Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Automatic Text Recognition for Burgenland Croatian Printed Materials: Developing a Custom Model on Transkribus

Marija Đokić Petrović

marija.djokic795@gmail.com

Union University, Belgrade, Serbia https://orcid.org/0000-0002-0568-5219

Marko Simonović

University of Graz, Austria https://orcid.org/0000-0002-9651-6399

Research Article

Keywords: Burgenland Croatian, Automatic Text Recognition, Transkribus, Digital Humanities, Low-resource Language, Under-resourced Language

Posted Date: June 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9957704/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Automatic Text Recognition for Burgenland Croatian Printed Materials: Developing a Custom Model on Transkribus

Marija Đokić Petrović[1*] and Marko Simonović[2†]

[1*]Department of Computer Science, The School of Computing, Union University, Kneza Mihaila VI/6, Belgrade, 11000, Serbia.
[2]Department of Slavic Studies, The University of Graz, Merangasse 70/I, Graz, 8010, Styria, Austria.

*Corresponding author(s). E-mail(s): mdjokicpetrovic@raf.rs;
Contributing authors: marko.simonovic@uni-graz.at;
†These authors contributed equally to this work.

Abstract

This paper presents the development and evaluation of an Automatic Text Recognition (ATR) model for printed Burgenland Croatian, implemented using the Transkribus platform. The *Burgenland Croatian Typewritten 2010–2019* model was trained on issues of the newspaper *Hrvatske novine*, the main regular publication in the Burgenland Croatian community, and accessed via the AustriaN Newspapers Online (ANNO) digital newspaper repository of the Austrian National Library. Focusing on a period marked by intensified standardisation, the model is optimised for Burgenland Croatian printed texts published in post-2009. To our knowledge, it is the first ATR model tailored to Burgenland Croatian, a low-resource and under-resourced language variety. Validation results show a Character Error Rate (CER) of 2.48%, while evaluation yielded a CER below 2.25%, demonstrating the model's robustness and practical usability. Our model outperformed some multilingual models supporting Slavic languages, as well as one super model available on Transkribus, in 94% of the evaluated cases. This work contributes a reusable language resource to the Digital Humanities community while demonstrating that high-quality ATR models can be developed for low-resource and under-resourced languages. The model is publicly available via Transkribus and will be continuously updated to improve accuracy further.

Keywords: Burgenland Croatian, Automatic Text Recognition, Transkribus, Digital Humanities, Low-resource Language, Under-resourced Language

1 Introduction

The digitization of printed and handwritten documents has become an important concern in Digital Humanities (DH), as it enables sustainable access to cultural heritage collections, supports large-scale text analysis, and provides the foundation for a wide range of downstream tasks (Hossain, 2025). A key technology in the context of digitization is Automatic Text Recognition (ATR), an umbrella term encompassing Optical Character Recognition (OCR) for printed text and Handwritten Text Recognition (HTR) for handwritten materials (Keinan-Schoonbaert, 2025). By applying ATR to transform documents into machine-readable text, many historical or literary sources can be searched, analyzed, and integrated into digital corpora and research infrastructures (Nockels, 2025).

Despite significant advances in ATR for high-resource languages, minority, low-resource, and under-resourced language varieties continue to face substantial barriers (Aepli, 2024), primarily due to limited access to annotated data, standardized tools, and linguistic resources (Pakray, Gelbukh, & Bandyopadhyay, 2025). Addressing these limitations is increasingly recognized as essential for reducing digital inequalities in DH and for enabling the meaningful integration of such language heritage into computational research environments.

Within this context, one such language is Burgenland Croatian (Gradišćanski hrvatski; Glottolog: burg1244; IETF: ckm-AT)[1]. It is a regional variety of Croatian, primarily based on the Čakavian dialect, spoken in Austria, Hungary, Slovakia, and the Czech Republic (Ščukanec, 2014). The variety emerged from migrations of Croatian-speaking populations to the border regions of the Habsburg monarchy in the 16$^{\text{th}}$ century and has since developed independently from the standard Croatian norm, in contact with German and Hungarian (Ščukanec, 2014). From a language-technology perspective, Burgenland Croatian remains severely underrepresented, with a notable lack of dedicated research in DH (Cajka, Jouitteau, & Vetter, 2025). A major limiting factor is the scarcity of annotated training data, which significantly constrains the development of computational methods for this language variety. In this context, the weekly newspaper *Hrvatske novine* (Croatian Weekly Newspaper)[2], which reflects the public use of Burgenland Croatian, could serve as an ideal foundation for developing language resources and DH tools, including ATR models.

This paper presents the development and evaluation of an ATR model for printed Burgenland Croatian within the Transkribus platform. The model, named *Burgenland Croatian Typewritten 2010–2019*[3], is trained on selected issues of *Hrvatske novine* and targets post-2009 printed materials, a period marked by intensified standardisation and representative of the contemporary written norm of the Burgenland Croatian community.

The contributions of this paper are as follows:

- We develop *Burgenland Croatian Typewritten 2010–2019*, the first dedicated ATR model for Burgenland Croatian printed materials post-2009.

[1] https://glottolog.org/resource/languoid/id/burg1244
[2] https://hrvatskenovine.at/
[3] https://app.transkribus.org/models/text/442685

- We describe the creation of a training dataset from issues of *Hrvatske novine* used to train the model.
- We demonstrate the practical usability of the model through rigorous empirical evaluation.
- The trained model and it's training dataset are publicly available on the Transkribus platform at https://app.transkribus.org/models/text/442685.

