Language Documentation and Standards in Digital Humanities: TEI and the documentation of Mixtepec-Mixtec

Jack Bowers

Language Documentation and Standards in Digital Humanities: TEI and the documentation of Mixtepec-Mixtec

Jack Bowers

jack.bowers@oeaw.ac.at

https://github.com/iljackb/Mixtepec_Mixtec

Austrian Center for Digital Humanities (ACDH)

Inria

École Pratique des Hauts Études - Paris

Working draft –version 2019/02/20

INVENTEURS DU MONDE NUMÉRIQUE

ÖAW AUSTRIAN ACADEMY OF SCIENCES

Mixtepec-Mixtec (Sa'an Savi)

- Sa'an Savi '*rain language*'
- ISO 639-3 code: 'mix'
- (Family): Oto-Manguean, Mixtecan, Mixtec-Cuicatec, Mixtepec-Mixtec

• San Juan de Mixtepec Juxtlahuaca district (Oaxaca, MEX) (also spoken in Puebla & Guerrero states)

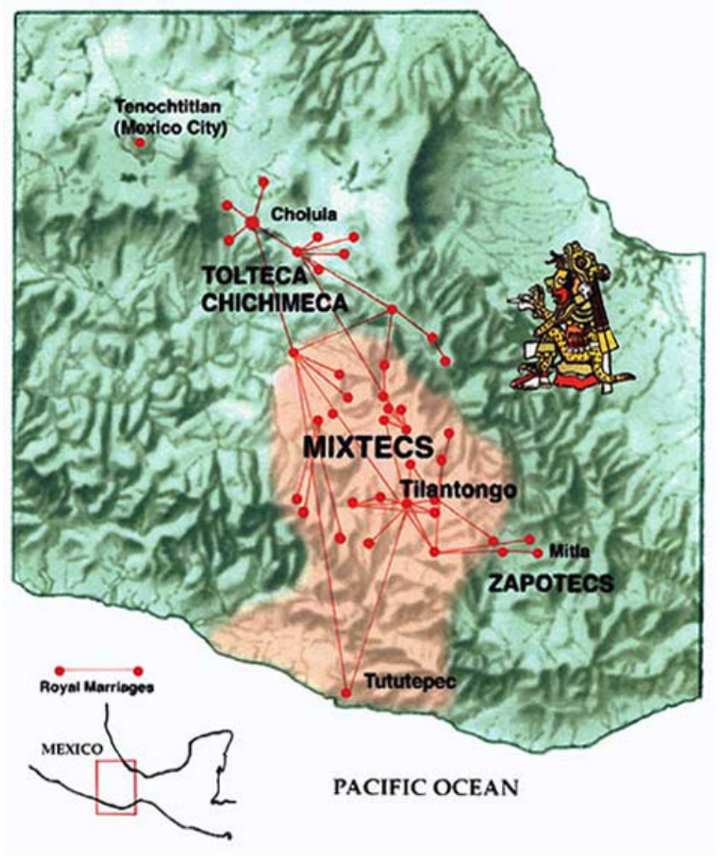
• Spoken data mostly collected in sessions working with speakers from a small village called Yucunani in the San Juan Mixtepec municipality

•status "vigorous" (source: Ethnologue 21st edition)

•Estimated +-7,611 speakers; *Source: INEGI (2010);(though probably several thousand more when considering speakers in US)*

Has been studied by:

- *Pike and Ibach (1978); Paster and Azcona (2004-2007); Beckman and Nieves-SIL (2005-current)*



On Mixtec Languages

- 52-85 Mixtec varieties! (Padgett, 2017); (Simons & Fennig 2017) & (INALI 2015:132-147)
- Unclear (possibly undefinable) boundaries between linguistic variant typology
- Tonal, most have at least 3 level tones
  - MIX has: L, M, H, F, R, *FR, *RF
- Varieties of Mixtec polysemy, spatial semantics and body part terms provide key examples to cogntive linguistics theory, particularly for language change and it's link to conceptualization (cf: Brugman, 1983; Brugman and Macaulay, 1986; Hollenbach,1995; Johnson, 1987; Lakoff and Johnson, 1989; Langacker, 2002; Bowers, 2016 (http://bit.ly/2FnsKPU); Bowers (forthcoming))

Desired Outcomes

- Create an open source body of reusable and extensible collection of multimedia language resources in the Mixtepec-Mixtec language
- Further the knowledge of all aspects of the language itself
- Demonstrate and evaluate the application of encoding and standards on an under-resourced non-Indo-European language
- Produce and publish empirical corpus-based descriptions and analyses of various aspects language's features
- Demonstrate and test the application and utility of descriptive features from cognitive linguistics such as those used to describe Mixtec in the literature in the annotation of the corpus
- Collect enough data so that it can be of help for speakers and potentially learners in using their language and creating more content
- Compatibility w/: LMF reserialization; TEI Lex-0; Ontolex-Lemon

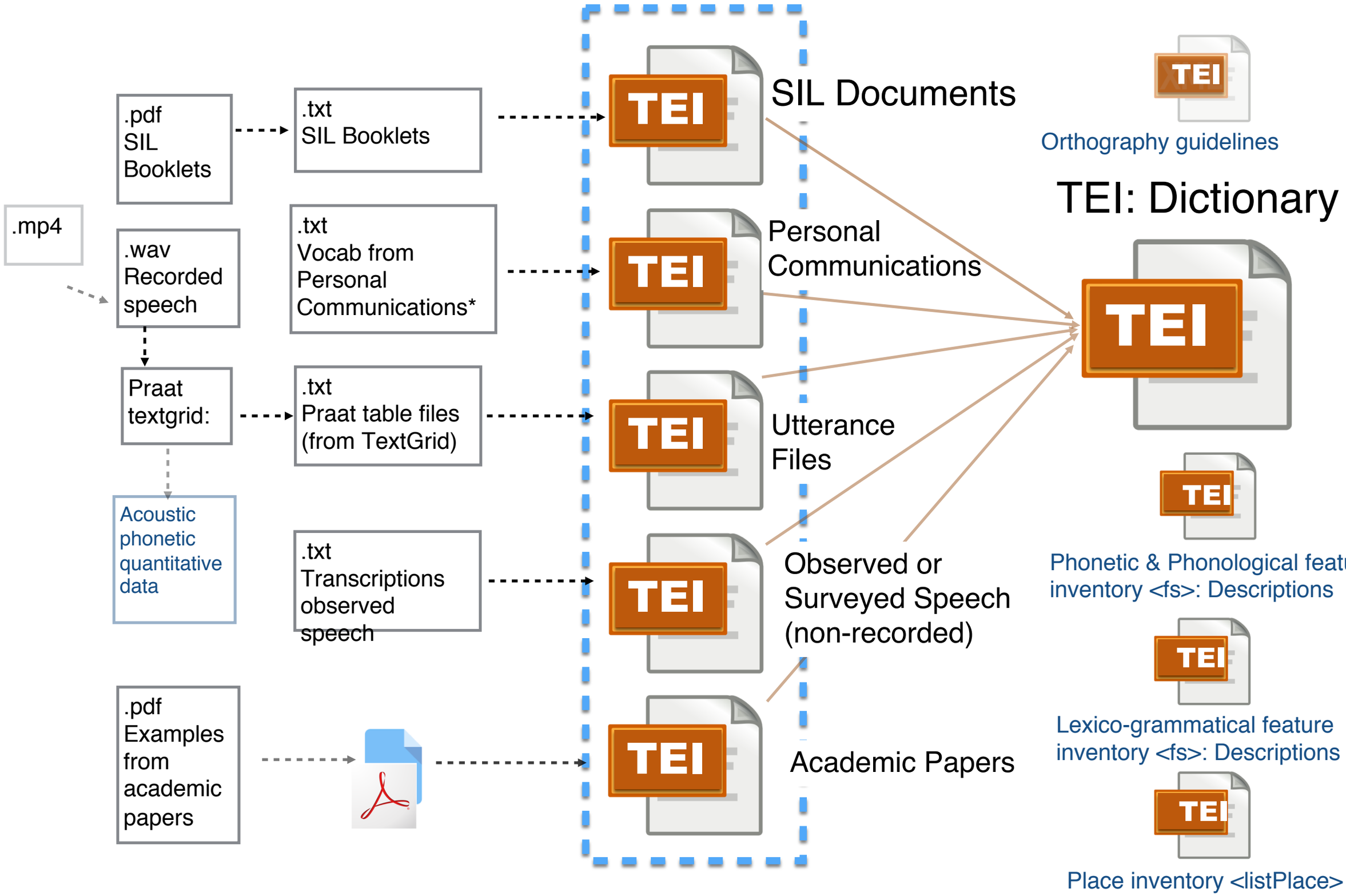
Primary Sources of Mixtepec-Mixtec Language Data

- Consultation w/ Speakers *(+- 600 recordings, written content)*
- Recordings made by speakers with other speakers
- Written content from speakers
- +-36 Children's Booklets *(Summer Institute of Linguistics Mexico)*
- Public Sources *(YouTube,etc.)*
    - Small number of papers *(phonology, some morphology)*
- Personal communications
- Public information pamphlets by Mexican government (new!)
- Videos by Conserva México Facebook page (new!)

