TEI and Bavarian dialect resources in Austria: updates from the DBÖ and WBÖ

Jack Bowers, Philip Stöckle

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.), Second workshop on Corpus-Based Research in the Humanities (CRH-2). Vienna, Austria: Gerastree Proceedings, GTP 1.

TEI and Bavarian dialect resources in Austria: updates from the DBÖ and WBÖ

Abstract

In our paper, we present a large historical database of Bavarian dialects (from the *Dictionary of Bavarian Dialects in Austria*) and its conversion from hand-written paper slips via TUSTEP into TEI-XML while elaborating on the topics discussed by Bowers [2] with regards to enhancement of its contents. While the original purpose of the digitalization was to facilitate the writing of dictionary articles, our current TEI database will be used as a corpus from which the materials are being gathered to both write print dictionary articles as well as serving as a basis for a web-based lexicographic information system. Herein we trace the different steps that have already been taken to create our current digital database from a legacy data collection, discuss the challenges we are still facing, and describe the approaches we are taking and considering to address such challenges.

1 Introduction: A short history of the WBÖ

The *Dictionary of Bavarian Dialects in Austria* (Wörterbuch der bairischen Mundarten in Österreich 'WBÖ') is a long-term project, whose main goal is the comprehensive lexicographic documentation of the manifold Bavarian base dialects in Austria and South Tyrol. Shortly after its initiation in 1911, the language data in this collection was obtained either indirectly with the help of so-called collectors ("*Sammler*") on the basis of questionnaires ("*Fragebücher*") sent out by mail, or directly during field explorations ("*Kundfahrten*"), and was further complemented with excerpts from specialized literature. All data were written down on paper slips and collected in the main catalog ("*Hauptkatalog*"), which contains approximately 3.6 million entries. To date, five volumes of the WBÖ have been published, covering the entries *A-Ezzes*.[1]

With the main purpose of facilitating and accelerating the process of writing dictionary articles, the hand-written paper slips were entered manually into a TUSTEP system in the 1990's (Barabas et al. [1]) and, subsequently, converted into TEI-XML.

After the relocation of the WBÖ to the department 'Variation and Change of German in Austria' at the Austrian Academy of Sciences (ÖAW) – Austrian Centre for Digital Humanities (ACDH) in December 2016, a new

[1] For more information on the history of the WBÖ cf. Geyer [4] and Reiffenstein [5].

team is working on a revised and modernized conception of the dictionary, which will include a continuation of writing dictionary articles as well as the creation of a web based research platform.

2 Database & Content Description

The TUSTEP database system[2] played a major, and beneficial role in the evolution of the DBÖ (Datenbank der bairischen mundarten in Österreich) project contents. However, this system was reliant on an antiquated and complex database structure which required its own software and less than trivial programming language to search and extract the data. Additionally, results of these searches can often be inexplicably incomplete or inconsistent. Given that TUSTEP is self contained and can only be accessed internally (using the system's native programming language), such errors cannot easily be investigated or resolved in a system-independent manner. Moreover, the system and previous practices were carried out prior to the widespread availability of Unicode leaving the data in serious need of modernization in order to properly represent and make full use of its linguistic contents. Thus given the renewed need to more efficiently access, reuse and preserve this data, as well as to bring it more into line with contemporary principles for best practice in language markup, it was necessary to extract the data out of the increasingly obsolete system and convert it into a format that will both help ensure that moving forward, we were able to meet these needs.

Therefore in order to achieve this, the data was converted to TEI (TEI Consortium [6]) which is widely accepted in the digital lexicographic community as the de facto standard for the encoding of both retro-digitized and born digital dictionaries. As we describe below, the TEI has the capacity to encode the entirety of the legacy existing dataset and all its various data fields.

Conversion: Over a period of a year the database was converted in stages using a series of transformation processes using the XSLT language in which certain aspects of the data structure were addressed sequentially. Between each stage of transformation, both the effects of the transformations and the remaining contents of the data were thoroughly checked semi-manually which allowed us to encounter and investigate and log the

[2] TUSTEP is a set of word processing programs, a tool for scientific processing of text data (http://www.tustep.uni-tuebingen.de/and the Handbuch TUSTEP 2017). It was (ab)used by the DBÖ team as a database because of its macro capabilities.

remaining flaws in the content needing to be addressed in future stages of the transformation.

Improvement to the Data Structure: The conversion process did not only involve the interchange from one data format to the other, nor were the benefits limited to issues related to data access issues endemic to TUSTEP. The improvements were achieved and permitted by: the correction of human errors; enhancements in the data structural efficiency inherent to TEI XML markup vocabulary (in contrast to TUSTEP); technological advancements since the first digitization in the 1990's and the refinement of certain flaws in the content structure from the original project guidelines.

