Automated Transcription of Mixed-Script Dialectal Materials

Markus Kunzmann

Austrian Centre for Digital Humanities (ACDH), Austrian Academy of Sciences (ÖAW),
Bäckerstraße 13, 1010 Wien, Austria
E-mail: markus.kunzmann@oeaw.ac.at

Abstract

The project Dictionary of Bavarian Dialects in Austria "Wörterbuch der bairischen Mundarten in Österreich" (WBÖ) project maintains an archive of approximately 3.6 million handwritten dialectal paper slips documenting dialectal evidence. While 2.4 million entries have been manually digitized and converted to TEI format, the remaining 1.2 million paper slips from sections A-C require automated processing. This paper presents a novel three-stage workflow concept combining Handwritten Text Recognition (HTR) technology with existing digitized holdings to overcome the challenges posed by heterogeneous writing systems, multiple scribes, and poor material condition. Initial tests with existing HTR models yielded unsatisfactory results. The proposed solution leverages the existing Database of Bavarian Dialects "Datenbank der bairischen Mundarten in Österreich" (DBÖ) to automatically correct HTR transcription errors through similarity-based alignment and N-gram matching algorithms. The corrected transcriptions serve as a gold standard or a kind of ground truth for training a specialized HTR model tailored to historical dialect materials. This methodology enables the creation of substantial training datasets without manual transcription, potentially generating 33.6 million words for model training. The approach promises complete digital access to the WBÖ archive and provides a transferable template for similar lexicographic projects with historical slip collections.

Keywords: handwritten text recognition; dialect lexicography; digital humanities; historical paper slips; workflow proposal

1. Introduction

Dialects represent a particularly valuable object of study for linguistics. As uncodified and primarily spoken language forms that are largely not subject to normative forces, they allow insights into socially conditioned language change processes. While standard languages are also subject to these forces, they are affected to a much lesser degree, as they achieve a certain stability through their codification on linguistic levels such as orthography. However, no empirical science like linguistics can do without a data foundation to be investigated, and in this regard, dialect is at a disadvantage compared to standard language. While data sources for investigating written standard language are already available as text, dialectal sources have already undergone many individual work steps before they can be used for research in this state.

Long-term projects such as the Dictionary of Bavarian Dialects in Austria "Wörterbuch der Bairischen Mundarten in Österreich" (WBÖ)[1] also utilize such extensive data treasures. It

[1] The WBÖ is situated within the project cluster of the same name and processes the Bavarian dialects in Austria and South Tyrol: https://www.oeaw.ac.at/de/acdh/wboe-projektcluster (accessed: 2025-09-30)

is being developed at the Austrian Centre for Digital Humanities (ACDH)[2] at the Austrian Academy of Sciences "Österreichische Akademie der Wissenschaften" (ÖAW)[3]. This dialect dictionary has access to a collection of around 3.6 million handwritten paper slips with dialect evidence. To simplify and accelerate work on dictionary articles, around 2.4 million entries were entered into a TUSTEP system[4] and converted to TEI format from the 2010s onwards (Barabas et al. 2010; Stöckle, 2021: 14-16). The part of the alphabetical section A to C (Kranzmayer, 1970, 1976; Hornung & Bauer, 1983), for which dictionary volumes have already been published, did not consider the corresponding paper slips in this first manual digitization step.

Since the resulting database is also used as a corpus, these remaining, not yet digitized holdings should also be recorded retrospectively. This is to be done using automated HTR[5] technology. However, the HTR approach is challenging due to the heterogeneous nature of this material. They were not only recorded by different writers, but are also different in terms of their writing system, as a mix of typewriting, Latin cursive, and Kurrent script can be contained on one and the same paper slip. In addition, the paper slips are in poor external condition due to their age, which makes automatic text recognition more difficult.

This paper presents a three-stage, largely automated workflow that combines HTR technology with the existing data holdings of the Database of Bavarian Dialects in Austria "Datenbank der bairischen Mundarten in Österreich" (DBÖ) of the WBÖ. The focus is on building a tailored HTR model that is to be trained with data where erroneous transcription results from the application of existing HTR models are to be compared and corrected using the already digitized holdings.

2. Background

The WBÖ data material exists today in three different forms of representation: The paper slips themselves, the so-called main catalog "Hauptkatalog", which comprises about 3 million physical paper slips and serves as the primary data source of the dictionary (section 2.1). Second, there are 2.4 million already manually digitized entries that were first recorded in TUSTEP from the 1990s (section 2.2) and later converted to TEI-XML-compliant datasets (section 2.3) and today form the Database of Bavarian Dialects in Austria (DBÖ). Third, since 2017, high-resolution scans of the physical paper slips have been systematically created, with currently about 1.6 million paper slips available as digital image data (2.4). This triple representation of the material creates the prerequisites for digitization approaches, as both the visual original data and the already structured contents are available for automated linking procedures.

