LUCAS SULZBACH

BUILDING A DATASET OF LATE MODERN GERMAN-BRAZILIAN NEWSPAPER ADVERTISEMENT PAGES FOR LAYOUT AND FONT RECOGNITION

(*pre-defense version, compiled at May 26, 2022*)

Trabalho apresentado como requisito parcial à conclusão do Curso de Bacharelado em Ciência da Computação, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Ciência da Computação*.

Orientador: Eduardo Todt.

CURITIBA PR

2022

RESUMO

Em anos recentes, muitos projetos de humanidades digitais foram iniciados com o objetivo de extrair informações de recursos culturais digitalizados por meio de técnicas de processamento de imagem. Com o advento e a difusão de redes neurais, modelos de *machine learning* se tornam mais robustos e abrangentes a cada dia, sendo utilizados para automatizar tarefas de reconhecimento de padrões cada vez mais complexos encontrados em imagens. Na área de processamento de documentos históricos, técnicas de *deep learning* são usadas para transcrever e extrair informações visuais de grandes coleções de documentos antigos e degradados de forma automática. Entretanto, alguns desafios persistem, já que métodos de análise de layout por vezes têm dificuldades para reconhecer arranjos e estruturas complexas. Adicionalmente, ferramentas de Reconhecimento Óptico de Caracteres (OCR) não estão preparadas para lidar com grandes variedades de fontes históricas, especialmente tratando-se de tipos góticos. Por esses motivos, tentativas de reconhecimento de layout e texto de periódicos teuto-brasileiros não obtiveram sucesso até o momento. A imprensa étnica teuto-brasileira foi responsável pela impressão de páginas com layouts extremamente complexos e contendo muitas fontes diferentes, tanto de tipografias góticas quanto de tipos latinos, especialmente em suas seções de anúncios. Com o objetivo de abrir o caminho para o reconhecimento de layout e fonte de documentos brasileiros em língua alemã, milhares de imagens de documentos digitalizados foram analisadas, e centenas destas selecionadas e disponibilizadas através da API de Imagens IIIF para a compilação do *German-Brazilian Newspaper Advertisement Pages Dataset* (*gbn-ads*) (*Dataset* de Páginas de Anúncios de Jornais Teuto-Brasileiros). Uma vez completo, este *dataset* conterá anotações para o treinamento de modelos de *deep learning* para segmentação de blocos de anúncio e classificação de fonte a nível de palavra, ambas técnicas que podem não apenas possibilitar o reconhecimento de texto de periódicos teuto-brasileiros complexos, mas também contribuir para todo o cenário de processamento de documentos históricos em língua alemã. Além disso, a infraestrutura para obtenção de imagens construída durante esta pesquisa lança as bases para uma biblioteca digital aberta da imprensa teuto-brasileira.

Palavras-chave: Documentos Históricos. Humanidades Digitais. Deep Learning.

ABSTRACT

In recent years, many digital humanities projects have been launched with the objective of retrieving information from digitized cultural heritage resources by means of image processing techniques. With the emergence and diffusion of deep neural networks, machine learning models become more robust and comprehensive every day, being used to automate recognition of increasingly complex patterns found in image data. In the field of historical document processing, deep learning techniques have been used to automatically transcribe and extract visual information of massive collections of old, degraded document pages. Still, some challenges persist, as layout analysis methods can struggle to recognize complex arrangements and structures and Optical Character Recognition (OCR) engines are not prepared to handle large varieties of historical fonts, specially blackletter types. For these reasons, attempts of recognizing layout and text of German-Brazilian periodicals have been unsuccessful so far. The late modern German-Brazilian ethnic press was responsible for printing pages with extremely complex layouts and many fonts of blackletter and Latin typefaces, specially in their advertisement sections. With the objective of paving the way for layout and font recognition of German-language Brazilian documents, thousands of digitized document images have been analysed, and hundreds of them selected and made available through the IIIF Image API for the compilation of the German-Brazilian Newspaper Advertisement Pages Dataset (*gbn-ads*). Once completed, this dataset will contain ground truth for the training of deep learning models for advert block segmentation and word-level font classification, techniques which can not only help in achieving text recognition of complex German-Brazilian periodicals, but also contribute to the whole scenario of historical German-language document processing. Furthermore, the image retrieval infrastructure set up during this research has laid the foundations for an open digital library of the German-Brazilian press.

Keywords: Historical Documents. Digital Humanities. Deep Learning.

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

bpt.	Brazilian Portuguese
hdt.	*Hochdeutsch* (High/Standard German)
hrs.	*Riograndenser Hunsrückisch*
OCR	Optical Character Recognition
OLR	Optical Layout Recognition
HTR	Handwritten Text Recognition
GT	Ground Truth
IIIF	International Image Interoperability Standard
PAGE	Page Content and Ground-truth Elements
METS	Metadata Encoding and Transmission Standard
OCR-D	DFG-funded Initiative for OCR Development
GT4HistOCR	Ground Truth for Historical OCR
LAREX	Layout Analysis and Region EXtraction
GB	GigaBytes
TB	TeraBytes
URI	Uniform Request Identifier
RGB	Red, Green, Blue
API	Application Programming Interface
VGG	Visual Geometry Group
VIA	VGG Image Annotation
TIFF	Tag Image File Format
PNG	Portable Network Graphics
JPEG	Joint Photographic Experts Group
GIF	Graphics Interchange Format
JSON	JavaScript Object Notation
XML	Extensible Markup Language
CSV	Comma Separated Values
DPI	Dots Per Inch
dbp	*Deutschsprachige Brasilianische Presse*
GBN	German Brazilian Newspapers
gbn-ads	German-Brazilian Newspaper Advertisement Pages
UFPR	*Universidade Federal do Paraná*
UNESP	*Universidade Estadual Paulista*
IECLB	*Igreja Evangélica de Confissão Luterana no Brasil*
SPK	*Stiftung Preußischer Kulturbesitz*

ICPR	International Conference on Pattern Recognition

CONTENTS

1 INTRODUCTION

During the nineteenth and first half of the twentieth centuries, many German-language periodicals used to circulate in Brazilian land. Ranging from local newspapers and church publications to advertisement magazines from overseas, these periodicals provide an unique perspective on the foundation and consolidation of German-speaking communities by immigrants and descendants in late modern Brazil. For this reason, the German-Brazilian periodicals are objects of interest in numerous research fields, such as history, linguistics and typography. Although, not many studies reviewing this documentation are carried out given the difficulty of getting access to these periodicals, since they are spread in multiple libraries and collections around Brazil. Besides, German is not widely spoken outside these specific, small and relatively isolated German-Brazilian communities, and therefore, there is a considerable language barrier for Brazilian audiences. In summary, the amount of researchers and enthusiasts able to access, read and extract meaningful information from late modern German-Brazilian periodicals is very low.

With the technological advancements from past decades, many libraries and archives started to make their contents available through web interfaces known as "digital libraries". This kind of platform can be used to make the German-Brazilian periodicals easier to be accessed by people not only from Brazil, but from all over the world. Nowadays, digital libraries are not only tools for viewing document images, but also offer useful features such as highlighting text lines and searching by term, optimizing research on large document collections. In order to implement this kind of functionality, it is required that the page images to had been previously segmented and transcribed, be it manually, by a human being, or automatically, by a machine (computer). The former approach might be feasible when working with documents with simple layouts and little text content (such as letters), but the latter is often necessary in more complicated cases, since the amount of time and resources needed to manually segment and transcribe such documents, specially when handling large collections, can be massive. The German-Brazilian periodicals, as it will be exposed in more detail in the following chapters, are many, and have particularly complex characteristics when it comes to layout structure and typesetting, enough reasons to dismiss entirely manual approaches.

In order to automatically process document images, computer vision and image processing techniques come into play. For the text extraction, *Optical Character Recognition* (OCR) methods are employed. Although, a complete "OCR" workflow often involves operations that precede the text recognition itself, such as binarization, skew and orientation correction and layout segmentation, which are used to identify and prepare the parts of the image which contain text so it can actually be extracted. State-of-the-art OCR suites usually rely on deep neural networks to provide accurate and precise results when performing these actions, and therefore a subset of document images must be manually annotated so a machine can learn by example and automatically process the entirety of the document pages. This subset of images plus their respective *ground truth* annotations is referred to as a *dataset*, which can be used for training, testing or evaluating machine learning models. Given the peculiar nature of the German-Brazilian periodicals, available models do not work as intended out-of-the-box, since they are not designed nor trained to work with German-Brazilian periodicals, but rather with contemporary documents. Even the few that aim at historical document processing are usually fine-tuned for other niches, often of documents of different periods, languages or types (e.g. books), not being appropriate to handle the complex layouts and typographical variety of the German-Brazilian ethnic press.

On the mission of improving layout and font recognition of German-Brazilian periodicals for better OCR results, a dataset of German-Brazilian Newspaper Advertisement Pages (*gbn-ads*) is proposed. The focus on advertisements is due to the fact that this sort of content is an often neglected, yet extremely valuable source of information about the everyday life in German-Brazilian communities, while composing the pages on which image processing workflows have the poorest performance, and therefore the poorest content retrieval. These pages stand out for their unorthodox layout arrangement and typesetting, often containing complex separation patterns which cannot be described by a fixed number of rows and columns, and text in multiple font types, sizes and orientations. By compiling a granular and multimodal dataset, that is, a dataset containing ground truth for different layout hierarchies such as pages, adverts, text regions, lines and words (granular) and different purposes such as layout analysis and font classification (multimodal), the author hopes that better deep learning models can be trained and used to improve the scenario of OCR for documents of the late modern German-language Brazilian press. Due to the amount of time and effort required to build such a dataset, the scope of this work is limited to reviewing existing datasets, sampling the images to be included and making plans for future annotation pipelines and training procedures.

The work is structured as follows: In Chapter 2, a historical review of German immigration and colonization in Brazil can be found, providing context on the emergence and consolidation of the German-Brazilian ethnic press. In Chapter 3, the basic document image processing and machine learning concepts addressed in the following chapters are explained. The previous works performed by the author, which led to the conception of the proposed dataset are the topic of Chapter 4. The review of related datasets can be found in Chapter 5. The methodology used to build the dataset is explained in Chapter 6, and the resulting products are outlined in Chapter 7. Chapter 8 concludes this work and outlines the objectives of future research on layout analysis and font classification of German-Brazilian periodicals.

2 HISTORICAL BACKGROUND

In this chapter, literature of historical research on German-Brazilian communities and heritage are reviewed in order to contextualize the German-Brazilian periodicals. In Section 2.1, the arrival and settlement of German colonists in late modernity is addressed, while in Section 2.2, an overview of the German-Brazilian press is provided.

2.1 GERMAN IMMIGRATION AND COLONIZATION

In 1824, the colony of *São Leopoldo* was settled in the *Sinos* river valley, in the southernmost Brazilian state of *Rio Grande do Sul*. While other colonies had been settled before, *São Leopoldo* is considered the first mark of German immigration because of its early prosperity and attraction of immigrants along the years. Until 1830, the settlement had received thousands of German-speaking immigrants of different origins, most of them native of the *Hunsrück* region in western Germany (Hunsche, 1975, 1977). Over the decades, colonization advanced to other valleys in the region, and was boosted with new waves of colonists when immigration resumed in 1845 (Altenhofen et al., 2018). By 1889, these valleys and whereabouts were location of over 80 different settlements[1] (Seyferth, 2010), from which further expansions to the west would originate in the following decades[2]. Over the course of the 19th century, incoming settlers from overseas would also settle in other regions, establishing colonies such as *Blumenau* and *Dona Francisca* (*Joinville*), while others formed urban communities within cities such as *Curitiba*. Until 1952, the country had received around 350,000 people from German-speaking locations, making it the primary destination of German immigrants in South America, and second only to the USA in a global scope (Soethe, 2020).

The pioneers, when arriving at the land to be colonized, composed of vast, mostly untouched forests, had to build their new *Heimat*[3] from scratch, as Altenhofen et al. (2018) describes in detail. The first step was to open the *Pikood*[4], where each family would occupy a piece of land, narrow in width and long in height. A typical property consisted of the family house, usually built near the access of the property, surrounded by other buildings such as warehouses and barns, while most of the plantations and pastures were located in the back of the property. This shape and structure ensures a proximity to the neighbors, on which the settlers would count in case of emergencies. Buildings such as a church, a school, a cemetery, an hall and a general store would emerge, usually next to each other, composing a “town square” of

[1] The “Old Colonies” of *Rio Grande do Sul*

[2] These movements would originate the “New Colonies” of the state and eventually spread to other locations in south- and mid-west and north of Brazil and even in Paraguay and Argentina (Altenhofen et al., 2018)

[3] *Heimat* (hdt.): Home, lair

[4] *Pikood* (hrs.) aka. *Picada* (bpt.), *Linha* (bpt.): Main trail, pathway of a community of colonists, often a long and straight line with properties on both sides

the community. Slowly, the *Pikood* becomes more self-sufficient, with other businesses such as sawmills, carpentries and forges. Associations such as the *Tanzvereine*, the *Sängervereine* and the *Schützenvereine*[5] start to appear, becoming important hubs of social interaction. During events and festivals such as the *Kerb*[6], members of other communities would come to visit and celebrate. The urban communities would be organized in a similar form, also with many businesses and associations. These environments constituted cultural strongholds, where the German language was spoken as a mother tongue and preserved for generations, until today in many of such places.

While the day-to-day interaction in the churches, schools, businesses, clubs and festivals can be considered responsible for maintaining the German language in the spoken form, be it through variants of High German (*Hochdeutsch*) or the *Hunsrückisch*, Pomeranian and Westphalian dialects[7], particularly the church and school provided support for the written tradition, almost always consisting of the High German (Altenhofen et al., 2018). Over time, the specificities and demands of the German-speaking communities would be attended by a local press in German language: Text books would be printed and used for teaching at the community schools[8]. *Kirchenblätter*[9] would feature prayers, articles, important dates, obituaries and relations of baptized and confirmed children, both of the Catholic and Evangelic (Lutheran) communities. The almanac or *Kalender*[10], described by Schappelle (1917) as the "colonist's encyclopedia", contained miscellaneous information, ranging from agricultural and medical advice to poetry and short stories. Eventually, newspapers would also circulate, bringing regional, national and international news, advertising goods and services and announcing events organized by the clubs and associations.