Specific TEI Output

- New Mixtec language content
- Searchable TEI corpus
- TEI dictionary
- Time aligned utterance annotated files
- Annotated TEI files of SIL booklets
- Lexical feature inventory
- Phonetic feature inventory
- Concepts inventory
- Place inventory
- Person list

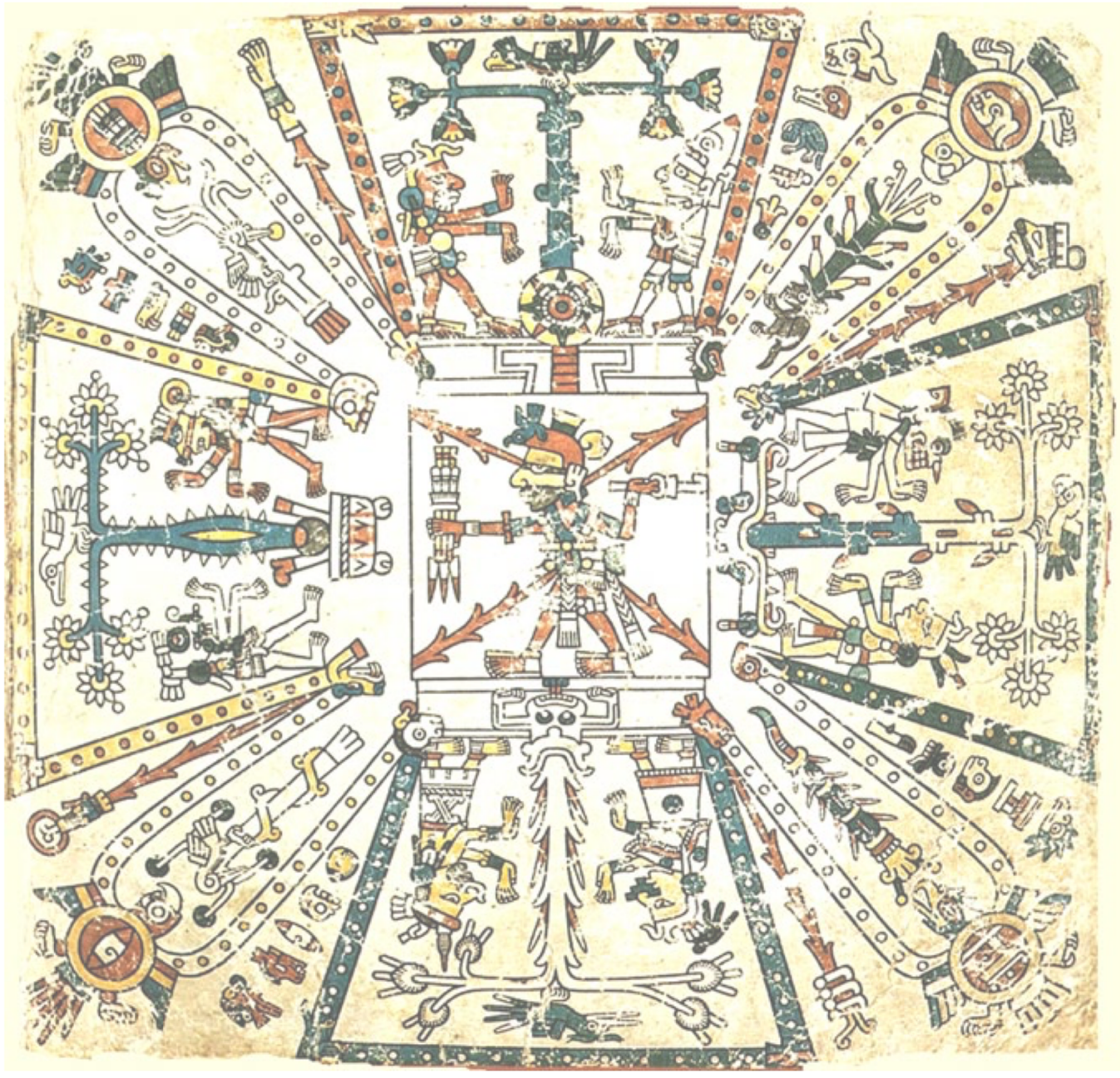
Mixtec Data: Sources, Links, Output



Intrinsic Challenges in Studying Mixtepec-Mixtec

- Lack of existing resources (under-resourced language)
- Lack of established linguistic description
- SIL Researchers working on the language in Mexico have (mostly) not shared their data
- Related language descriptions are old, syntax based, scanned documents
- Speaker consultants work full time, often don't have time to consistantly help edit, gloss text
- Orthography not fully conventionalized, still changes, speakers often not aware of/don't use the standards *(requires significant normalization in markup*)
- IPA also has too many different ways to transcribe (*especially tones*), normalization still needed
- Lexical tone, adds complexity to characterization and it is (mostly) not represented in the orthography (*lot's of homographs*)
- Not enough data to automate annotation! (*I am making the basis of any training set*)

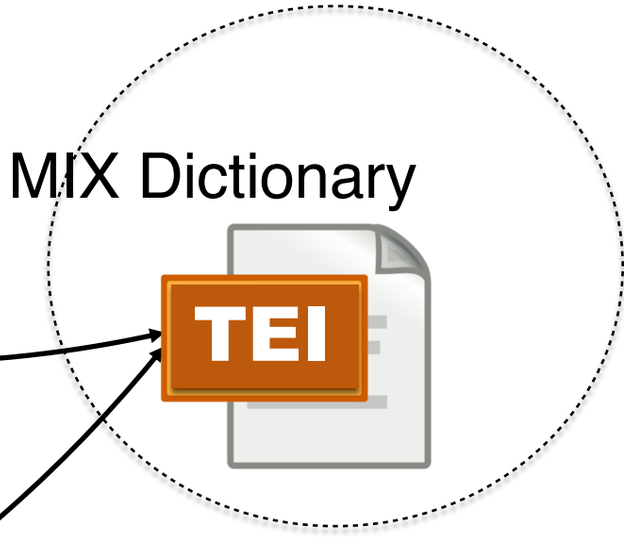
(I) Project Metadata



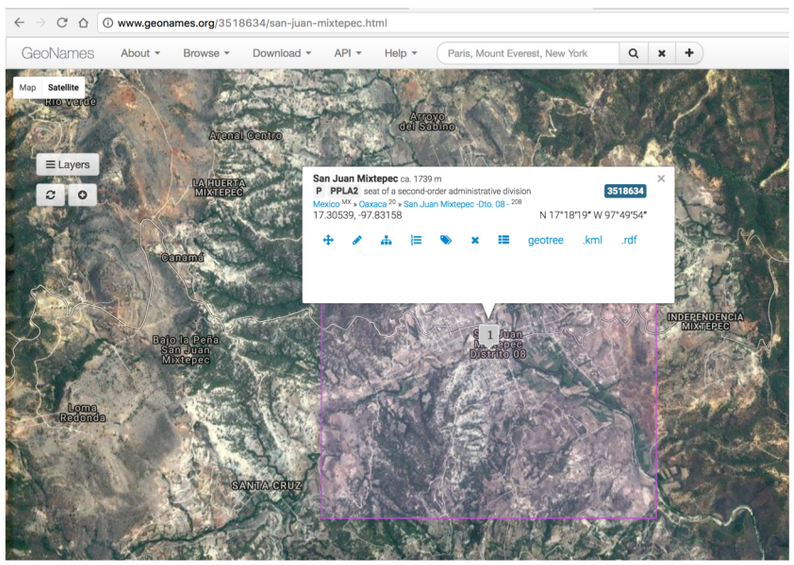
Mixtec Borgia Codex

Metadata: Places

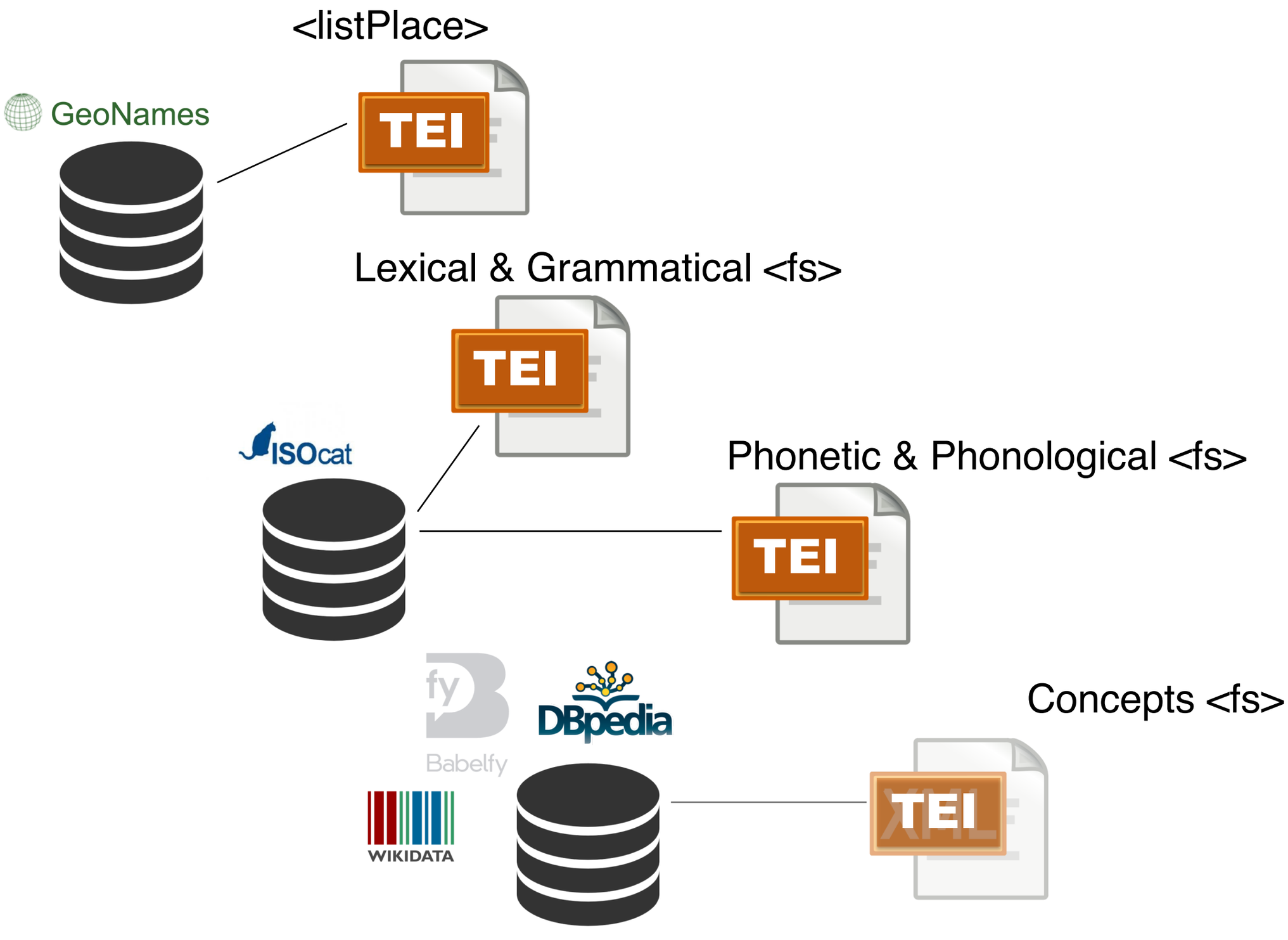
```
<listPlace>
....
  <place xml:id="Yucunany" corresp="http://www.geonames.org/8880392">
    <placeName xml:lang="es">Yucunany</placeName>
    <placeName xml:lang="en">Yucanany</placeName>
    <placeName xml:lang="en">Yucanani</placeName>
    <placeName xml:lang="mix" cert="medium">Yukunani</placeName>
    <location>
      <geo>17.30083, -97.89389</geo>
    </location>
  </place>

  <place xml:id="SanJuanMixtepec" corresp="http://www.geonames.org/3518634">
    <placeName xml:lang="es">San Juan de Mixtepec</placeName>
    <placeName xml:lang="es">San Juan Mixtepec</placeName>
    <placeName xml:lang="mix">Snuviko</placeName>
    <placeName xml:lang="mix">Xnuviko</placeName>
    <location>
      <geo>17.30539, -97.83158</geo>
    </location>
    <note resp="JB">Mixtec place name added to geonames</note>
  </place>
....
</listPlace>
```