Human Error: Because of the particularities in the TUSTEP data structure and labelling, the size of the database, the duration of the project and the large number of different individuals who worked on the process of digitizing the entries from the notecards into the original database, there was a large degree of irregularity due to idiosyncratic practices as well as simple typing errors. Of the 510 data field tags present in the initial export from TUSTEP, 197 of them were found to be due to human error (either by way of typos or non-adherence to the project guidelines). However, whereas the sheer number of these incorrect tags is large, with a few exceptions, the vast majority of them had less than 10 instances.

TEI XML improvement to TUSTEP-Inherent Structural Flaws: In addition to the structural errors identified and corrected due to human error, the contents adhering to the project guidelines comprised of 313 unique field tags occuring in hundreds of thousands of entries. The sole reason for this extremely high number of different tags was due to the nature of the TUSTEP database structure, which is a flat sequence of unique fields and that has no means of pointing between or expression relations between different specific instances of different fields outside of the name of the field itself. One of the foremost benefits of using TEI, as given that it is an XML vocabulary, it can readily solve this issue by making use of attributes, which can be used for labelling and/or pointing and nested data structure to reduce the excess data complexity necessitated by TUSTEP.

Numbering: In a TUSTEP entry the data field tags are simply a single string of uppercase letters (and possibly digits) encased in asterisks, e.g. "*HL*" is the "Hauptlemma"; ("*headword*") tag. Numerous different fields often could occur more than once, for example an entry could have up to ten dialect forms, and even though the content was the same in nature, and there is no reason a user would ever specifically want to search for a specific

numbered instance of the category, in TUSTEP they were required to have unique tags, e.g. *LT1*, *LT2*, *LT3*, etc.

Example 1. Numbering tag labels in TUSTEP

Thus, what was in TUSTEP *LT1*...*LT10* (all distinct tags), is in TEI represented as <form type=”lautung”> and each unique number be expressed using the number attribute: @n.

```
<form type="lautung" n="1">
  <pron notation="tustep">Züglen</pron>
</form>
<form type="lautung" n="2">
  <pron notation="tustep">zigeln</pron>
</form>
<form type="lautung" n="3">
  <pron notation="tustep">zigln</pron>
</form>
```

Example 2. TEI version of entry with multiple dialect forms

Nesting: Where in TUSTEP, there is a complementary or supplementary category that modifies or adds to a field above (e.g. *translations of example sentences, comments, location, miscellaneous notes, references*, among others) these categories need to specify the tag they pertain to within their tag name as well, e.g. “bedutung kontext 1” (*“meaning usage context 1”*) would be *BD/KT1*.

Example 3. TUSTEP entry with complimentary field *BD/KT1*

In TEI these relations are encoded as nested elements with the complimentary content of the main field nested within the latter. Given the fact that the sub-ordinate relationships between nested elements and their parent are

defined as part of the fundamental data model of XML, the TEI conversions of these contents do not need to maintain and further reference to the target element. In the TEI version, the translations of a usage example is encoded in the definition element <def> and labelled with the language attribute the value of which is the ISO 639-2 code for High German "de".

```
<cit type="kontext" n="1">
  <quote>in Ausdzu¡g [m,sg4] g;ei¡n</quote>
  <def xml:lang="de">in den Auszug gehen</def>
</cit>
```

Example 4. TEI translation of usage context example

Pointers: Many fields such as: *meaning, usage context, context-specific sense, references, notes*, *etymology,* can apply to one or all of the fields given in the entry, for each specific variant, a unique field tag was required, so there existed tags such as: *BD/LT2/LT3*. Any one form could have (1..n) meanings as well which would be represented along the same lines: e.g. *BD2/LT2/LT3*[3].

```
====
*LT1* Hamperl [D,m]
*LT2* Hamperl [D,n]
*BD/LT1/LT2* allzu nachgiebiger Mensch; Siemannldumme
Person; Blödling
*BD2/LT1/LT2* dumme Person; Blödling
**********
```

Example 5. TUSTEP entry with complex referential tags

In certain contexts it is not appropriate to nest in TEI and instead it is better to use pointers to express relations between content. In such cases, this was done in TEI using a pointer attribute-value combination with the @corresp making use of the TEI prefix definition (<prefixDef>[4]) scheme to point to the specific corresponding content in a predefined structure within an entry. Note also in the example below that in the TEI version the content in brackets from TUSTEP which is the grammatical information ([D,m] 'diminutive,

[3] The initial export had 123 variants of the "bedeutung" ("*meaning*") field and TUSTEP does not allow for partial string searches of tag content, which means that to search for specific content within all of those variants, one would have to specify each one, or just run a string search of the desired contents without specifying the fields resulting in a large number of false positives and greatly increasing run-time.