Unfortunately, most of these periodicals disappeared as a result of the 1937 "nationalization campaign". This campaign consisted of a series of impositions aiming to coerce the *Deutschbrasilianer*[11] and other ethnic groups[12] into abandoning their foreign identities and assimilating the Brazilian culture and language. First, the schools were subject of a series of restrictions which forced many of them to cease operations. Next, the act of speaking foreign languages in public places was prohibited. Finally, the impositions were extended to the press, interfering in the publications by censoring or demanding the inclusion of determined articles and requiring the periodicals to be published in both languages. Eventually, the foreign language

[5] *Tanz-, Sänger-, Schützenverein* (hdt.): Dance, singing, shooting club

[6] *Kerb* (hrs.): Traditional festival, celebrated in the local church anniversary (hdt. *Kirchweih*)

[7] High German, despite being always present, prevailed in urban environments, while the others were usually associated with the rural colonies. *Riograndenser Hunsrückisch* is the most common variation, being spoken in all southern states and in the mid-west of the country, and also in Paraguay and Argentina. Many other smaller language islands exist or existed until some point, as it can be verified in the works of Fausel (1959) and Altenhofen et al. (2018)

[8] For example, the "*Lese- und Uebungsbuch zur Erlernung der portugiesischen Sprache für die Deutschbrasilischen Siedlungsschulen*" by Schäfer (1925)

[9] *Kirchenblatt* (hdt.): Magazine or newsletter issued by a church

[10] *Kalender* (hdt.): Almanac published usually in a yearly basis

[11] *Deutschbrasilianer* (hdt.): German-Brazilian, as the German descendants started to identify themselves, specially in the German-language press (Seyferth, 1999)

[12] Such as the Italian-, Polish- and Japanese-Brazilians

press was completely shut down (Seyferth, 1999). Despite the existence of German-language prints after the end of the nationalization campaign, the vast majority of the legacy of the German-Brazilian press is dated from before the prohibition.

2.2 THE GERMAN-LANGUAGE BRAZILIAN PRESS

Probably the most important statistics of the late modern German-language press in Brazil are provided by Hans Gehse in his "*Die deutsche Presse in Brasilien von 1852 bis zur Gegenwart*", published in 1931. In his work, Gehse provides a map with the numbers of German-language newspapers by municipality, which can be found in Figure 2.1. In the map, newspapers still in circulation at the time are represented by a "+" sign and the publications which were already defunct by then are represented by a "*x*" sign.

By analysing the map, it can be observed that most of the publications were concentrated in the big, urban centers such as *Porto Alegre*, *Curityba* — currently *Curitiba* — and *São Paulo*. The older colonies of *Rio Grande do Sul* (*São Leopoldo*, *Montenegro*, *Santa Cruz*) and the colonies of *Blumenau* and *Joinville* and surroundings also constitute important hubs of the German-Brazilian press.

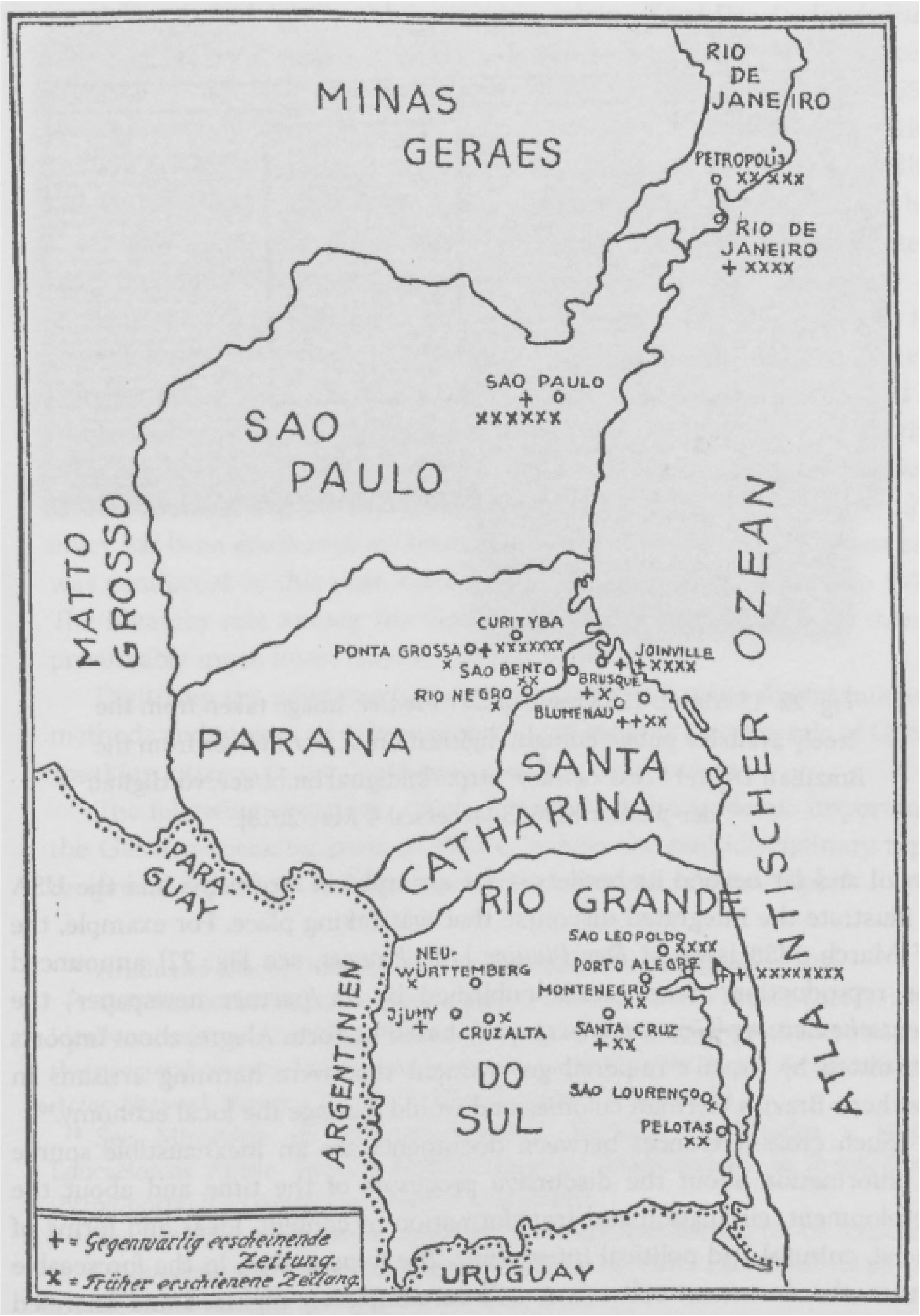


Figure 2.1: German-Brazilian newspapers published by municipality until 1931 according to Gehse (obtained from Soethe (2020))

3 CONCEPTUAL BACKGROUND

In order to better understand the contents of this work, some key concepts must first be explained in detail. In section 3.1, an overview of the image processing techniques used to extract information from digitized historical document images is provided. In section 3.2, deep learning and related concepts are introduced. Finally, the main industry standards for encoding, storage and representation of images, metadata and ground truth are addressed in section 3.3.

3.1 HISTORICAL DOCUMENT PROCESSING

“Historical document processing” refers to a collection of image processing techniques used to extract information from digital historical document images. This information can be either visual or textual. It should be noticed that “visual information” can mean not only pictorial elements of a document page, such as images, illustrations and decorations, but any imagery extracted from a digitized document image. This imagery can be of structural components, such as separators, and even images including text such as advertisements end article blocks. Textual information, on the other hand, refers specifically to machine-readable text recognized from images.

Text recognition techniques are divided in two different fields: Handwritten Text Recognition (HTR) and **Optical Character Recognition** (**OCR**) (Philips and Tabrizi, 2020). The former, as the name says, is used for manuscripts, while the latter is used for printed documents, which is the case of the German-Brazilian periodicals. Visual content retrieval, on the other hand, is conventionally performed before the OCR process, with the objective of identifying and segmenting components such as image, graphical, separator and text regions, in a procedure named **Optical Layout Recognition** (**OLR**) or **layout analysis**. Figure 3.1, replicated from Philips and Tabrizi (2020), illustrates a simple historical document processing workflow, where the images are submitted to a preprocessing phase, consisting of binarization[1], layout analysis and text line segmentation. A more elaborate example is the functional model of the OCR-D project (Herrmann, 2017), which can be found in Figure 3.2. In this representation, OLR operations are used to obtain pages, then text and finally lines, with preprocessing operations being executed before each of these segmentation steps. Once the lines are segmented, further preprocessing steps are executed, then, finally OCR is performed at line level.

[1] Binarization: Conversion of a color or gray image to a binary, black-and-white one. Binarization is historically a very important step in document image processing, since it is essentially a classification of each pixel into two classes: Foreground and background. Algorithms would use this information to perform OLR and OCR. When processing historical document images, often noisy given the aging and degradation, binarization can be a challenging step, strongly influencing the results of subsequent operations. After the advent and popularization of neural networks for historical document processing, virtually replacing conventional algorithms, OLR and OCR models started to be trained directly with images in RGB or grayscale qualities, dismissing the need for binarization in some workflows.

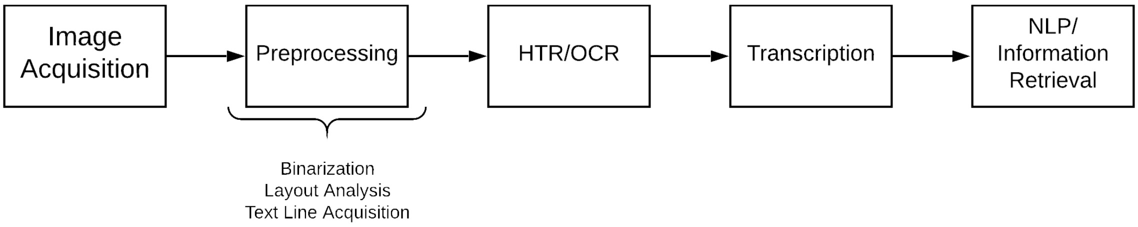


Figure 3.1: Historical document processing phases according to Philips and Tabrizi (2020)

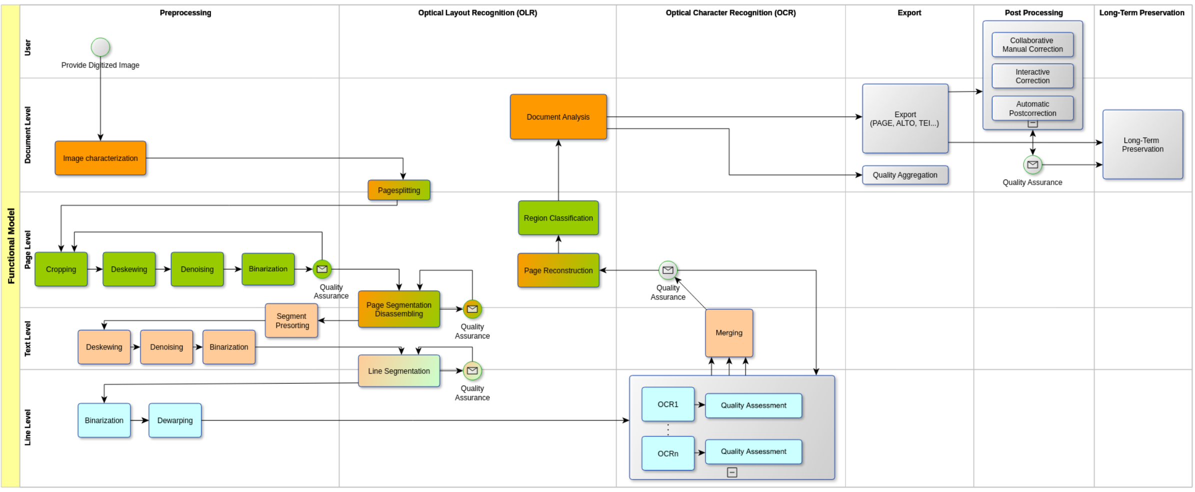


Figure 3.2: OCR-D project functional model (Herrmann, 2017)

In preprocessing, besides OLR and binarization, important concepts are **skew correction** and **orientation correction**, which are represented as *deskewing* in Figure 3.2. For the purposes of this work, skew correction means the adjustment of slight skews or inclinations of a document page in relation to the image, which are usually introduced during scanning. Orientation correction, on the other hand, means the normalization of pages or segments rotated in multiples of 90º. While this also has to do with digitization when occurring at page level, segments such as adverts and text regions were often printed in different orientations, so it is common for a page image to feature content in orientations that are other than the page's, and this must be adjusted before the OCR phase.

Historical document layout analysis constitutes a particularly challenging step, due to the noisy, degraded pages and complex layouts that are often featured in antique documents. Optical character recognition is also complicated in some cases, since old typefaces are often harder to recognize than contemporary fonts, which is the case of the blackletter types found in historical German-language documents (Neudecker et al., 2019). These two challenges are present in the German-Brazilian press, specially when it comes to advertisement pages, which have the most complex layouts and are not only typeset in many different blackletter types, but also share room with Latin types.

3.2 DEEP LEARNING, DATASETS AND GROUND TRUTH

Most techniques for both layout and text recognition nowadays are based on **deep learning**, which is a sub-field of machine learning whose scope is deep neural networks. A **dataset** consists of a set of document images and annotations in respect to such images. These annotations constitute the **ground truth**. In the context of OLR, the ground truth is usually encoded as polygon- or pixel-wise labels that describe the layout components of a page image, and for OCR, it is the machine-encoded representations of text characters. The ground truth is used for training deep learning models to be able to **predict** such labels when processing document images, and, if verified that the model learned how to identify such patterns without producing many errors, it can be used to automate the annotation process for unseen images, which are not included in the dataset. The dataset is usually divided into a **training set**, which is supposed to be used for training models, and an **evaluation or test set**, which is used to assess the performance of the model and do the necessary adjustments during training.

3.3 IMAGE, METADATA AND GROUND TRUTH STANDARDS

In order to represent, encode and distribute images, metadata and ground truth, there are many available standards. For distributing document images, the **Image API**[2] of the **International Image Interoperability Framework (IIIF)**[3] is a common method. By manipulating

[2] https://iiif.io/api/image/3.0/

[3] https://iiif.io/

parameters of a HTTP request (URI), images, or regions of images, can be retrieved in different sizes, rotations, qualities (colorspaces) and formats from an IIIF image server. These features are usually used by image viewers for an optimized, lightweight and versatile presentation of document images, and for this reason it is used by digital libraries.