Note: also included as entries in Mixtec Dictionary

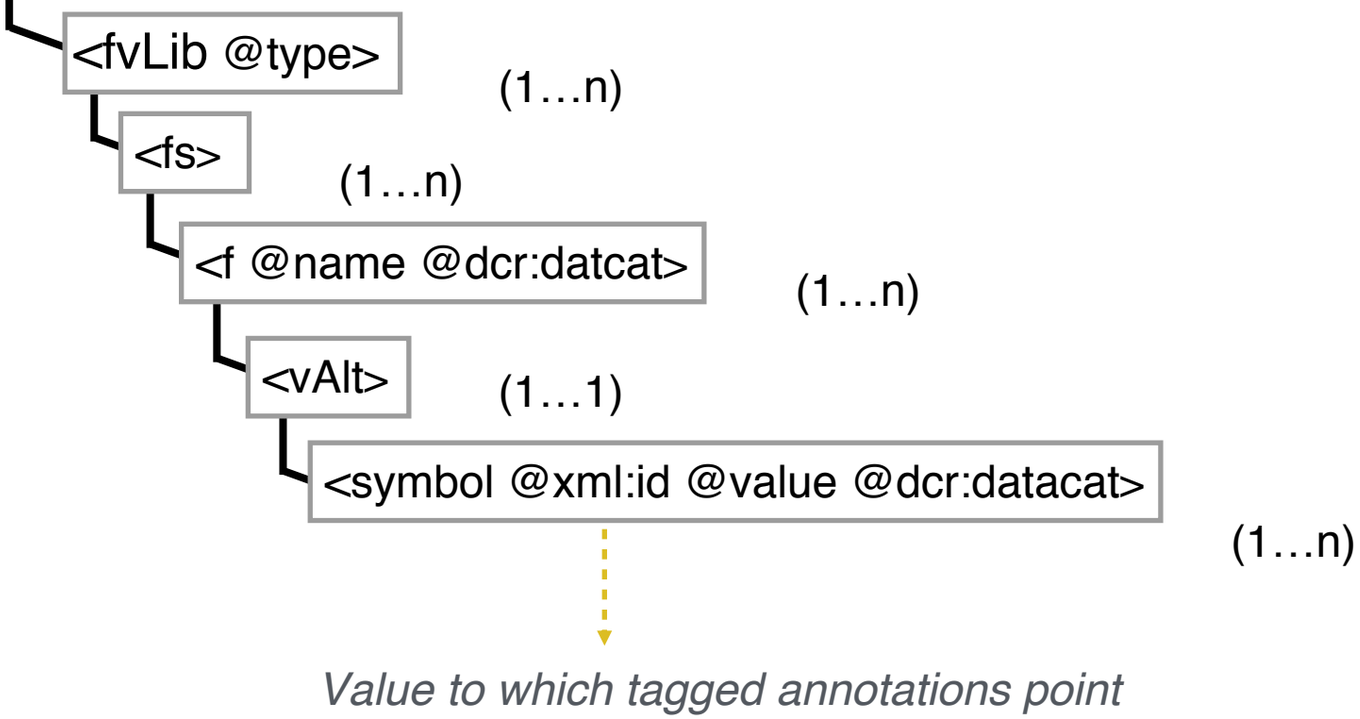


TEI Feature Structures & Standardized Resources



Linguistic Annotation: TEI Feature Structures

Inventory of MIX linguistic features kept in feature structures




```
<fvLib>
    <fs>
      <f name="number" xmlns:dcr="http://www.isocat.org/ns/dcr" dcr:datcat="http://www.isocat.org/datcat/DC-3351">
       <vAlt>
         <symbol xml:id="SG" value="singular" dcr:datcat="http://www.isocat.org/datcat/DC-252"/>
         <symbol xml:id="PL" value="plural" dcr:datcat="http://www.isocat.org/datcat/DC-253"/>
       </vAlt>
      </f>
    </fs>
    <!-- other feature structures here -->
  </fvLib>
```

Linguistic Annotation: TEI Feature Structures

Inventory of MIX linguistic features kept in feature structures

```
<fs><!--  Declerck, Thierry; The number of arguments controlled by a verbal predicate. -->
  <f name="valency" xmlns:dcr="http://www.isocat.org/ns/dcr" dcr:datcat="http://www.isocat.org/datcat/DC-1410">
    <vAlt>
      <symbol value="VAL-GRM0" xml:id="VAL-0"/><!-- Contains neither Actor or Undergoer; corresponds w/ “#ATRANS” -->
      <symbol value="VAL-GRM1" xml:id="VAL-1"/><!-- Contains only Actor or Undergoer -->
      <symbol value="VAL-GRM1" xml:id="VAL-2"/><!-- Contains Actor and Undergoer -->
      <symbol value="VAL-GRM1" xml:id="VAL-3"/><!-- Contains Actor, Undergoer and Oblique -->
    </vAlt>
  </f>
</fs>
```

```
<fs>
  <f name="transitivity">
    <vAlt>
      <symbol xml:id="TRANS" value="transitive" xmlns:dcr="http://www.isocat.org/ns/dcr" dcr:datcat="http://www.isocat.org/datcat/DC-3275"/>
      <symbol xml:id="INTRANS" value="intransitive" xmlns:dcr="http://www.isocat.org/ns/dcr" dcr:datcat="http://www.isocat.org/datcat/DC-3275"/>
      <symbol xml:id="DITRANS" value="ditransitive" xmlns:dcr="http://www.isocat.org/ns/dcr" dcr:datcat=“http://www.isocat.org/datcat/DC-1275"/>
      <symbol xml:id="ATRANS" value="a-transitive"/><!-- corresponds w/ “#VAL-0” Contains neither Actor or Undergoer -->
    </vAlt>
  </f>
</fs>
```