[4] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-prefixDef.html

masculine' & [D, n] 'diminutive, neuter') for each form is moved from the line with the form itself in the TUSTEP to its own element block <gramGrp><gram>.

```
<form type="lautung" n="1">
  <pron notation="tustep">Hamperl</pron>
  <gramGrp><gram>[D,m]</gram> </gramGrp>
</form>
<form type="lautung" n="2">
  <pron notation="tustep">Hamperl</pron>
  <gramGrp><gram>[D,n]</gram></gramGrp>
</form>
<sense n="1" corresp="this:LT1 this:LT2">
  <def xml:lang="de">allzu nachgiebiger Mensch; Siemannldumme Person;
Blödling</def>
</sense>
<sense n="2" corresp="this:LT1 this:LT2">
  <def xml:lang="de">dumme Person;Blödling</def>
</sense>
```

Example 6. Pointing to non-adjacent contents in TEI

Attribution: Additionally, certain feilds distinguished the editorial responsibility of its contents e.g. *ANMO* "anmerkung original" ("*comment by the original editor"*) and *ANMB* "anmerkung bearbeiter" ("*comment by the editor"*). It was common and possible to have combinations of many of these complex tags as well e.g. *VRWO/BD/LT1* ("*reference - original editor - meaning form one*"). In TEI this tag feature was converted to the responsibility attribute (@resp).

```
<sense corresp="this:LT1">
  <def xml:lang="de">abfärben; die Farbe abgeben, verlieren</def>
</sense>
<note type="anmerkung" resp="O" corresp="this:BD">(z.B.: die Wand färbt
ab)</note>
```

Example 7. TEI @resp

Conclusion on Conversion: Thus, among the most significant achievements of the conversion of the database is the reduction of the number of data field tags from 510 to 37. This, in combination with the use of a BaseX database system using the XQuery language (which are both open source and have extensive online user resources) the conversion to TEI allows us to greatly improve: the process of searching and manipulating the

contents of the data are greatly simplified with a greatly improved level of granularity over those of the TUSTEP internal database search system; and our ability to accurately document the contents of our database for new users. Additionally, the TEI community is constantly growing and more and more projects are adopting conversions to and from it thus its use helps in reaching a wider audience. Given the structural improvements to the data, and the fact that the XML markup language and the TEI guidelines are open source, systematically documented, our data now has a much higher degree of long-term sustainability as well as compatibility with potential partner projects.

Remaining Issues and Ongoing Work in the DBÖ: As described above, to this point, the contents of DBÖ have been greatly improved in structure, consistency, accessibility have been greatly improved in the conversion. However, due to the legacy data structure, a number of significant issues which inhibit the quality of the resource remain. Some of the most notable of which are as follows.

The transcription notation of both the headwords and similar forms, as well the phonetic forms do not correspond to any standard and in many cases they are not entirely human readable and/or are complex to search for directly, often requiring the use of regular expressions. These are in the process of being normalized. Such changes will allow us to both maintain the linguistically significant morphological segmentation information while allowing users to search and retrieve the contents. The example below shows a complex compound headword what was previously just the form in <orth type=”orig”> which will be enhanced as follows:

```
<form type="hauptlemma">
 <orth type="orig">(Amts─pflicht)for─halt</orth>
 <orth type="parsed">
   <seg>
     <seg>Amts</seg><seg>pflicht</seg>
   </seg>
   <seg>for</seg><seg>halt</seg>
 </orth>
 <orth type="normalized">Amtspflichtforhalt</orth>
</form>
```

Example 8. Normalized and segmented headwords in TEI

The phonetic dialect forms will be converted from a TUSTEP interpretation of the original Teutonista script and characters to fully Unicode Teutonista: for example, what is currently: "d-.es" will be converted to: "dė¯s".

Loanwords (which are in fact dialect forms), are expressed in a different category from the rest of the forms. These will be converted and the loanword information will be included in <eytm type=”loanword”>. Several categories containing multiple distinct fields of information have not been entirely decomposed. There are several thousand entries with no headword, many dialect forms are not directly accessible as many entries have only an example in contextual usage within which the dialect form is not explicitly tagged. Many areas of the data structure are not in line with various international standards for language markup. Finally because of the lack of consistency in both the form-related contents, sense related contents, and the nature of the questionnaires used to elicit the data, there remains a significant gap in the means in which users can search for both semasiological (form-based) and onomasiological (concept-based) contents. To alleviate this we are working on creating a normalized inventory of semantic labels.