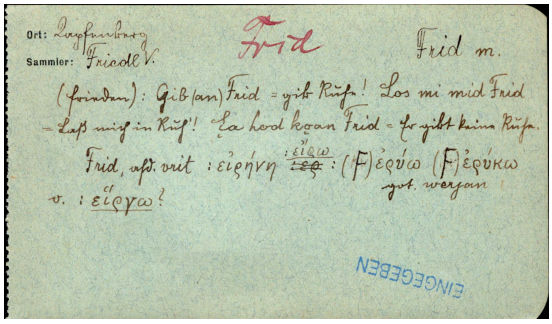
2.1 The Historical Dialect Paper Slips

The WBÖ is a long-term project of the Austrian Academy of Sciences (ÖAW) and documents as a dictionary in its current form the lexicon of the parts of Austria where

[2] https://www.oeaw.ac.at/acdh/ (accessed: 2025-09-30)
[3] https://www.oeaw.ac.at/
[4] Tuebingen System of Text Processing tools "Tübinger System von Textverarbeitungs-Programmen" https://www.tustep.uni-tuebingen.de/tustep_eng.html (accessed: 2025-09-30)
[5] Handwritten Text Recognition. In relation to all writing systems, also referred to as *Automatic Text Recognition* (ATR).

Bavarian serves as a regional vernacular, and South Tyrol, where Bavarian dialects are also spoken. As with other regional dictionaries, e.g., the Bavarian Dictionary "Bayerisches Wörterbuch"[6] (BWB) or the Swiss Idiotikon "Schweizerisches Idiotikon"[7], an archive of paper slips serves as the data basis for dictionary work. About half of the dialect examples comes from voluntary collectors, a third was determined via so-called field trips "Kundfahrten", i.e., via direct surveys, the rest consists of excerpts from literature such as other regional dictionaries, dissertations, or historical sources (Stöckle, 2021: 12).

On these paper slips, the most essential information is documented on the basis of which the dictionary entry is then created. Even with the current WBÖ, the dictionary articles are based exclusively on the contents of this extensive collection. The paper slips contain the dialect word itself in its sound and pronunciation as well as the lemma that is set for it. In addition, grammatical information and, if applicable, references to the etymology of the word are also recorded on the paper slip. Sometimes the meaning is supplemented or illustrated by an example sentence. In addition, survey location, questionnaire number, information about the collector point to the origin of the term.



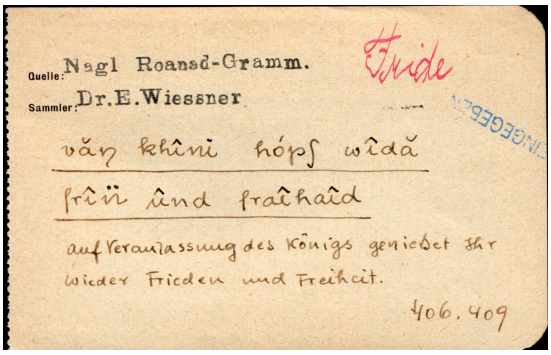


Figure 1: Examples of paper slips containing phonetic transcription, Latin cursive, and Kurrent script.

The paper slips themselves are highly heterogeneous both in their structure and in terms of written documentation.

Although the lemma is almost consistently written in red ink and prominently placed on the paper slip, a fixed location where it can be found is not provided. The phonetic transcription and meaning are also not embedded in a fixed grid but can be located in different places. Only the information about the location or source is found, insofar as it concerns the preprinted forms, in the upper left corner.

Heterogeneity is also shown in the use of different writing systems. A variety of different writing systems can be used on one paper slip and is even the rule. A mix of different scripts, from typewritten content alongside handwritten notes in Latin cursive, can be found as well as those in Kurrent script. The phonetically transcribed dialect word as well as the example sentences, insofar as they are present, are written in Teuthonista. This phonetic transcription system is very common especially in dialectology of Upper German varieties and is comparable to the Böhmer-Ascoli system, which is also used for the writing

[6] https://bwb.badw.de/ (accessed: 2025-09-30)
[7] https://www.idiotikon.ch/ (accessed: 2025-09-30)

of sound values (Wiesinger, 1964). In addition to this proliferation of writing systems, it is also different writers who, with their diverse handwritings and writing habits, additionally create variance within the paper slips.

2.2 First Digitization in TUSTEP

One of the most time-intensive preliminary works for actually writing the dictionary articles was searching for and sorting the respective paper slips from the main catalog comprising several million paper slips. This was one of the main reasons why it was decided at the beginning of the 1990s to convert the paper slip catalog into electronic form. However, this project also thought beyond the editorial work on the WBÖ. Through the new sorting possibilities that come with this digitization, new approaches to the material and thus also other research questions should be able to be answered (Bauer & Kühn, 1998).
From 1993, the paper slips were entered into the TUSTEP system, this project was completed in 2010. Figure 2 shows an example of such a paper slip, Figure 3 the structure of the corresponding TUSTEP dataset.