In order to represent layout ground truth, usually polygons associated with a label (e.g. text, separator, graphical region, etc.) are used to describe the components. This is the case of the **PAGE** format[4] (Pletschacher and Antonacopoulos, 2010). It consists of a XML file which references and describes a page image. Besides the layout, described as XML elements, the specification also supports annotating orientations (and skews), text transcriptions and text styles (e.g. font size and family). To describe collections of pages, the structural fields of a **METS**[5] file are often used. This is the case of the **OCR-D**[6] framework, which uses the METS file groups[7] to organize not only the PAGE ground truth of a set of images, but also results of document processing operations, usually encoded as images (the case of binarization) or even as PAGE files. This is part of a collection of specifications that allow the implementation of modular image processing tools, which can be used to build customized historical document processing workflows.

[4] https://www.primaresearch.org/schema/PAGE/gts/pagecontent/2019-07-15/pagecontent.xsd

[5] https://www.loc.gov/standards/mets/

[6] https://ocr-d.de/

[7] https://www.loc.gov/standards/mets/METSOverview.v2.html#filegrp

4 PREVIOUS WORKS

In this chapter, related activities and contributions made before the work on the *gbn-ads* dataset are outlined. Experiments conduced with available ground truth of German-Brazilian document images and development of related software are the subject of Section 4.1. In Section 4.2, software developed for visualization of PAGE annotations is addressed. In Section 4.3, the main contributions made to the open-source OCR-D community are outlined.

4.1 LAYOUT ANALYSIS OF GERMAN-BRAZILIAN NEWSPAPERS

In the field of historical document processing, specially when it comes to German prints, a technique for layout analysis that became popular in recent years is the usage of models that perform pixel-wise predictions. This means that models are trained with and predict patches of images where each pixel is labeled according to its respective layout component, which can be, for example, none (background), text region, separator region, graphical region, etc. The *Qurator* team (Rehm et al., 2020), in a partnership with the Prussian Cultural Heritage Foundation of the Berlin State Library[1], has developed and distributed outstanding software and deep learning models for this purpose. As examples, the *sbb-pixelwise-segmentation*[2] for training pixel-wise segmentation models and the *sbb-textline-detector*[3] for the actual segmentation/prediction (later superseded by the more complete *eynollah*[4]) can be brought up. In early experiments with German-Brazilian periodical page images from the GBN dataset (Araújo, 2019) (mostly non-advert pages), the results of layout and text line segmentation of such tools were impressive, specially when compared to other established segmentation utilities.

In order to provide a modular interface to the *Qurator*/SPK tools, the *ocrd-gbn* project[5] was developed. Following the OCR-D specifications, the features of layout and text line segmentation, as well as cropping and binarization[6] provided by the *Qurator* tools were re-implemented as separate command lines. This allowed the composition of customized workflows, which were used for other experiments. An example of such experiences can be found in Figure 4.1, where the results of predictions of models provided by *Qurator* can be verified, both built with a ResNet U-Net 50 architecture for the neural networks.

[1] *Stiftung Preußischer Kulturbesitz — Staatsbibliothek zu Berlin*

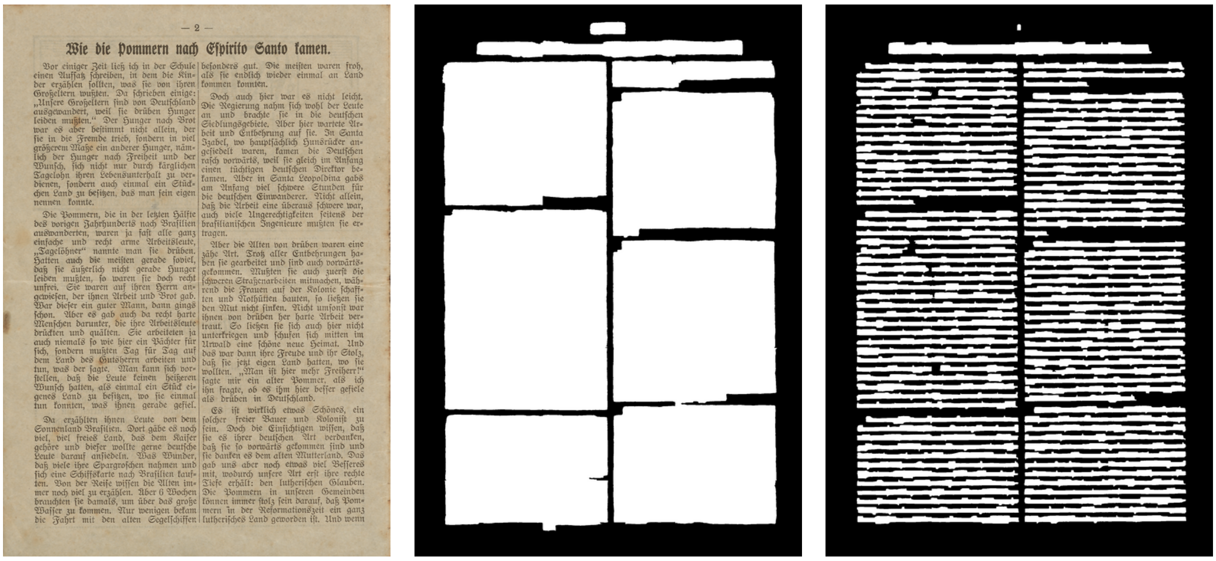
[2] https://github.com/qurator-spk/sbb_pixelwise_segmentation

[3] https://github.com/qurator-spk/sbb_textline_detection

[4] https://github.com/qurator-spk/eynollah

[5] https://github.com/GBN-DBP/ocrd-gbn/tree/dev

[6] https://github.com/qurator-spk/sbb_binarization



— 2 —

Wie die Pommern nach Espirito Santo kamen.

Vor einiger Zeit ließ ich in der Schule einen Aufſatz ſchreiben, in dem die Kinder erzählen ſollten, was ſie von ihren Großeltern wußten. Da ſchrieben einige: „Unſere Großeltern ſind von Deutſchland ausgewandert, weil ſie drüben Hunger leiden mußten.“ Der Hunger nach Brot war es aber beſtimmt nicht allein, der ſie in die Fremde trieb, ſondern in viel größerem Maße ein anderer Hunger, nämlich der Hunger nach Freiheit und der Wunſch, ſich nicht nur durch kärglichen Tagelohn ihren Lebensunterhalt zu verdienen, ſondern auch einmal ein Stückchen Land zu beſitzen, das man ſein eigen nennen konnte.

Die Pommern, die in der letzten Hälfte des vorigen Jahrhunderts nach Braſilien auswanderten, waren ja faſt alle ganz einfache und recht arme Arbeitsleute, „Tagelöhner“ nannte man ſie drüben. Hatten auch die meiſten gerade ſoviel, daß ſie äußerlich nicht gerade Hunger leiden mußten, ſo waren ſie doch recht unfrei. Sie waren auf ihren Herrn angewieſen, der ihnen Arbeit und Brot gab. War dieſer ein guter Mann, dann gings ſchon. Aber es gab auch da recht harte Menſchen darunter, die ihre Arbeitsleute drückten und quälten. Sie arbeiteten ja auch niemals ſo wie hier ein Pächter für ſich, ſondern mußten Tag für Tag auf dem Land des Gutsherrn arbeiten und tun, was der ſagte. Man kann ſich vorſtellen, daß die Leute keinen heißeren Wunſch hatten, als einmal ein Stück eigenes Land zu beſitzen, wo ſie einmal tun konnten, was ihnen gerade gefiel.

Da erzählten ihnen Leute von dem Sonnenland Braſilien. Dort gäbe es noch viel, viel freies Land, das dem Kaiſer gehöre und dieſer wollte gerne deutſche Leute darauf anſiedeln. Was Wunder, daß viele ihre Spargroſchen nahmen und ſich eine Schiffskarte nach Braſilien kauften. Von der Reiſe wiſſen die Alten immer noch viel zu erzählen. Über 6 Wochen brauchten ſie damals, um über das große Waſſer zu kommen. Nur wenigen bekam die Fahrt mit den alten Segelſchiffen beſonders gut. Die meiſten waren froh, als ſie endlich wieder einmal an Land kommen konnten.

Doch auch hier war es nicht leicht. Die Regierung nahm ſich wohl der Leute an und brachte ſie in die deutſchen Siedlungsgebiete. Aber hier wartete Arbeit und Entbehrung auf ſie. In Santa Izabel, wo hauptſächlich Hunsrücker angeſiedelt waren, kamen die Deutſchen raſch vorwärts, weil ſie gleich im Anfang einen tüchtigen deutſchen Direktor bekamen. Aber in Santa Leopoldina gabs am Anfang viel ſchwere Stunden für die deutſchen Einwanderer. Nicht allein, daß die Arbeit eine überaus ſchwere war, auch viele Ungerechtigkeiten ſeitens der braſilianiſchen Ingenieure mußten ſie ertragen.

Aber die Alten von drüben waren eine zähe Art. Trotz aller Entbehrungen haben ſie gearbeitet und ſind auch vorwärtsgekommen. Mußten ſie auch zuerſt die ſchweren Straßenarbeiten mitmachen, während die Frauen auf der Kolonie ſchafften und Nothütten bauten, ſo ließen ſie den Mut nicht ſinken. Nicht umſonſt war ihnen von drüben her harte Arbeit vertraut. So ließen ſie ſich auch hier nicht unterkriegen und ſchufen ſich mitten im Urwald eine ſchöne neue Heimat. Und das war dann ihre Freude und ihr Stolz, daß ſie jetzt eigen Land hatten, wo ſie wollten. „Man iſt hier mehr Freiherr!“ ſagte mir ein alter Pommer, als ich ihn fragte, ob es ihm hier beſſer gefiele als drüben in Deutſchland.

Es iſt wirklich etwas Schönes, ein ſolcher freier Bauer und Koloniſt zu ſein. Doch die Einſichtigen wiſſen, daß ſie es ihrer deutſchen Art verdanken, daß ſie ſo vorwärts gekommen ſind und ſie danken es dem alten Mutterland. Das gab uns aber noch etwas viel Beſſeres mit, wodurch unſere Art erſt ihre rechte Tiefe erhält: den lutheriſchen Glauben. Die Pommern in unſeren Gemeinden können immer ſtolz ſein darauf, daß Pommern in der Reformationszeit ein ganz lutheriſches Land geworden iſt. Und wenn

(a) Original page (b) Text regions prediction (c) Text lines prediction

Figure 4.1: Experiments of text region and line predictions using a ResNet U-Net 50

4.2 VISUALIZATION OF PAGE XML ANNOTATIONS

In order to provide customizable and scalable visualization options for layout analysis results and ground truth, a powerful command-line tool was developed to parse PAGE annotations and draw them on top of the page image files as overlays. Named *page-xml-draw*[7], this tool is useful for viewing specific annotations in arbitrary colors and opacities. An example of a visualization image obtained with this tool can be found in Figure 4.2.

4.3 CONTRIBUTIONS TO THE OCR-D COMMUNITY

During this prior experimentation phase, a few bug fixes and improvements were submitted and accepted by the open-source projects of the OCR-D community. They are enumerated and briefly described below.

OCR-D/olena Fix *python* interpreter version

cisogroup/ocrd_cis Fix metadata annotation (indicating that *deskewing* was performed)

qurator-spk/sbb_textline_detection Fix metadata compliance

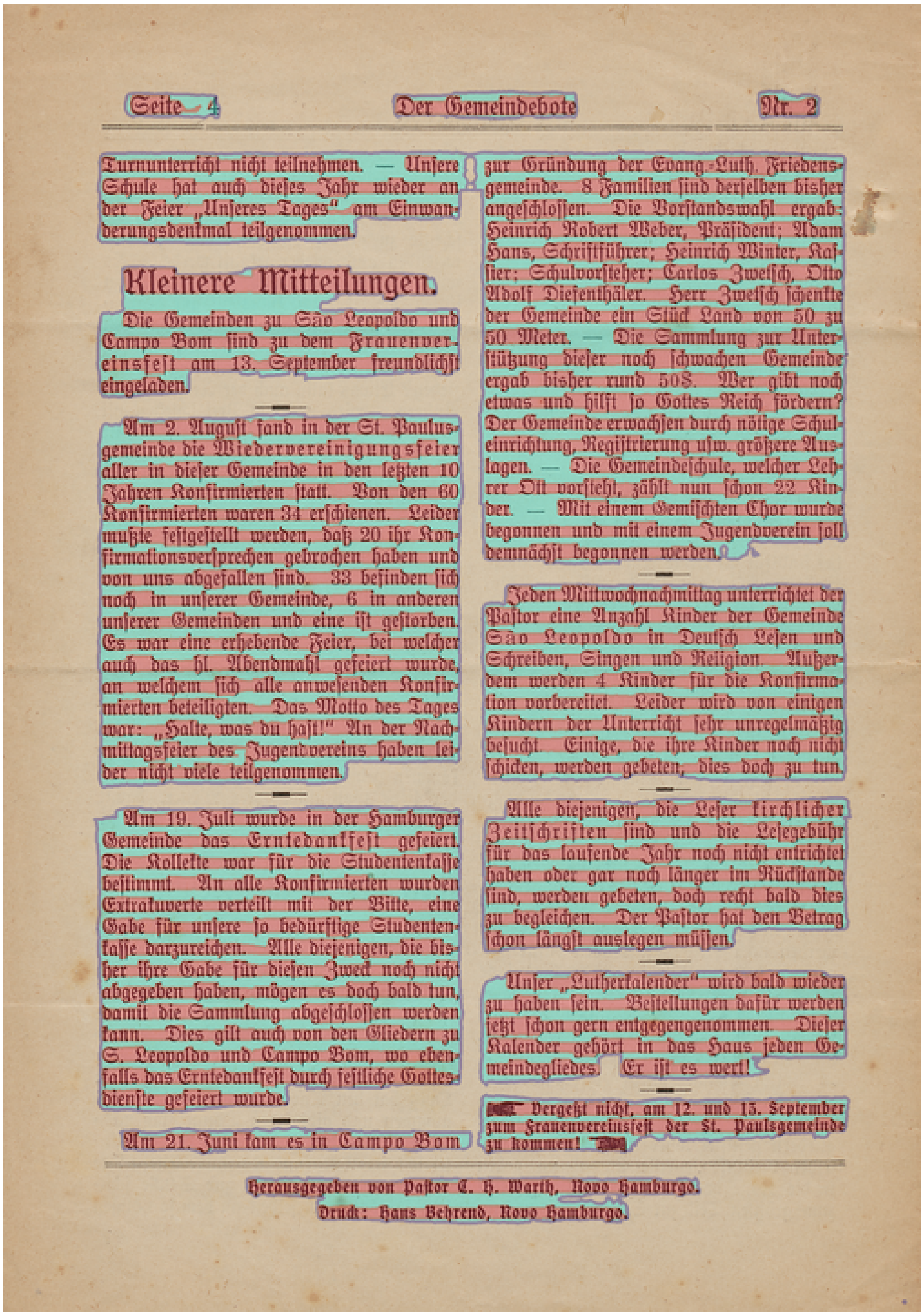
UB-Mannheim/ocrd_pagetopdf Update dependencies

OCR-D/ocrd_all Update dependencies

hnesk/browse-ocrd Add support for viewing images with transparency

qurator-spk/sbb_binarization Fix memory (space) leak

[7] https://github.com/GBN-DBP/page-xml-draw



Seite 4 Der Gemeindebote Nr. 2

Turnunterricht nicht teilnehmen. — Unſere Schule hat auch dieſes Jahr wieder an der Feier „Unſeres Tages“ am Einwanderungsdenkmal teilgenommen.