Linguistic Annotation: TEI Feature Structures

Inventory of MIX linguistic features kept in feature structures

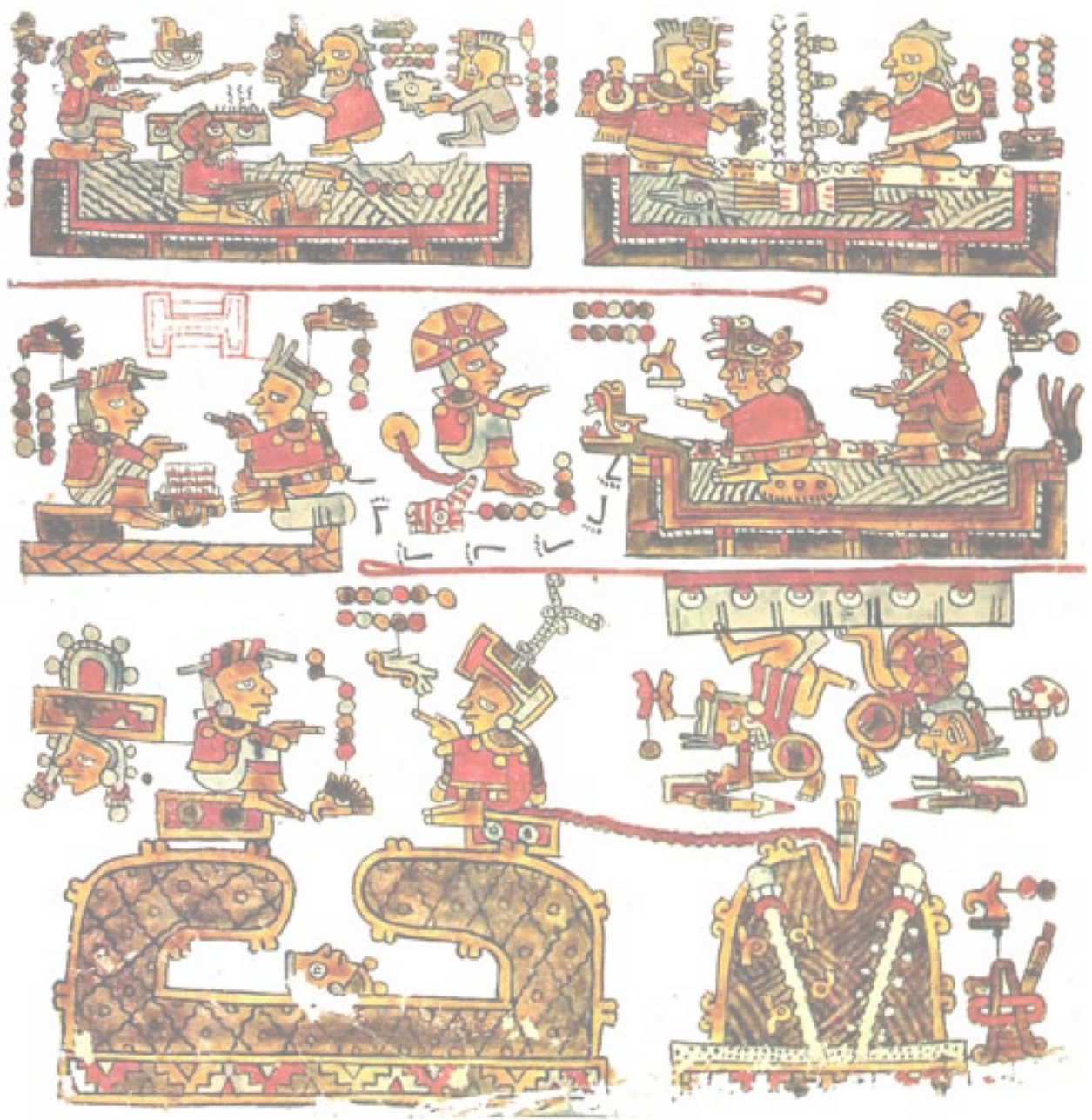
```
<fs>
  <f name="Trajector" xml:id="TR">
    <vAlt>
      <symbol value="StaticTrajector" xml:id="TR-static"/>
      <symbol value="DynamicTrajector" xml:id="TR-dynam"/>
      <symbol value="Person-Object" xml:id="TR-pers-obj"/>
      <symbol value="EventTrajector" xml:id="TR-event"/>
    </vAlt>
  </f>
</fs>
<fs>
  <f name="Landmark" xml:id="LM">
    <vAlt>
      <symbol value="personLM" xml:id="LM-PERS"/>
      <symbol value="objectLM" xml:id="LM-OBJ"/>
      <symbol value="eventLM" xml:id="LM-EVNT"/>
    </vAlt>
  </f>
</fs>

<fs>
  <f name="frameOfReference" xml:id="FoR">
    <vAlt>
      <symbol value="viewpoint-centeredFoR" xml:id="VPTC-FoR"/>
      <symbol value="relativeFoR" xml:id="REL-FoR"/>
      <symbol value="intrinsicFor" xml:id="INTR-FoR"/>
    </vAlt>
  </f>
</fs>
```

external ontologies:

- GUM?
- Eagles?

(II) Source Documents

i. SIL Booklets



Mixtec Codex Seldon

Source Data: SIL Documents

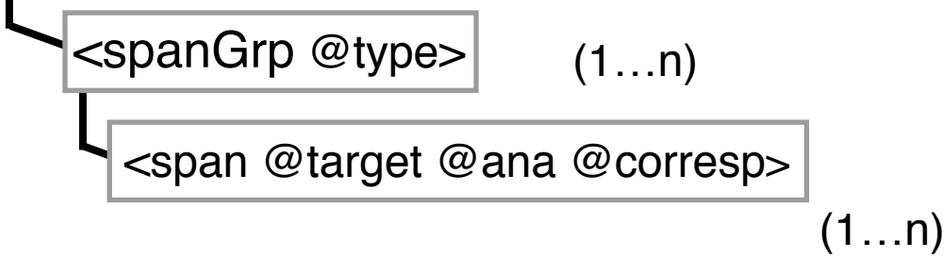
The Summer Institute of Linguistics (SIL) documents all have an intended audience of children, there are several different document types which have different formats:

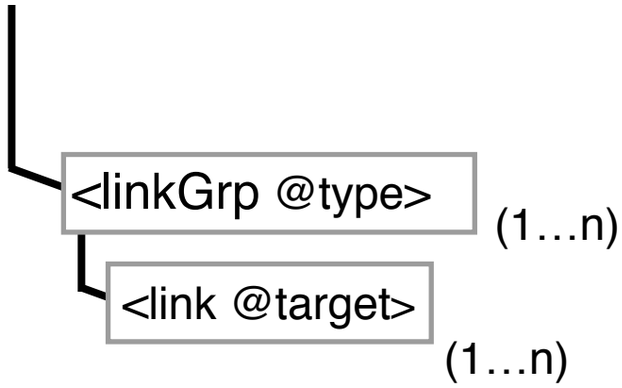
- Prose *(short stories, legends, etc.)*
- Activity/Workbooks *(picture-based excercises, crossword puzzles, mazes, etc.)*
- Vocabulary & Basic Pedigogical Reference

Current document taxonymy contains the following classifications:

- Pedagogical
  - Interactive
  - Referential
- Fiction
  - Fantasy
  - Realistic
- Folklore

Standoff Annotation in TEI: <spanGrp> & <linkGrp>:





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

- Points to 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>

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

SIL Documents: Basic Vocabulary

chumi xini ka'nu
tecolote
búho cornado

chumi lunchi
tecolote llanero
tecolote zancón

chumi sai
tecolotito

```
<item>
    <graphic url="Aves-01.png"/>
    <w xml:id="d1e35" xml:lang="mix" type="compound">
      <w xml:id="d1e36">chumi</w> <w xml:id="d1e38">lunchi</w>
    </w>
    <w xml:id="d1e40" xml:lang="es" type="compound">
      <w xml:id="d1e41">tecolote</w> <w xml:id="d1e43">llanero</w>
    </w>
    <w xml:id="d1e45" xml:lang="es" type="compound">
      <w xml:id="d1e46">tecolote</w> <w xml:id="d1e48">zancón</w>
    </w>
  </item>
  <item>
    <graphic url="Aves-02.png"/>
    <w xml:id="d1e53" xml:lang="mix" type="compound">
      <w xml:id="d1e54">chumi</w> <w xml:id="d1e56">xini</w> <w xml:id="d1e58">ka'nu</w>
    </w>
    <w xml:id="d1e60" xml:lang="es">
      <w xml:id="d1e61">tecolote</w>
    </w>
    <w xml:id="d1e63" xml:lang="es" type="compound">
      <w xml:id="d1e64">búho</w> <w xml:id="d1e66">cornado</w>
    </w>
  </item>
  <item>
    <graphic url="Aves-03.png"/>
    <w xml:id="d1e71" xml:lang="mix" type="compound">
      <w xml:id="d1e72">chumi</w> <w xml:id="d1e74">sai</w>
    </w>
    <w xml:id="d1e76" xml:lang="es">
      <w xml:id="d1e77">tecolotito</w>
    </w>
  </item>
```

SIL Documents: Basic Vocabulary Annotation

```
<item>
    <graphic url="Aves-02.png"/>
    <w xml:id="d1e53" xml:lang="mix" type="compound">
      <w xml:id="d1e54">chumi</w>
      <w xml:id="d1e56">xini</w>
      <w xml:id="d1e58">ka'nu</w>
    </w>
    <w xml:id="d1e60" xml:lang="es-MEX">
      <w xml:id="d1e61">tecolote</w>
    </w>
    <w xml:id="d1e63" xml:lang="es" type="compound">
      <w xml:id="d1e64">búho</w>
      <w xml:id="d1e66">cornado</w>
    </w>
</item>
```