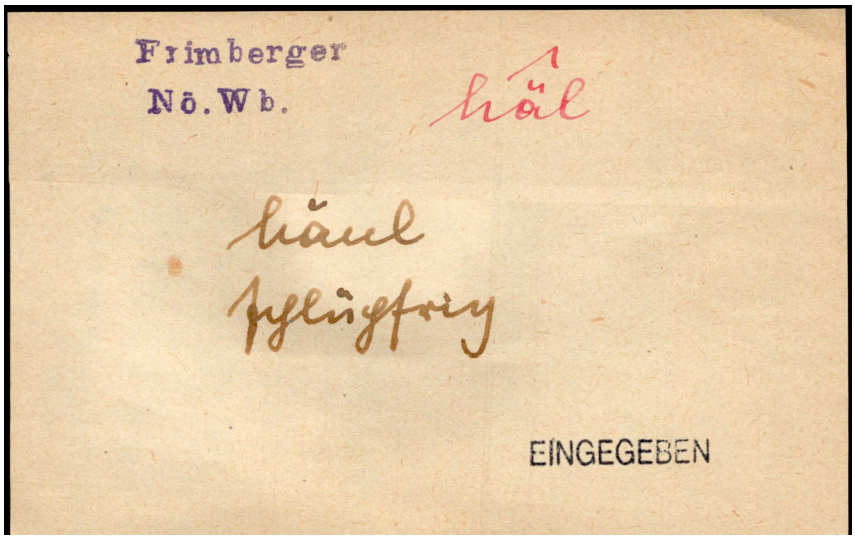


Figure 2: paper slip from the Frimberger collection for the lemma *hâl*

```
*A* HK 364, h364 = h3640715.pir#42.1
*HL* h<^al:2
*QU* Frimb. N^o.Wb.
*QDB* {6.4j,6.4k,6.4l,6.4m,6.4n,6.4o}
Marchfd.:s^oWeinv. *@ Slg.FRIMBERGER^. (u.1890) o.S. Arch.29/1-
6
===
*LT1* h)aul [P]
*BD/LT1* schl^upfrig
```

Figure 3: TUSTEP formatting of a WBÖ dataset with field identifications

The information was entered using standardized scheme. The fields were marked by asterisks and had specific meanings. Some of these are meta-information such as `*A*` for the archival ordering system or `*QDB*` for the reference to the source database. Other information can be read on the paper slip, such as those for the main lemma `*HL*`, the stamp with the reference to the source `*QU*` as well as the sound of the dialect word `*LT1*` and the associated meaning `*BD/LT1*`. Characters that are not part of the ASCII character set, including e.g., the umlauts in *häul* or *schlüpfrig*, but also the phonetic transcription in Teuthonista were encoded with non-alphabetic ASCII characters.

However, this first digitization step did not cover the entire holdings of the main catalog. Since at the time of deciding on this process, the volumes for letters A, B/P, and C had already been published, the paper slips were only recorded from letter D onwards.

2.3 Conversion into TEI XML

In 2014, work began on converting the TUSTEP data into a TEI-compliant XML version (Stöckle, 2021: 16). The reasons for this were both technical and content-related. From a technical side, the TUSTEP format was outdated and maintaining the technical infrastructure was costly. In addition, search queries within the material were not entirely simple, as a separate query language had to be used for this. A conversion to TEI XML not only eliminated this problem, as one could thereby use already existing infrastructure and know-how, but also made it possible to make the data accessible to third-party users. The transcriptions could now also be displayed using Unicode in Teuthonista, a possibility that TUSTEP did not yet have at that time (Bowers & Stöckle, 2018: 46).

In terms of content, the conversion achieved a major step toward simplifying the data model. Through the combination of individual fields in the TUSTEP version, e.g., the meaning for specific word like `*BD/LT1*`, `*BD/LT2*`, etc., one potentially had around 500 fields. Through the conversion, the number was reduced to 37 fields, which allowed for easier overview of the material and thereby also minimized documentation effort (Bowers & Stöckle, 2018: 51).

The conversion was carried out using XSLT in several passes. The transformation results were repeatedly checked semi-automatically to identify further steps and possibly also gaps in the material.

```
<entry xmlns="http://www.tei-c.org/ns/1.0"
       xml:id="h364_qdb-d1e1124"
       n="543911"
       source="#orig-h364_qdb-d1e1124"
       facs="H_364%2FH3640715.pir%2FH3640715.pir_00044_00044_a.jp2">

  <form type="hauptlemma" xml:id="tu-85950.56">
    <orth>hâl</orth>
  </form>

  <gramGrp>
    <pos>Adj</pos>
  </gramGrp>

  <form type="lautung" n="1" xml:id="tu-85950.60">
    <pron xml:lang="bar" notation="teuthonista">hãul</pron>
    <gramGrp>
      <gram> [P]</gram>
    </gramGrp>
  </form>

  <sense corresp="this:LT1" xml:id="tu-85951.1">
    <def xml:lang="de">schlüpfrig</def>
  </sense>

  <ref type="archiv" xml:id="tu-85950.55">HK 364, h364 = h3640715.pir#42.1</ref>
  <ref type="quelle" xml:id="tu-85950.57">Frimb. Nö.Wb.</ref>
  <ref type="quelleBearbeitet" xml:id="tu-85950.58">
    {6.4j,6.4k,6.4l,6.4m,6.4n,6.4o} Marchfd.:söWeinv.
  </ref>
  <ref type="bibl" corresp="this:QDB">
    <bibl>Slg.FRIMBERGER· (u.1890) o.S. Arch.29/1-6</bibl>
  </ref>

</entry>
```

