Quantifying Page Segmentation Quality in Historical Job Advertisements Retrieval

Klara Venglarova, Raven Adam, Saranya Balasubramanian, Georg Vogeler

Quantifying Page Segmentation Quality in Historical Job Advertisements Retrieval

Klára Venglařová*[1], ORCID: 0009-0007-6441-7795,
Raven Adam*, ORCID: 0000-0001-7841-2601,
Saranya Balasubramanian*, ORCID: 0000-0001-7516-7671,
Georg Vogeler*, ORCID: 0000-0002-1726-1712
* University of Graz, Graz, Austria
[1] Corresponding author: Klára Venglařová <klara.venglarova@uni-graz.at>

Abstract

This paper addresses the question of which metric is the most suitable for evaluating page segmentation in the context of extracting historical job advertisements in digitized newspapers. Accurate page segmentation is essential for high-quality Optical Character Recognition (OCR) results, yet the methodology for comparing and evaluating segmentation algorithms has received limited attention in Digital Humanities. The paper presents an evaluation framework developed within the JobAds Project, focusing on textual congruence between predicted and ground-truth regions. This is important for an evidence-based selection of the best-performing segmentation algorithm, and offers insights into the resulting segmented data quality, which in turn impacts research outcomes. The paper examines three evaluation features: intersection area, text similarity based on Levenshtein distance, and text presence/absence in non-intersecting parts of the predicted region and its ground truth, revealing their effectiveness through logistic regression models. The method involves manual ground-truth creation, aiming for an automatic metric to quantify textual congruence. Results show that combining the text presence/absence feature with Hausdorff distance achieves the highest performance, reaching an F1 score of 0.957 on the testing subset. The study emphasizes the need for tailored evaluation metrics in Digital Humanities according to the specific needs and goals. The proposed evaluation framework offers insights for segmentation assessment in historical newspapers, with further application beyond the specific dataset and use case.

Keywords

Digitized Newspapers, Historical Advertisements, Layout Analysis Evaluation, Page Segmentation Evaluation

1. Introduction

Page segmentation is the process by which regions in a page are identified. For a page containing predominantly text, it would be blocks, paragraphs, lines and words. This step precedes text recognition, and hence, the quality of page segmentation directly impacts the quality of the text recognized by Optical Character Recognition (OCR) algorithms (Yanikoglu and Vincent, 1998; Can and Kabadayi, 2020; Ma et al., 2020; Martínek et al., 2020; Barman et al., 2021; Liebl and Burghardt, 2021). It is important to measure the quality of the segmentation; however, limited attention has been paid to the methodology for automatic comparison and evaluation of different page segmentation algorithms (Zhang and Gerbrands, 1994; Mao and Kanungo, 2001; Jiang et al., 2006).

The JobAds Project (FWF P35783) aims to explore the historical development of the labor market by extracting and analyzing job advertisements from digitized newspapers from 1850-1950. Accurate segmentation of job advertisements from historical newspapers presents unique difficulties due to their irregular layout, as newspapers use space very efficiently and not necessarily in neat rectangular blocks or recognizable border margins (see Fig. 1). Challenges further arise from the presence of rotated content, a combination of diverse fonts, occasional incorporation of decorative elements, and varying scan quality. Consequently, achieving accurate segmentation becomes an essential yet demanding endeavor.

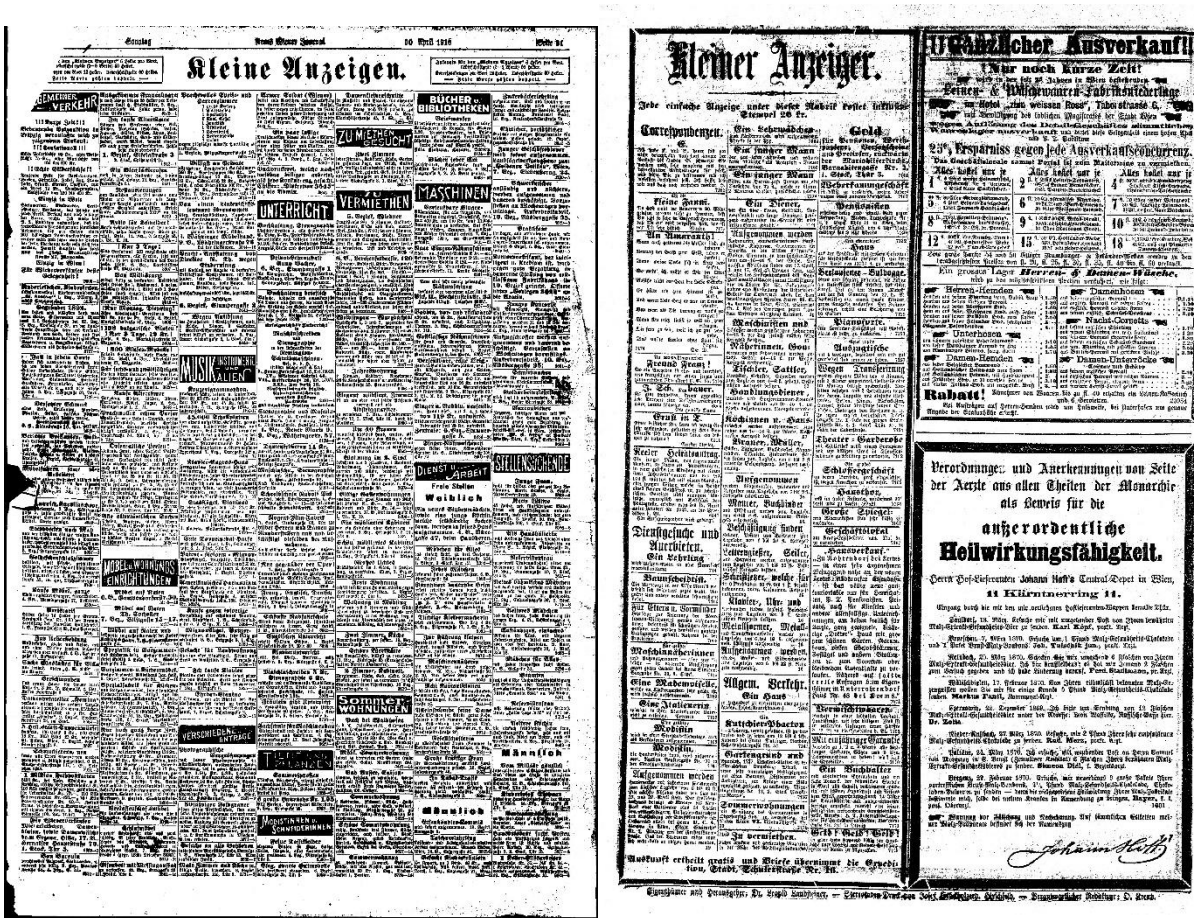


***Fig. 1**: Examples of pages from ANNO Corpus containing job advertisements. The layout of advertisements can be very tight and space-saving (left image) or not always visually separated by lines (right image). Left image: Neues Wiener Journal, 30.4.1916, p. 31, https://anno.onb.ac.at/cgi-content/anno?aid=nwj&datum=19160430&seite=31 [7.11.2023], right image: Morgen-Post, 3.4.1870, p. 8, https://anno.onb.ac.at/cgi-content/anno?aid=mop&datum=18700403&seite=8 [7.11.2023].*