Beyond its immediate contributions, this work establishes a foundation for future DH research on Burgenland Croatian and comparable language varieties. By making both the model and curated training dataset publicly available via Transkribus and by documenting the data preparation and training workflow in detail, the study supports reproducibility and provides a basis for further model refinement and adaptations. Additionally, the developed model could be applicable to related South Slavic language varieties and could serve as a baseline for training new ATR models within the Transkribus platform, while the training data can be reused for other Natural Language Processing (NLP) and DH tasks.

The rest of the paper is organized as follows: Section 2 provides an overview of relevant research on Burgenland Croatian linguistic approaches, language resources, and computational techniques; Section 3 describes the training data; Section 4 outlines the model training procedure within Transkribus; Section 5 reports, while Section 6 discusses the evaluation results; Section 7 concludes the paper and outlines future work.

2 Related Work

Burgenland Croatian is a South Slavic minority language variety spoken primarily in eastern Austria (Burgenland and Lower Austria), as well as in Hungary, Slovakia, and the Czech Republic (Ščukanec, 2014), with recent estimates placing the number of speakers at approximately 65,000 (Tyran, 2022). In Austria, Burgenland Croatian is recognized as one of six autochthonous national minority languages and enjoys formal protection under the Framework Convention for the Protection of National Minorities (FCNM 1995) [4], the European Charter for Regional or Minority Languages (Charter RML 1992)[5], and subsequent legislation[6]. Linguistically, Burgenland Croatian is primarily based on the Čakavian dialect, while also incorporating Kajkavian and Štokavian features[7]. Compared with Standard Croatian, it exhibits differences in phonology, morphology, lexicon, and idiomatic usage, reflecting centuries of separate development and language contacts with German and Hungarian. Nevertheless, mutual intelligibility with standard Croatian remains generally high.

Although Burgenland Croatian has a long history of written and public use with some level of spontaneous standardisation/koineisation, the publication of standard works explicitly codifying a standard Burgenland Croatian variety happened mostly in the 21$^{\text{th}}$ century. A central role in this process has been played by the

[4] https://www.bundeskanzleramt.gv.at/en/topics/ethnic-groups.htm
[5] https://rm.coe.int/austria-evall-5-iria-en/1680b516bb
[6] https://minorityrights.org/communities/croats-of-burgenland/
[7] https://grokipedia.com/page/Burgenland_Croatian

Znanstveni institut Gradišćanskih Hrvatov - ZIGH[8] (The Scholarly Institute of Burgenland Croats), founded in 1994 and headquartered in Trajštof (Trausdorf) today. Within ZIGH, the Language Commission has met regularly since 1995 to research Burgenland Croatian and to coin and standardise terminology, feeding its decisions into codificatory resources and lexicography, including the online dictionary *rjecnik.at*[9]. Major milestones of this codification include the *Gramatika gradišćanskohrvatskoga jezika* (Grammar of the Burgenland Croatian language) (2003) (Kinda-Berlakovich et al., 2003) and the *Pravopis gradišćanskohrvatskoga književnoga jezika* (Orthography of the Buregenland Croatian literary language) (2009) (Benčić, Csenar-Schuster, Rotter, Sučić, & Vlašić, 2009), alongside bilingual and multilingual dictionaries and other manuals of the standard variety. The post-2009 period reflects intensified standardisation and the contemporary written norm of the Burgenland Croatian community, as the principal codificatory resources became available for systematic use by authors and editors of printed materials from this time onward.

Linguistic research on Burgenland Croatian has focused primarily on contact phenomena, dialect description, and the sociolinguistic situation of the community. From a structural perspective, several studies address specific aspects of Burgenland Croatian grammar. For example, the author in (Browne, 2016) compared clitics in Burgenland Croatian in Austria and in Pannonian Rusyn in northern Serbia, and discussed the relationship between Burgenland Croatian and the Croatian language. Other works examine verb aspect and lexical borrowing under intense contact with German and Hungarian, documenting the emergence of mixed constructions and hybrid patterns at the morphosyntactic level (Berghaus, 2023; Breu, Berghaus, & Scholze, 2016). A detailed overview of the most important linguistic publications on the Burgenland Croatian dialects is provided by Houtzagers (2013), who surveys the leading publications on these dialects and emphasises that our picture of them is still far from complete. Earlier work on isoglosses and dialect classification likewise highlights both the internal diversity of Burgenland Croatian and its relevance for reconstructing pre-migration Croatian dialect continua (Houtzagers, 2008). Sociolinguistic studies consistently describe Burgenland Croatian as an endangered minority language. Author in (Pawischitz, 2012), for instance, documents "first signs of language decay", including massive lexical borrowing and negative attitudes towards the variety, particularly among younger speakers. The analysis of Burgenland Croatian identity in (Ščukanec, 2014) further shows how local, regional, and ethnic identities are intertwined with language use, and how the weakening of Burgenland Croatian in everyday domains affects community cohesion. Work in (Csenar-Schuster, 2014) confirms that Buregenland Croatian is no longer reliably transmitted to children and is frequently perceived as old-fashioned or of limited practical value.

