ÖAW ACDH

Austrian Centre for Digital Humanities

dbo@TEI

New access to dialectal language data: From *TUSTEP* to <xml>

Introduction

The Wörterbuchkanzlei at the Austrian Academy of Sciences was founded to issue an Austrian dialectal dictionary (WBÖ). The collection of the Bavarian dialectal data, written on paper slips, started in 1911. From 1993 to 2011 these paper slips (approx. 3.5 million in 733 drawers) were manually entered in a database (DBÖ) based on the programme TUSTEP, a smaller part was put in, or rather transferred to a MySQL-database (dbo@ema). In line with the project exploreAT! these data are to be converted to a common format, compatible with one another and with the requirements of LOD.

```
**********
*A* HK 220, e220#234.1 = enke0527.pir#51.1
*HL* Enkel#I:1
*QU* Fusch Huber
*QDB* {4.2c01} Fuschert.:UPinzg.:Pinzg.:Sa. *@ FbB.HUBER·
(1913-30) [SFb./EFb./Mtlg.] *O* Fusch Sa.
===
*NR* 32A14: Gelenk als Gw. (Hand-, Fuß-, Fingergelenk)
*LT1* Enk¡hl
*BD/LT1* Fußgelenk
**********
```

Fig. 1. TUSTEP-entry with filled in data fields (field labels enclosed in asterisks)

Linguistic data / Metadata

- Location data for each record
- Source data for excerpts
- Phonetic representation(s) of dialectal forms
- Grammatical information, e. g. gender, case etc.
- Definition of meaning
- Etymological information
- Folkloristic information, e. g. old traditions

Challenges in converting

- Heterogeneous formatting in the records, e. g. data of the same category in different data fields
- Invalid non-Unicode characters in the text output
- Phonetic transcriptions are not in consistent notation due to different handling of the collectors
- Data scientists used different conventions when entering the information into the database

```
<record n="234">
  <field name="A">HK 220, e220^#234.1 = enke0527.pir^#51.1</field>
  <field name="HL">Enkel^#I:1</field>
  <field name="QU">Fusch Huber</field>
  <field name="QDB">{4.2c01} Fuschert.:UPinzg.:Pinzg.:Sa. FbB.HUBER·
(1913-30)  [SFb./EFb./Mtlg.]</field><field name="O">Fusch Sa.</field>
  <field name="NR">32A14: Gelenk als Gw. (Hand-, Fuß-, Fingergelenk)</field>
  <field name="LT1">Enk¡hl</field>
  <field name="BD/LT1">Fußgelenk</field>
  <orig>*A* HK 220, e220^#234.1 = enke0527.pir^#51.1
        *HL* Enkel^#I:1
        *QU* Fusch Huber
        *QDB* {4.2c01} Fuschert.:UPinzg.:Pinzg.:Sa. FbB.HUBER· (1913-30)
        [SFb./EFb./Mtlg.] *O* Fusch Sa.
         ===
        *NR* 32A14: Gelenk als Gw. (Hand-, Fuß-, Fingergelenk)
        *LT1* Enk¡hl
        *BD/LT1* Fußgelenk
</orig></record>
```

Fig. 2. Basic xml-record. The TUSTEP data fields are transferred into xml-attributes.

```
<item xml:id="d1e437" n="21">
  <ref type="archive">HK 220, e220^#234.1 = enke0527.pir^#51.1</ref>
  <cit type="etymon">
    <pRef>Enkel</pRef>
    <gramGrp><pos sameAs="lemma_wortat-tei-fs.xml#Sub">Sub</pos></gramGrp>
  </cit>
  <ref type="source">Fusch Huber</ref>
  <bibl>
    <ref>4.2c01</ref> Fuschert.:UPinzg.:Pinzg.:Sa. FbB. HUBER·
(<date notBefore="1913" notAfter="1930">1913-30</date>) [SFb./EFb./Mtlg.]
  </bibl>
  <placeName>Fusch Sa.</placeName>
  <ref type="questionnaire">32A14: Gelenk als Gw. (Hand-, Fuß-, Fingergelenk)</ref>
  <cit type="lautung" xml:id="d1e451" n="1">
    <pRef>Enk¡hl</pRef>
  </cit>
  <cit type="translation" corresp="#d1e451">
    <oRef xml:lang="de">Fußgelenk</oRef>
  </cit>
</item>
```

Fig. 3. TEI-list version of annotated xml-record

Defined objectives

- Offering means to base semantics for a given concept in predefined knowledge hubs
- Access to dialectal data according to different features (lexical, semantic, geographic, temporal etc.)

Poster creators: Jack.Bowers@oeaw.ac.at
Ingeborg.Geyer@oeaw.ac.at
Barbara.Piringer@oeaw.ac.at

Austrian Centre for Digital Humanities
Austrian Academy of Sciences

acdh@oeaw.ac.at
acdh.oeaw.ac.at