Several tools explicitly tailored for newspaper or historical document segmentation have been developed, which are employed independently or as part of an OCR process, e.g. (Reul et al., 2017;

Oliveira et al., 2018; Rezanezhad et al., 2023). However, comparing their performance automatically has proven to be a non-trivial task. Yet, such a comparison and evaluation are essential not only to make an evidence-based choice of the most suitable tool but also to quantify the final quality of the segmentation and understand how it affects our ability to answer research questions.

In the project, we are primarily interested in correctly segmented text regions that yield valid text recognition, with only marginal emphasis on identifying decorations, frames, images, types of other segments or reading order of advertisements, treating them as independent units. Hence, in our analysis of page segmentation, the only criterion is that the text in the predicted region matches exactly with the text in the annotated ground-truth region. It is of no consequence for us if one region contains the border or frame or additional whitespace as long as the texts are identical. Consequently, many approaches commonly used in Digital Humanities or Computer Vision, as described in section 2, do not match our needs because of their complex or general focus, which is not tailored for comparing textual content of an image.

In this paper, we are therefore proposing and evaluating three easy-to-implement yet effective features relevant to segmentation comparison, as well as their combinations: the intersection area, text similarity based on the Levenshtein distance (Levenshtein, 1965), and the presence or absence of text in non-intersecting parts of the predicted region and its ground truth. We also address the challenges posed by OCR errors and incorrect text presence identification in this paper. The aim of the proposed method is to obtain a metric that can automatically quantify whether the textual content of the predicted region and its ground truth are identical or not.

To develop the evaluation framework, we sampled 2900 job advertisements across newspaper titles and years. These job advertisements were manually annotated in the course of the project. We obtained predicted coordinates and manually labeled the predictions as correctly or incorrectly segmented based on the textual content. The details of this process and information about the dataset are described in section 3 of this paper. Section 4 describes how we calculated the intersection area, the Levenshtein distance and derived similarity score, and identified the presence/absence of text in the non-intersecting parts of corresponding regions. We then compared the effectiveness of these features through training several logistic regression models and their evaluation on a testing dataset in sections 5 and 6. Section 7 concludes the paper, while section 8 highlights the open questions.

2. Related Work

Several relevant metrics and evaluation strategies have been proposed in the context of evaluating page segmentation. Randriamasy and Vincent (1994) proposed a region and bitmap-based approach, which shall be independent of other steps of the OCR process. In contrast, Kanai et al. (1995) considered OCR results and the cost of human editing operations to assess the accuracy of a page decomposition into text zones. Yanikoglu and Vincent (1998) developed an environment for benchmarking page segmentation, distinguishing between the type and severity of errors in page segmentation. Liang et al. (1998) evaluated layout analysis using a performance evaluation protocol based on the area of zones overlap.

Antonacopoulos and Bridson (2007) propose a goal-oriented evaluation framework delivering information at the dataset, page and region levels and offer a comprehensive overview of approaches for evaluating layout analysis/page segmentation. They discuss the evaluation methods that involve measuring the distance between OCR results of the ground truth and predicted region, highlighting their limitations due to OCR errors. Examining geometric comparison of regions is considered efficient with limitations of regions of complex shapes. The pixel-based comparison of regions, as used, e.g., by Shafait et al. (2006), is considered very accurate but less efficient than the

geometric comparison. Liebl and Burghardt (2021) employ the Matthews Correlation Coefficient to assess pixel segmentation, offering a single value based on the number of true positives, true negatives, false positives and false negatives.

The importance of layout segmentation evaluation and its evolution are also documented by the ICDAR competitions, e.g. ("ICDAR 2023," 2023). The First International Newspaper Segmentation contest, which aimed to compare different algorithms for newspaper segmentation, evaluated their performance based on matches between annotated and predicted entities, such as text, title or photo (B. Gatos et al., 2001). This pixel-based approach was used in subsequent years of the competition (Antonacopoulos et al., 2007, 2005), evaluating matches based on the intersection between the ground truth and the prediction. Since (Antonacopoulos et al., 2009), the competition used an interval representation and considered the type and severity of errors (Antonacopoulos et al., 2011, 2013, 2015; Clausner et al., 2017, 2019). More recent evaluations (Auer et al., 2023) used metrics like Mean Average Precision, incorporating the Intersection over Union, as used in the COCO object detection competition (Lin et al., 2014).

A segmentation evaluation has also been addressed in the context of computer vision, with a great number of different approaches and metrics proposed (Zhang and Gerbrands, 1994; Zhang, 1996, 2001; Jiang et al., 2005, 2006; Feng Ge et al., 2007; Zhang et al., 2008; Wang et al., 2020). The segmentation algorithms can be evaluated analytically or empirically (Zhang, 1996), employing supervised or unsupervised methods across various domains (Wang et al., 2020). However, they yielded insufficient results with our data as they do not reflect our need to compare the textual content only. For instance, considering the intersection area of bounding boxes, as presented by Phillips and Chhabra (1999), might fit our use case. However, defining thresholds for classification into one of the two classes without further tests and considerations might systematically fail in evaluating segmentation of certain advertisement types, such as those outside the very structured "Kleine Anzeige" section (see Fig. 1), as they often contain more decorations or illustrations.

While segmentation evaluation has a rich history, especially in computer vision, tailoring the metrics to the specific challenges posed by historical newspaper layout analysis requires careful consideration and adaptation. The selection of segmentation evaluation indicators "depends on the particular application" (Wang et al., 2020), which in our context can be summarized as a need for a metric quantifying the congruence of textual content of two regions.