In terms of concrete language resources, one of the substantial digital corpora of Burgenland Croatian is the *Croate du Burgenland subcorpus* hosted by the Pangloss collection (CNRS)[10] (Michailovsky et al., 2014). This corpus consists of twelve recorded oral texts from several villages in northern and central Burgenland, accompanied by transcriptions and linguistic metadata. It is explicitly designed

[8] https://www.zigh.at/
[9] https://rjecnik.at/
[10] https://pangloss.cnrs.fr/corpus/Croate_du_Burgenland?lang=en

as a documentation resource for an endangered Slavic variety and is discussed in the context of methodologies for documenting endangered languages. The CROATINA CLARIN Knowledge Centre for Croatian Language[11] lists Burgenland Croatian dialects among its areas of expertise and points to existing lexicographic and corpus materials for Croatian and its regional varieties. The GRALIS Korpusarium[12] at the University of Graz, which serves as a platform for Slavic corpora, also includes Burgenland Croatian among the documented varieties, but primarily as part of broader Slavic corpus initiatives rather than as a stand-alone resource. In ethnomusicology, the University of Music and Performing Arts, Vienna, Austria, maintains an archive of Burgenland Croatian music[13], providing decades of field recordings. Turning to the print media domain, the weekly newspaper of the Burgenland Croatian community *Hrvatske novine*[14], founded in 1910 and published by the Hrvatsko štamparsko društvo (Croatian Press Association) in Željezno (Eisenstadt), occupies a central position as both a cultural institution and a valuable textual resource. Through sustained coverage of local cultural, educational, and political affairs, it reflects public language use and contributes to the preservation of linguistic continuity and community identity. Media studies and minority-language reports consistently identify *Hrvatske novine* as the leading medium for Burgenland Croats, emphasizing its role in identity construction and language maintenance[15] (Cajka et al., 2025). In addition to *Hrvatske novine*, several other printed media play an important role in the Burgenland Croatian public sphere, including *Crikveni Glasnik* (Croatian Church Newspaper)[16], the weekly church newspaper of the Diocese of Željezno; *Glasilo*[17], the quarterly publication of the *Hrvatsko Kulturno Društvo u Gradišću - HKD* (Croatian Cultural Association in Burgenland); and *Novi Glas* (The New Voice)[18], the journal of the *Hrvatski Akademski Klub - HAK* (Croatian Academic Club), which addresses minority politics, literature, and socio-cultural topics and increasingly appears online. Further publications such as *Gradišće Kalendar* (Burgenland Calendar)[19] and *Panonska ljetna knjiga* (Pannonian Yearbook) provide annual forums for literary, historical, and scholarly contributions, collectively reflecting the diversity of minority-language print media within the Burgenland Croatian community.

Compared with the extensive linguistic and sociolinguistic literature, explicitly DH research on Burgenland Croatian remains limited. One notable exception is the study in (Narančić Kovač & Alardović Slovaček, 2020), which presents a digital comparative analysis of the Croatian original and two Burgenland Croatian editions of the children's classic *Čudnovate zgode šegrta Hlapića* by Ivana Brlić-Mažuranić. By combining methods from digital linguistics and translation studies, this publication demonstrates how digital approaches can be applied to closely related language varieties to examine

[11] https://www.clarin.eu/k-centres/croatina
[12] https://gralis2.at/gralis/
[13] https://www.mdw.ac.at/ive/research/research-areas/ethnomusicological-research-on-music-and-minorities/burgenland-croatian-music/
[14] https://hrvatskenovine.at/
[15] https://search.coe.int/cm?i=090000168078bb08
[16] https://www.martinus.at/glasnik/hr/informacija/ogradianskihrvati
[17] https://hkd.at/hrvatski/veze/biblioteka/glasilo/
[18] https://noviglas.online/
[19] https://hrvatskenovine.blog/kalendar-gradisce-2025/

linguistic complexity, stylistic variation, and the treatment of culture-specific elements. Beyond that text analysis, applied DH initiatives include the development of ČaBot[20], a custom GPT-based conversational agent designed to support children in practicing Burgenland Croatian, as well as projects such as RISE UP (Cajka et al., 2025), whose leaders explicitly emphasize the importance of Artificial Intelligence (AI) methods for the preservation of the language. More generally, Burgenland Croatian appears only sporadically in broader surveys of endangered Slavic varieties and digital documentation initiatives, typically as part of multi-language projects that combine corpus creation with NLP tasks or annotation workflows. To date, however, no ATR models have been published that are specifically trained on Burgenland Croatian. This gap is particularly striking given the growing interest in ATR for low-resource, under-resourced, and minority languages. This paper addresses the identified gap by developing and evaluating an ATR model for printed Burgenland Croatian.

3 Training Data

This study employs training data drawn from selected issues of *Hrvatske novine*, the principal printed newspaper in the Burgenland Croatian community. The issues are available in digitized form through the *AustriaN Newspapers Online* (ANNO)[21] platform of the Austrian National Library, which provides publicly accessible high-resolution newspaper scans (300 dpi). All material used from the ANNO platform was obtained under Open Access, and no additional licensing restrictions apply.

The training data were selected from three post-2009 issues of *Hrvatske novine*, a period characterized by intensified linguistic and orthographic standardization of contemporary Burgenland Croatian. The selected issues - dated 18 June 2010, 17 October 2014, and 11 January 2019 - span different publication years and cover a wide range of topics, thereby providing a temporally distributed and thematically diverse sample of newspaper text. The resulting dataset comprises 88 pages and 64,744 words in total. An overview of the training data is provided in Table 1, and representative sample pages are shown in Figure 1.