Figure 4: TEI-XML structure of a WBÖ dataset after conversion from TUSTEP

2.4 Scans

The recording of paper slips in the TUSTEP system and the subsequent processing in TEI are steps that make a valuable contribution to data transparency. However, in the two decades of manual recording, scanning of the paper slips themselves was omitted. With its approximately 3 million paper slips, the main catalog is the largest collection of dialectal lexicon of Bavarian in Austria. To sustainably secure the holdings, scans of these paper slips have been made since 2017. Several reasons justify this undertaking.

The paper slips are subject to an aging process. This manifests itself, for example, through increasing fading of some handwritten notes, but the paper itself is also affected, becoming increasingly brittle in part. However, reasons of scientific transparency also speak for the continuous optical digitization of the paper slips. Even though the typing of the paper slips followed a strict system, the step from paper slip to digital dataset remains an interpretive process. Making scans of the paper slips available enables comparison with the data currently available in TEI format in case of doubt.

An assignment of the image material to the individual datasets appears trivial at first glance. However, this is made more difficult by the fact that there is no 1:1 relationship between the paper slips and the datasets of the DBÖ. A paper slip can be the starting point for several datasets, and there are also bundles of paper slips that have been combined into one entry. This assignment step is therefore currently done manually, which is why only a small proportion of the approximately 1.6 million paper slips (as of 08/2024) have been linked to datasets so far.

3. Digitization

3.1 The Intention

The project involves the integration of the not yet digitized holdings of the main catalog into the project database (DBÖ). For a non-manual approach, this includes scanning the remaining paper slips with the initial letters A, B/P, and C. The actual indexing is done through the use of HTR procedures.

3.2 The Core Problem

To index the contents of the not yet recorded paper slips of the main catalog, tests were conducted in July and August 2024 using HTR methods. The transcription platform Transkribus was chosen as the application. This product was selected because it offers a comparatively low-threshold access to HTR methods. In addition, Transkribus has been in use for several years in various projects[8] at the ÖAW, so a large internal expert community can be relied upon. Transkribus also has a number of models that seemed suitable for application to the paper slips. Most models in Transkribus are based on PyLaia[9] (READ-Coop, 2022), a successor of Laia by Joan Puigcerver (2017), an HTR engine that implements deep learning technologies based on *Convolutional Recurrent Neural Networks* (CRNN) combined with *Connectionist Temporal Classification* (CTC) (Kahle et al., 2017). This architecture first extracts visual features through convolutional layers, i.e., layers optimized for pattern recognition in image data, and subsequently processes them through recurrent layers – typically bidirectional *Long Short-Term Memory* (LSTM) units – which are needed to capture sequential dependencies between characters. The CTC algorithm enables end-to-end training, allowing the network to automatically learn how to align image segments with corresponding characters without requiring manual specification of each character's exact location in the image during training – a particular advantage for cursive scripts like Kurrent with fluid character boundaries. In 2023, Transkribus introduced so-called Transformer-based "Super Models" such as *The Text Titan I*, which significantly improve recognition quality. However, these cannot currently be trained by end users and are therefore not an option for domain-specific model building in this project (READ-Coop, 2023). Models like *Transkribus German Kurrent*[10] show good results when applied to German-language Kurrent script. The *Character Error Rate* (CER), which reflects the percentage of incorrectly recognized characters, is 5.4% for this model. But

[8] A list of current ÖAW projects using Transkribus can be found on the ACDH pages: https://www.oeaw.ac.at/de/acdh/what-we-offer/transkribus-1 (accessed: 2025-09-30)

[9] PyLaia on GitHub: https://github.com/jpuigcerver/PyLaia (accessed: 2025-09-30)

[10] HTR model *Transkribus German Kurrent*: https://app.transkribus.org/models/public/text/german-kurrent-and-sutterlin-17th-20th-century (accessed: 2025-09-30)

universal models like *The German Giant I*[11], which was trained on a broad basis of documents in Kurrent script and Latin handwriting, also seemed promising for conducting initial tests. In applying these models to a test sample - 20 representative paper slips were selected - the results were not acceptable. The error rate was 82.06% for the *Transkribus German Kurrent* model, and for the *The German Giant I* model it was 69.91%. Regarding the transcription of the paper slips, it is problematic to use CER as a reference value. The contents of the paper slips are, unlike other documents, not continuous text. Since the text is compared sequentially, some transcriptions in the gold standard are at different line positions than in the comparison text, which can result in high CER values. Nevertheless, these two CER values speak for an unusable result. Only a structured CER, which is determined only within the text regions, could remedy this. However, the prerequisite for this would also be that text regions exist that contain the same text in the original scan.