3. Dataset

The dataset used in our project comprises 29 titles from the ANNO corpus, a collection of digitized newspapers of the Austrian National Library (Österreichische Nationalbibliothek, 2021). We are interested in issues from the defined span of 1850-1950. The newspapers from our dataset are in German but also contain job advertisements in other languages (e.g., French, English, Hungarian, Italian, Czech). We preselected over 4 million pages potentially containing job advertisements based on a manual check of the regularity of the appearance of the advertisement sections. In order to obtain ground-truth data, we randomly sampled one page per year for every newspaper title available for that year, resulting in 3 300 sampled pages. We manually annotated all job advertisements on these pages using doccano software (Nakayama et al., 2018), obtaining 14 985 annotated job ads. Not all of the preselected pages contained advertisements. During annotation, we drew single bounding boxes that included all texts from one job ad. We did not consider the precision of inclusion or exclusion of eventual separator lines between columns or surrounding ads necessary for our task. The manual annotation included basic classification as indicators (like headings), job offers, job searches, collective ads, and service offers. We did not distinguish between these classes in the evaluation.

For the task of segmentation evaluation, we randomly sampled 50 pages with annotated ads per newspaper title, or the maximum possible if fewer pages with ads were available, and from each of these sampled pages, we randomly sampled one job advertisement, resulting in 1024 sampled advertisements. We obtained its corresponding region prediction from the ANNO Corpus for every sampled advertisement. The corresponding region from the ANNO Corpus annotation was defined as the region from the page with the highest intersection with our annotated region. These ads were manually categorized as correctly (386 ad pairs) or incorrectly (495 ad pairs) segmented, forming our training and testing data.

4. Proposed Evaluation Features

Aiming to find a metric to quantify whether the textual content of a predicted region and its ground truth are identical, we will examine the following features and their combinations:

1. Intersection area of the annotated and predicted region,
2. text similarity based on Levenshtein distance of OCRed text within the two regions and
3. presence or absence of text in non-intersecting parts of the two regions.

These features were chosen based on human considerations of the definition of correct segmentation, each having specific advantages and disadvantages. While calculating the intersection does not face technical obstacles, it does not focus on the text but on the region area regardless of its content. The text similarity and text identification provide key information regarding the textual content but face technical issues regarding OCR errors or text presence misrecognition. It is, therefore, necessary to find a balanced solution between the conceptual and technical benefits and disadvantages. For clarity, Table 1 offers a short overview of the features used.

Table 1: Features overview

Feature	Implemented through	Short description
Intersection area	Relative Intersection	The intersection area of the annotated and predicted region divided by the total area of the larger region.
	Intersection over Union	The intersection area of the annotated and predicted region divided by their union.
	Hausdorff distance	Maximum distance from any point of the annotated region to its nearest point of the predicted region.
Levenshtein distance	Texts similarity	Text similarity based on Levenshtein distance.
Text Presence in non-intersecting parts	Text Presence	Binary value representing whether the non-intersecting parts of the predicted and annotated region contain text or not.

***Table 1**: Overview of features used in the segmentation evaluation with their short description.*

4.1 Intersection Area

Calculating the intersection area of annotated and predicted regions is an intuitive and efficient feature commonly used in segmentation evaluation, which can be measured in several ways. We use relative intersection, defined as the absolute intersection divided by the maximum of the areas of the annotated and predicted regions. This relative measure accounts for differences in region sizes, penalizing cases when, e.g., the whole page is predicted as a single region and thus fully contains the annotated region. We also use the Intersection over Union (IoU) metric (or also *Jaccard coefficient*, (Jaccard, 1901)), well-established in image segmentation evaluation, which returns the ratio between the intersection and union area of two regions (Jobin and Jawahar, 2017; Simistira et al., 2017; Can and Kabadayi, 2020), and Hausdorff distance (used implementation: (Damian Eads, 2007)) which indicates the maximum distance from any point of the ground-truth region to its nearest point in the predicted region (Huttenlocher et al., 1993; Beauchemin et al., 1998).

Using an intersection area for segmentation evaluation faces a conceptual disadvantage. A segmentation algorithm might produce tight results, leaving as little white space as possible or be greedy, producing as large a region as possible that is considered a single unit. Maintaining the focus on textual content, we consider segmentation correct regardless of frames or the area of white space included (see Fig. 2). The segmentation is correct as long as the textual content is identical, which may, in some cases, lead to lower relative intersection than e.g., when a heading is missing (Fig. 3). Consequently, this approach may classify some segmented regions as false negatives if the decision threshold is set too high, potentially systematically affecting ads outside structured sections (the 'Kleine Anzeigen' section shown in Fig. 1), but also false positives if the decision threshold is set too low (Fig. 3).

***Fig. 2**: Example of a correct segmentation with a low intersection value. While the textual content is identical, the relative intersection is only 50.94%. The predicted region (right, prediction by eynollah software (Rezanezhad et al., 2023) is fully contained by the annotated region (left). Image: Grazer Tagblatt, 9.8.1902, p. 12, https://anno.onb.ac.at/cgi-content/anno?aid=gtb&datum=19020809&seite=12 [21.11.2023].*

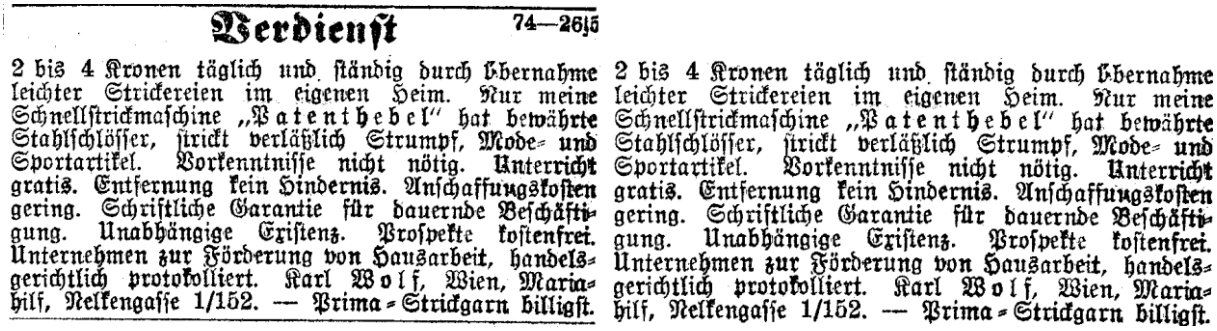


***Fig. 3**: Example of an incorrect segmentation with a relative intersection of 78.19%. The predicted region (right, prediction by eynollah software) is fully contained by the annotated region (left) but is missing the heading and advertisement identification number. Image: Innsbrucker Nachrichten, 15.6.1912, p. 25, https://anno.onb.ac.at/cgi-content/anno?aid=ibn&datum=19120615&seite=25 [23.8.2023].*