chumi xini ka'nu

tecolote

búho cornado

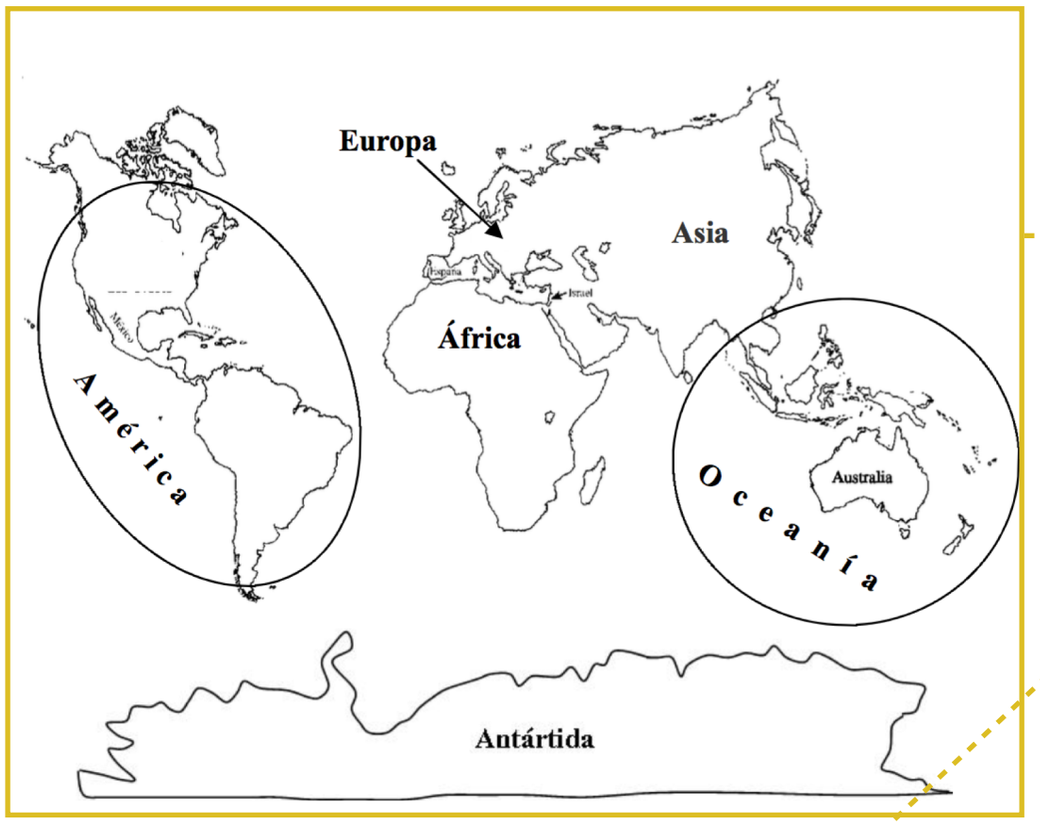
Annotations: (pre-existing) Translations & Sense (concept);

```
<linkGrp type="translation">
   <link target="#d1e53 #d1e60"/>
   <link target="#d1e53 #d1e63"/>
</linkGrp>

<spanGrp type="semantics">
  <span type="sense" target="#d1e53" corresp="https://www.wikidata.org/wiki/Q81515"/>
  <span type="sense" target="#d1e53" corresp="http://dbpedia.org/resource/Great_horned_owl"/>
  <span type="domain" target="#d1e53" corresp="http://dbpedia.org/resource/Bird"/>
  <span type="domain" target="#d1e53" corresp="http://dbpedia.org/resource/Animal"/>
</spanGrp>
```

SIL Documents: Prose

PDF Source

Ñu'u Ncha'i ka



Yee ñu'u tsi chikuii nuu Ñu'u Ncha'i. Yee kua'a ka chikuii cha xoo ka ñu'u. Yee ñu'u luu ka nania "islas", cha inkai ma'i chikuii. Cha ñu'u ka'nu ka nania "continente". Yee iñu "continente" nania: África, América, Antártida, Asia, Europa tsi Oceanía .

TEI

```
<div xml:id="L145-13">
  <head>
    <seg xml:id="L145-13-00" type="subject">
      <w xml:id="d1e1437">
        <w xml:id="d1e1438">Ñu'u</w>
        <w xml:id="d1e1441">Ncha'i</w>
      </w>
      <w xml:id="d1e1444">ka</w>
    </seg>
  </head>
  <head><graphic url="L145_10.jpeg"/></head>
  <p>
    <seg xml:id="L145-13-01" type="S">
      <w xml:id="d1e1458">Yee</w>
      <w xml:id="d1e1461">ñu'u</w>
      <w xml:id="d1e1464">tsi</w>
      <w xml:id="d1e1467">chikuii</w>
      <w xml:id="d1e1470">nuu</w>
      <w xml:id="d1e1471">
        <w xml:id="d1e1473">Ñu'u</w>
        <w xml:id="d1e1477">Ncha'i</w>
      </w>
      <pc>.</pc>
    </seg>
    ....

</div>
```

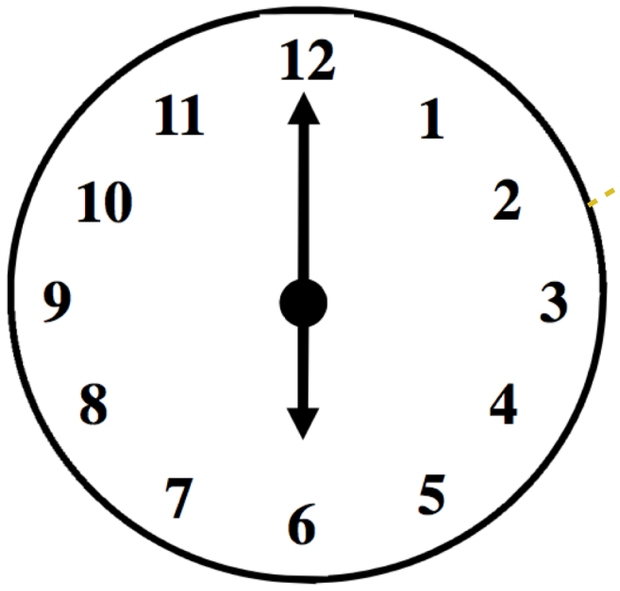
SIL Documents: Prose annotation

```
<div xml:id="L145-13">
  …
 <s xml:id="L145-13-01" type="declarative">
  <w xml:id="d1e1458">Yee</w>
  <w xml:id="d1e1461">ñu'u</w>
  <w xml:id="d1e1464">tsi</w>
  <w xml:id="d1e1467">chikuii</w>
  <w xml:id="d1e1470">nuu</w>
  <w xml:id="d1e1471">
    <w xml:id="d1e1473">Ñu'u</w>
    <w xml:id="d1e1477">Ncha'i</w>
  </w>
  <pc>.</pc>
 </s>
  …
</div>
```

Annotations: Translations

```
<spanGrp type="translation">
  <span target="#L145-13-01"xml:lang="en">There is land and water on the Earth.</span>
  <span target="#L145-13-01" xml:lang="es">Hay tierra y agua en la Tierra.</span>

  <span target="#d1e1458" xml:lang="en">there is</span>
  <span target="#d1e1458" xml:lang="es">hay</span>

  <span target="#d1e1461" xml:lang="en">land</span>
  <span target="#d1e1461" xml:lang="es">tierra</span>

  <span target="#d1e1464" xml:lang="en">and</span>
  <span target="#d1e1464" xml:lang="es">y</span>

  <span target="#d1e1467" xml:lang="en">water</span>
  <span target="#d1e1467" xml:lang="es">agua</span>

  <span target="#d1e1470 #d1e1471" xml:lang="en">on Earth</span>
  <span target="#d1e1470 #d1e1471" xml:lang="es">en la tierra</span>

  <span target="#d1e1471" xml:lang="en">Earth</span>
  <span target="#d1e1471" xml:lang="es">tierra</span>
</spanGrp>
```

SIL Documents: Workbook
(reference version w/answers)



¿Nchii hora kui?

Ka iñu ntaa.


```
<div xml:id="L093-01">
  <head>
    <graphic url="L093-1-what_time_is_it-6.jpg"/>
  </head>
  <label>
    <time>6:00</time>
  </label>
  <lb/>
  <p>
    <s xml:id="d1e160" type=“interrogative">
      <pc>¿</pc>
      <w xml:id=“d1e163" orig="Nchii">Nchi</w>
      <w xml:id="d1e165">hora</w>
      <w xml:id="d1e167">kui</w>
      <pc>?</pc>
    </s>
    <lb/>
    <s xml:id="d1e174" type=“declarative">
      <w xml:id=“d1e175" orig="Ka">Kaa</w>
      <w xml:id="d1e177">iñu</w>
      <w xml:id="d1e179">ntaa</w>
      <pc>.</pc>
    </s>
  </p>
</div>
```

SIL Documents: Workbook (*reference version w/answers*) annotation

```
<div xml:id="L093-01">
…..
 <p>
    <seg xml:id="d1e160" type=“S">
      <pc>¿</pc>
      <w xml:id="d1e163" orig="Nchii">Nchi</w>
      <w xml:id="d1e165">hora</w>
      <w xml:id=“d1e167">kui</w>
      <pc>?</pc>
    </seg>
    <lb/>
    <seg xml:id="d1e174" type=“S">
      <w xml:id=“d1e175" orig="Ka">Kaa</w>
      <w xml:id="d1e177">iñu</w>
      <w xml:id="d1e179">ntaa</w>
      <pc>.</pc>
    </seg>
  </p>
</div>
```