Based on this outcome, further attempts were made to train custom models using gold standard data. Based on around 400 paper slips that were manually transcribed, models were trained in several iterative steps with 60, 120, 180, 240, 270, and 300 paper slips. Between the individual training processes, the CER was recorded and the transcriptions were checked on a random basis (see Table 1).

Model	**Pages**	**Words**	**CER Train (%)**	**CER Validation (%)**
HK_HTR_60	60	904	32.78	38.72
HK_HTR_120	120	1,775	31.10	29.74
HK_HTR_180	180	2,412	25.39	33.88
HK_HTR_240	240	3,228	34.80	36.89
HK_HTR_270	270	3,580	24.12	33.39
HK_HTR_300	300	3,857	24.63	30.95

Table 1: HTR model experiments with WBÖ training data

These CER values were also not satisfactory, and the increasing amount of training material showed no significant improvement in the models. The finding that these manually created gold standard data are not sufficient to generate robust HTR models made a solution strategy necessary that operates without this type of training data.

4. Proposed Solution

Since preparing this gold standard data manually takes a lot of time, but large amounts of it would be needed for training to generate stable HTR models, a solution approach should be pursued that makes use of the approximately 2.4 million datasets of the DBÖ. The concept involves correcting the erroneous results of an initial HTR process using the DBÖ. This first requires an assignment of the scan of a paper slip to its digital counterpart in the DBÖ (4.1). When this connection exists, the transcribed contents can be identified based on a similarity algorithm and replaced by the character string from the dataset

[11] HTR model *The German Giant I*: https://app.transkribus.org/models/public/text/the-german-giant-i (accessed: 2025-09-30)

(4.2). If this is done with a sufficient number of scans or datasets, gold standard can be generated based on this (4.3) and an HTR model can be trained (4.4).

4.1 Scan to Database Alignment

The goal is to find the most similar dataset from the database DBÖ to the error-prone transcriptions of the paper slips from the HTR process. Unlike the database contents, which are equipped with content categories such as main lemma, phonetics, or meaning, the result of the HTR process represents an uncategorized sequence of characters. A similarity assessment therefore cannot be made at a category level such as the lemma but must use the entire text as a reference. A corresponding flattening of the DBÖ datasets is therefore sensible.

The approach therefore first includes the step of normalizing the DBÖ entries. In this process, those elements from a DBÖ dataset are extracted that represent the information actually written on a paper slip. In the example given above, these would be the categories main lemma (`<form type="hauptlemma">/<orth>`), phonetics (`<form type="lautung">/<pron>`), meaning (`<sense>`), and the source reference (`<ref type="quelle>`). This database entry would be concatenated into a linear character string "`hâl hăul schlüpfrig Frimb. Nö.Wb.`".

In the second step, a similarity calculation takes place between the HTR results and the normalized DBÖ entries. Since correct recognition of individual words is not given in the HTR results, comparison procedures at the token level are excluded. Character-based approaches should prove more robust here, which are thereby also more tolerant with regard to typical HTR errors. This could be accomplished with procedures based on N-grams or also the proven Levenshtein distance. A combination of these or other matching approaches and the summation of the scores of the individual procedures could also increase the accuracy of automatic assignment.

The procedure can be evaluated based on already manually correctly linked scans and adjusted if necessary, e.g., through further normalizations (case insensitive, removal of special characters and diacritics).

4.2 Correction

In this step, the results of the HTR process are compared and corrected with the contents of the corresponding dataset(s) from the DBÖ. With the step of aligning scan and dataset, the prerequisites for this have been created. The procedure for this step is similar to that of assigning the paper slip scans to the datasets. While previously the assignment took place at the level of the datasets of the two holdings (HTR result, DBÖ entry), similar procedures now take place within a dataset.

Here again, the character string that was already formed from the contents of the DBÖ in the previous assignment step is used, except that this time it is definitely not subjected to normalizing steps such as the elimination of special characters or diacritics. This is important because the DBÖ contains the correct transcriptions and these should later serve as the basis for training an HTR model.

One problem encountered in the automatic correction of HTR results is that the recognized words can often be fragmented. Thus, distances between individual characters interpreted as spaces can produce two substrings, although it is actually one string (e.g., `"schl" "uprig"` instead of `"schlüpfrig"`). In contrast, the correct strings in the DBÖ are present as one connected string. A similarity-based assignment at the token level is therefore excluded.