4.2 Levenshtein Distance

Levenshtein distance (Levenshtein, 1965) measures the difference between two string sequences, indicating the minimum number of editing operations required to obtain identical sequences. We used the implementation from the Levenshtein Python library (Bachmann, 2023). From the Leveshtein distance, we can derive the strings' similarity score as

Texts Similarity = 1 - (Levenshtein distance/max (length(text1), length(text2))),

see e.g. (Zhang et al., 2017). Levenshtein distance and derived similarity score are illustrated in the example of text retrieved from the two images in Fig. 3:

Text 1:

- Vendée-:| 744655

2 bis 4 Kronen täglieb und, ftändig ßurch übernahme (Renter Strickereien lm Werren Herm. Nur meine Söhnellftrickmafénne „Vatentbebel“ hat, bewährte Stahlfchlöffeß kuckt verxäßlrQStrumpf- Mode- und" Sportartikel. orkenntmfje mehr nöttg. Unterricht gratis. Entfernnng kem Hrnderms. Anfchaffnugskofien gering. Schriftlrche G-aranrre für dauernde Befchäftit-!gung. Unabhängige Exrftenz. Vrofpekte' koftenfrei.] Unternehmen zur Friede-rung von Hausarbezt- handels- 'g-err'chtlich vrotobollrert. Karl jWolf- .Wten- Marm- hilf- Nelkengaffe 1/152. *- Vrtma-Smckgarn billigft.

Text 2:

2 bis 4 Kronen täglick) und, ftändig ßurch übernahme (Renter Strickereien lm Werren Herm. Nur meine Söhnellftrickmafénne „Barentbebel“ hat, bewährte Stahlfchlöffeß kuckt verxaßUQStrumpß Mode- und" Sportartikel. orkenntmfje mehr nöttg. Unterricht gratis. Entfernnng kem Hrnderms. Anfchaffnugskofien gering. Schriftlrche G-aranrre für dauernde Befchäftit-gung. Unabhängige Exrftenz. Vrofpekte' koftenfrei. Unternehmen zur Friede-rung von Hausarbezt- handels-genehtlich vrotobollrert, Karl jWolf, “Wren- Marra- hilf- Nelkengaffe 1/152. *- Vrtma-Smckgarn billigft.

Levenshtein Distance = 45

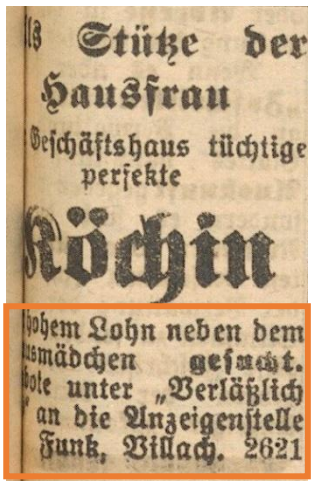
Texts Similarity = 1 - (45/max(591, 564)) = 0.924

This approach considers only text, in contrast to calculating regions overlap. However, it is sensitive to OCR errors and may produce biased results when the text is not recognized correctly (Antonacopoulos and Bridson, 2007), as the OCR errors can be inconsistent. For the OCR, we used the frak2021 (Mannheim University Library, 2021) model, which yielded the best OCR results on our data.

4.3 Text Presence/Absence in the Non-intersecting Parts

The absence of text in non-intersecting parts of the ground truth and its corresponding predicted region defines correct segmentation. If a non-intersecting part of an annotated region contains text, then not all parts of that region were identified in the predicted region, and the segmentation is incorrect. Similarly, if a non-intersecting part of a predicted region contains text, the predicted region includes content from another region, and the segmentation is also incorrect. This feature will thus penalize the algorithm, which includes the area with text from other regions or is, on the contrary, missing text. For illustration, see Fig. 4 a) and b).

a)



b)

***Fig. 4 a) and b)**: Illustration of the Text Presence feature. Images pair a): A part of the predicted region that does not lie in the intersection contains a text (orange frame), and the segmentation is incorrect. Image: Freie Stimmen, 24.8.1920, p.5, https://anno.onb.ac.at/cgi-content/anno?aid=fst&datum=19200824&seite=5&zoom=33 [23.8.2023]. Images pair b): None of the non-intersecting parts of the two regions contains text (illustrated by orange frames). For our use case, we consider the segmentation correct, even when the relative intersection is lower. Image: Grazer Tagblatt, 9.8.1902, p. 12, https://anno.onb.ac.at/cgi-content/anno?aid=gtb&datum=19020809&seite=12 [21.11.2023].*

Despite its apparent simplicity, text presence identification faces challenges. On the one hand, lower scan quality or usage of diverse fonts pose problems for the OCR model, whose responsibility is to identify the presence/absence of the text. On the other hand, some cases are ambiguous, even from the human perspective, especially in edge cases, as in the examples in Fig. 5 and Fig. 6.

Fig. 5

Fig. 6

***Fig. 5 and 6**: Examples of edge cases, annotated as 'containing text'. Image in Fig. 6: Arbeiter Zeitung, 8.8.1920, p. 14, https://anno.onb.ac.at/cgi-content/anno?aid=aze&datum=19200808&seite=14 [23.8.2023]. Image in Fig. 7: Grazer Volksblatt, 21.8.1908, p. 9, https://anno.onb.ac.at/cgi-content/anno?aid=gre&datum=19080821&seite=9 [23.8.2023].*

For the text presence identification, we have chosen a variety of pre-trained OCR models and compared their performance. We considered only models tailored to the language and/or font of our data, leaving out, e.g., text identification in scenery tools (Naosekpam and Sahu, 2022; Zhou et al., 2017). Tesseract is considered, together with Transkribus, the best-performing OCR tool (Martínek et al., 2020). We compared the ability of text identification of the 'deu' (tesseract-ocr, 2018a), 'frk' (tesseract-ocr, 2018b) and 'deu_frak' (tesseract-ocr, 2018c) models. Additionally, we included the GT4HistOCR (Springmann et al., 2018) and the frak2021 (Mannheim University Library, 2021) models which perform the best on our data. To compare the ability of these models to identify text presence, we randomly sampled 100 non-intersecting parts of corresponding regions (either upper, lower, left or right edge) per each newspaper title in our dataset, resulting in 2900 segments in total, and manually classified them as containing text (570 segments) or not containing text (2330 segments), and let all the above-mentioned models identify, whether they contain text or not. For

the comparison of models' performance, see Table 2. Note that we only considered alphanumeric characters, as dot decorations tended to be identified as interpunction.