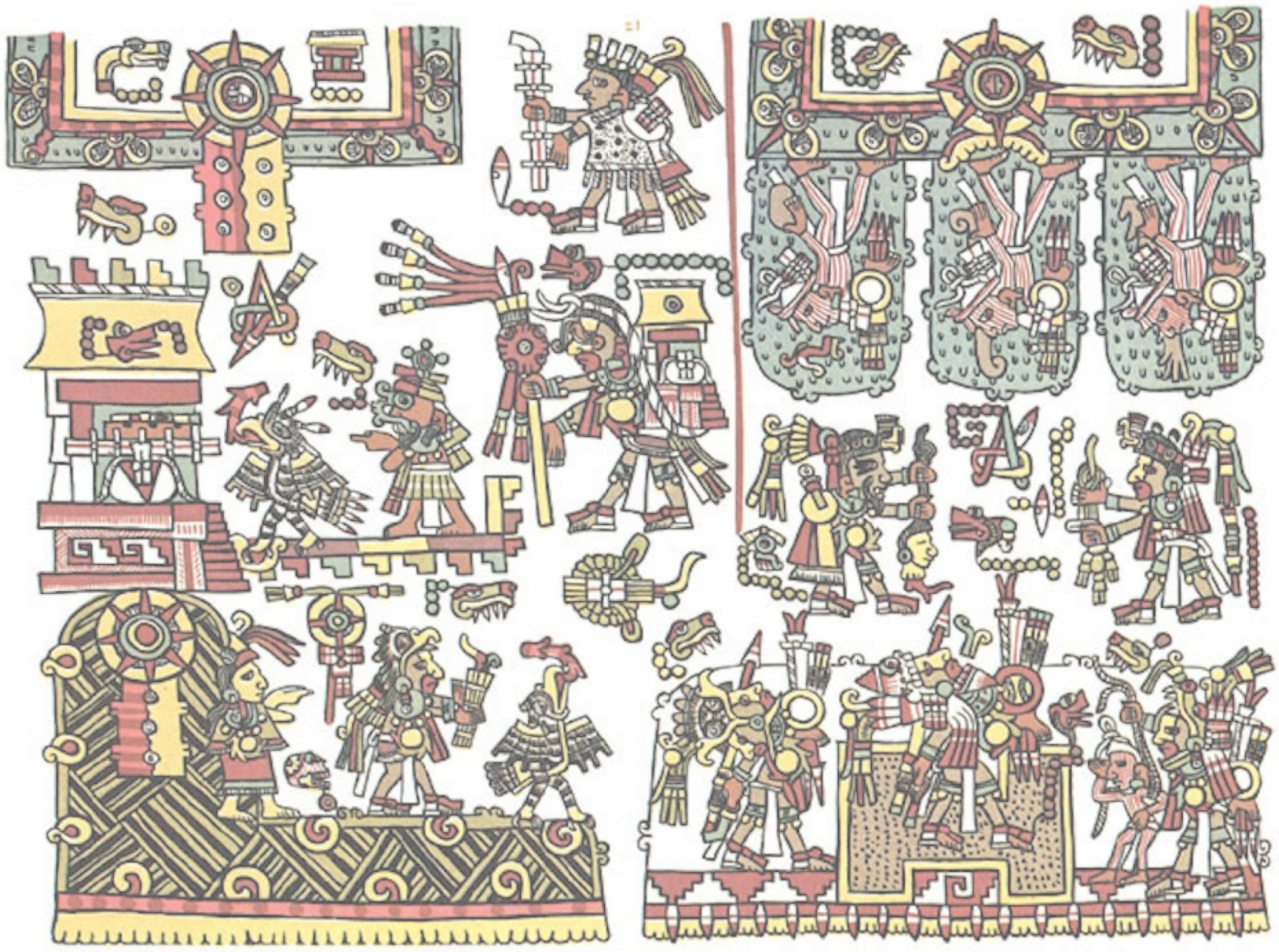
Annotations: Grammar

```
<spanGrp type="gram">
  <span type="sentence" target="#d1e160" ana="#Q #WH #TEMP"/>
  <span type="phrase" target="#d1e163 #d1e165" ana="#ADVP #WH #TEMP"/>
  <span type="pos" target="#d1e167" ana="#COP #INCMPL”/>
  <span type="aspect" target="#d1e167" ana=“#INCMPL"/>
  <span type="person" target="#d1e169" ana=“#3PERS"/>
  <span type="number" target="#d1e169" ana=“#SG"/>
</spanGrp>
```

```
<spanGrp type="gram">
  <span type="sentence" target="#d1e174" ana="#RESP #Q #WH #TEMP"/>
  <span type="pos" target="#d1e175" ana="#COP #REAL"/>
  <span type="phrase" target="#d1e177 #d1e179" ana="#ADVP #TEMP"/>
</spanGrp>
```

(II) Source Documents

ii. Spoken Language Resources



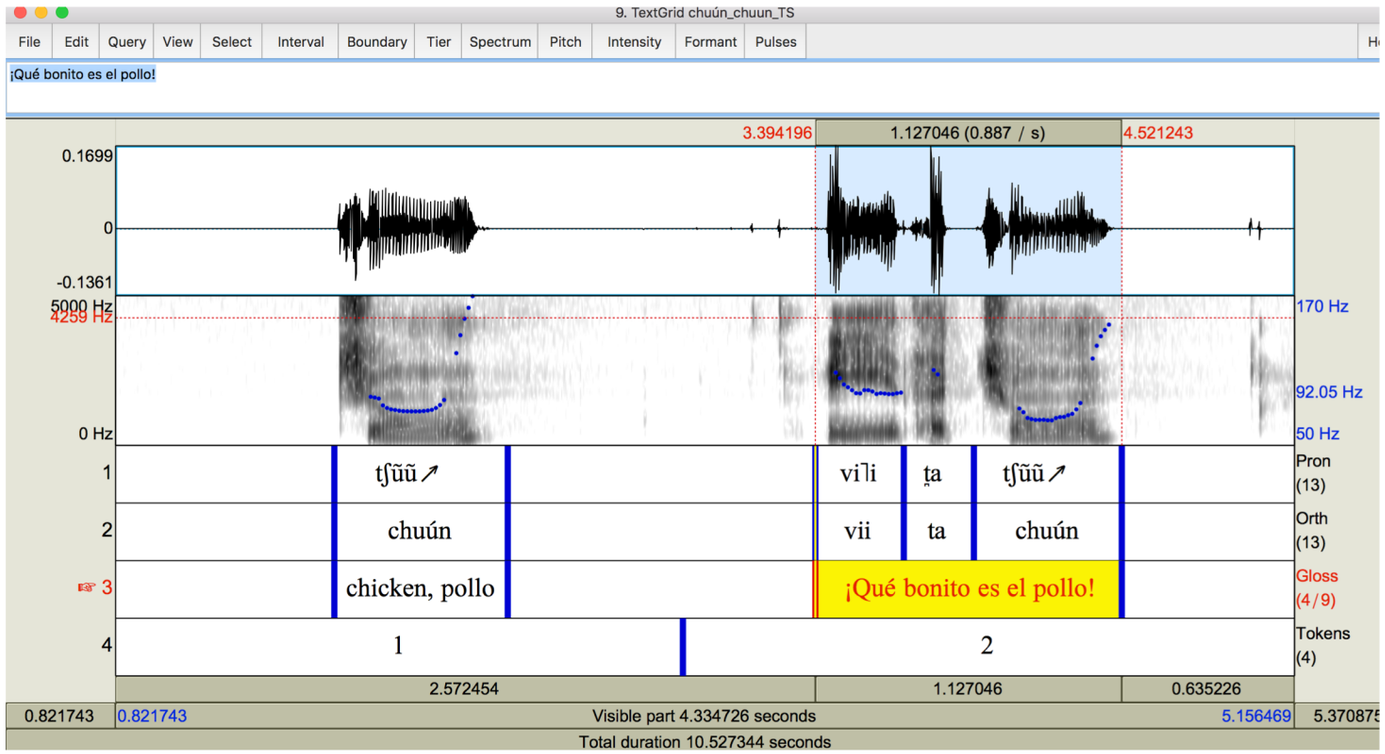
Codex Zouche-Nuttall, British Museum.

Speech Annotation: Toolkits & Features

	Praat	Exmaralda
metadata*	no	yes
spectrogram view	yes	no
XML/TEI output option	no	yes
tiered/ time aligned segmentation	yes	yes
scripting	yes	no
TEI/XML export	no	yes
corpus managment, searching	*(via scripting)*	yes *(text based only)*
video annotation	no	yes
visualization	yes*	yes
quantitivative data extraction	yes	no
pitch (F0) view/analysis	**yes**	no

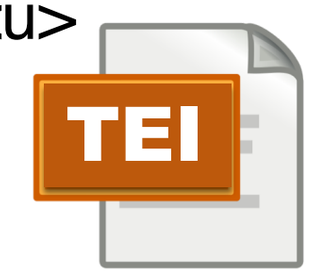
Speech Annotation: Praat

(basic transcription method)



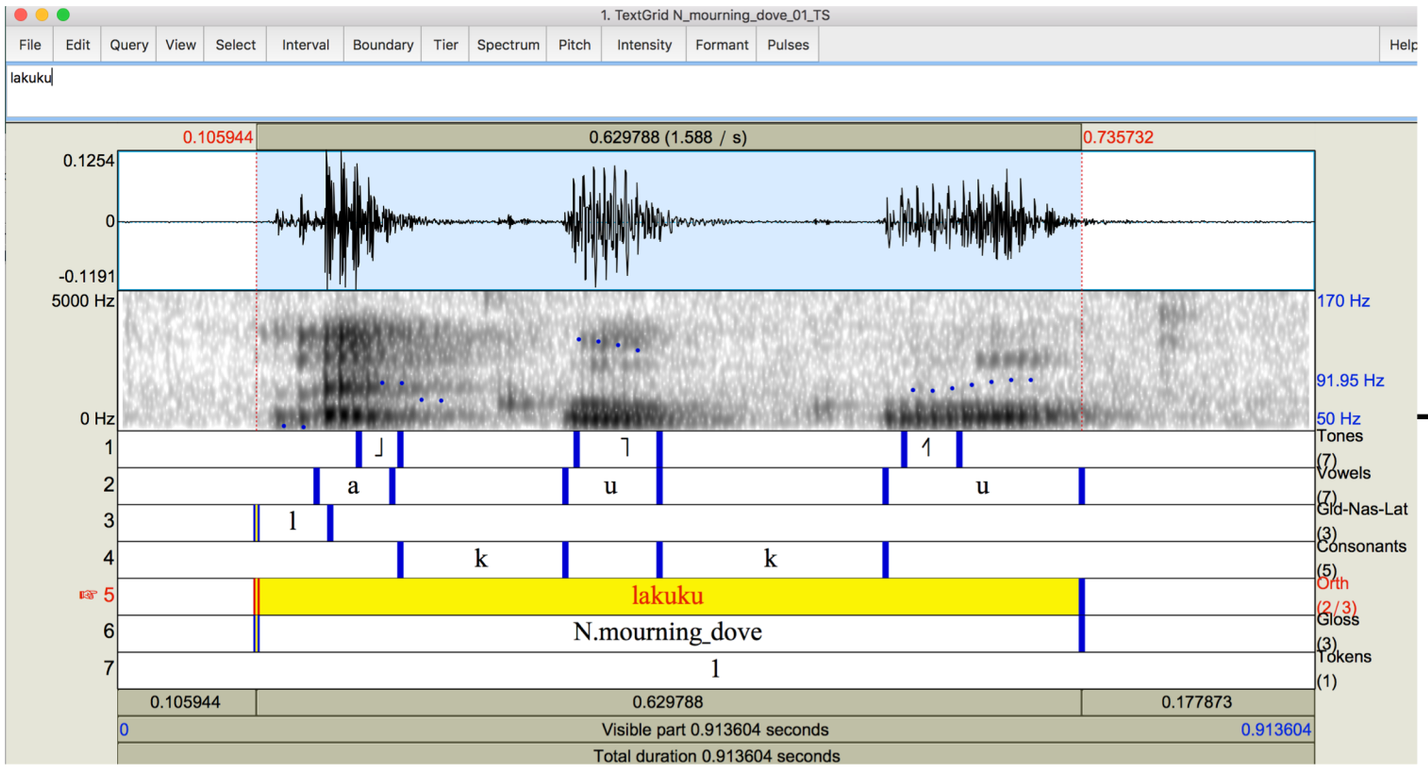
tmin	tier	text	tmax
0	Tokens	1	2.91
1.63	Gloss	chicken, pollo	2.26
1.63	Pron	tʃũũ↗	2.26
1.63	Orth	chuún	2.26
2.91	Tokens	2	5.18
3.39	Orth	vii	3.72
3.39	Pron	vi˥i	3.72
3.39	Gloss	¡Qué bonito es el pollo!	4.52
3.72	Pron	ṭa	3.98
3.72	Orth	ta	3.98
3.98	Orth	chuún	4.52
3.98	Pron	tʃũũ↗	4.52