Table 1: Composition of the training data

Issues of *Hrvatske novine*	**Number of pages**	**Words count**
18 June 2010	28	28,826
17 October 2014	20	15,565
11 January 2019	40	20,353
Total	**88**	**64,744**

The typography across the training data is consistent. Body text is primarily printed in a serif Latin typeface typical of traditional newspaper layouts, with stable character shapes and consistent use of Burgenland Croatian diacritics. Headings and

[20] https://chatgpt.com/g/g-679259f04b208191b601c542539c01d2-cabot
[21] https://anno.onb.ac.at/



(a) 18 June 2010, Page 11 (b) 17 October 2014, Page 18 (c) 11 January 2019, Page 1

Fig. 1: Representative pages from the training data (sourced from ANNO/Austrian National Library, newspaper Hrvatske Novine)

section titles are clearly distinguished through increased font size, bold weight, and colour, while column layouts remain generally stable across issues. Although pages include non-textual elements such as images, coloured backgrounds, and decorative separators, the text regions exhibit uniform line spacing, regular baseline alignment, and consistent character proportions. In several cases, pages were cropped to remove margins. No additional image enhancement or binarisation was applied.

Generally, the selected training data exhibits consistent printing quality and stable typographic features, making it an appropriate source for developing a custom ATR model intended for contemporary Burgenland Croatian printed materials.

4 Ground Truth Preparation and Model Training within the Transkribus Platform

The *Burgenland Croatian Typewritten 2010–2019* model was developed in Transkribus using semi-automatic layout analysis and *PyLaia* text recognition engine (Puigcerver & Mocholí, 2018). First, page segmentation was performed using the Transkribus *Baroness of Blocks*[22] field-recognition model, suitable for newspapers, which generally produced accurate text region boundaries. However, many pages required manual correction due to overlapping or incorrectly detected regions (Figure 2). Additionally, the text regions containing images were manually deleted and marked as image regions. The text placed on the images was not taken into account.

Following this, automatic layout recognition was applied with the *Universal Lines*[23] model to detect text lines, while preserving the existing text regions. In this case, some segmentation errors, such as the overlapping of text lines between regions

[22] https://app.transkribus.org/models/text/60963
[23] https://app.transkribus.org/models/text/51962



Fig. 2: Example of overlapping of incorrectly detected layout regions

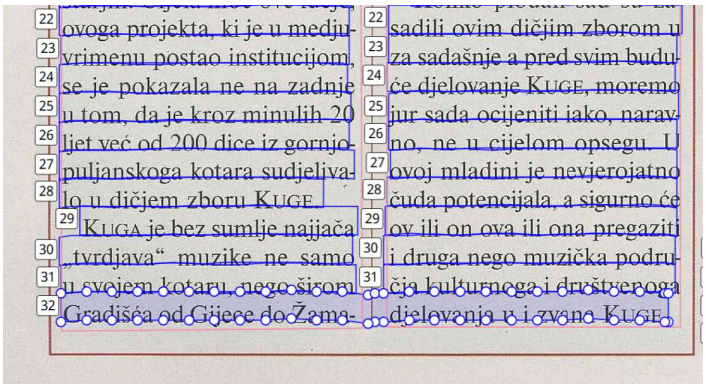


Fig. 3: Example of overlapping of lines between regions

(Figure 3), were manually corrected. In this step, we also observed issues with line polygon detection due to their oversize, as shown in Figure 4. We believe that these inaccuracies did not critically affect the current training cycle, as the focus was on producing consistent line-level ground-truth transcriptions rather than on refining polygon geometry. Nevertheless, the imperfect line polygons introduced latent noise, which affected recognition performance during evaluation. Addressing these segmentation issues (either through improved automatic line detection or targeted manual correction) will therefore be an important focus of the future model refinement.

Afterward, accurate line-level transcripts were created manually to serve as Ground Truth (GT) data. For training our custom model, the GT data were automatically split into a development set and a validation set at a 90:10 ratio, totaling 58,405



Fig. 4: Example of oversized line polygons (line polygons are in blue, while baseline is in red)

words in the development set and 6,339 in the validation set. The model was trained for 100 epochs, totaling 4 h 23 m 38 s. Table 2 summarizes the performance of the *Burgenland Croatian Typewritten 2010–2019* model under the final training configuration. Throughout training, both the training and validation Character Error Rate (CER) decreased rapidly during the first epochs and stabilised below 3% after approximately 30 epochs. The model achieved a best validation CER of 2.34%, with a final CER of 2.48%. The training and validation curves (Figure 5) show consistent convergence without evidence of overfitting. These results indicate that carefully created GT data enable the development of a reliable ATR model for contemporary Burgenland Croatian printed materials.

Table 2: Burgenland Croatian Typewritten 2010–2019 model performance

Words count in the GT data	Epochs	CER (Training set)	CER (Validation set)
58 405	100	1.5%	2,48%

5 Model Evaluation

The performance of the *Burgenland Croatian Typewritten 2010–2019* model was evaluated on a set of documents not included in the training data[24]. This evaluation was designed to assess the model's effectiveness on post-2009 materials, as well as its ability to generalize across varying topics and typographic conditions. We should note that, following Transkribus guidelines[25] for working with newspaper material, where layout recognition can be particularly challenging, the evaluation was preceded by a field recognition model, combined with advanced line-detection configuration settings, to ensure reliable segmentation and transcription results.

