TEI and the Mixtepec-Mixtec corpus: data integration, annotation and normalization of heterogeneous data for an under-resourced language
Jack Bowers, Laurent Romary

To cite this version:
Jack Bowers, Laurent Romary. TEI and the Mixtepec-Mixtec corpus: data integration, annotation and normalization of heterogeneous data for an under-resourced language. 6th International Conference on Language Documentation and Conservation (ICLDC), Feb 2019, Honolulu, United States. hal-02075475

HAL Id: hal-02075475
https://inria.hal.science/hal-02075475
Submitted on 21 Mar 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
TEI and the Mixtepec-Mixtec corpus: data integration, annotation and normalization of heterogeneous data for an under-resourced language

Jack Bowers 1,2,3
iljackb@gmail.com
https://github.com/iljackb/Mixtepec_Mixtec

Laurent Romary 2,3,4

1 Austrian Center for Digital Humanities (ACDH)
2 Inria – Team ALMAnaCH
3 École Pratique des Hauts Études – Paris
4 Berlin Brandenburgische Akademie der Wissenschaften (BBAW)
Mixtepec-Mixtec (Sa'an Savi)

• *Sa’an Savi* ‘rain language’
• ISO 639-3 code: ‘mix’
• Oto-Manguean
• San Juan de Mixtepec, Juxtlahuaca district (Oaxaca, MEX)
• Estimated 9-10,000 speakers; +~3000 in California, Oregon, Washington, Arkansas
• Status “vigorous” (Ethnologue 21st edition) but “threatened” (ELDP via Ethnologue 17th edition)
• Has been studied by: Pike and Ibach (1978); Paster and Beam de Azcona (2004-2007); Beckman and Nieves-SIL (2005-current)

Primary Sources of Language Data

• Consultation with Speakers
• Recordings made by speakers of others
• ~500 recordings, ~10hrs
• ~ 40 Booklets by SIL Mexico
• Public sources (Facebook, YouTube, etc.)
• Academic papers on language (5)
• Pamphlets from Mexican Government
• Bible (1 book) by Scripture Earth (+1hr spoken; 34,000 words)
• Personal Communication (texts, emails, facebook messages, etc.)
Desired Outcomes

• Create an open source collection of reusable and extensible LR
• Produce corpus-based descriptions and analyses of various aspects language’s features to further the knowledge of all aspects of the language
• Demonstrate and evaluate the application of encoding and standards (particularly TEI) on an under-resourced non-Indo-European language
• Collect enough data so that it can be re-used for creating learning material
Specific Output

• Searchable (TEI) corpus from text- and time-aligned spoken sources
• Multilingual Digital (TEI) Dictionary (Mixtec, English, Spanish)
• Annotated (TEI) files of SIL booklets
• Lexical feature inventory
• Phonetic feature inventory
• Open Archive of all media and other files (Dataverse, OLAC)
• Concepts inventory
• Place inventory
• Person list
Challenges Intrinsic to Studying Mixtepec-Mixtec

- Lack of existing resources (under-resourced)
- Need to gather from many heterogeneous sources
- Consultants limited availability to edit, gloss text, etc.
- Data exchange and communication with SIL researchers working on the language in Mexico have has been limited (though I do have agreement to reuse SIL publications)
- Orthography (in development by SIL) not fully conventionalized, still changes, speakers often not aware of/don’t use the standards (requires significant normalization in markup)
- Lexical tone, adds complexity to characterization and it is (mostly) not represented in the orthography (lot’s of homographs), creates irregularity in transcriptions
- Not enough data to automate annotation!
Overview of the Text Encoding Initiative (TEI)

• XML vocabulary began in 1987
• **Digital Humanities de facto standard** (open source consortium), used in lexicography, library sciences, digitizing archives, linguistic corpora, time-aligned speech
• **Dictionary well developed** (and increasingly widely adopted)
• Pre-existing tag vocabulary, schemas, documentation, (w/tools Oxygen XML editor) automatic validation, content completion, templates
• Enables **common corpus structure** for all levels of language (makes searching and retrieving easier)
• **Indepth metadata** available in header
• No limits on number of annotation categories
• Only need to document practices (not likely unique)
• **Freely available documentation**, mailing list, stable community with journal (jTEI), growing active users
• **Evolving to fit needs of users**, can submit issues, proposals to change (via GitHub)
• Easy to generate **human readable formats** (html, and/or using CSS)
• **Not software or platform dependant** (data-centered not platform-centered)
TEI for LD and Indigenous Languages

- **Underdeveloped** usage of standard
- **Mostly European languages**, increasing in Asia (esp. Japan) (only Czaykowska-Higgins et al. 2014 & Bowers and Romary 2018)
- **Already has appropriate components to handle all major aspects of LD project** (however nobody has previously used all), thus refinement needed
- **Enables integrated treatment** of variants, normalization, conflicting description, speaker attribution, etc.
- **Allows customization**