Table 2: Comparison of models' ability to identify text

Model	True Positives	True Negatives	False Positives	False Negatives
Frk	482	2302	28	88
Deu	490	2301	29	80
Deu_frak	491	2298	32	79
GT4HistOCR	474	2312	18	96
Frak2021	483	2303	27	87

Table 2*: Performance of different pre-trained models to identify the presence of text in non-intersecting parts of 2900 regions. True positives stand for regions containing text that were correctly predicted as containing text; true negatives stand for regions not containing text that were correctly predicted as non-containing text; false positives stand for regions not containing text that were incorrectly predicted as containing text, and false negatives stand for regions containing text that were incorrectly predicted as not-containing text.*

Based on these results, we chose the 'GT4HistOCR' model because it has the lowest number of false positives. Risking not penalizing a model for saying that an incorrect segmentation was correct (not identifying text when there was one) is a minor problem compared to not penalizing a model for stating that the correct segmentation is incorrect (identifying text when there was none). Also, the edge cases (as above in Fig. 5 and 6) tended to be labeled rather as 'containing text'. However, some cases might be considered ambivalent, making the false negatives less severe error. The result of text identification is saved as a binary True/False value.

5. Evaluation Methodology

In this section, we present the methodology employed to evaluate the contribution of individual features (or *predictors*) and their combinations to the model's prediction of whether a segment was correctly or incorrectly segmented. The features, encompassing intersection area, text similarity based on Levenshtein distance, and text presence in non-intersecting parts, are obtained for every pair of regions (predicted region and its ground truth) specified in the Dataset section.

To evaluate segmentation quality, we construct a classifier to distinguish between correctly and incorrectly segmented regions based on the alignment of textual content in the ground truth and predicted region. We employ logistic regression as the classification algorithm, as it is well suited for binary classification tasks (Cabrera, 1994; Peng et al., 2002). The interpretability of logistic regression allows insights into individual variables' contribution to the classification decision. The model returns probability, ranging from 0 to 1, indicating whether a region was correctly or incorrectly segmented. The probability higher than a set threshold, typically 0.5, is turned into 1; otherwise, it becomes 0.

We use the statsmodels.api.Logit[1] implementation (Seabold and Perktold, 2010), as it offers a range of statistical descriptions of the trained model. Default settings and a decision threshold of 0.5 are applied. Additionally, a constant is included to add an intercept, which is not included by default.

We fit and evaluate a separate model for every feature and their combinations, maintaining the same random states, which ensures that any difference in results is accountable to the features used. As the classes in our dataset are not perfectly balanced, we undersample the more populated class to ensure that the model is not biased to predict one of the two classes more often. Afterwards, we split our labeled data into training and testing subsets, with 70% of the data allocated for training and the remaining 30% for testing. The split is stratified, maintaining a similar positivity rate in the

[1] https://tedboy.github.io/statsmodels_doc/generated/generated/statsmodels.api.Logit.html [4.1.2024]

training and testing subsets. The assignment to the training/testing subset is random and ensures that the two subsets are mutually exclusive.

As logistic regression is used as a classifier and the stated goal is the classification task, it is appropriate to evaluate its performance by the metrics assessing the predictions (Hosmer Jr et al., 2013). Therefore, we use metrics commonly used in machine learning. Namely, the model's performance is measured by accuracy, expressing the ratio of correct answers to all answers, and the F1 score, a harmonic mean of precision and recall (Powers, 2011).

Additionally, we can also assess the ability of the models to distinguish between the positive and negative classes across different threshold values using the Receiver Operating Characteristic Area Under the Curve (ROC AUC) (Peterson et al., 1954; Zou et al., 2007), without depending on a specific decision threshold. The evaluation is only conducted on the testing subset, i.e., the data that the model did not see during the training.

In addition to testing the predictive power of the model, we can also explore the model fit and the relationship between predictors and the classifier's decision. For this, we report the LLR p-value, indicating whether the parameters significantly improve the model fit compared to the null model. We set the significance level at 0.05 as a common practice. LLR p-value lower than this significance threshold indicates that the model fits the data significantly better than the null model and vice versa. We also report the P > |z| value for every model parameter, indicating whether the given parameter significantly influences the classifier's decision.

6. Results

Table 3 presents results for all trained Logistic Regression models on the testing subset, as described in the Evaluation Methodology section above. From the classification point of view, the model achieves the best results using Hausdorff distance and Text Presence as predictors, with an F1 score of 0.957. Combining Intersection and Text Presence reaches almost identical results. As for alone-standing features, the model trained on the Text Presence feature reaches the F1 score of 0.952 on the testing subset, indicating that this feature offers more information than the other stand-alone features. The Levenshtein distance proved to be the most problematic, reaching the F1 score of 0.848, due to OCR errors and resulting recognition inconsistencies. We cannot generalize that this feature is useless per se, but rather that it is highly dependent on the quality of the recognized text.

In some cases, no text was recognized at all (e.g., because of very low scan quality), making it impossible to evaluate that segment using Levenshtein distance. The Intersection and Text Presence features would still yield valid predictions even in such cases. The scores reached by different representations of intersection areas are similar. The lowest F1 score of 0.917 is for Hausdorff distance, and the highest F1 score of 0.93 for IoU.