[24] https://github.com/marijadjokic/automatic-text-recognition-for-burgenland-croatian-printed-materials/

[25] https://help.transkribus.org/newspapers-layout-recognition

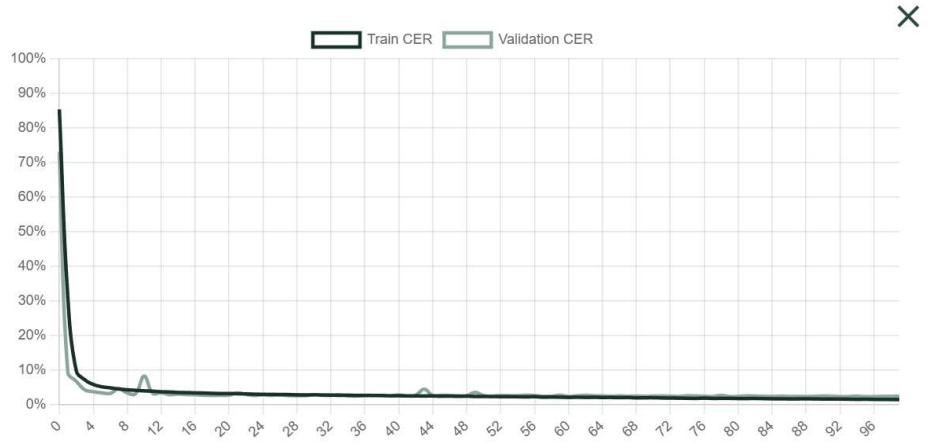


Fig. 5: Learning curve of *Burgenland Croatian Typewritten 2010–2019* model in Transkribus

In the first experiment, we evaluated the model on several pages from selected issues of *Hrvatske novine* (8 April 2011 and 25 October 2013). The evaluation results are presented in Table 3. On these previously unseen pages, the model consistently produced high-quality transcriptions, achieving low CER values - typically below 2.25% - outperforming the domain validation results.

Table 3: Recognition performance of the Burgenland Croatian Typewritten 2010–2019 model on selected pages of the newspaper Hrvatske novine

Newspaper issue (pages)	**CER per page in %**	**Processing time**
8 April 2011 (3/4/8)	2.25 / 0.75 / 0.50	14s 982ms / 23s 600ms/ 19s 830ms
25 October 2013 (2/14/18)	0.35 / 0.44 / 0.97	21s 545ms / 17s 757ms / 21s 398ms

To assess the robustness of the proposed model under out-of-domain conditions, we evaluated its performance on additional Burgenland Croatian documents. These comprised an issue of the newspaper *Galsilo*[26] dated from 2020 avaliable at the website of *HKD*[27], and a publication *Gradišćanskohrvatski franjevci – za svoj narod* (Burgenland-Croatian Franciscans – for their people)[28] dated from 2025 and published by the *ZIGH*[29]. These documents differ from *Hrvatske novine* in formatting, stylistic conventions, and subject matter, offering a complementary test of domains suitable

[26] https://hkd.at/wp-content/uploads/2022/03/glasilo202004.pdf
[27] https://hkd.at/
[28] https://www.zigh.at/wp-content/uploads/2025/05/Ivan-Karall-Gh-franjevci-V.c.pdf
[29] https://www.zigh.at/

for evaluation. The evaluation results are presented in Table 4. During this evaluation, our model achieved CER values below 1,47%, confirming that its performance generalises well to other contemporary printed Burgenland Croatian materials.

Table 4: Recognition performance of the Burgenland Croatian Typewritten 2010–2019 model on documents from the HKD and ZIGH websites

Document (date; pages)	CER per page in %	Processing time
Glasilo newspaper issue (December 2020; 2/3/5)	1.46, 1.18, 1.47	12s 167ms / 15s 345ms / 19s 530ms
ZIGH document (May 2025; 1/2/3)	0.91, 1.12, 0.82	19s 756ms / 19s 482ms / 17s 873ms

To contextualise these results and justify the development of the *Burgenland Croatian Typewritten 2010–2019* model, we compared its performance with two publicly available multilingual ATR models, available on Transkribus, that support Slavic languages, such as *Transkribus Print M1*[30] (supports Polish, Czech, Slovak, and Slovenian) and *Czech, Slovak Print model M1*[31], both trained on printed material from the 21th century, as well as the super model *The Text Titan I*[32]. The evaluation results are presented in Table 5. In this evaluation step, the *Burgenland Croatian Typewritten 2010–2019* model achieved lower CER values in 94% of cases.

Given that Burgenland Croatian is linguistically closest to standard Croatian, we also evaluated several publicly available online OCR tools for Croatian. For instance, i2OCR[33] performs reasonably well on plain text with simple layouts, but its accuracy declines substantially when applied to newspaper material, while our model demonstrates reliability across diverse printed formats. By contrast, Convertio[34] produces good results on both text types but shows limitations in recognizing German- and Hungarian-specific characters, whereas our model successfully handles these character sets.