Kleinere Mitteilungen.

Die Gemeinden zu São Leopoldo und Campo Bom ſind zu dem Frauenvereinsfeſt am 13. September freundlichſt eingeladen.

Am 2. Auguſt fand in der St. Paulusgemeinde die Wiedervereinigungsfeier aller in dieſer Gemeinde in den letzten 10 Jahren Konfirmierten ſtatt. Von den 60 Konfirmierten waren 34 erſchienen. Leider mußte feſtgeſtellt werden, daß 20 ihr Konfirmationsverſprechen gebrochen haben und von uns abgefallen ſind. 33 befinden ſich noch in unſerer Gemeinde, 6 in anderen unſerer Gemeinden und eine iſt geſtorben. Es war eine erhebende Feier, bei welcher auch das hl. Abendmahl gefeiert wurde, an welchem ſich alle anweſenden Konfirmierten beteiligten. Das Motto des Tages war: „Halte, was du haſt!“ An der Nachmittagsfeier des Jugendvereins haben leider nicht viele teilgenommen.

Am 19. Juli wurde in der Hamburger Gemeinde das Erntedankfeſt gefeiert. Die Kollekte war für die Studentenkaſſe beſtimmt. An alle Konfirmierten wurden Extrakuverte verteilt mit der Bitte, eine Gabe für unſere ſo bedürftige Studentenkaſſe darzureichen. Alle diejenigen, die bisher ihre Gabe für dieſen Zweck noch nicht abgegeben haben, mögen es doch bald tun, damit die Sammlung abgeſchloſſen werden kann. Dies gilt auch von den Gliedern zu S. Leopoldo und Campo Bom, wo ebenfalls das Erntedankfeſt durch feſtliche Gottesdienſte gefeiert wurde.

Am 21. Juni kam es in Campo Bom zur Gründung der Evang.-Luth. Friedensgemeinde. 8 Familien ſind derſelben bisher angeſchloſſen. Die Vorſtandswahl ergab: Heinrich Robert Weber, Präſident; Adam Hans, Schriftführer; Heinrich Winter, Kaſſier; Schulvorſteher; Carlos Zweſch, Otto Adolf Dieſenthäler. Herr Zweſch ſchenkte der Gemeinde ein Stück Land von 50 zu 50 Meter. — Die Sammlung zur Unterſtützung dieſer noch ſchwachen Gemeinde ergab bisher rund 50$. Wer gibt noch etwas und hilft ſo Gottes Reich fördern? Der Gemeinde erwachſen durch nötige Schuleinrichtung, Regiſtrierung uſw. größere Auslagen. — Die Gemeindeſchule, welcher Lehrer Ott vorſteht, zählt nun ſchon 22 Kinder. — Mit einem Gemiſchten Chor wurde begonnen und mit einem Jugendverein ſoll demnächſt begonnen werden.

Jeden Mittwochnachmittag unterrichtet der Paſtor eine Anzahl Kinder der Gemeinde São Leopoldo in Deutſch Leſen und Schreiben, Singen und Religion. Außerdem werden 4 Kinder für die Konfirmation vorbereitet. Leider wird von einigen Kindern der Unterricht ſehr unregelmäßig beſucht. Einige, die ihre Kinder noch nicht ſchicken, werden gebeten, dies doch zu tun.

Alle diejenigen, die Leſer kirchlicher Zeitſchriften ſind und die Leſegebühr für das laufende Jahr noch nicht entrichtet haben oder gar noch länger im Rückſtande ſind, werden gebeten, doch recht bald dies zu begleichen. Der Paſtor hat den Betrag ſchon längſt auslegen müſſen.

Unſer „Lutherkalender“ wird bald wieder zu haben ſein. Beſtellungen dafür werden jetzt ſchon gern entgegengenommen. Dieſer Kalender gehört in das Haus jeden Gemeindegliedes. Er iſt es wert!

☞ Vergeßt nicht, am 12. und 13. September zum Frauenvereinsfeſt der St. Paulsgemeinde zu kommen! ☜

Herausgegeben von Paſtor C. H. Warth, Novo Hamburgo.
Druck: Hans Behrend, Novo Hamburgo.

Figure 4.2: Visualization of predicted text regions (cyan) and lines (red)

5 RELATED WORKS

In this chapter, the most relevant research datasets and ground truth corpora for the challenges of 1. layout analysis and 2. recognition of text in multiple fonts are reviewed. By reviewing these corpora, it is intended to identify the most appropriate standards, technologies and strategies for the German-Brazilian Newspaper Advertisement Pages Dataset (*gbn-ads*). In Section 5.1, layout ground truth specifically is addressed and in Section 5.2, font-related ground truth corpora are reviewed.

5.1 LAYOUT GROUND TRUTH CORPORA

Due to the absence — at least to the author's knowledge — of datasets composed specifically of German historical advertisements or advertisement pages, datasets of German historical newspapers in general were looked up for this review. The selected corpora follow the criteria of 1. being open for academic research purposes, 2. being easy to access and download, 3. being at least partially composed of historical German-language newspaper images and 4. containing layout annotations (ground truth).

Opening the list of reviewed works, the **Impresso (Black Letter) OCR Ground Truth**[1] features digitized page images of the Swiss newspaper "*Neue Zürcher Zeitung*" (Ströbel and Clematide, 2019; Ehrmann et al., 2020). The images are encoded as grayscale TIFF files, with varying resolutions. The ground truth, consisting of PAGE files containing region, line and word annotations (with a few lacking annotations for words)[2], was produced using the *Transkribus* platform[3]. The corpus is composed exclusively of front pages of the newspaper, and does not seem to contain advertisements.

Next, the German-Brazilian Newspapers dataset (**GBN**) (Araújo, 2019) can be brought up. Its images come from images gathered by the *dbp digital* project and is apparently the single ground truth corpus containing images of German-Brazilian periodicals. Not only newspapers are featured in GBN, but also magazines and church newsletters. The ground truth consists of text, image, graphical and separator regions annotations, produced with the Aletheia document analysis system[4]. Only the training set is public[5], being the images distributed both as bitonal (black-and-white) and color PNG files, and the ground truth as PAGE and also as plain text files containing polygon coordinates and labels. Some pages have a resolution of 300 DPI, while others were scanned at 600 DPI. A few advertisement pages can be found, but the majority of the collection consists of article pages.

[1] https://doi.org/10.5281/zenodo.3333626

[2] https://github.com/impresso/NZZ-black-letter-ground-truth

[3] https://readcoop.eu/transkribus/?sc=Transkribus

[4] https://www.primaresearch.org/tools/Aletheia

[5] https://web.inf.ufpr.br/vri/databases/gbn/

The **NewsEye**'s **READ OCR**[6] and **ICPR 2020**[7] datasets come next in the list, featuring region-, line- and word-level annotations of Austrian newspaper pages. Despite one containing pages that are not present in the other and vice-versa, there seems to exist a significant overlap of images and ground truth files. Both datasets feature images in multiple resolutions, both in the TIFF and JPEG file formats, and the ground truth stored as PAGE files with word-level annotations, compiled with *Transkribus*. The READ OCR corpus has both gray and RGB images, while the ICPR 2020 set has bitonal and RGB files. Advertisements can be found in many pages of both corpora.

In order to improve text recognition and enable sentiment analysis[8] research on digitized pages of the *Berliner Börsen-Zeitung* newspaper, Liebl and Burghardt (2020) present the **Origami** OCR Pipeline. Besides the actual tool[9], the ground truth used to train the models was made available[10]. Both transcriptions and region- and line-level layout annotations are included in a custom format[11]. The images consist of grayscale JPEG files digitized at 300 DPI. Advert pages are also common in this collection.

Closing the list of layout ground truth corpora, figures the **Impresso Article Segmentation Ground Truth**[12] (Ehrmann et al., 2020; Barman et al., 2021), with pages of the French-language Swiss newspapers *Journal de Genève*, *Gazette de Lausanne* and *L'Impartial*, as well of the German-, French- and Luxembourgish-language *Luxemburger Wort*. This dataset has a significantly different ground truth composition than the other corpora presented here: Instead of providing annotations for every region, only certain content blocks are segmented and labeled either as serials, weather forecasts, obituaries or stock exchange tables. This approach allows training semantic segmentation models, and since only few specific regions are segmented, a significantly large corpus has been compiled. The images of the Swiss newspapers are distributed as JPEG grayscale files at 72 DPI, and the *Luxemburger Wort* images, while following the same specifications, are provided by an IIIF image server, and only referenced in the ground truth files. The region polygons and labels were annotated with the VGG Image Annotator[13] and stored as their dedicated VIA JSON format.

A summary of these corpora, with their respective image and ground truth information can be found in Table 5.1. The proprietary *Transkribus* figures as the favorite annotation tool, and PAGE as the main standard for ground truth encoding. Despite the support provided by PAGE specification to advert annotations (*AdvertRegion*), all the works reviewed here seem to make use of only a small set of basic layout classes: Text, image, graphic and separator regions. The *Origami* GT and *Impresso* Article Segmentation GT are exceptions, featuring annotations

[6] https://doi.org/10.5281/zenodo.4943581

[7] https://doi.org/10.5281/zenodo.4943582

[8] http://media-sentiment.uni-leipzig.de/

[9] https://github.com/poke1024/origami

[10] https://github.com/poke1024/origami_models

[11] https://github.com/poke1024/origami/blob/master/docs/formats.md

[12] https://doi.org/10.5281/zenodo.3706862

[13] https://www.robots.ox.ac.uk/~vgg/software/via/

of “composed” types of visual content, which are tables in the case of *Origami*, and semantic types in the *Impresso* GT. However, no annotation effort makes a distinction between the text, separators and graphics of advert blocks and those of articles, headers, and non-advertisement content in general.

Dataset	**Images**			**Ground Truth**			**# Pages**
	Quality	**DPI**	**Format**	**Level**	**Tool**	**Format**	
Impresso BlaLet GT	Gray	224..521	TIFF	Line Word	Transkribus	PAGE	167
GBN	Bitonal RGB	300 600	PNG	Region	Aletheia	PAGE Custom	*102
NewsEye READ OCR	Gray RGB	194..460	TIFF JPEG	Word	Transkribus	PAGE	161
NewsEye ICPR 2020	Bitonal RGB	194..473	TIFF JPEG	Word	Transkribus	PAGE	100
Origami GT	Gray	300	JPEG	Line	Origami	Custom	1182
Impresso ArtSeg GT	Gray	72	JPEG	Region	VIA	JSON	23577

* Only the 102 pages of the training set are public

Table 5.1: Inventory of reviewed layout ground truth and datasets

5.2 FONT AND TEXT GROUND TRUTH CORPORA

There are not many works addressing recognition of text typeset in multiple fonts, at least when it comes to historical documents. The few that exist will be classified here in two categories: Font-specific and font-generic. Font-specific approaches consist of classifying images of text by font type so it can be transcribed by models trained specifically to recognize its respective font type, while font-generic approaches consist of training OCR models with multiple font data so that a single model can transcribe text of different font types.

Since layout is not relevant for this topic, the reviewed datasets do not necessarily contain newspaper images, but are at least partially composed of historical German-language document images in general. Some sort of font ground truth is required for the font-specific approaches, while for font-generic approaches the text transcription alone is enough. The presented corpora are also open for academic purposes and easily retrievable.

In the font-specific front, figures the **font groups** dataset presented by Seuret et al. (2019). In this corpus, pages of early modern books from different European libraries were classified according to the typefaces found in them. The *Labelbox*[14] software was used to classify fonts at page-level, and the ground truth is distributed as a two-column CSV file associating each image file to the assigned label. While page-level font classification is not sufficient to describe the typographical variety and complexity of German-Brazilian advertisement pages, this

[14] https://labelbox.com/

technique, if combined to granular layout annotations, could be used to label smaller segments such as words.

Examples of font-generic approaches are the **Archiscribe** corpus[15] and the **GT4HistOCR** dataset[16] of (Springmann et al., 2018). Both collections are composed of early modern books, featuring Latin and Gothic typefaces. The datasets are composed of line images encoded as PNG files, with plain text files containing the transcriptions for each line. This approach allows training of very comprehensive and robust OCR models, although, font ground truth could be useful for evaluating the performance of such models in different font groups, providing feedback that could help in the achievement even more robust and comprehensive models.

[15] https://github.com/jbaiter/archiscribe-corpus.git

[16] https://doi.org/10.5281/zenodo.1344131

6 DATA AND METHODS

In this chapter, both the corpus where the images of the German-Brazilian Newspaper Pages Dataset (*gbn-ads*) come from and the methodology used to sample, handle and distribute such images are described. The corpus is addressed in Section 6.1. In Section 6.2, the selection criteria and procedures undertook for sampling, handling and distributing the images are exposed. In Section 6.3, the approach for producing ground truth for the dataset is outlined, and the development of an annotation tool to help compile such ground truth is the topic of Section 6.4.

6.1 THE *DBP DIGITAL* CORPUS

Nowadays, most remaining specimen of the German-language Brazilian press can be found in libraries and collections around Brazil. In a few cases, these organizations undertook digitization efforts, and some made the images available online. The foreign language newspapers collection of the *São Paulo* State University (UNESP) Library[1], the Brazilian Lutheran Confession Church (IECLB) collection[2] and the German-language periodicals collection of the Brazilian National Library[3] are examples of available collections. The *dbp digital* project[4], aiming to build a digital library of the German-Brazilian press, has compiled and digitized a considerable amount of German-language periodicals from different sources, from which a corpus of 6190 page images have been gathered, organized and stored in a server with remote access and back-up routines for carrying out research on layout and font recognition. An inventory of the document images included in the *dbp digital* corpus can be found in Table 6.1.