3 Resuming WBÖ publication & Creation of an Online DB

The planned output of the future WBÖ work is twofold: On the one hand, the WBÖ staff will continue to write “classical” dictionary articles, which will, however, appear in a revised and modernized form. These revisions include a more standardized structure of the articles, a modernized layout, more condensed and generalized information about pronunciation, etymology and geography. For each ‘Hauptlemma’ which will enter the dictionary as a headword, the semantic information is categorized, as well as phonetic variants, geographic distribution and more information essential to the dictionary articles.

Another goal for this project is to create a comprehensive online lexicographic information system, i.e. a (re)search tool for professional linguists and the general public, where users will be able to perform queries regarding different aspects of the database, both linguistic, (i.e. lemma, sense, etc.), metalinguistic (i.e. geographic location/region) as well as legacy materials such as scans of the original paper slips or scans of questionnaires. This lexicographic information system will be integrated into the SFB

research platform (DiÖ [3]), thus providing a multi-perspective approach to language variation in Austria.

Moreover, the articles will be accessible via the online lexicographic information system, which makes them directly linkable with the different types of information stored in the database. In addition to semasiological research which is characteristic for dictionaries (i.e. different meanings connected with one lemma), it will allow users also to perform onomasiological queries, i.e. different linguistic forms connected with the same semantic concept (such as '*Fasching*', '*Fastnacht*' and '*Fasnacht*' meaning carnival).

Interface of Classical Dictionaries and Digital Humanities: On the back-end, the data structure that will be used will also be TEI, though, it will involve the creation of a much more complex set of entry templates in order to accommodate the various different data fields common in dictionary articles. While the use of the TEI dictionary module is of course well established in accommodation of both retro-digitized and born digital dictionary content, this usage will represent a rather novel usage of the standard in a two ways.

First, while it is of course common for print dictionaries to be retro-digitized, it is less common (perhaps even unprecedented) for the print dictionary to be compiled, generated and edited first in TEI. Second, given the inherent complexity of the contents of a dialect dictionaries, this usage of the TEI as a digital template (or templates) for the creation of such information provides an opportunity to balance out a number of issues that often are in conflict in such projects. Such issues include: the structural and content demands of print dictionaries and those of the digital data structure (*i.e. best practice in TEI markup*); editorially, the potential conflict in the usability for non-experts tools needed to edit and create articles in TEI directly versus those used in traditional practice (*e.g. basic word processing*).

Conclusion

In our paper we present a large historical database of Bavarian dialects (DBÖ) and give an example-based overview of the TEI structure, contents and remaining issues pertaining to the revised TEI-XML dataset. Additionally we introduce the plans already underway for a revived print version of the WBÖ and the creation of an online publicly searchable version of the database which will both be structured and edited within task-specific TEI templates. Finally, we discuss the challenges we are still facing, and the

approaches we are taking and considering in order to address such challenges. Our project offers potential insights for the use of the TEI vocabulary for such tasks.

References

[1] Barabas, B., Hareter-Kroiss, C., Hofstetter, B., Mayer, L., Piringer, B. and Schwaiger, S. (2010) Digitalisierung handschriftlicher Mundartbelege. Herausforderungen einer Datenbank. In *Germanistische Linguistik* 199-201, pp. 47–64.

[2] Bowers, J. (2017) TEI conversion of Bavarian dialect lexical *resources: insights, observations, challenges, next steps. Presented at the COST-ENeL Meeting*, Budapest.

[3] DiÖ (2017) Task Cluster E: Collaborative Online Research Platform. In *DiÖ-Online*. Available at: https://dioe.at/en/article-details/ [Accessed on 14th October 2017]

[4] Geyer, I. (2004) Arbeitsbericht zum Wörterbuch der bairischen Mundarten in Österreich. In: Gaisbauer, S. and H. Scheuringer eds. *Linzerschnitten. Beiträge zur 8. Bayerisch-österreichischen Dialektologentagung, zugleich 3. Arbeitstagung zu Sprache und Dialekt in Oberösterreich, in Linz, September 2001*, pp. 583–588, Linz.

[5] Reiffenstein, I. (2004) Die Geschichte des "Wörterbuchs der bairischen Mundarten in Österreich" (WBÖ). Wörter und Sachen im Lichte der Kulturgeschichte. In: Hausner, I. and Wiesinger, P. eds. *Deutsche Wortforschung als Kulturgeschichte. Beiträge zum Symposium "90 Jahre Wörterbuchkanzlei" der Österreichichischen Akademie der Wissenschaften, Wien, 25-27. September 200*3, pp. 1–13, Wien.

[6] TEI Consortium, eds. (2016) *TEI P5: Guidelines for Electronic Text Encoding and Interchange. [Version 3.1.0]*. [Last updated on 15th December 2016]. TEI Consortium. Available at: http://www.tei-c.org/Guidelines/P5/ [accessed on 13th February 2017].