6 Discussion

The analysis of the evaluation results in Tables 3 and 4 shows that the CER values ranged from 0.35% to 2.25%. Qualitative inspection of the recognition outputs indicates that the majority of errors are layout-driven, primarily due to inaccuracies in line polygon segmentation. These issues most commonly affect line-break handling, leading to unintended word splits or merges (e.g., missing or retained hyphens at line endings), as well as punctuation normalization problems, such as confusion between commas and periods. In addition, a number of isolated character-level substitutions were observed, particularly involving diacritics (č/ć/š/ž), uppercase-lowercase mismatches, and characters influenced by special typographic features (e.g., italics, superscripts, or colored and highlighted text). Such errors occur predominantly in

[30] https://app.transkribus.org/models/text/39995
[31] https://app.transkribus.org/models/text/42352
[32] https://app.transkribus.org/models/text/51170
[33] https://www.i2ocr.com/free-online-croatian-ocr
[34] https://convertio.co/ocr/croatian/

Table 5: Performance comparison of the Burgenland Croatian Typewritten 2010–2019 model with public multilingual ATR models on Transkribus

Testing material (date; pages)	**Model**	**CER per page in (%)**	**Processing time**
Hrvatske novine (8 April 2011; 3/4/8)	Text Titan I	4.46 / 4.32 / 4.30	45s 726ms / 36s 54ms / 1m 11s 54ms
	Transkribus Print M1	2.16 / 2.40 / 2.00	12s 710ms / 17s 629ms / 13s 351ms
	Czech, Slovak Print model M1	3.73 / 5.66 / 3.88	19s 666ms / 9s 963ms / 18s 695ms
Hrvatske novine (25 October 2013; 2/14/18)	Text Titan I	4.31 / 5.01 / 4.38	1m 6s 961ms / 1m 14s 370ms / 56s 935ms
	Transkribus Print M1	1.74 / 2.86 / 2.52	13s 5ms / 17s 590ms / 20s 539ms
	Czech, Slovak Print model M1	4.62 / 5.49 / 5.25	16s 294ms / 3s 725ms / 13s 304ms
Glasilo newspaper issue (December 2020; 2/3/7)	Text Titan I	4.43 / 4.13 / 3.66	48s 675ms / 1m 7s 747ms / 34s 670ms
	Transkribus Print M1	3.91 / 2.40 / 2.35	12s 24ms /14s 995ms / 11s 126ms
	Czech, Slovak Print model M1	5.31 / 4.53 / 4.38	14s 803ms/ 14s 930ms / 16s 178ms
ZIGH document (May 2025; 1/2/3)	Text Titan I	2.13 / 2.98 / 3.07	26s 449ms / 31s 237ms / 29s 648ms
	Transkribus Print M1	0.87 / 0.91 / 1.27	11s 265ms / 12s 306ms / 11s 928ms
	Czech, Slovak Print model M1	3.66 / 3.67 / 5.54	12s 986ms / 9s 869ms / 13s 235ms

acronyms, foreign words, and proper names. Representative examples of the identified error categories are provided in Table 6.

The comparative evaluation of the results, presented in Table 3, 4 and 5, highlights clear performance differences between the *Burgenland Croatian Typewritten 2010–2019* model and selected publicly available Transkribus models when applied to Burgenland Croatian printed materials from 2009 onward. In terms of text recognition accuracy, our model achieved better results in 94% of the cases. Across all tested materials, models such as *Text Titan I* and *Czech, Slovak Print model M1* consistently yield higher CER values, typically ranging from 2.13% to 5.66%. Model *Transkribus Print M1* performs more robustly than the other generic models, achieving lower CER values across all test pages. In comparison with our model, *Transkribus Print M1* achieved lower CER in only two cases. In these cases, the *Burgenland Croatian Typewritten 2010–2019* model failed to detect certain characters, whereas *Transkribus Print M1* detected their presence but misrecognized them. In terms of transcription speed, the *Text Titan I*'s slower work is noticeable. This is expected, given that super models are optimised for broader language coverage and trained on extensive data. The transaction speeds of the *Transkribus Print M1* and *Czech, Slovak Print model M1* models are similar to those of our model, as expected, given that these models

Table 6: Representative examples of error types observed in the evaluation of the Burgenland Croatian Typewritten 2010-2019 model

Error type	**Observed output**	**Correct form**
Wrong diacritics	Zamar; Zarnovica; dvojezićne; Cirilsko-Metodska	Žamar; Žarnovica; dvojezične; Ćirilsko-Metodska
Wrong characters	Provincja; penašnjega; rodjo	Provincia; današnjega; rodio
Punctuation	17.5%; subotu. 2. aprila; e tačaj	17,5%; subotu, 2. aprila; e-tačaj
Uppercase/lowercase mismatches	žablje selo; offrovana; čenča	Žablje selo; Offrovana; Čenča
Foreign words	Abfangiäger; voutube	Abfangjäger; youtube
Acronyms	OVP	ÖVP
Proper names	Ostermaverovi; Anöđelke; Sárkóőzy; Kanizsaj; Seraphini Gvivkovics	Ostermayerovi; Anđelke; Sárkőzy; Kanizsai; Šeraphini Gyivkovics

are relatively small. Generally, these results show that a relatively modest, but carefully curated training corpus can outperform super models for low-resource language varieties. Overall, the findings underscore the importance of dedicated ATR model development and show that reliance on generic models may lead to systematically higher error rates, thereby limiting the usability of digitized texts for downstream linguistic and historical research.