Periodical	**Location**	**Period**	**DPI**	**# Pages**
Colonie- / Kolonie-Zeitung	Joinville	1862-1900	600	4529
Deutsches Wochenblatt	Curitiba	1886	300	84
Der Erzähler*	Curitiba	1886	300	36
Deutsches Echo	Curitiba	1886	300	64
Der Pionier	Curitiba	1887-1891	300/600	1309
Neuer Deutscher Kolonie-Anzeiger**	Wiesbaden	1887	300	4
Brusquer Zeitung	Brusque	1912	300/600	108
Evangelisch-Lutherisches Kirchenblatt für Süd-Amerika	Porto Alegre	1916-1919	300	56
		1862-1919	**300/600**	**6190**

* *Deutsches Wochenblatt* supplement
** *Der Pionier* supplement (printed in Germany)

Table 6.1: *dbp digital* collection

[1] https://bibdig.biblioteca.unesp.br/handle/10/8047 (in bpt.)

[2] https://www.luteranos.com.br/conteudo_organizacao/jornais/a-cruz-no-sul-kreuz-im-suden (in bpt.)

[3] http://memoria.bn.br/docreader/docmulti.aspx?bib=ger (in bpt.)

[4] https://dokumente.ufpr.br/en/dbpdigital.html

The most expressive periodical of the collection is the *Colonie-Zeitung* (published as *Kolonie-Zeitung* from 1869 onwards), a newspaper with thousands of pages from an almost 40-year range of publishing dates. The short-lived *Deutsches Wochenblatt* and *Deutsches Echo*, followed by *Der Pionier* constitute a set of newspapers published in *Curitiba*, containing supplements such as *Der Erzähler*, which feature content such as serials and humorous sections, and the *Neuer Deutscher Kolonie-Anzeiger*, featuring advertisements from German producers and traders. Another newspaper is the *Brusquer Zeitung*, published in the early 20th century. Closing the list, figures the *Evangelisch-Lutherisches Kirchenblatt für Süd-Amerika*, which used to be the main publication of the Lutheran Missouri Synod in South America (Weiduschadt, 2015). A few examples of these publications can be found in Figure 6.1.

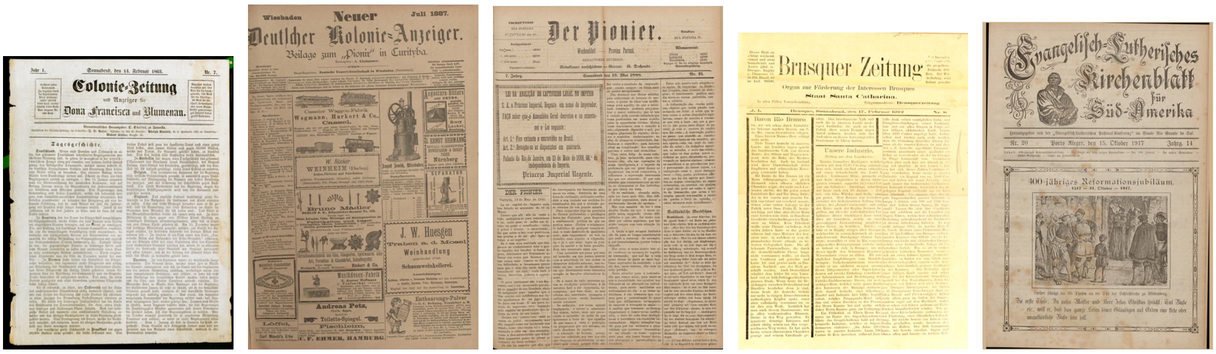
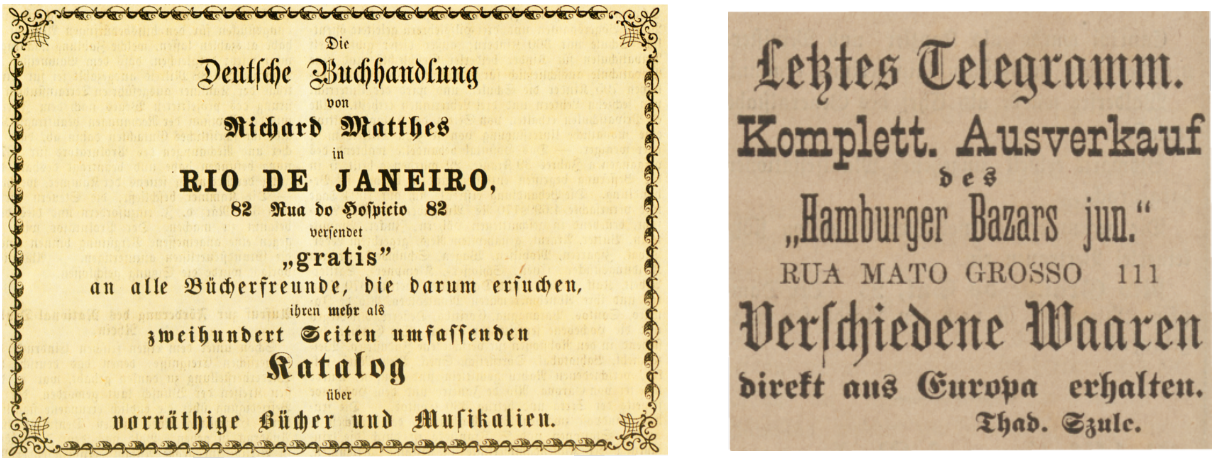


Figure 6.1: Example front pages from the *dbp digital* collection

The documents, being from different sources, were digitized in different resolutions: While the resolution of the *Kolonie-Zeitung*, *Der Pionier* and *Brusquer Zeitung* periodicals is 600 DPI (dots, or pixels, per inch), the remaining titles were scanned in 300 DPI. Exceptions are the 1887 editions of the *Der Pionier* and the first one of the *Brusquer Zeitung*, which also have a resolution of 300 DPI. Small manual corrections were performed, such as rotating the images in cases of "landscape" orientations and splitting multi-page images. All the files are stored as RGB TIFF files.

An outstanding characteristic of the periodicals is the prevalence of blackletter, or Gothic typefaces, coexisting with Roman/Latin types. Conventionally, German text is typeset in blackletter, most cases in variants of *Fraktur*, while Latin fonts — usually *Antiqua* — are used for Portuguese and other languages in general. This contrast is not only observed in complete text blocks and sentences written in different languages, but also in individual words of a same sentence. For example, it is common to find person and place names of German origin typeset in Gothic even though the rest of the sentence is redacted in Portuguese and typeset in Latin, and vice versa. Exceptions can be found in early editions of the *Kolonie-Zeitung*, where blackletter types are often used also for Portuguese text, and the *Brusquer Zeitung* pages, which are entirely typeset in Latin types. Advertisements also do not often follow such convention, frequently resorting to Latin typesetting for German words, either to create contrasts between blackletter and Roman types. Figure 6.2 shows some great examples of these contrasts. It is also common to find adverts featuring the same text content on both languages and typefaces, either side-by-side or in Z, similarly to what has been observed in German-Polish adverts of Silesian Catholic periodicals (Haładewicz-Grzelak and Lubos-Kozieł, 2013).



Die
Deutſche Buchhandlung
von
Richard Matthes
in
RIO DE JANEIRO,
82 Rua do Hoſpicio 82
verſendet
„gratis"
an alle Bücherfreunde, die darum erſuchen,
ihren mehr als
zweihundert Seiten umfaſſenden
Katalog
über
vorräthige Bücher und Muſikalien.

(a) *Fraktur* (many different forms) and *Antiqua*

Letztes Telegramm.
Komplett. Ausverkauf
des
„Hamburger Bazars jun.“
RUA MATO GROSSO 111
Verſchiedene Waaren
direkt aus Europa erhalten.
Thad. Szulc.

(b) *Textur*, *Fraktur* and *Antiqua*

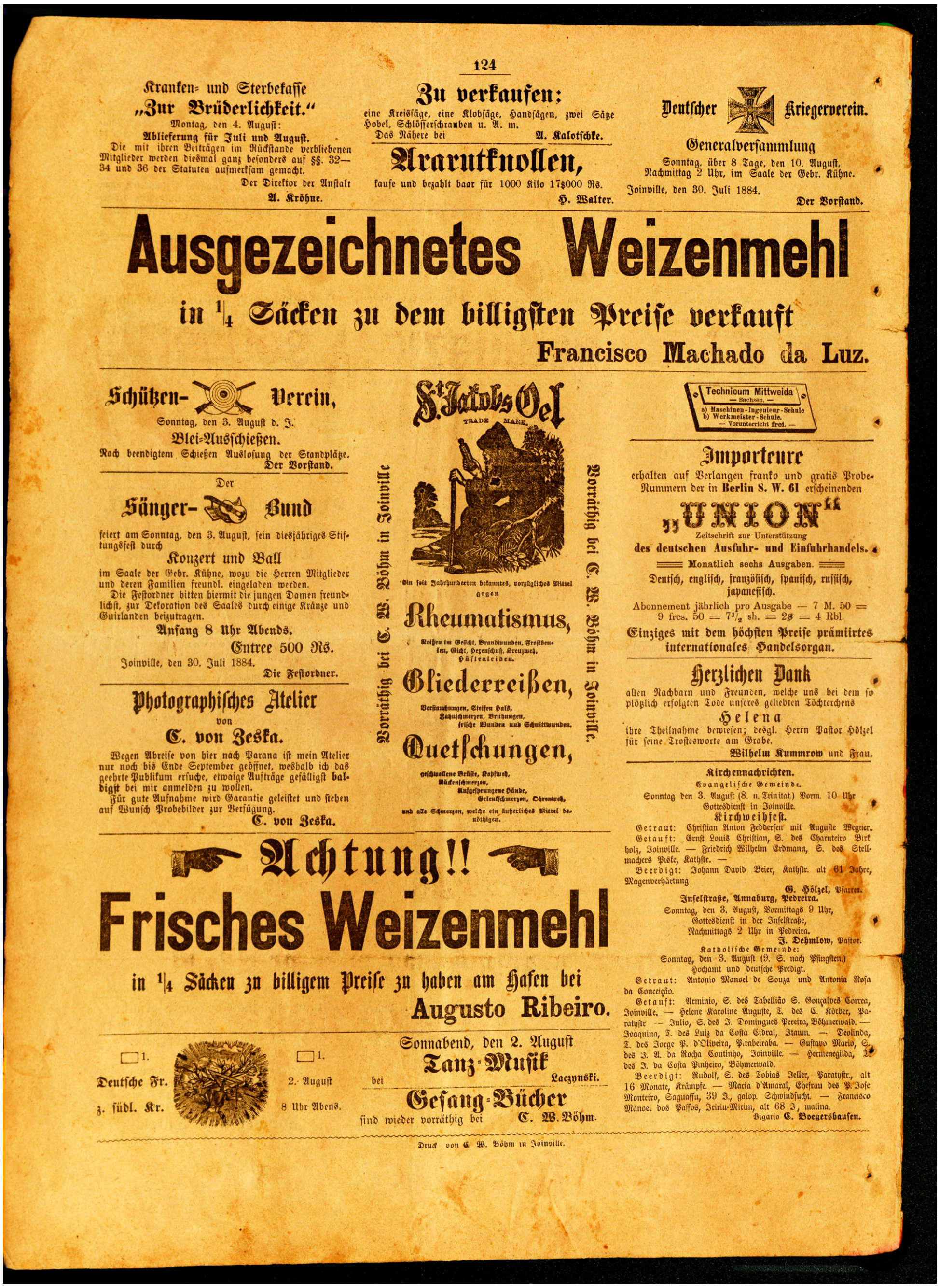
Figure 6.2: Examples of adverts with font variation

When it comes to newspaper advertisements, not only a large variety of fonts come to attention, but also the layouts in which content is arranged in some pages, often not respecting a fixed number of rows and columns. These unusual structures, combined with the typographical peculiarities mentioned above, make the German-Brazilian newspapers, and specially advertisement pages, substantially harder to extract information from than letters, books and even other kinds of periodicals such as magazines. A great example of a page with complex layout can be observed in Figure 6.3. To sum it up, many of the pages are in poor condition: Besides the yellow- and brownish colors of the aged pages, stains, tears, ink fades and bleed-throughs[5] are frequent kinds of degradation. Some teared pages were also taped back together, creating occlusions. Some distortions were introduced during scanning, causing some pages to be slightly skewed or warped. When sampling the images, these problems must be taken into account for the composition a representative dataset, which will enable training of models that can recognize layout and text despite the difficulties.

6.2 IMAGE SELECTION AND INTEROPERABILITY

Before sampling, the images of pages containing adverts were identified. For this task, any visually distinctive and delimited block of content where one or more products, services, requests or events are announced was considered an "advert". After this pre-selection, the actual sampling took place, aiming to compose a smaller, yet representative set of advert pages. In order to achieve this goal, the most important variables of digitization quality, physical condition, layout and typography were identified and taken into account during sampling. In respect to

[5] Bleed-through: Leak of ink to the back side of a page



124

Kranken- und Sterbekaſſe „Zur Brüderlichkeit.“

Montag, den 4. Auguſt:

Ablieferung für Juli und Auguſt.

Die mit ihren Beiträgen im Rückſtande verbliebenen Mitglieder werden diesmal ganz beſonders auf §§. 32—34 und 36 der Statuten aufmerkſam gemacht.

Der Direktor der Anſtalt
A. Kröhne.

Zu verkaufen:

eine Kreisſäge, eine Klobſäge, Handſägen, zwei Sätze Hobel, Schlöſſerſchrauben u. A. m.
Das Nähere bei **A. Kalotſchke.**

Ararutknollen,

kaufe und bezahlt baar für 1000 Kilo 17$000 Rs.

H. Walter.

Deutſcher Kriegerverein.

Generalverſammlung

Sonntag, über 8 Tage, den 10. Auguſt, Nachmittag 2 Uhr, im Saale der Gebr. Kühne.

Joinville, den 30. Juli 1884.

Der Vorſtand.

Ausgezeichnetes Weizenmehl

in $^1|_4$ Säcken zu dem billigſten Preiſe verkauft

Francisco Machado da Luz.

Schützen-Verein,

Sonntag, den 3. Auguſt d. J.

Blei-Ausſchießen.

Nach beendigtem Schießen Ausloſung der Standplätze.

Der Vorſtand.

Der

Sänger-Bund

feiert am Sonntag, den 3. Auguſt, ſein diesjähriges Stiftungsfeſt durch

Konzert und Ball

im Saale der Gebr. Kühne, wozu die Herren Mitglieder und deren Familien freundl. eingeladen werden.

Die Feſtordner bitten hiermit die jungen Damen freundlichſt, zur Dekoration des Saales durch einige Kränze und Guirlanden beizutragen.

Anfang 8 Uhr Abends.

Entree 500 Rs.

