reuse of text in the Finnish press, 1771–1920. *CEUR Workshop Proceedings*, 2364:394–544.
- Schofield, A., Thompson, L., and Mimno, D. (2017). Quantifying the effects of text duplication on semantic models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2737–2747, Copenhagen, Denmark, September. Association for Computational Linguistics.
- Smith, D. A. and Cordell, R. (2018). A Research Agenda for Historical and Multilingual Optical Character Recognition. <https://ocr.northeastern.edu/>.
- Smith, D. A., Cordell, R., and Mullen, A. (2015). Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers. *American Literary History*, 27(3):E1–E15, sep.
- Sporleder, C. (2010). Natural Language Processing for Cultural Heritage Domains. *Language and Linguistics Compass*, 4(9):750–768.
- Ströbel, P. B. and Clematide, S. (2019). Improving OCR of Black Letter in Historical Newspapers: The Unreasonable Effectiveness of HTR Models on Low-Resolution

- Images. In *Proceedings of the Digital Humanities 2019, (DH2019)*. CLARIAH.
- Terras, M. M. (2011). The Rise of Digitization. In Ruth Rikowski, editor, *Digitisation Perspectives*, pages 3–20. SensePublishers, Rotterdam.
- Vilain, M., Su, J., and Lubar, S. (2007). Entity Extraction is a Boring Solved Problem: Or is It? In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers, NAACL-Short '07*, pages 181–184. Association for Computational Linguistics. event-place: Rochester, New York.
- Weidemann, M., Michael, J., Grüning, T., and Labahn, R. (2018). HTR Engine Based on NNs P2 Building Deep Architectures with TensorFlow. Technical report.
- Wevers, M. (2019). Using word embeddings to examine gender bias in Dutch newspapers, 1950-1990. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 92–97, Florence, Italy, August. Association for Computational Linguistics.
- Yang, T.-I., Torget, A., and Mihalcea, R. (2011a). Topic modeling on historical newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 96–104, Portland, OR, USA, June. Association for Computational Linguistics.
- Yang, T.-I., Torget, A. J., and Mihalcea, R. (2011b). Topic modeling on historical newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTech)*, pages 96–104.

10. Language Resource References

- Ströbel, Phillip Benjamin and Clematide, Simon. (2019). *NZZ Black Letter Ground Truth*. University of Zurich, 1.0, ISLRN 855-418-004-842-0.