Overall, the performance of our model, achieving CER values lower than 2,25% on unseen material and outperforming some other public models, demonstrates its effectiveness and reliability. The uniform printing style of *Hrvatske novine* and the carefully prepared GT data contributed decisively to this outcome. Crucially, the model delivers good results for other Burgenland Croatian sources, indicating that it is suitable for different types of formats. Additionally, although explicitly designed for Burgenland Croatian, the model could also serve as a useful ATR model for related South Slavic languages written in the Latin script or as a baseline model for developing new custom models on the Transkribus platform, while the training data can be reused for other NLP and DH tasks.

7 Conclusions and Future Work

This paper presented the development and evaluation of *Burgenland Croatian Typewritten 2010-2019*, a dedicated ATR model for contemporary printed Burgenland Croatian, implemented within the Transkribus platform using the PyLaia engine. The model was trained on 88 pages (64,744 words) from three post-2009 issues of the weekly newspaper *Hrvatske novine*, obtained from the Austrian National Library's ANNO repository. The selected training period reflects a phase of intensified linguistic and orthographic standardization, providing relatively stable typographic conditions and enabling the creation of high-quality line-level GT data.

The evaluation results confirm the model's practical usability and strong generalisation capability. Validation yielded a CER of 2.48%, while evaluation on previously unseen pages from *Hrvatske novine* and on additional Burgenland Croatian print

sources consistently achieved CER values below 2.25%. Comparative experiments with publicly available multilingual models on the Transkribus platform (Transkribus Print M1 and Czech–Slovak Print M1) supporting Slavic languages, as well as with a super model (Text Titan I), demonstrate that our model achieved higher recognition accuracy in 94% of the evaluated cases, while also maintaining relatively efficient transcription speed. These results indicate that even a relatively modest but carefully curated training dataset can support the development of high-performing ATR models for low-resource and under-resourced languages.

Qualitative analysis indicates that recognition errors are predominantly layout-driven, arising primarily from inaccuracies in line-polygon segmentation. These issues mainly affect line-break processing, leading to unintended word splits or merges (e.g., incorrect hyphenation at line endings) and punctuation normalization errors, such as confusion between commas and periods. Beyond these systematic effects, a number of isolated character-level substitutions were observed, chiefly involving diacritics (č/ć/š/ž), uppercase–lowercase mismatches, and characters affected by non-standard typographic features (e.g., italics, superscripts, or colored and highlighted text). Such cases occur primarily in acronyms, foreign words, and proper names.

Building on these results, several directions emerge for extending the model's coverage and robustness. Expanding the training data with additional issues of *Hrvatske novine* and other Burgenland Croatian print sources would increase typographic diversity and further stabilize text recognition. Also, extending the model to pre-2009 material would enable coverage of a broader historical span of Burgenland Croatian print culture, though such texts present additional challenges due to greater orthographic variability and less consistent typographic conventions. Improvements in layout processing, particularly in line polygon detection, are also planned to reduce latent noise and support more accurate text recognition.

Overall, this study presents the first publicly available ATR model tailored explicitly to Burgenland Croatian post-2009 printed materials. The developed model provides a reusable resource with potential applicability to related South Slavic language varieties and could serve as a baseline for training new ATR models within the Transkribus platform. In addition, the training dataset developed for this work represents a valuable language resource that can be reused for further NLP and DH applications. In general, with this paper we introduce Burgenland Croatian into the wider DH research context.

Declarations

- **Funding**: This research was supported by the Transkribus CLARIAH-AT Scholarship. The second author's work was funded by the Austrian Science Fund (FWF), grant DOI: 10.55776/I6258.
- **Competing interests**: The author declares that there are no competing interests.
- **Ethics approval**: Not applicable. This study does not involve human participants, personal data, or experimental interventions.
- **Consent**: Not applicable.

- **Data availability**: The ATR model developed in this study, Burgenland Croatian Typewritten 2010–2019, together with its corresponding training dataset, are publicly available on the Transkribus platform. The training data comprise digitized issues of the newspaper *Hrvatske novine* (Croatian Weekly Newspaper) obtained from the ANNO digital newspaper repository of the Austrian National Library. The evaluation corpus was collected from ANNO as well as from publicly accessible materials provided by the Znanstveni institut Gradišćanskih Hrvatov - ZIGH and the Hrvatsko kulturno društvo u Gradišću - HKD. All evaluation results are publicly available on GitHub.
- **Code availability**: No custom code was developed for this study. Model training and evaluation were conducted using the Transkribus platform and its integrated PyLaia text recognition engine.
- **Authors' contribution statements**: All authors contributed to the conception and design of the study. Data curation was carried out by both authors. Model training, model evaluation, and analysis of the results were performed by the first author, M. Đ. P. The first draft of the manuscript was written by the first author, M. Đ. P. Review and editing were performed by both authors.

References

Aepli, N. (2024). *There Is Plenty of Room at the Bottom: Challenges & Opportunities in Low-Resource Non-Standardized Language Varieties* (Doctoral dissertation, University of Zurich). https://doi.org/10.5167/uzh-262877

Benčić, N., Csenar-Schuster, A., Rotter, I., Sučić, I., Vlašić, J. (2009). *Pravopis gradišćanskohrvatskoga književnoga jezika*. Eisenstadt, Austria: HKDC - Kroatisches Kultur- und Dokumentationszentrum.