Joinville, den 30. Juli 1884.

Die Feſtordner.

Photographiſches Atelier

von

C. von Zeska.

Wegen Abreiſe von hier nach Parana iſt mein Atelier nur noch bis Ende September geöffnet, weshalb ich das geehrte Publikum erſuche, etwaige Aufträge gefälligſt baldigſt bei mir anmelden zu wollen.

Für gute Aufnahme wird Garantie geleiſtet und ſtehen auf Wunſch Probebilder zur Verfügung.

C. von Zeska.

St. Jakobs Oel

TRADE MARK.

Vorräthig bei C. W. Böhm in Joinville

Ein ſeit Jahrhunderten bekanntes, vorzügliches Mittel gegen

Rheumatismus,

Reißen im Geſicht, Brandwunden, Froſtbeulen, Gicht, Hexenſchuß, Kreuzweh, Hüftenleiden.

Gliederreißen,

Verſtauchungen, Steifen Hals, Zahnſchmerzen, Brühungen, friſche Wunden und Schnittwunden.

Quetſchungen,

geſchwollene Brüſte, Kopfweh, Rückenſchmerzen, Aufgeſprungene Hände, Gelenkſchmerzen, Ohrenweh,

und alle Schmerzen, welche ein äußerliches Mittel benöthigen.

Vorräthig bei C. W. Böhm in Joinville.

Technicum Mittweida
— Sachsen. —
a) Maschinen-Ingenieur-Schule
b) Werkmeister-Schule.
— Vorunterricht frei. —

Importeure

erhalten auf Verlangen franko und gratis Probe-Nummern der in **Berlin S. W. 61** erſcheinenden

„UNION“

Zeitschrift zur Unterstützung
des deutschen Ausfuhr- und Einfuhrhandels.

Monatlich sechs Ausgaben.

Deutſch, engliſch, franzöſiſch, ſpaniſch, ruſſiſch, japaneſiſch.

Abonnement jährlich pro Ausgabe — 7 M. 50 = 9 frcs. 50 = $7^1/_2$ sh. = 2$ = 4 Rbl.

Einziges mit dem höchſten Preiſe prämiirtes internationales Handelsorgan.

Herzlichen Dank

allen Nachbarn und Freunden, welche uns bei dem ſo plötzlich erfolgten Tode unſeres geliebten Töchterchens

Helena

ihre Theilnahme bewieſen; desgl. Herrn Paſtor Hölzel für ſeine Troſtesworte am Grabe.

Wilhelm Kummrow und Frau.

Kirchennachrichten.

Evangeliſche Gemeinde.

Sonntag den 3. Auguſt (8. n. Trinitat.) Vorm. 10 Uhr Gottesdienſt in Joinville.

Kirchweihfeſt.

Getraut: Chriſtian Anton Fedderſen mit Auguſte Wegner.

Getauft: Ernſt Louis Chriſtian, S. des Charuteiro Birkholz, Joinville. — Friedrich Wilhelm Erdmann, S. des Stellmachers Piske, Kathſtr. —

Beerdigt: Johann David Beier, Kathſtr. alt 61 Jahre, Magenverhärtung

G. Hölzel, Pfarrer.

Inſelſtraße, Annaburg, Pedreira.

Sonntag, den 3. Auguſt, Vormittags 9 Uhr, Gottesdienſt in der Inſelſtraße, Nachmittags 2 Uhr in Pedreira.

J. Dehmlow, Paſtor.

Katholiſche Gemeinde:

Sonntag, den 3. Auguſt (9. S. nach Pfingſten.) Hochamt und deutſche Predigt.

Getraut: Antonio Manoel de Souza und Antonia Roſa da Conceição.

Getauft: Arminio, S. des Tabellião S. Gonçalves Correa, Joinville. — Helene Karoline Auguſte, T. des C. Körber, Paratyſtr — Julio, S. des J. Domingues Pereira, Böhmerwald. — Joaquina, T. des Luiz da Coſta Cidral, Itaum. — Deolinda, T. des Jorge P. d'Oliveira, Pirabeiraba. — Guſtavo Mario, S. des J. A. da Rocha Coutinho, Joinville. — Hermenegilda, T. des J. da Coſta Pinheiro, Böhmerwald.

Beerdigt: Rudolf, S. des Tobias Jeller, Paratyſtr., alt 16 Monate, Krämpfe. — Maria d'Amaral, Ehefrau des P. Joſe Monteiro, Saguaſſu, 39 J., galop. Schwindſucht. — Francisco Manoel dos Paſſos, Iririu-Mirim, alt 68 J, malina.

Vigario **C. Boegershauſen.**

Achtung!!

Friſches Weizenmehl

in $^1|_4$ Säcken zu billigem Preiſe zu haben am Hafen bei

Augusto Ribeiro.

□ 1.
Deutſche Fr.
z. ſüdl. Kr.

□ 1.
2. Auguſt
8 Uhr Abens.

Sonnabend, den 2. Auguſt

Tanz-Muſik

bei **Laczynski.**

Geſang-Bücher

ſind wieder vorräthig bei **C. W. Böhm.**

Druck von C. W. Böhm in Joinville.

Figure 6.3: Example of newspaper page with complex layout

digitization quality, page images in both resolutions (300 and 600 DPI) were included, as well as pages with distortions introduced during scanning, such as slightly skews/inclinations and warps. The variation of physical condition was also contemplated, with the inclusion pages with different background (paper) colors and ink levels and also pages with degradations such as stains, tears and bleed-throughs. When it comes to layout and structure, pages with different row and column numbers, separator styles, advert styles, sizes and orientations and text orientations were selected. Typographical attributes were also observed, with images containing text typeset in different font families and sizes being sampled.

Since the images have been digitized in high resolutions, they are stored in relatively large files. The complete *dbp digital* corpus has 1,1 TB of size, with the resulting sample, alone, having 57 GB in total. This is a huge limitation in storage and distribution of the dataset, and can also slow down the annotation process since the images take a long time to render. Most of the works reviewed in Chapter 5 seem to deal with this problem by either compiling smaller datasets, resampling the images into lower resolutions, or resorting to image formats with compression, be it lossless (PNG) or lossy (JPEG). On the other hand, most of these methods (except lossless image formats) consist of somehow reducing the amount of information included in the images, which can impact the accuracy of neural networks trained with such data.

To solve this dilemma, it was opted for the implementation of an IIIF image server. The original files are stored in a web server, and, through the IIIF Image API, images can be retrieved in different qualities, formats and sizes, according to the needs of the user. With this approach, the dataset does not have to follow storage constraints, as the images are not provided, only referenced, similarly to what is done to the *Luxemburger Wort* pages of the *Impresso* Article Segmentation Ground Truth. When accessing the images for viewing, they can be requested in smaller sizes and/or divided in tiles and rendered more fluidly. On the other hand, the images can still be retrieved in its original sizes, qualities and formats, without information loss. Other functions provided by the API can also be used to automate image transformations such as cropping and rotation.

For the implementation of the IIIF image server, the *Cantaloupe*[6] server was chosen. After installing this software and its dependencies in a remote server, the original image files selected for the dataset were copied and exposed to the image server daemon. This setup was enough for the web server to work, implementing the IIIF Image API functionalities. The *Cantaloupe* server supports features[7] such as arbitrary rotations, retrieval of images in color, gray and bitonal qualities (being color the default quality) and in TIFF, JPEG, GIF and PNG formats.

[6] https://cantaloupe-project.github.io/

[7] https://iiif.io/api/image/3.0/#57-extra-functionality

6.3 GROUND TRUTH COMPILATION

As discussed in Chapter 5, there is a lack of detailed, granular ground truth when it comes to both advert layout annotations and font classification of historical German-language newspapers. Aiming to fulfill this gap, the following approach is chosen for ground truth compilation of the *gbn-ads* dataset: First, the angle of each page is adjusted, correcting any skew introduced during digitization. These adjustments can make manual segmentation easier, specially when drawing rectangles, speeding up the annotation process. Next, the advert blocks of each page are segmented. The segmentation of other components such as headers and articles is not so important for the specified purposes, because 1. in order to detect adverts, only an "advert" class and a "non-advert" class are needed and 2. the font variety in non-advertisement content is negligible. By concentrating the annotation efforts in adverts alone, a lot of time and energy can be saved. Moving on, the orientation of each advert is corrected. In theory, the skew/inclination of advert blocks should be the same of their respective pages, and therefore already normalized at this point, so only rotations by multiples of 90º should be needed. Basic region types, such as text, separators and graphical elements are then segmented and labeled accordingly afterwards. Some adverts contain text regions in distinct skews and orientations, so another angle correction is performed. For more granularity, the resulting text regions have their lines segmented, which subsequently have their words split. No further rotations are necessary, since skew and orientation variations do not seem to exist between lines in a same region nor words in a same line. The words are then classified according to their font types, and finally, the text of each word transcribed. An illustration of this annotation pipeline can be found in Figure 6.4.

This methodology essentially breaks down the conventional layout analysis process of segmenting whole pages into basic region types into a two-step approach where advert blocks are identified first, regardless of their compositions in terms of texts, graphics or separators, and later the layout analysis of basic region types is performed individually for each advert, which pose as significantly simpler challenges in comparison to layout analysis of complex advertisement pages. The word-level font classification can not only be used to train granular font classification models and font-specific OCR models, but also to assess the effectiveness of font-generic methods for font groups. Given the modular, step-by-step structure of the annotation workflow, other training and annotation pipelines can be executed concurrently. For example, once the advertisements are segmented, advert segmentation model training can already take place, even though the annotation flow is still being followed. Other annotation activities such as assigning semantic labels to adverts can also be performed during the proposed pipeline execution.

For ground truth encoding, the PAGE standard is used. For the implementation, a PAGE file for every image is created, containing a *Page* annotation pointing to an IIIF image URI. According to the schema[8], a *Page* element can contain multiple elements of *Region* types,

[8] https://www.primaresearch.org/schema/PAGE/gts/pagecontent/2019-07-15/pagecontent.xsd

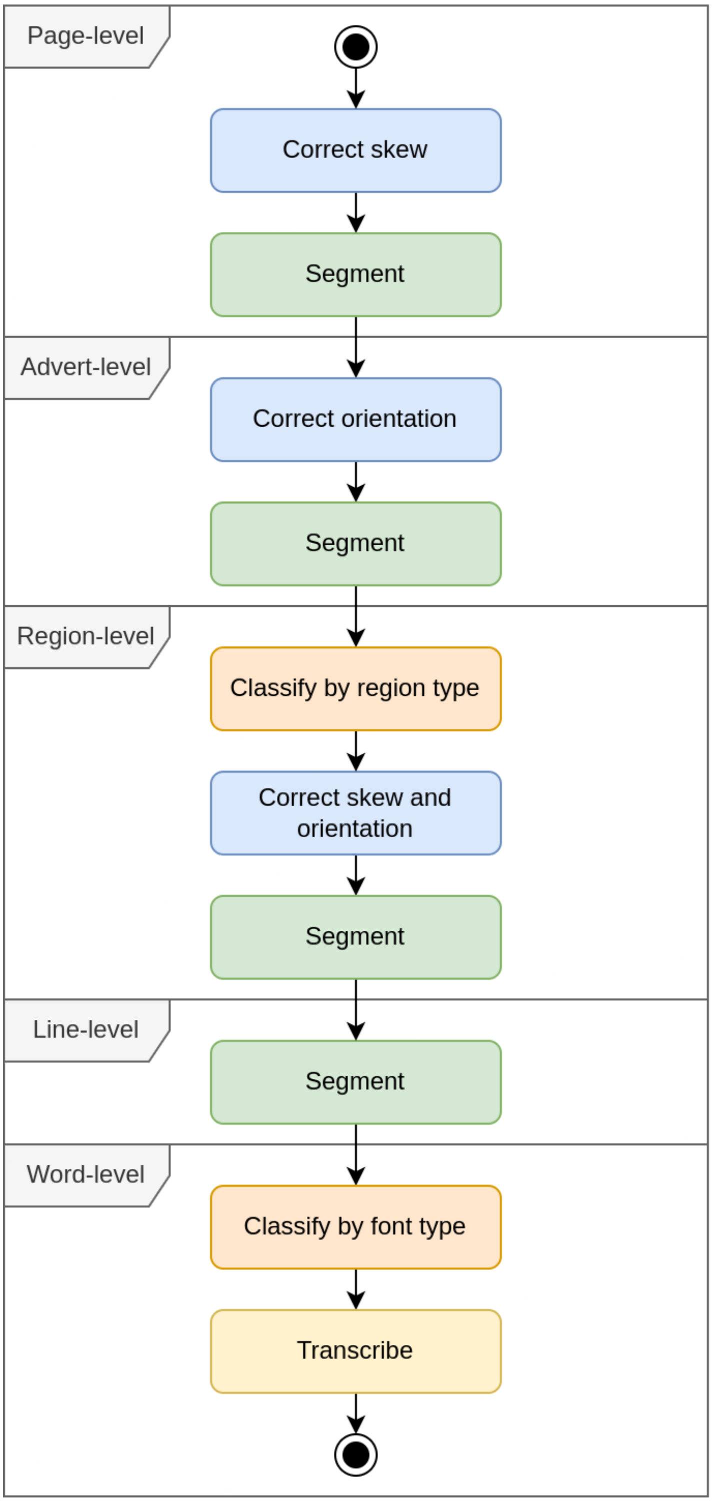


Figure 6.4: Proposed annotation pipeline

which in turn can also recursively contain *Region* elements. Following the annotation pipeline, *AdvertRegion* elements can be appended to each *Page*, and then elements such as *TextRegion*, *GraphicRegion* and *SeparatorRegion* can be appended to each *AdvertRegion*. Each *TextRegion* can contain *TextLine* elements, and each *TextLine* can contain *Word* elements. To encode the skew and orientation angles at page-, advert- and text-region-level, the *orientation* attribute can be used. A *TextStyle* element can be used to encode the word font, and a *TextEquiv* component to encode the transcriptions. An example of what these annotations should look like can be found in Figure 6.5. In order to facilitate the storage and compilation of intermediary results of the annotation flow, the METS file groups can be used to distinguish and organize the PAGE files produced in different steps of the pipeline, similarly to what OCR-D recommends for their automatic workflows.

```
<Page imageFilename="http://server/id/full/full/0/default.png" orientation="0.0">
   <AdvertRegion orientation="0.0">
      <Coords points="x0,y0 x1,y0 x1,y1 x0,y1"/>
      <TextRegion orientation="0.0">
         <Coords points="x0,y0 x1,y0 x1,y1 x0,y1"/>
         <TextLine>
            <Coords points="x0,y0 x1,y0 x1,y1 x0,y1"/>
            <Word>
               <Coords points="x0,y0 x1,y0 x1,y1 x0,y1"/>
               <TextEquiv>
                  <Unicode>Beispieltext</Unicode>
               </TextEquiv>
               <TextStyle fontFamily="Fraktur"/>
            </Word>
         </TextLine>
      </TextRegion>
   </AdvertRegion>
</Page>
```

Figure 6.5: Example PAGE ground truth

Given time constraints, the compilation of ground truth could not be performed, and is therefore proposed as a future work.