Mixtec Data: Sources, Workflow, Output

**TEI: Corpus**

- **SIL Documents**
- **Misc Sources**
- **Utterance/Time aligned speech**
- **Academic Papers**
- **Original content by speakers**

**TEI: Dictionary**

- **Phonetic & Phonological feature inventory**
  - Descriptions
- **Lexico-grammatical feature inventory**
  - Descriptions
- **Speaker inventory**
  - listPerson
- **Place inventory**
  - listPlace

**Sources**

- **SIL Booklets**
- **Mexican Gov**
- **Academic papers**
- **Recorded speech**
  - Praat textgrid
  - (praat) tsv
- **Scripture Earth**
- **Quantitative Phonetic Data**
- **Daily Journals, etc.**
A insect was stuck on his hat.

Se pegó un insecto en el sombrero de él.

An insect was stuck on his hat.

Se pegó un insecto en el sombrero de él.

 Tokens 3

 Tokens 4

 Tokens 1

 Tokens 4

 Tokens 1

 Tokens 1

 Tokens 1

 Tokens 1

 Tokens 1

 Tokens 1

 Tokens 1

 Tokens 1
<span xml:lang="es" target="#T-seg-orth-6.13 #T-seg-pron-6.13">Se pegó un insecto en el sombrero de él.</span>

<span xml:lang="en" target="#T-seg-orth-6.13 #T-seg-pron-6.13">An insect sat on his hat.</span>
Standoff Annotation in TEI: <spanGrp> & <linkGrp>:

**<spanGrp>** is used to annotate: Translations (*English, Spanish*), grammar, Semantics (multiple aspects), Interlinear glossed text, General editorial notes

**<linkGrp>** links (via **<link @target>**) pre-existing translation content

- Point to separated language content (usually `<w>` or `<seg>`)  
- Requires `@xml:id` for all values to be annotated  
- Can be included in most TEI be inserted close to target content  
- Structure and tag content correspond to feature structure inventory `<fs>`
Nu'u Ncha'i ka

Yee nu'u tsi chikuii nuu Nu'u Ncha'i. Yee kua'a ka chikuii cha xoo ka nu'u. Yee nu'u luu ka nania "islas", cha inkai ma'i chikuii. Cha nu'u ka'nu ka nania "continente". Yee iñu "continente" nania: África, América, Antárctica, Asia, Europa tsi Oceanía.
There is land and water on the Earth.

Hay tierra y agua en la Tierra.

there is

hay

land

tyerra

and

y

water

aguad

Earth
Sil Documents: Prose annotation

The text contains annotations for grammar, with spans indicating parts of speech and other grammatical information.
<entry xml:id="bird-mourning_dove">
  <form type="lemma">
    <orth xml:lang="mix">lakuku</orth>
  </form>
  <pron xml:lang="mix" notation="ipa" cert="medium">la\'ku\'ku</pron>
  <ptr type="sound" target="/media/N_mourning_dove.wav"/>
  <gramGrp>
    <pos>noun</pos>
  </gramGrp>
  <sense corresp="http://dbpedia.org/resource/Mourning_dove">
    <usg type="domain" corresp="http://dbpedia.org/resource/Animal">Animal</usg>
    <xr type="hyponymOf">
      <ref corresp="#bird" xml:lang="en">bird</ref>
      <ref corresp="#bird" xml:lang="mix">saa</ref>
    </xr>
    <ref type="sense" corresp="http://dbpedia.org/resource/Bird"></ref>
    <cit type="translation">
      <form><orth xml:lang="en">mourning dove</orth></form>
    </cit>
    <cit type="translation">
      <form><orth xml:lang="es">tortolita</orth></form>
    </cit>
    <cit type="example" corresp="/SIL_docs/L152/L152-tok.xml#L152-01-01">
      <quote xml:lang="mix">In kii ra in <oRef>lakuku</oRef> kunia tanta'i tsi in ncho'o, cha koo xu'in sa'i viko.</quote>
    </cit>
  </sense>
</entry>
**Lakuku** (noun)  
[laJku1ku1]  
(Animal) type of Bird in Saa  

*En.* mourning dove  
*Es.* tortolita  

**Example:**  
In kii ra in *lakuku* kunia tanta'i tsi in ncho'o, cha koo xu'in sa'i viko.  
  
  *En.* One day a mourning dove wanted to get married to a hummingbird, but didn’t have the money for the party.  
  *Es.* Un día una tortolita quería casarse con un colibrí, pero no tenía dinero para la boda.