Table 3: Results for Logistic Regression

Features	Accuracy	F1	ROC AUC	LLR p-value	P >|z|
Intersection	0.92	0.922	0.92	2.792×10^{-117}	[0]
Intersection - IoU	0.929	0.93	0.929	1.517×10^{-118}	[0]
Hausdorff distance	0.912	0.917	0.912	1.008×10^{-117}	[0]
Levenshtein	0.844	0.848	0.845	8.422×10^{-77}	[0]
Text Presence	0.951	0.952	0.951	1.530×10^{-125}	[0]
Intersection + Levenshtein	0.924	0.923	0.924	1.982×10^{-119}	[0; 0.001]
IoU + Levenshtein	0.929	0.928	0.929	6.714×10^{-120}	[0; 0.002]
Hausdorff + Levenshtein	0.924	0.925	0.925	3.923×10^{-129}	[0; 0]

Intersection + Text Presence	0.956	0.956	0.956	2.558×10^{-140}	[0; 0]
IoU + Text Presence	0.947	0.947	0.947	5.626×10^{-141}	[0; 0]
Hausdorff + Text presence	**0.956**	**0.957**	**0.956**	$\mathbf{1.603 \times 10^{-137}}$	**[0; 0]**
Levenshtein + Text Presence	0.929	0.929	0.929	4.632×10^{-127}	[0.001; 0]
Intersection + Levenshtein + Text Presence	0.947	0.946	0.947	1.013×10^{-139}	[0; 0.499; 0]
IoU + Levenshtein + Text Presence	0.942	0.942	0.942	3.799×10^{-140}	[0; 0.813; 0]
Hausdorff + Levenshtein + Text Presence	0.929	0.928	0.929	1.495×10^{-140}	[0; 0.017; 0]

Table 3: *Accuracy, F1 score, ROC AUC score, LLR p-value and P > |z| of individual parameters of Logistic Regression models trained on different features and their combinations for the testing subset. The results are rounded to three decimal places. In the case of multiple parameters, the results in the P > |z| column are reported in the respective order as in the first column ('Features').*

Inspecting LLR p-values, we conclude that all the trained models fit the data significantly better than the null model, as all of them are smaller than the threshold value of 0.05. As for the significance of individual features, all of them significantly influence the classification decision, except for Levenshtein distance in the case of the Intersection + Levenshtein + Text Presence model and the IoU + Levenshtein + Text Presence model. The latter two have a p-value higher than 0.05, and the result is thus statistically insignificant.

7. Conclusion

In this paper, we addressed the metrics for evaluating page segmentation when extracting job advertisements from historical newspapers, which is an important step for an evidence-based choice of the most suitable tool, as well as for a quantitative evaluation of the resulting data quality. Our definition of correct segmentation was that the textual content of a predicted region and its ground truth are identical. Thus, the goal was to find a way to measure whether the text within two images is identical, regardless of frames, blank space, or reading order.

We evaluated the efficiency of several features for this evaluation task, namely the area of intersection, the similarity derived from Levenshtein distance, and the presence/absence of text in non-intersecting parts of corresponding regions. These features and their combinations were assessed by fitting a logistic regression model and its evaluation on the testing dataset. The best results were achieved by combining the presence/absence of text in non-intersecting parts with Hausdorff distance as predictors, reaching the F1 score of 0.957 on the testing subset. Using intersection is inefficient in some cases due to its dependency on including elements other than the text, e.g., frames or blank space. The text presence in non-intersecting parts is conceptually sufficient as a stand-alone feature but faces technical obstacles in terms of correct text presence identification. The Levenshtein distance is even more dependent on the high quality of the OCR, otherwise proves to be less efficient because of the inconsistencies in the recognition.

While our study's immediate relevance lies in the specific context of the JobAds Project, its broader implications extend beyond our dataset and use case. Firstly, the presence/absence of text in the non-intersecting parts of the corresponding regions could be a helpful feature in evaluating any text page segmentation algorithm where researchers are only interested in the textual content, and we have not found this concept in the literature. Secondly, we show the usefulness and efficiency of evaluation features tailored to address specific needs in contrast to generic approaches. Given the unique challenges posed by historical newspaper layouts and various use case-dependent needs in the field of Digital Humanities, it can be more suitable to establish hand-crafted features and adjust the evaluation framework with the help of the domain expert insight. Thirdly, transparency and methodological rigor remain at the core of all scientific endeavors. By examining the performance of various segmentation features and documenting their strengths and limitations, our work serves as a

reminder that the choice of evaluation criteria should align with the specific goals and challenges of the research.

With the accuracy achieved by the combined Hausdorff + Text presence method, we feel confident that we can find a suitable method for unsupervised image segmentation in our project, resulting in a reliable selection of job ads in the entire corpus.

8. Open Questions

Further research work shall concentrate on developing an automatic approach for segmentation evaluation that would not require the manual creation of ground-truth data, as this task is time-consuming. In addition, we were only dealing with a model predicting binary answer correct/incorrect segmentation because the language of advertisements is very concise and missing even one word can lead to losing information about, e.g., the name of the position. However, for some use cases, it might be useful to measure the extent of correctness in, e.g., percentage, according to the amount of text that was incorrectly segmented.

Acknowledgements

The work presented here was supported by the FWF project FWF P35783, PI Jörn Kleinert, doi:10.55776/P35783. Ground truth was created in cooperation with our project colleagues Wiltrud Mölzer and Jörn Kleinert. We would like to thank the ÖNB for providing the image, in particular Christoph Steindl for his efficient support.

Code and Data Availability

The code is available at https://github.com/JobAds-FWFProject/SegmentationEvaluation. Data were obtained from the ANNO corpus available at https://anno.onb.ac.at/ [accessed 13.10.2023]. To obtain data from the ANNO corpus on a larger scale, please refer to the Austrian National Library, or see data under a Public Domain Mark at https://labs.onb.ac.at/en/datasets/anno/.

Bibliography

Antonacopoulos, A., Bridson, D., 2007. Performance Analysis Framework for Layout Analysis Methods. https://doi.org/10.1109/ICDAR.2007.4377117

Antonacopoulos, A., Clausner, C., Papadopoulos, C., Pletschacher, S., 2015. ICDAR2015 competition on recognition of documents with complex layouts - RDCL2015, in: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). pp. 1151–1155. https://doi.org/10.1109/ICDAR.2015.7333941

Antonacopoulos, A., Clausner, C., Papadopoulos, C., Pletschacher, S., 2013. ICDAR 2013 Competition on Historical Newspaper Layout Analysis (HNLA 2013), in: 2013 12th International Conference on Document Analysis and Recognition. pp. 1454–1458. https://doi.org/10.1109/ICDAR.2013.293

Antonacopoulos, A., Clausner, C., Papadopoulos, C., Pletschacher, S., 2011. Historical Document Layout Analysis Competition, in: Proceedings of the 2011 International Conference on Document Analysis and Recognition, ICDAR '11. IEEE Computer Society, USA, pp. 1516–1520. https://doi.org/10.1109/ICDAR.2011.301