6.4 ANNOTATION TOOL DEVELOPMENT

In order to implement the annotation pipeline described in Section 6.3, annotation tools such as the previously addressed *Transkribus* and *Aletheia* platforms, as well as the LAREX[9] software (Reul et al., 2017) could be used. Despite supporting the PAGE format, and being useful for collaborative annotations (except *Aletheia*), many problems arise when trying to perform the annotation activities, of which the following can be enumerated:

1. No optimized display, taking long for the images to render
2. None of the tools seem to support page-, advert- nor text-region-level rotations for skew and orientation correction

[9] https://github.com/OCR4all/LAREX

3. There is no feature for displaying regions such as adverts or text alone, with only a page-level view being available[10]

4. There is apparently no support for text style (font) annotations

5. Apparently none of them allow annotation of recursive regions

6. *Transkribus* and *Aletheia* specifically are proprietary software with licensing limitations

To solve these issues, the IIIF Image API features could be taken advantage of to build a web-based, lightweight application for image annotation. For optimized rendering and zooming, a tile-based IIIF image viewer can be used. For skew and orientation angle annotation, the image can be rotated in real time simply by manipulating the *rotation* parameter of the IIIF Image API. By setting the *region* parameter, previously segmented elements (adverts, text regions, lines or words) can be retrieved and displayed separately. For transcribing and classifying segments, something similar to the interface of the *neat* annotation tool[11], which displays lines retrieved from image servers, could be implemented.

In the scope of this work, an annotation tool has been developed for skew and orientation correction, on top of *Leaflet-IIIF*[12] (which in turn is built on top of *Leaflet.js*[13]). *Leaflet.js* is a minimalist, extensible image viewer, conventionally used to display map images retrieved from image server as tiles, in a fluid and lightweight manner. *Leaflet-IIIF* is a plugin that makes it easy to display an image of an IIIF image server. With the implementation of a customizable overlay grid and an angle selector, the skew/orientation of an image can be easily verified and corrected. An overview of the tool functionality and interface is provided in Chapter 7.

[10] In this context, display of smaller segments would be helpful for filtering and focusing on the content to be annotated

[11] https://github.com/qurator-spk/neat

[12] https://github.com/mejackreed/Leaflet-IIIF

[13] https://leafletjs.com/

7 RESULTS AND DISCUSSION

In this chapter, the results obtained from the activities performed and described in Chapter 6 are outlined. Section 7.1 addresses the resulting image sample and how it is organized, while in Section 7.2, an overview of the annotation tool developed for skew and orientation correction is provided.

7.1 DATASET IMAGES AND METADATA

From the 6190 page images of the *dbp digital* collection, adverts were identified in 2023 of them. From this amount, 311 images were sampled for the German-Brazilian Newspaper Advertisement Pages Dataset (*gbn-ads*). The number of advert pages and selected page images per publication can be found in Table 7.1. To organize this set of images, the METS standard is used. A file group stores the IIIF Information Request URI[1] of each selected image, and another file group references local PAGE files representing each page image. The PAGE files do not contain any ground truth as of now, only the reference to their respective images as IIIF Image Request URIs[2] and image metadata such as width and height. These image requests should return the images in their full regions and sizes, without any rotation, in their original colors (RGB), and encoded as PNG files for a lossless compression. The same URIs are also registered in the METS. This METS/PAGE representation allows an easy integration with OCR-D workflows in the future, facilitating the execution of OCR-D tools using the *gbn-ads* images or ground truth as input.

Title	**# Advert Pages**	**# Selected Pages**
Colonie- / Kolonie-Zeitung	1295	201
Deutsches Wochenblatt	22	5
Deutsches Echo	11	3
Der Pionier	640	85
Neuer Deutscher Kolonie-Anzeiger	4	4
Brusquer Zeitung	51	13
	2023	**311**

Table 7.1: Pages selected for the *gbn-ads* dataset

As specified in Chapter 6, not only structural and typographical characteristics were considered during the sampling process, but also many physical- and digitization-related degradations were included. This property, on one side, likely makes it more difficult to produce layout ground truth and to train effective segmentation models with than the average dataset, but, on the other hand, the models trained with such data will develop robustness and hopefully achieve good prediction results on pages other models will suffer to extract significant information.

[1] https://iiif.io/api/image/3.0/#22-image-information-request-uri-syntax

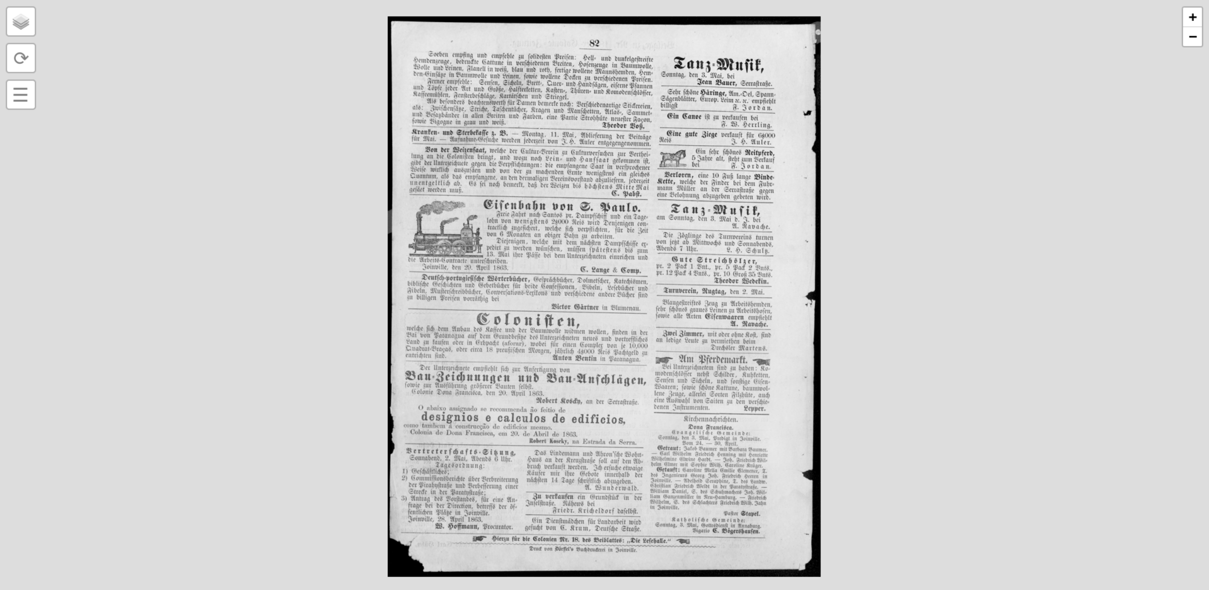
[2] https://iiif.io/api/image/3.0/#21-image-request-uri-syntax

Many of the selected advert pages contain significant amounts of non-advert content, which constitute an important attribute for training data, since neural network should also learn with examples of what is not considered an advert. On a side note, the absence of completely advertisement-free pages is probably not ideal when following the same line of thought. For example, while occurrences of article and advertising sections in a same page are not rare, the same cannot be said about headers, always found in front pages, and adverts, typically featured in the last pages of an edition. Since the dataset does not contain many front pages, the behavior of a model trained with such data will be uncertain when predicting pages with headers. This configures a limitation of the proposed *gbn-ads* dataset, and an extra sample of the remaining 4167 non-advert pages might be needed in order to accomplish consistent and comprehensive results.

Once the ground truth is compiled, the *gbn-ads* dataset will not only be the single ground truth corpora of advert regions of historical German newspapers, but also the most granular dataset of historical font groups. From the advert ground truth perspective, this will enable the conception of innovative approaches for layout analysis and information retrieval, and from the font ground truth point of view, comprehensive training data for font and text recognition of late modern German typesetting will be made available.

7.2 ANNOTATION TOOL

A screenshot showing the interface of the developed annotation tool can be found in Figure 7.1. It has a minimalist interface, which consists of a wide canvas where a zoomable and draggable page image is displayed in the center, with a few buttons called “controls” in the corners of the canvas.



82

Soeben empfing und empfehle zu ſolideſten Preiſen: Hell- und dunkelgeſtreifte Hemdenzeuge, bedruckte Cattune in verſchiedenen Breiten, Hoſenzeuge in Baumwolle, Wolle und Leinen, Flanell in weiß, blau und roth, fertige wollene Mannshemden, Hemden-Einſätze in Baumwolle und Leinen, ſowie wollene Decken zu verſchiedenen Preiſen.

Ferner empfehle: Senſen, Sicheln, Brett-, Quer- und Handſägen, eiſerne Pfannen und Töpfe jeder Art und Größe, Halfterketten, Kaſten-, Thüren- und Komodenſchlöſſer, Kaffeemühlen, Fenſterbeſchläge, Kartätſchen und Striegel.

Als beſonders beachtenswerth für Damen bemerke noch: Verſchiedenartige Stickereien, als: Zwiſchenſätze, Striche, Taſchentücher, Kragen und Manſchetten, Atlas-, Sammet- und Beſatzbänder in allen Breiten und Farben, eine Partie Strohhüte neueſter Façon, ſowie Vigogne in grau und weiß.

Theodor Voß.

Kranken- und Sterbekaſſe z. B. — Montag, 11. Mai, Ablieferung der Beiträge für Mai. — Aufnahme-Geſuche werden jederzeit von J. H. Auler entgegengenommen.

Von der Weizenſaat, welche der Cultur-Verein zu Culturverſuchen zur Vertheilung an die Coloniſten bringt, und wozu noch Lein- und Hanfſaat gekommen iſt, gibt der Unterzeichnete gegen die Verpflichtungen: die empfangene Saat in verſprochener Weiſe wirklich auszuſäen und von der zu machenden Ernte wenigſtens ein gleiches Quantum, als das empfangene, an den dermaligen Vereinsvorſtand abzuliefern, jederzeit unentgeltlich ab. Es ſei noch bemerkt, daß der Weizen bis höchſtens Mitte Mai geſäet werden muß.

C. Pabſt.

Eiſenbahn von S. Paulo.

Freie Fahrt nach Santos pr. Dampfſchiff und ein Tagelohn von wenigſtens 2$000 Reis wird Denjenigen contractlich zugeſichert, welche ſich verpflichten, für die Zeit von 6 Monaten an obiger Bahn zu arbeiten.

Diejenigen, welche mit dem nächſten Dampfſchiffe expedirt zu werden wünſchen, müſſen ſpäteſtens bis zum 13. Mai ihre Päſſe bei dem Unterzeichneten einreichen und die Arbeits-Contracte unterſchreiben.

Joinville, den 29. April 1863.

C. Lange & Comp.

Deutſch-portugieſiſche Wörterbücher, Geſprächbücher, Dolmetſcher, Katechismen, bibliſche Geſchichten und Gebetbücher für beide Confeſſionen, Bibeln, Leſebücher und Fibeln, Muſterſchreibbücher, Converſations-Lexikons und verſchiedene andere Bücher ſind zu billigen Preiſen vorräthig bei

Victor Gärtner in Blumenau.

Coloniſten,

welche ſich dem Anbau des Kaffee und der Baumwolle widmen wollen, finden in der Bai von Paranagua auf dem Grundbeſitze des Unterzeichneten neues und vortreffliches Land zu kaufen oder in Erbpacht (aforar), wobei für einen Complex von je 10,000 Quadrat-Braças, oder circa 18 preußiſchen Morgen, jährlich 4$000 Reis Pachtgeld zu entrichten ſind.

Anton Bentin in Paranagua.

Der Unterzeichnete empfiehlt ſich zur Anfertigung von

Bau-Zeichnungen und Bau-Anſchlägen,

ſowie zur Ausführung größerer Bauten ſelbſt.

Colonie Dona Franciſca, den 20. April 1863.

Robert Koſcky, an der Serraſtraße.

O abaixo assignado se recommenda ão feitio de

designios e calculos de edificios,

como tambem a construcção de edificios mesmo.

Colonia de Dona Francisca, em 20. de Abril de 1863.

Robert Kosckу, na Estrada da Serra.

Vertreterſchafts-Sitzung,

Sonnabend, 2. Mai, Abends 6 Uhr.

Tagesordnung:

1) Geſchäftliches;
2) Commiſſionsberichte über Verbreiterung der Pirahyſtraße und Verbeſſerung einer Strecke in der Paratyſtraße;
3) Antrag des Vorſtandes, für eine Anfrage bei der Direction, betreffs der öffentlichen Plätze in Joinville.

Joinville, 28. April 1863.

W. Hoffmann, Procurator.

Das Lindemann und Ahron'ſche Wohnhaus an der Kreuzſtraße ſoll auf den Abbruch verkauft werden. Ich erſuche etwaige Käufer mir ihre Gebote innerhalb der nächſten 14 Tage ſchriftlich abzugeben.

A. Wunderwald.

Zu verkaufen ein Grundſtück in der Inſelſtraße. Näheres bei

Friedr. Kricheldorf daſelbſt.

Ein Dienſtmädchen für Landarbeit wird geſucht von C. Krum, Deutſche Straße.

Tanz-Muſik,

Sonntag, den 3. Mai, bei

Jean Bauer, Serraſtraße.

Sehr ſchöne **Häringe,** Am.-Oel, Spann-Sägenblätter, Europ. Leim ꝛc. ꝛc. empfiehlt billigſt

F. Jordan.

Ein Canoe iſt zu verkaufen bei

F. W. Herrling.

Eine gute Ziege verkauft für 6$000 Reis

J. H. Auler.

Ein ſehr ſchönes **Reitpferd,** 5 Jahre alt, ſteht zum Verkauf bei

F. Jordan.

Verloren, eine 10 Fuß lange **Binde-Kette,** welche der Finder bei dem Fuhrmann Müller an der Serraſtraße gegen eine Belohnung abzugeben gebeten wird.

Tanz-Muſik,

am Sonntag, den 3. Mai d. J. bei

A. Ravache.

Die Zöglinge des Turnvereins turnen von jetzt ab Mittwochs und Sonnabends, Abends 7 Uhr.