Utterance File <u>



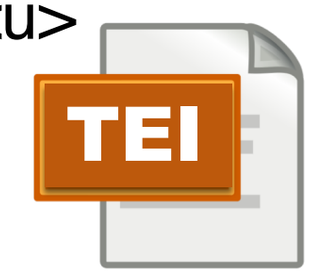
Speech Annotation: Praat

(phonetic focus transcription)



tmin	tier	text	tmax
0	Tokens	1	0.91
0.11	Gld-Nas-Latl		0.16
0.11	Orth	lakuku	0.74
0.11	Gloss	N.mourning_dove	0.74
0.15	Vowels	a	0.21
0.18	Tones	˩˥	0.22
0.22	Consonantsk		0.34
0.34	Vowels	u	0.41
0.35	Tones	˥	0.41
0.41	Consonants	k	0.59
0.59	Vowels	u	0.74
0.60	Tones	˧	0.64

Utterance File <u>



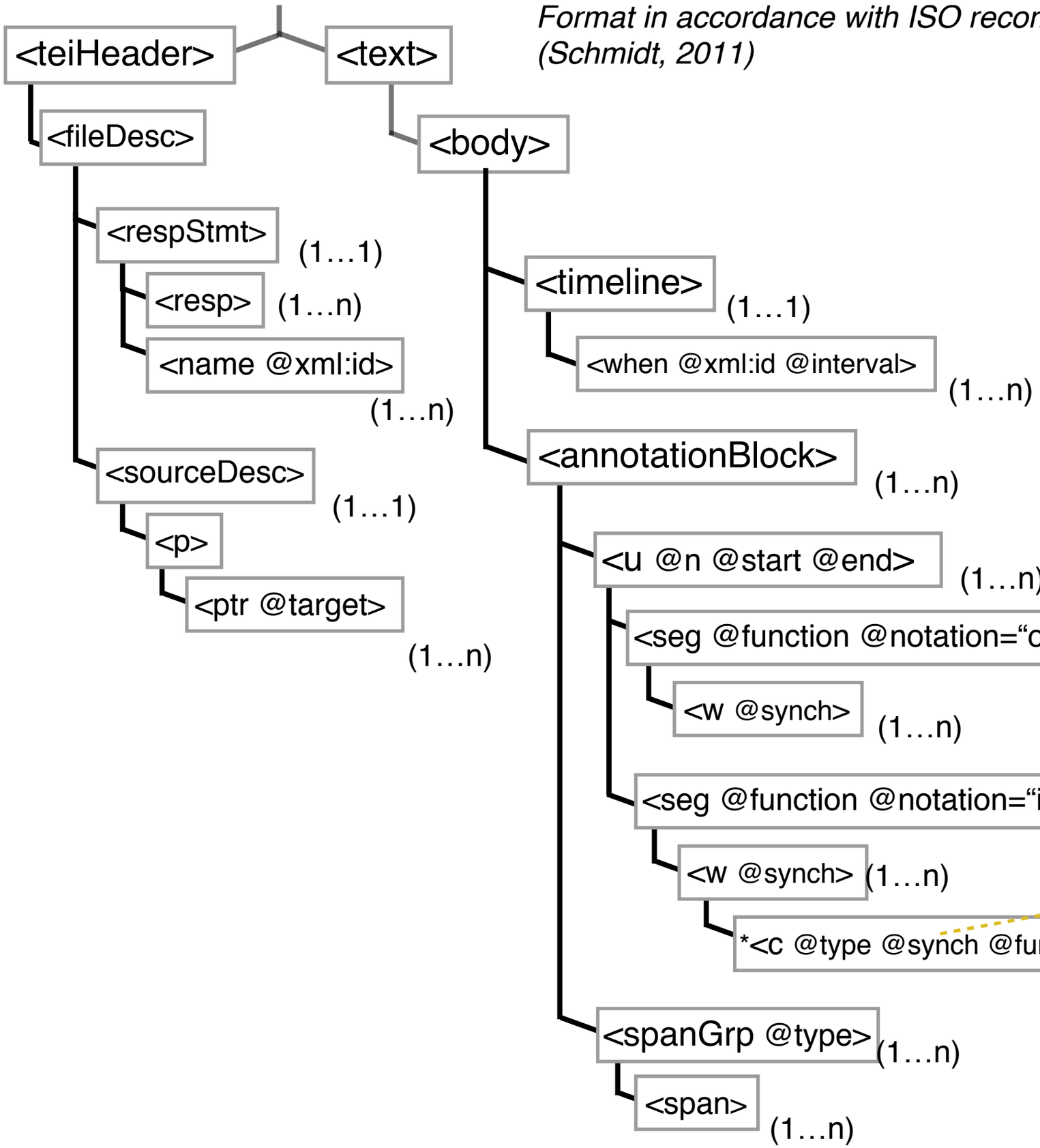
Praat (phonetic focus transcription): Available Data

Acoustic data available potentially allows for:

- Quantitative, acoustic evidence for phonetic and phonological linguistic descriptions of language;
- Fine grained tests of existing hypotheses about phonology
  - tone patterns, sandhi, etc.;
- Train HMM models for ASR (applied to: Automatic Annotation of spoken data);
- Comparative studies of different speaker groups (e.g. *from different villages, those who live in the US and those that don't*)

TEI Utterance files (from Praat)

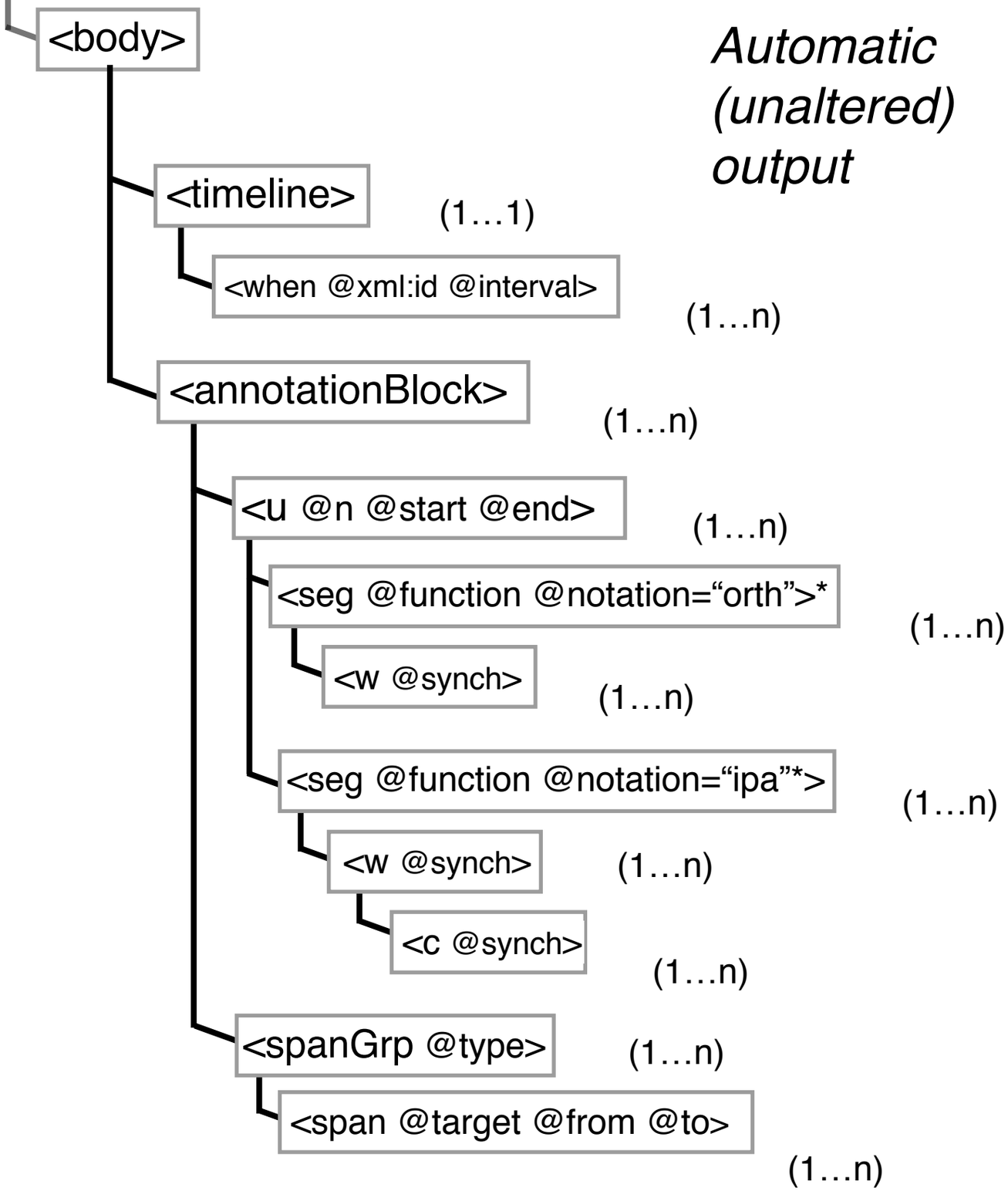
Format in accordance with ISO recommendation for speech transcription (Schmidt, 2011)