Antonacopoulos, A., Gatos, B., Bridson, D., 2007. Page Segmentation Competition, in: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007). pp. 1279–1283. https://doi.org/10.1109/ICDAR.2007.4377121

Antonacopoulos, A., Gatos, B., Bridson, D., 2005. ICDAR2005 page segmentation competition, in: Eighth International Conference on Document Analysis and Recognition (ICDAR'05). pp. 75-79 Vol. 1. https://doi.org/10.1109/ICDAR.2005.184

Antonacopoulos, A., Pletschacher, S., Bridson, D., Papadopoulos, C., 2009. ICDAR 2009 Page Segmentation Competition, in: 2009 10th International Conference on Document Analysis and Recognition. pp. 1370–1374. https://doi.org/10.1109/ICDAR.2009.275

Auer, C., Nassar, A., Lysak, M., Dolfi, M., Livathinos, N., Staar, P., 2023. ICDAR 2023 Competition on Robust Layout Segmentation in Corporate Documents, in: Fink, G.A., Jain, R., Kise, K., Zanibbi, R. (Eds.), Document Analysis and Recognition - ICDAR 2023. Springer Nature Switzerland, Cham, pp. 471–482.

B. Gatos, S. L. Mantzaris, A. Antonacopoulos, 2001. First International Newspaper Segmentation contest, in: Proceedings of Sixth International Conference on Document Analysis and Recognition. Presented at the Proceedings of Sixth International Conference on Document Analysis and Recognition, pp. 1190–1194. https://doi.org/10.1109/ICDAR.2001.953973

Bachmann, M., 2023. python-Levenshtein: Python extension for computing string edit distances and similarities.

Barman, R., Ehrmann, M., Clematide, S., Oliveira, S., Kaplan, F., 2021. Combining Visual and Textual Features for Semantic Segmentation of Historical Newspapers. J. Data Min. Digit. Humanit. HistoInformatics. https://doi.org/10.46298/jdmdh.6107

Beauchemin, M., Thomson, K.P.B., Edwards, G., 1998. On the Hausdorff Distance Used for the Evaluation of Segmentation Results. Can. J. Remote Sens. 24, 3–8. https://doi.org/10.1080/07038992.1998.10874685

Cabrera, A., 1994. "Logistic Regression Analysis in Higher Education: An Applied Perspective, in: Higher Education: Handbook of Theory and Research(225-256). pp. 225–256.

Can, Y., Kabadayi, M., 2020. CNN-Based Page Segmentation and Object Classification for Counting Population in Ottoman Archival Documentation. J. Imaging 6, 32. https://doi.org/10.3390/jimaging6050032

Clausner, C., Antonacopoulos, A., Pletschacher, S., 2019. ICDAR2019 Competition on Recognition of Documents with Complex Layouts - RDCL2019, in: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1521–1526. https://doi.org/10.1109/ICDAR.2019.00245

Clausner, C., Antonacopoulos, A., Pletschacher, S., 2017. ICDAR2017 Competition on Recognition of Documents with Complex Layouts - RDCL2017, in: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). pp. 1404–1410. https://doi.org/10.1109/ICDAR.2017.229

Damian Eads, 2007. scipy.spatial.distance.directed_hausdorff.

Feng Ge, Song Wang, Tiecheng Liu, 2007. New benchmark for image segmentation evaluation. J. Electron. Imaging 16, 033011. https://doi.org/10.1117/1.2762250

Hosmer Jr, D.W., Lemeshow, S., Sturdivant, R.X., 2013. Applied logistic regression. John Wiley & Sons.

Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J., 1993. Comparing images using the Hausdorff distance. IEEE Trans. Pattern Anal. Mach. Intell. 15, 850–863. https://doi.org/10.1109/34.232073

ICDAR 2023 [WWW Document], 2023. . ICDAR 2023 Compet. URL https://icdar2023.org/program/competitions/ (accessed 12.5.23).

Jaccard, P., 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull Soc Vaudoise Sci Nat 37, 547–579.

Jiang, X., Marti, C., Irniger, C., Bunke, H., 2006. Distance Measures for Image Segmentation Evaluation. EURASIP J. Adv. Signal Process. 2006, 035909. https://doi.org/10.1155/ASP/2006/35909

Jiang, X., Marti, C., Irniger, C., Bunke, H., 2005. Image Segmentation Evaluation by Techniques of Comparing Clusterings, in: Roli, F., Vitulano, S. (Eds.), Image Analysis and Processing – ICIAP 2005. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 344–351.

Jobin, K.V., Jawahar, C.V., 2017. Document Image Segmentation Using Deep Features, in: National Conference on Computer Vision, Pattern Recognition, Image Processing, and Graphics.

Kanai, J., Rice, S.V., Nartker, T., Nagy, G., 1995. Automated Evaluation of OCR Zoning. Pattern Anal. Mach. Intell. IEEE Trans. On 17, 86–90. https://doi.org/10.1109/34.368146

Levenshtein, V.I., 1965. Binary codes capable of correcting deletions, insertions, and reversals. Sov. Phys. Dokl. 10, 707–710.

Liang, J., Phillips, I., Haralick, R., 1998. Performance evaluation of document layout analysis algorithms on the UW data set. Proc. SPIE - Int. Soc. Opt. Eng. https://doi.org/10.1117/12.270067

Liebl, B., Burghardt, M., 2021. An Evaluation of DNN Architectures for Page Segmentation of Historical Newspapers, in: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 5153–5160. https://doi.org/10.1109/ICPR48806.2021.9412571

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: Common Objects in Context, in: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (Eds.), Computer Vision – ECCV 2014. Springer International Publishing, Cham, pp. 740–755.

Ma, W., Zhang, H., Jin, L., Wu, S., Wang, J., Wang, Y., 2020. Joint Layout Analysis, Character Detection and Recognition for Historical Document Digitization. https://doi.org/10.48550/arXiv.2007.06890

Mannheim University Library, 2021. frak2021.

Mao, S., Kanungo, T., 2001. Empirical performance evaluation methodology and its application to page segmentation algorithms. IEEE Trans. Pattern Anal. Mach. Intell. 23, 242–256. https://doi.org/10.1109/34.910877

Martínek, J., Lenc, L., Král, P., 2020. Building an efficient OCR system for historical documents with little training data. Neural Comput. Appl. 32, 17209–17227. https://doi.org/10.1007/s00521-020-04910-x

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., Liang, X., 2018. doccano: Text Annotation Tool for Human.