L. H. Schultz.

Gute Streichhölzer,

pr. 2 Pack 1 Vnt., pr. 5 Pack 2 Vnts., pr. 12 Pack 4 Vnts., pr. 10 Groß 35 Vnts.

Theodor Wedekin.

Turnverein, Rugtag, den 2. Mai.

Blaugeſtreiftes Zeug zu Arbeitshemden, ſehr ſchönes graues Leinen zu Arbeitshoſen, ſowie alle Arten **Eiſenwaaren** empfiehlt

A. Ravache.

Zwei Zimmer, mit oder ohne Koſt, ſind an ledige Leute zu vermiethen beim

Drechsler Martens.

Am Pferdemarkt.

Bei Unterzeichnetem ſind zu haben: Komodenſchlöſſer nebſt Schilder, Kuhketten, Senſen und Sicheln, und ſonſtige Eiſen-Waaren; ſowie ſchöne Kattune, baumwollene Zeuge, allerlei Sorten Filzhüte, auch eine Auswahl von Saiten zu den verſchiedenen Inſtrumenten.

Lepper.

Kirchennachrichten.

Dona Franciſca.

Evangeliſche Gemeinde:

Sonntag, den 3. Mai, Predigt in Joinville.

Vom 24. — 30. April.

Getraut: Jakob Baumer mit Barbara Baumer. — Carl Wilhelm Friedrich Henning mit Henriette Wilhelmine Elwine Hardt. — Joh. Friedrich Wilhelm Elmer mit Sophie Wilh. Caroline Krüger.

Getauft: Caroline Mella Emilie Clemence, T. des Ingenieurs Georg Joh. Friedrich Heeren in Joinville. — Adelheid Seraphine, T. des Landw. Chriſtian Friedrich Weldt in der Paratyſtraße. — William Daniel, S. des Schuhmachers Joh. William Ganzenmüller in Neu-Hamburg. — Friedrich Wilhelm, S. des Schlachters Friedrich Wilh. Jahn in Joinville.

Paſtor **Stapel.**

Katholiſche Gemeinde:

Sonntag, 3. Mai, Gottesdienſt in Annaburg.

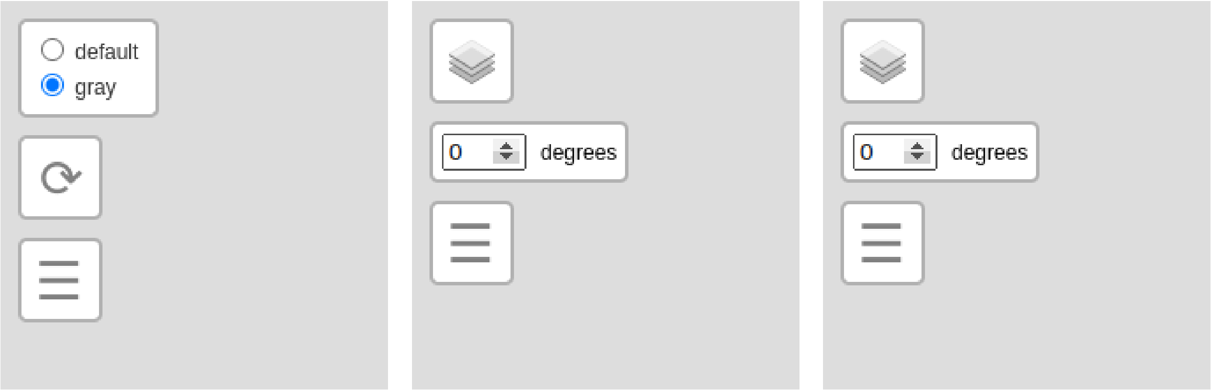
Vigario **C. Bögershauſen.**

Hierzu für die Colonien Nr. 18. des Beiblattes: „Die Leſehalle.“

Druck von Dörffel's Buchdruckerei in Joinville.

Figure 7.1: Annotation tool prototype

In the *Leaflet* architecture, “controls” are used to perform determined actions on the current visualization. Some controls are provided by default, which is the case of the zoom control (on the top right corner of Figure 7.1) and the layers control (on the top left corner). For IIIF images, the layers control can be used to select between qualities provided by the image server in real time, as it can be observed in Figure 7.2(a), which shows the behavior of the control when the mouse cursor is over the element. Following the same visual style, two custom controls were implemented for this protoype: The rotation control (Figure 7.2(b)), which can be used to select a rotation angle for the image being viewed, and the grid control (Figure 7.2(c)), which can be used to draw an overlay grid on top of the image. The rotation control, combined with a few changes in the *Leaflet-IIIF* source code, manipulates the *rotation* field of the IIIF image requests, then requests rotated tiles and assemble them together nicely in real time, a feature which is neither supported by *Leaflet.js* nor *Leaflet-IIIF*. The grid control uses an overlay vector layer to draw a grid on demand, only in the region comprehended by the canvas. The X and Y offsets and steps can be manipulated to customize the format of the grid in real time.



(a) Layers control

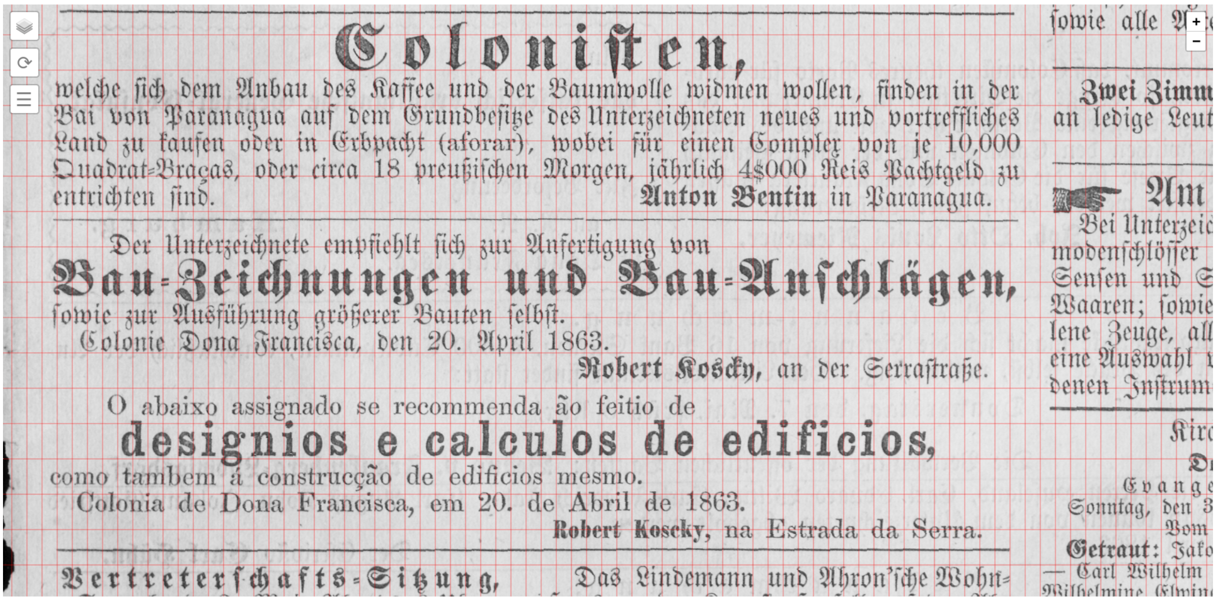
(b) Rotation control

(c) Grid control

Figure 7.2: Options of each control on mouse over

While the rotation control is naturally supposed to be used for skew or orientation correction, the purpose of the grid is to provide a baseline to help the user determine which angle is right. An example of what this grid looks like can be seen in Figure 7.3.

In the future, a control for providing the IIIF URI to be annotated should be implemented. At this point, the tool will be production-ready for annotating skew and orientation. Polygon drawing utilities for segmentation are also a planned feature, which will be used to compile layout ground truth.



Coloniſten,

welche ſich dem Anbau des Kaffee und der Baumwolle widmen wollen, finden in der Bai von Paranagua auf dem Grundbeſitze des Unterzeichneten neues und vortreffliches Land zu kaufen oder in Erbpacht (aforar), wobei für einen Complex von je 10,000 Quadrat=Braças, oder circa 18 preußiſchen Morgen, jährlich 4$000 Reis Pachtgeld zu entrichten ſind.

Anton Bentin in Paranagua.

Der Unterzeichnete empfiehlt ſich zur Anfertigung von

Bau=Zeichnungen und Bau=Anſchlägen,

ſowie zur Ausführung größerer Bauten ſelbſt.

Colonie Dona Francisca, den 20. April 1863.

Robert Koscky, an der Serraſtraße.

O abaixo assignado se recommenda ão feitio de

designios e calculos de edificios,

como tambem á construcção de edificios mesmo.

Colonia de Dona Francisca, em 20. de Abril de 1863.

Robert Koscky, na Estrada da Serra.

Vertreterſchafts=Sitzung,

Das Lindemann und Ahron'ſche Wohn=

ſowie alle A

Zwei Zimm
an ledige Leut

Am

Bei Unterzeic
modenſchlöſſer
Senſen und S
Waaren; ſowie
lene Zeuge, all
eine Auswahl v
denen Inſtrum

Kirc

D

Evange
Sonntag, den 3
Vom
Getraut: Jako
— Carl Wilhelm
Wilhelmine Elwin

Figure 7.3: Example of the grid view

8 CONCLUSION

While the work done and presented here is just a small step in a larger, longer project, it is an important building block not only for further research on layout and font recognition of German-Brazilian periodicals, but also for the publication and distribution of said documents in the digital form. Once ground truth for advert regions and font types is available, deep learning techniques can be used to experiment new strategies of information extraction from historical German-language documents. In the meantime, the choice of the IIIF framework for image retrieval and visualization has revealed a promising technology for the implementation of a digital library of the German-Brazilian press.

REFERENCES

Altenhofen, C. V., Morello, R., Winckelmann, A. C., Seiffert, A. P., Prediger, A., Schmitt, G., Bergmann, G. L., Habel, J. M., dos Santos Souza, L. C., Kohl, S. F., and Godoi, T. G. (2018). *Hunsrückisch: Inventário de uma Língua do Brasil*. Garapuvu.

Araújo, A. B. (2019). Análise de layout de página em jornais históricos germano-brasileiros.

Barman, R., Ehrmann, M., Clematide, S., Oliveira, S. A., and Kaplan, F. (2021). Combining visual and textual features for semantic segmentation of historical newspapers. *Journal of Data Mining & Digital Humanities*.

Ehrmann, M., Romanello, M., Clematide, S., Ströbel, P., and Barman, R. (2020). Language resources for historical newspapers: the impresso collection.

Fausel, E. (1959). *Die deutschbrasilianische Sprachmischung: Probleme, Vorgang und Wortbestand*. Erich Schmidt Verlag.

Haładewicz-Grzelak, M. and Lubos-Kozieł, J. (2013). Boundary mechanisms in adverts from silesian catholic periodicals from the second half of the 19th and early 20th centuries. *Sign Systems Studies*, 41(1):42–68.

Herrmann, E. (2017). Ocr-d – koordinierte förderinitiative zur weiterentwicklung von ocr-verfahren. *Bibliotheksdienst*, 52:34–41.

Hunsche, C. H. (1975). *O Biênio 1824/1825 da Imigração e Colonização Alemã no Rio Grande do Sul (Província de São Pedro)*. A Nação / DAC / SEC.

Hunsche, C. H. (1977). *O Ano 1826 da Imigração e Colonização Alemã no Rio Grande do Sul*. Metrópole.

Liebl, B. and Burghardt, M. (2020). From historical newspapers to machine-readable data: The origami ocr pipeline. In *CEUR Workshop Proceedings*, pages 351–373.

Neudecker, C., Baierer, K., Federbusch, M., Boenig, M., Würzner, K.-M., Hartmann, V., and Herrmann, E. (2019). OCR-D: An end-to-end open source OCR framework for historical printed documents. In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*. ACM.

Philips, J. P. and Tabrizi, N. (2020). Historical document processing: A survey of techniques, tools, and trends.

Pletschacher, S. and Antonacopoulos, A. (2010). The page (page analysis and ground-truth elements) format framework. *Proceedings of the 20th International Conference on Pattern Recognition*, pages 257–260.

Rehm, G., Bourgonje, P., Hegele, S., Kintzel, F., Schneider, J. M., Ostendorff, M., Zaczynska, K., Berger, A., Grill, S., Räuchle, S., Rauenbusch, J., Rutenburg, L., Schmidt, A., Wild, M., Hoffmann, H., Fink, J., Schulz, S., Seva, J., Quantz, J., Böttger, J., Matthey, J., Fricke, R., Thomsen, J., Paschke, A., Qundus, J. A., Hoppe, T., Karam, N., Weichhardt, F., Fillies, C., Neudecker, C., Gerber, M., Labusch, K., Rezanezhad, V., Schaefer, R., Zellhöfer, D., Siewert, D., Bunk, P., Pintscher, L., Aleynikova, E., and Heine, F. (2020). QURATOR: Innovative technologies for content and data curation.

Reul, C., Springmann, U., and Puppe, F. (2017). LAREX: A semi-automatic open-source tool for layout analysis and region extraction on early printed books. In *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*. ACM.

Schappelle, B. F. (1917). *The German element in Brazil: colonies and dialecy*. Americana Germanica.

Schäfer, R. (1925). *Lese- und Uebungsbuch zur Erlernung der portugiesischen Sprache für die deutsch-brasilischen Siedlungsschulen*. Typographia do Centro, 3 edition.

Seuret, M., Limbach, S., Weichselbaumer, N., Maier, A., and Christlein, V. (2019). Dataset of pages from early printed books with multiple font groups. In *Proceedings of the 5th International Workshop on Historical Document Imaging and Processing*. ACM.

Seyferth, G. (1999). Os imigrantes e a campanha de nacionalização do estado novo. *Repensando o estado novo*, pages 199–228.

Seyferth, G. (2010). Deutsche einwanderung nach brasilien. *Brasilien heute. Geographischer Raum, Politik, Wirtschaft, Kultur*, pages 739–756.

Soethe, P. (2020). *On the Transience of (Latin) American German Identities*, pages 229–245. Liverpool University Press.

Springmann, U., Reul, C., Dipper, S., and Baiter, J. (2018). Ground truth for training ocr engines on historical documents in german fraktur and early modern latin.

Ströbel, P. and Clematide, S. (2019). Improving ocr of black letter in historical newspapers: The unreasonable effectiveness of htr models on low-resolution images.

Weiduschadt, P. (2015). Os caminhos do sínodo de missouri no rio grande do sul: Educação e religiosidade (1900-1910). *Hist. Educ.*, 19(47):249–269.