Therefore, a division of the substrings from the DBÖ such as lemma, phonetics, meaning, etc. into their possible N-grams also takes place here. These vary from individual characters to the entire word in their length. This segmentation follows a purely mechanical principle and does not consider linguistic segments such as individual morphemes.

The correction process can be illustrated using the example of the word `"schlüpfrig"`: First, the correct string `"schlüpfrig"` from the database is decomposed into all (1- until 10-grams) possible N-grams – from 1-grams (s, c, h, l, ü, p, f, r, i, g) via 4-grams (schl, chlü, hlüp, lüpf, üpfr, pfri, frig) up to the complete 10-gram `"schlüpfrig"`. Subsequently, the HTR token `"schl"` is compared with all N-grams and receives a similarity value for each. The 4-gram `"schl"` achieves a perfect match (100% similarity, position 0-3), while 3-rams such as `"sch"` (75% similarity) or `"sc"` (50% similarity) show lower values. Analogously, the HTR token `"uprig"` is compared with all available n-grams. Here, the 6-gram `"üpfrig"` (position 4-9) shows the highest similarity, although the characters do not match exactly. For each token, the n-gram with the highest similarity value is now selected, while checking whether the assigned positions overlap. In the present case, the positions 0-3 (`"schl"`) and 4-9 (`"üpfrig"`) do not overlap, so both assignments can be maintained. The HTR tokens are finally replaced by their assigned n-grams: `"schl"` remains `"schl"` and `"uprig"` becomes `"üpfrig"`. The combined result `"schl" + "üpfrig"` yields the correct word `"schlüpfrig"`. This comprehensive decomposition and stepwise assignment makes it possible to systematically correct even heavily fragmented HTR outputs.

Now the individual tokens that arose through the HTR process are compared to those of the N-grams and evaluated using a similarity-oriented algorithm. The score should be weighted by the length of these N-grams. This is intended to prevent, for example, monograms, which tend to have a higher probability of being found again in the complementary string, from being incorrectly assigned. In addition, it is implemented in the procedure that each database section – i.e., each possible replacement unit within a database entry – may only be used once for correction. Since multiple HTR tokens may reference the same N-gram area, the need for conflict resolution arises. The final selection is made via a procedure that selects the best still available N-gram candidate for each token. This means that for each token, the N-gram unit with the highest similarity value is selected from the remaining, not yet assigned units. If no sufficiently similar candidate can be identified for a token, it remains unchanged. This is intended to prevent erroneous replacements that would falsify the source material from being introduced through uncertain correction suggestions.

An illustrative example of such a conflict arises with the HTR output of the tokens `"schl"` and `"uprig"`. Both represent fragmentations of the originally intended word `"schlüpfrig"` and can each be mapped to different N-grams of this database entry – for example, `"schl"` to the word beginning and `"üpfrig"` to the word end. However, since both tokens reference the same database entry and the assigned N-grams overlap in content, the algorithm must decide which assignment to maintain. Through the conflict-resolving strategy, it is ensured

that "schl" and "uprig" can be converted together, but without mutual overlap, into a coherent correction – namely "schlüpfrig".

Through this step, which runs via the N-grams, it is also possible that a string schlüpfrig can be assigned to the HTR strings schlu and prig.

4.3 Generating Gold Standard Data

Through the two steps of assigning paper slips to scans and the correction algorithm, the prerequisites have now been established to build up a stock of gold standard data for training an adapted HTR model for the handwritten slips.

The practical implementation of this approach first requires systematic transcription of the handwritten slips using existing HTR models. Since the handwritten slips themselves are considerable in terms of script types, scribes, and linguistic aspects - a significant portion of the content is non-standard - the large, universal models should be suitable. These were often trained on the basis of multilingual training data and with different script types and should therefore provide the best prerequisite for creating a transcription that serves as the basis for corrections.

The correction is performed on the PAGE XML files of the generated HTR outputs. These should take place outside the training platforms. The PAGE XML contains the layout information, i.e., the coordinates of the text regions and the baselines, which should remain untouched during correction. The transcription of the above-mentioned handwritten slip, which was transcribed in Transkribus using the model *The German Giant I*, shows the typical HTR recognition errors (Figure 5).

```
<TextLine id="tr_1_tl_1">
    <Coords points="362,178 638,166 823,188 ..."/>
    <Baseline points="367,155 396,159 426,161 ..."/>
    <TextEquiv>
        <Unicode>Pyimberger</Unicode>
    </TextEquiv>
</TextLine>
<TextLine id="tr_1_tl_2">
    <Coords points="403,332 472,311 614,322 ..."/>
    <Baseline points="409,304 434,301 459,303 ..."/>
    <TextEquiv>
        <Unicode>Dö. WB</Unicode>
    </TextEquiv>
</TextLine>
<TextLine id="tr_1_tl_5">
    <Coords points="697,995 789,1002 842,967 ..."/>
    <Baseline points="705,902 738,906 772,908 ..."/>
    <TextEquiv>
        <Unicode>schlüpfrigh</Unicode>
    </TextEquiv>
</TextLine>
```

Figure 5: Example of a fully converted XML file, here shortened by the elements <usg type=”geo”> and <fs type=”change”>.