Naosekpam, V., Sahu, N., 2022. Text detection, recognition, and script identification in natural scene images: a Review. Int. J. Multimed. Inf. Retr. 11, 291–314. https://doi.org/10.1007/s13735-022-00243-8

Oliveira, S., Seguin, B., Kaplan, F., 2018. dhSegment: A generic deep-learning approach for document segmentation.

Österreichische Nationalbibliothek, 2021. ANNO Historische Zeitungen und Zeitschriften [WWW Document]. URL https://anno.onb.ac.at/

Peng, J., Lee, K., Ingersoll, G., 2002. An Introduction to Logistic Regression Analysis and Reporting. J. Educ. Res. - J EDUC RES 96, 3–14. https://doi.org/10.1080/00220670209598786

Peterson, W.W., Birdsall, T.G., Fox, W.C., 1954. The theory of signal detectability. Trans IRE Prof Group Inf Theory 4, 171–212.

Phillips, I.T., Chhabra, A.K., 1999. Empirical Performance Evaluation of Graphics Recognition Systems. IEEE Trans Pattern Anal Mach Intell 21, 849–870. https://doi.org/10.1109/34.790427

Powers, D.M.W., 2011. Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. J. Mach. Learn. Technol. 2, 37–63.

Randriamasy, S., Vincent, L., 1994. Benchmarking page segmentation algorithms. https://doi.org/10.1109/CVPR.1994.323859

Reul, C., Springmann, U., Puppe, F., 2017. LAREX - A semi-automatic open-source Tool for Layout Analysis and Region Extraction on Early Printed Books. CoRR abs/1701.07396.

Rezanezhad, V., Baierer, K., Gerber, M., Labusch, K., Neudecker, C., 2023. Document Layout Analysis with Deep Learning and Heuristics, in: Proceedings of the 7th International Workshop on Historical Document Imaging and Processing HIP 2023, San José, US, August 25-26, 2023, ACM.

Seabold, S., Perktold, J., 2010. statsmodels: Econometric and statistical modeling with python, in: 9th Python in Science Conference.

Shafait, F., Keysers, D., Breuel, T., 2006. Pixel-Accurate Representation and Evaluation of Page Segmentation in Document Images. https://doi.org/10.1109/ICPR.2006.934

Simistira, F., Bouillon, M., Seuret, M., Gygli, M., Alberti, M., Ingold, R., Liwicki, M., 2017. ICDAR2017 Competition on Layout Analysis for Challenging Medieval Manuscripts. https://doi.org/10.1109/ICDAR.2017.223

Springmann, U., Reul, C., Dipper, S., Baiter, J., 2018. Ground Truth for training OCR engines on historical documents in German Fraktur and Early Modern Latin.
tesseract-ocr, 2018a. deu.
tesseract-ocr, 2018b. frk.
tesseract-ocr, 2018c. deu_frak.
Wang, Z., Wang, E., Zhu, Y., 2020. Image segmentation evaluation: a survey of methods. Artif. Intell. Rev. 53, 5637–5674. https://doi.org/10.1007/s10462-020-09830-9
Yanikoglu, B.A., Vincent, L., 1998. Pink Panther: A Complete Environment for Ground-Truth and Benchmarking Doocument Page Segmentation. Pattern Recognit. 31, 1191–1204. https://doi.org/10.1016/S0031-3203(97)00137-4
Zhang, H., Fritts, J.E., Goldman, S.A., 2008. Image segmentation evaluation: A survey of unsupervised methods. Comput. Vis. Image Underst. 110, 260–280. https://doi.org/10.1016/j.cviu.2007.08.003
Zhang, S., Hu, Y., Bian, G., 2017. Research on string similarity algorithm based on Levenshtein Distance, in: 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). pp. 2247–2251. https://doi.org/10.1109/IAEAC.2017.8054419
Zhang, Y.J., 2001. A review of recent evaluation methods for image segmentation, in: Proceedings of the Sixth International Symposium on Signal Processing and Its Applications (Cat.No.01EX467). pp. 148–151 vol.1. https://doi.org/10.1109/ISSPA.2001.949797
Zhang, Y.J., 1996. A survey on evaluation methods for image segmentation. Pattern Recognit. 29, 1335–1346. https://doi.org/10.1016/0031-3203(95)00169-7
Zhang, Y.J., Gerbrands, J.J., 1994. Objective and quantitative segmentation evaluation and comparison. Signal Process. 39, 43–54. https://doi.org/10.1016/0165-1684(94)90122-8
Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., Liang, J., 2017. EAST: An Efficient and Accurate Scene Text Detector.
Zou, K.H., O'Malley, A.J., Mauri, L., 2007. Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models. Circulation 115, 654–657. https://doi.org/10.1161/CIRCULATIONAHA.105.594929

Image Sources

Fig. 1:
Neues Wiener Journal, 30.4.1916, p. 31, https://anno.onb.ac.at/cgi-content/anno?aid=nwj&datum=19160430&seite=31 [7.11.2023].
Morgen-Post, 3.4.1870, p. 8, https://anno.onb.ac.at/cgi-content/anno?aid=mop&datum=18700403&seite=8 [7.11.2023].

Fig. 2:
Grazer Tagblatt, 9.8.1902, p. 12, https://anno.onb.ac.at/cgi-content/anno?aid=gtb&datum=19020809&seite=12 [21.11.2023].

Fig. 3:
Innsbrucker Nachrichten, 15.6.1912, p. 25, https://anno.onb.ac.at/cgi-content/anno?aid=ibn&datum=19120615&seite=25 [23.8.2023].

Fig. 4:
Freie Stimmen, 24.8.1920, p.5, https://anno.onb.ac.at/cgi-content/anno?aid=fst&datum=19200824&seite=5&zoom=33 [23.8.2023].
Grazer Tagblatt, 9.8.1902, p. 12, https://anno.onb.ac.at/cgi-content/anno?aid=gtb&datum=19020809&seite=12 [21.11.2023].

Fig. 5:
Arbeiter Zeitung, 8.8.1920, p. 14, https://anno.onb.ac.at/cgi-content/anno?aid=aze&datum=19200808&seite=14 [23.8.2023].

Fig. 6:
Grazer Volksblatt, 21.8.1908, p. 9, https://anno.onb.ac.at/cgi-content/anno?aid=gre&datum=19080821&seite=9 [23.8.2023].