Berghaus, J. (2023). Adverbielle Präfigierung im Sprachkontakt: Partikelverben und ihre aktionale Rolle im Burgenlandkroatischen. In *L'aspettualità nel contatto linguistico: lingue slave e oltre* (p. 19–36). Firenze University Press. Retrieved from http://dx.doi.org/10.36253/979-12-215-0184-1.05

Breu, W., Berghaus, J., Scholze, L. (2016). Der Verbalaspekt im totalen Sprachkontakt: Moliseslavisch, Obersorbisch und Burgenland-Kroatisch im Vergleich. W. Breu & T. Reuther (Eds.), *Slavistische Linguistik 2014 : Referate des XL. Konstanzer Slavistischen Arbeitstreffens in Konstanz/Hegne, 1.-5. September 2014, pages=55–116.* Frankfurt am Main: Peter Lang. (Retrieved from http://nbn-resolving.de/urn:nbn:de:bsz:352-2-afbbsmzrsrpk3)

Browne, E. (2016). Burgenland Croatian: An Old Language on a Do-it-Yourself Border with a New Name. In T. Kamusella, M. Nomachi, & C. Gibson (Eds.), *The Palgrave Handbook of Slavic Languages, Identities and Borders* (pp. 347–359). London: Palgrave Macmillan UK. Retrieved from https://doi.org/10.1007/978-1-137-34839-5_17

Cajka, S., Jouitteau, M., Vetter, E. (2025, aug). A Hiking Path for Burgenland

Croatian. *Proceedings of EUROCALL 2025.* Milan, Italy. (Retrieved from https://hal.science/hal-05385510v1)

Csenar-Schuster, A. (2014). *Sprachplanung für die burgenlandkroatische Sprache in Österreich: ein soziolinguistischer Befund und Beitrag zur Sprachrevitalisierung* (Master's thesis, Universität Wien, Vienna, Austria). (University repository u:theses, ID 29156, AC12074975) https://doi.org/10.25365/thesis.32833

Hossain, A. (2025). The Impact of Digitalization on Humanities Research: A Critical Review. *International Journal of Original Recent Advanced Research*, *2*(2), 14, (Retrieved from https://www.ijorarjournal.com/Research-Paper.php?id=146)

Houtzagers, P. (2008). On Burgenland Croatian Isoglosses. *Studies in Slavic and General Linguistics*, *34*, 293–331, Retrieved from http://www.jstor.org/stable/40997588 (Retrieved from https :// api .semanticscholar.org/CorpusID:132557634)

Houtzagers, P. (2013). Burgenland Croats and Burgenland Croatian: some unanswered questions. *Rasprave Instituta za hrvatski jezik*, *39*(1), 253–269, (Retrieved from https://hrcak.srce.hr/113036)

Keinan-Schoonbaert, A. (2025). Automatic Text Recognition (OCR/HTR). *A LIBER Digital Scholarship & Data Science Topic Guide for Library Professionals*, 9, https://doi.org/10.23636/k8qs-wc65

Kinda-Berlakovich, A.Z., Bencsics, N., Agnes, C.-S., Ivo, S., Koschat, H., Znika, M. (2003). *Gramatika gradišćanskohrvatskoga jezika: Grammatik der burgenlandkroatischen Sprache.* Trausdorf, Austria: Znanstveni institut Gradišćanskih Hrvatov/Wissenschaftliches Institut der Burgenländischen Kroaten.

Michailovsky, B., Mazaudon, M., Michaud, A., Guillaume, S., François, A., Adamou, E. (2014). Documenting and researching endangered languages: the Pangloss Collection. *Language Documentation & Conservation*, *8*, 119-135, (Retrieved from https://shs.hal.science/halshs-01003734)

Narančić Kovač, S., & Alardović Slovaček, K. (2020). The Adventures of Hlapić in Burgenland Croatian. *Sic: časopis za književnost, kulturu i književno prevođenje*, *10*(3), 0–0, https://doi.org/10.15291/sic/3.10.lc.4

Nockels, J. (2025). *Making the Past Readable: A Study of the Impact of Handwritten Text Recognition (HTR) on Libraries and Their Users* (PhD thesis, The

University of Edinburgh). https://doi.org/http://dx.doi.org/10.7488/era/5988

Pakray, P., Gelbukh, A., Bandyopadhyay, S. (2025). Natural language processing applications for low-resource languages. *Natural Language Processing*, *31*(2), 183-197, https://doi.org/10.1017/nlp.2024.33

Pawischitz, S. (2012). Burgenland-Croatian – a language in contact – First signs of language decay. *Bosanskohercegovački slavistički kongres*, *1*(1), 575–585, (Retrieved from https://www.ceeol.com/search/article-detail?id=1116727)

Puigcerver, J., & Mocholí, C. (2018). *PyLaia.* 1.0.6. GitHub repository. GitHub. (https://github.com/jpuigcerver/PyLaia).

Ščukanec, A. (2014). In Search of Identity: The Burgenland-Croatian Perspective. *Croatian Studies Review - Časopis hrvatskih studija*(10), 35–56, (Retrieved from https://hrcak.srce.hr/143272)

Tyran, K. (2022). Zur Situation der kroatischen Volksgruppe und ihrer Sprache in Österreich. *Europäisches Journal für Minderheitenfragen*, *15*(3-4), 241–260, https://doi.org/10.35998/ejm-2022-00012