After applying the N-gram-based correction, the text contents in the <TextEquiv><Unicode> elements should be replaced using XSLT scripts, in the above example “schlüpfrigh” with

"schlüpfrig". This ensures a clean separation between geometric and textual data. The corrected PAGE XML can then be uploaded via the Transkribus API and used as gold standard for domain-specific HTR training.

An analogous approach is also feasible with the open-source platform eScriptorium, which also supports PAGE XML as a standard exchange format. eScriptorium enables the import of transcriptions created outside the platform via the "Import/Transcription (XML)" function in the Images tab. The assignment between XML files and document pages occurs automatically based on file names and metadata contained in the XML structures. After external correction of the Unicode elements, the modified PAGE XML files can be re-imported as a ZIP archive. This cross-platform compatibility of the PAGE XML standard ensures that the developed N-gram-based correction methodology can be used regardless of the chosen HTR training environment[12].

4.4 Model training

Based on the gold standard data generated in this way, an adapted model should be built entirely from scratch. Training from scratch which typically requires large amounts of tokens, is fulfilled by the expected amount of training data.

For model training, Transkribus recommends a number of 5,000 words for printed texts, and at least 10,000 words for handwritten documents. However, the latter value only refers to material from one scribe. For documents originating from multiple scribes - which applies to the handwritten slips - at least 100,000 words are recommended. From preliminary work, as already described above, it is known that a handwritten slip contains around 14 words. This means that training should begin with a database of at least around 7,200 handwritten slips[13].

After each training round, the result should be tested using a selection of manually transcribed handwritten slips. To assess the progress of the model's quality, the CER (Character Error Rate) will be used, and in subsequent steps, the training basis will be continuously increased. In parallel, the results will also be evaluated by experts, since the CER, as already described above, is not suitable on its own due to the distributed information on a handwritten slip. Particular attention should be paid to the transcription results of content recorded in Teuthonista, as these present a special challenge. Through this iterative process, the quality improvement with increasing training material should be estimated during the creation of new HTR models.

If we consider the number of around 2.4 million slips that can potentially be used for training, then in terms of the available token count, we reach ranges in which large models also operate[14]. With a training basis of an estimated 33.6 million words, the prospects for a suitable HTR model with which the main catalog of the WBÖ "Wörterbuch der bairischen Mundarten in Österreich" (Dictionary of Bavarian Dialects in Austria) can be accessed should be good.

[12] eScriptorium: Import segmentation and transcription: https://escriptorium.readthedocs.io/en/latest/import/ (accessed: 2025-09-30)

[13] Transkribus: Data Preparation. Guide to preparing training data. https://help.transkribus.org/data-preparation (accessed: 2025-09-30).

[14] The currently largest model at Transkribus *The Text Titan I* comprises approximately 31 million words, https://app.transkribus.org/models/public/text/the-text-titan-i-ter (accessed: 2025-09-30).

5. Challenges and Expected Benefits

The implementation of the presented approach for HTR transcription of historical WBÖ slips promises methodological and practical advances for digital lexicography, but also harbors specific technical and methodological challenges.

5.1 Challenges

A primary challenge is translating the proposed workflow into a robust, scalable implementation of the theoretical approach presented here for accessing the main catalog.

The systematic empirical evaluation of various similarity metrics plays a fundamental role in the realization of the project.

The adaptation of the Character Error Rate to historical archival holdings such as those of the WBÖ, where information about the document is individually distributed across different text regions, also represents a specific desideratum. CER procedures must be developed that are applicable to this distributed information on the handwritten slips and thus provide a reliable metric for the quality of HTR models. These structured CER metrics are methodologically demanding and require extensive empirical validation.

5.2 Expected Benefits

The primary goal of the project is the development of an HTR model for automated transcription of the remaining 1.2 million non-digitized slips of the WBÖ main catalog. This model will enable, for the first time, complete digital access to the historical slip archive without time-intensive creation of training data through manual means and thus lay the foundation for a comprehensive digital WBÖ database.

In addition to pure text recognition, the development of a procedure is also sought that automatically identifies the different information categories of the handwritten slips (lemma, pronunciation, meaning, bibliography, source reference, etc.) and thus enables the generation of structured data. This would allow the results of the HTR process to be directly implemented into the existing TEI XML structure.

A particular focus also lies on the precise recognition of the Teuthonista phonetic script used in historical dialectology with its specific diacritical marks. The resulting specialized model for this phonetic type of notation could be made available as a publicly accessible resource for other dialectological research projects. For example, the methodological approaches and technical solutions could also be transferred to the sister project of the Bavarian Dictionary "Bayerisches Wörterbuch" (BWB) at the Bavarian Academy of Sciences "Bayerische Akademie der Wissenschaften" (BAdW), which has a comparable slip archive.