Editing and Normalizing Speaker Authored Sources:

**Source written by speaker**

Tsika *kuan *yuu *ka *un treen ka

  kua’an  yu   kaa  nuu

**TEI Encoding**

<seg xml:id="d1e13171" xml:lang="mix" resp="#TS" type="S">
  <w xml:id="d1e13172">Tsika</w>
  <w xml:id="d1e13174" orig="kuan">kua’an</w>
  <w xml:id="d1e13176" orig="yuu">yu</w>
  <w xml:id="d1e13178" cert="high" orig="ka">kaa</w>
  <w xml:id="d1e13180" orig="un">nuu</w>
  <w xml:id="d1e13182">treen</w>
  <w xml:id="d1e13184">ka</w>
</seg>

- Tag speaker/author of MIX content using @resp
- Correct/normalize spelling, preserve original with @orig
- (also @norm available to keep original and stored normalized form in attribute)
- Editorial certainty indicated with @cert
Integrating Content from Academic Papers: Pike & Ibach (1978)

Paper is the original benchmark for the language’s tonal system

Contains a significant amount of vocabulary examples and their tones

However a lack of standardization in both their phonetic alphabet and their tone characterization creates an enormous amount of work to normalize and integrate into project’s data model (IPA & TEI)

PDF source scanned, OCR works on English but not Mixtec

Doesn’t use normal standards of phonetic representation!

### Non-IPA

<table>
<thead>
<tr>
<th>Non-IPA</th>
<th>Tone levels are reversed from conventional description</th>
<th>Strings are often interrupted</th>
</tr>
</thead>
<tbody>
<tr>
<td>ȼ</td>
<td>= ts</td>
<td>Š[i.]²ši³-ʧi² '</td>
</tr>
<tr>
<td>š</td>
<td>= Ŧ</td>
<td>K[o.]¹lo¹</td>
</tr>
<tr>
<td>č</td>
<td>= tʃ</td>
<td></td>
</tr>
<tr>
<td>ź</td>
<td>= tz</td>
<td></td>
</tr>
<tr>
<td>ĵ</td>
<td>= dʒ</td>
<td></td>
</tr>
<tr>
<td>ɐ</td>
<td>= ľ</td>
<td></td>
</tr>
</tbody>
</table>
In Mixteco Mixtec\(^1\) the minimal phonological word is made up of the sequence of two syllables – a couplet. This couplet is the complex nucleus of the word; the first syllable is marked phonologically by a lengthened vowel unless that vowel is preceding /ʔ/: k\(\text{[o.]}\)\(^{1}\)lo\(^{1}\)ko\(^{1}\) ‘our (excl.) male turkey’, s\(\text{[i.]}\)\(^{2}\)Si\(^{3}\)-e\(^{2}\) ‘his or her (child) aunt’, s\(\text{[o.]}\)\(^{3}\)ko\(^{3}\)-yu\(^{3}\) ‘my (polite)……..

**TEI Encoding**

```xml
<seg xml:id="d1e24" xml:lang="mix" notation="ipa">
  <w xml:id="d1e25" orig="k\[o.\]¹lo¹">ko˥lo˥</w> <w xml:id="d1e27" orig="ko¹">ko˥</w>
</seg>

'\n
<seg xml:id="d1e28" xml:lang="en" notation="orth">
  <w xml:id="d1e29a">our</w> <w xml:id="d1e29b">(excl.)</w>
  <w xml:id="d1e30a">male</w> <w xml:id="d1e30b">turkey</w>
</seg>',
```

**Source pdf (scanned)**
Challenges in integrating diverse resources

- “Lone wolf” & data logjam
- Adding new diverse data as found requires time and new workflows
- Where to stop!?
- Multi-layered linguistic annotations and translations time consuming;
- Normalization of extensive variation in orthography, and phonetic transcriptions from external sources;
- Moving targets (changing orthography) & as of last yr + nobody to make suggestions to regarding orthography
- Use of emerging technology no automatic tools, and ever increasing diversity of data structure input makes limits capacity for automation and make manual treatment necessary
Key Conclusions

- TEI has the capacity to handle needs of LD data, representation;
- Applying TEI to LD and indigenous language strengthens the standard;
- There is a need for existing tools (Flex etc.) to enable TEI output (currently only EXMARaLDA allows TEI output);
- LD projects would benefit from adopting TEI as working or output data format in the interest of compatibility, interchange and stability;
- Time needed to process and deal with all of the aforementioned issues makes my work less user friendly for the community.
Moving Forward

• Need to make data user friendly, and accessible to community (online portal)
• Make reusable for the creation of learning materials
• Find non-specialist tools for editing the data
• Pursue grants to continue project, hire speakers as editors and contributors
• Utilize phonetic material to train ML to automate annotation
• Bootstrap one translation language to produce the other (where only one is present)
• Automate collection of conceptual data (for annotations) w/ webscraping, other methods
• Need for documentation projects to have long-term view in the data format outputs (standards)
• TEI stabilization of annotations will give rise to tools to ease encoding process
Mahalo! Tatsa’vi kueni!

Special thanks to Geremaia and Tisu’ma Salazar for sharing their language with me!

iljackb@gmail.com
https://github.com/iljackb/Mixtepec_Mixtec