In general, the concept presented here and its future implementation should provide a template for lexicographic projects whose data basis consists of historical slip archives. This methodology is principally applicable to any dictionary project where, in addition to the actual analog slip archive, structured digital partial holdings already exist.

Software

Barabas, B., Hareter-Kroiss, C., Hofstetter, B., Mayer, L., Piringer, B. & Schwaiger, S. (2010). Digitalisierung handschriftlicher Mundartbelege. Herausforderungen einer Datenbank. In *Germanistische Linguistik 199–201. Fokus Dialekt. Festschrift für Ingeborg Geyer zum 60. Geburtstag*. pp. 47–60.

Bauer, W. & Kühn, E. (1998). Vom Zettelkatalog zur Datenbank. Neue Wege der Datenverwaltung und Datenbearbeitung im „Wörterbuch der bairischen Mundarten in Österreich". In C.J. Hutterer & G. Pauritsch (eds.) *Beiträge zur Dialektologie des ostoberdeutschen Raumes. Referate der 6. Arbeitstagung für bayerisch-österreichische Dialektologie, 20.–24.9.1995 in Graz*, number 636 in Göppinger Arbeiten zur Germanistik. Göppingen: Kümmerle Verlag, pp. 369–382.

Bowers, J. & Stöckle, P. (2018). TEI and Bavarian dialect resources in Austria: updates from the DBÖ and WBÖ. In A.U. Frank, C. Ivanovic, F. Mambrini, M. Passarotti & C. Sporleder (eds.) *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities (CRH-2)*, volume 1 of *Gerastree proceedings*. Wien, pp. 45–54.

eScriptorium Project (2025). *Import data into eScriptorium*. École Pratique des Hautes Études, AOROC. URL https://escriptorium.readthedocs.io/en/latest/import/. User documentation for data import in eScriptorium HTR platform.

Frank, A.U., Ivanovic, C., Mambrini, F., Passarotti, M. & Sporleder, C. (eds.) (2018). *Proceedings of the Second Workshop on Corpus- Based Research in the Humanities CRH-2*, volume 1 of *Gerastree proceedings*.

Hornung, M. & Bauer, W. (1983). *Wörterbuch der bairischen Mundarten in Österreich (WBÖ). Band 3: Pf – C*, volume 3. Wien: Verlag der Österreichischen Akademie der Wissenschaften.

Kahle, P., Colutto, S., Hackl, G. & Mühlberger, G. (2017). Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 04. pp. 19–24.

Kranzmayer, E. (1970). *Wörterbuch der bairischen Mundarten in Österreich (WBÖ). Band 1: A – Azor*, volume 1. Wien: Kommissionsverlag der Österreichischen Akademie der Wissenschaften.

Kranzmayer, E. (1976). *Wörterbuch der bairischen Mundarten in Österreich (WBÖ). Band 2: B(P) – Bezirk*, volume 2. Wien: Verlag der Österreichischen Akademie der Wissenschaften. Maria Hornung (Redaktion).

Lenz, A.N. & Stöckle, P. (eds.) (2021). *Germanistische Dialektlexikographie zu Beginn des 21. Jahrhunderts*. Stuttgart: Steiner. Unter Mitarbeit von A. Bergermayer, A. Gellan, S. Wahl, E. Wahlmüller & P. Zeitlhuber.

Puigcerver, J. (2017). Are Multidimensional Recurrent Layers Really Necessary for Handwritten Text Recognition? In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01. pp. 67–72.

READ-Coop (2022). The road to Handwritten Text Recognition – Part 1. URL https://blog.transkribus.org/en/insights/road-to-htr-1. Accessed: 2025-09-30.

READ-Coop (2023). Introducing Transkribus Super Models – Get access to the Text Titan I. URL https://blog.transkribus.org/en/introducing-transkribus-super-models-get-access-to-the-text-titan-i. Accessed: 2025-09-30.

Stöckle, P. (2021). Wörterbuch der bairischen Mundarten in Österreich (WBÖ). In A.N. Lenz & P. Stöckle (eds.) *Germanistische Dialektlexikographie zu Beginn des 21.*

Jahrhunderts. Stuttgart: Steiner, pp. 11–46. Unter Mitarbeit von A. Bergermayer, A. Gellan, S. Wahl, E. Wahlmüller & P. Zeitlhuber.

Transkribus (2025). Data Preparation. URL https://help.transkribus.org/data-preparation.

Wiesinger, P. (1964). Das Phonetische Transkriptionssystem Der Zeitschrift 'Teuthonista'. Eine Studie Zu Seiner Entstehung Und Anwendbarkeit in Der Deutschen Dialektologie Mit Einem Überblick Über Die Geschichte Der Phonetischen Transkription Im Deutschen Bis 1924. *Zeitschrift für Mundartforschung*, 31(1), pp. 1–20. URL http://www.jstor.org/stable/40500597.