- One utterance file per Praat TextGrid
- Source .wav and praat textgrid filenames in header <ptr @target> within <sourceDesc>
- Can generate speaker info in header from file name <respStmt>

<c>'s correspond to <fs> values for phonetic/phonological invetory (only included in output from fully segmented (phonetic focus) praat annotations)

TEI Utterance files (from Praat)



Automatic (unaltered) output

```
<body>
  <timeline>
    <when xml:id="T1" interval="0.11"/>
    <when xml:id="T2" interval="0.15"/>
    <when xml:id="T3" interval="0.18"/>
    <when xml:id="T4" interval="0.22"/>
    <when xml:id="T5" interval="0.34"/>
    <when xml:id="T6" interval="0.35"/>
    <when xml:id="T7" interval="0.41"/>
    <when xml:id="T8" interval="0.59"/>
    <when xml:id="T9" interval="0.60"/>
    <when xml:id="T10" interval="0.74"/>
  </timeline>
  <annotationBlock>
    <u xml:id="d1e39" n="1" start="0" end="0.91">
      <seg xml:id="d1e40" function="utterance" notation="orth">
        <w xml:id="d1e41" synch="#T1">lakuku</w>
      </seg>
      <seg xml:id="d1e44" function="utterance" notation="ipa">
        <w xml:id="d1e45" synch="#T1">
          <c>l</c>
          <c>a</c>
          <c function="tone">˩</c>
          <c>k</c>
          <c>u</c>
          <c function="tone">˥</c>
          <c>k</c>
          <c>u</c>
          <c function="tone">˩</c>
        </w>
      </seg>
    </u>
    <spanGrp type="praatGloss">
      <span from="#T1" to="#T10">N.mourning_dove</span>
    </spanGrp>
    .....
  </annotationBlock>
</body>
```

TEI Utterance files (from Praat):Annotated

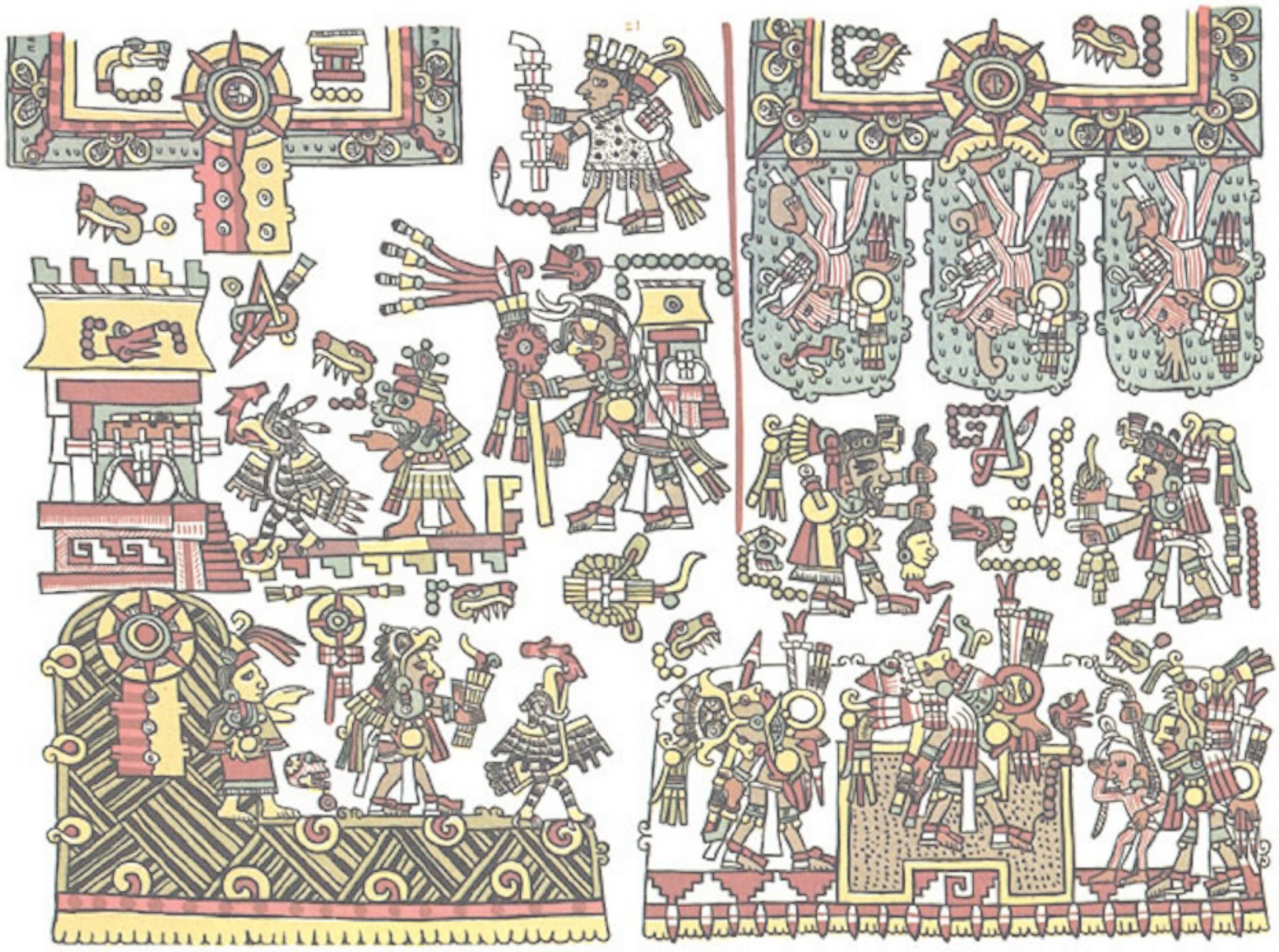
```
<timeline>
……
</timeline>
<annotationBlock>
  <u xml:id="d1e39" n="1" start="0" end="0.91">
    <seg xml:id="d1e40" function="utterance" notation="orth">
      <w xml:id="d1e41" synch="#T1">lakuku</w>
    </seg>
    <seg xml:id="d1e44" function="utterance" notation="ipa">
      <w xml:id="d1e45" synch="#T1">
        <c>l</c>
        <c>a</c>
        <c function="tone">˩</c>
        <c>k</c>
        <c>u</c>
        <c function="tone">˥</c>
        <c>k</c>
        <c>u</c>
        <c function="tone">˥</c>
      </w>
    </seg>
  </u>
  <spanGrp type="praatGloss">
    <span from="#T1" to="#T10">N.mourning_dove</span>
  </spanGrp>
  <spanGrp type="gram">
    <span type="pos" target="#d1e41 #d1e45" ana="#N"/>
  </spanGrp>
  <spanGrp type="semantics">
    <span type="sense" target="#d1e41 #d1e45" corresp="http://dbpedia.org/resource/Mourning_dove"/>
    <!-- is_a:Bird -->
    <span type="domain" target="#d1e41 #d1e45" corresp="http://dbpedia.org/resource/Bird"/>
  </spanGrp>
  <spanGrp type="translation">
    <span target="#d1e41 #d1e45" xml:lang="en" corresp="https://en.wiktionary.org/wiki/mourning_dove">mourning dove</spa
    <span target="#d1e41 #d1e45" xml:lang="es" corresp="https://es.wiktionary.org/wiki/tortolita">tortolita</span>
  </spanGrp>
</annotationBlock>
```

to TEI Dictionary: (value of) //form[@type="lemma"]/orth

to Dictionary: (value of) //form[@type="lemma"]/pron[notation="ipa"]

Manually added in Oxygen

(II) Source Documents
iii. Data From Academic Papers



Codex Zouche-Nuttall, British Museum.

Integrating Content from Academic Papers: Pike & Ibach (1978)

Paper is the 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)

The amount of data is too small to justify automation but it is important content both for comparing my results and in maximizing the quantity of data collected

Non-IPA

ȼ = t͡s
š = ʃ
č = t͡ʃ
ʒ = t͡ʒ
ǰ = d͡ʒ
ɡ = k͡

Tone levels are reversed from conventional description

^{3} = Low
^{2} = Mid
^{1} = High

Strings are often interrupted

š[i.]^{2}ši^{3}-ȼi^{2} '

Integrating Content from Academic Papers: Pike & Ibach (1978)

Source pdf (scanned)

1. PHONOLOGICAL WORD

In Mixtepec 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*[*o*.]^{1}*lo*^{1}*ko*^{1} 'our (excl.) male turkey', *š*[*i*.]^{2}*ši*^{3}-*ɛi*^{2} 'his or her (child) aunt', *s*[*o*.]^{3}*ko*^{3}-*yu*^{3} 'my (polite)

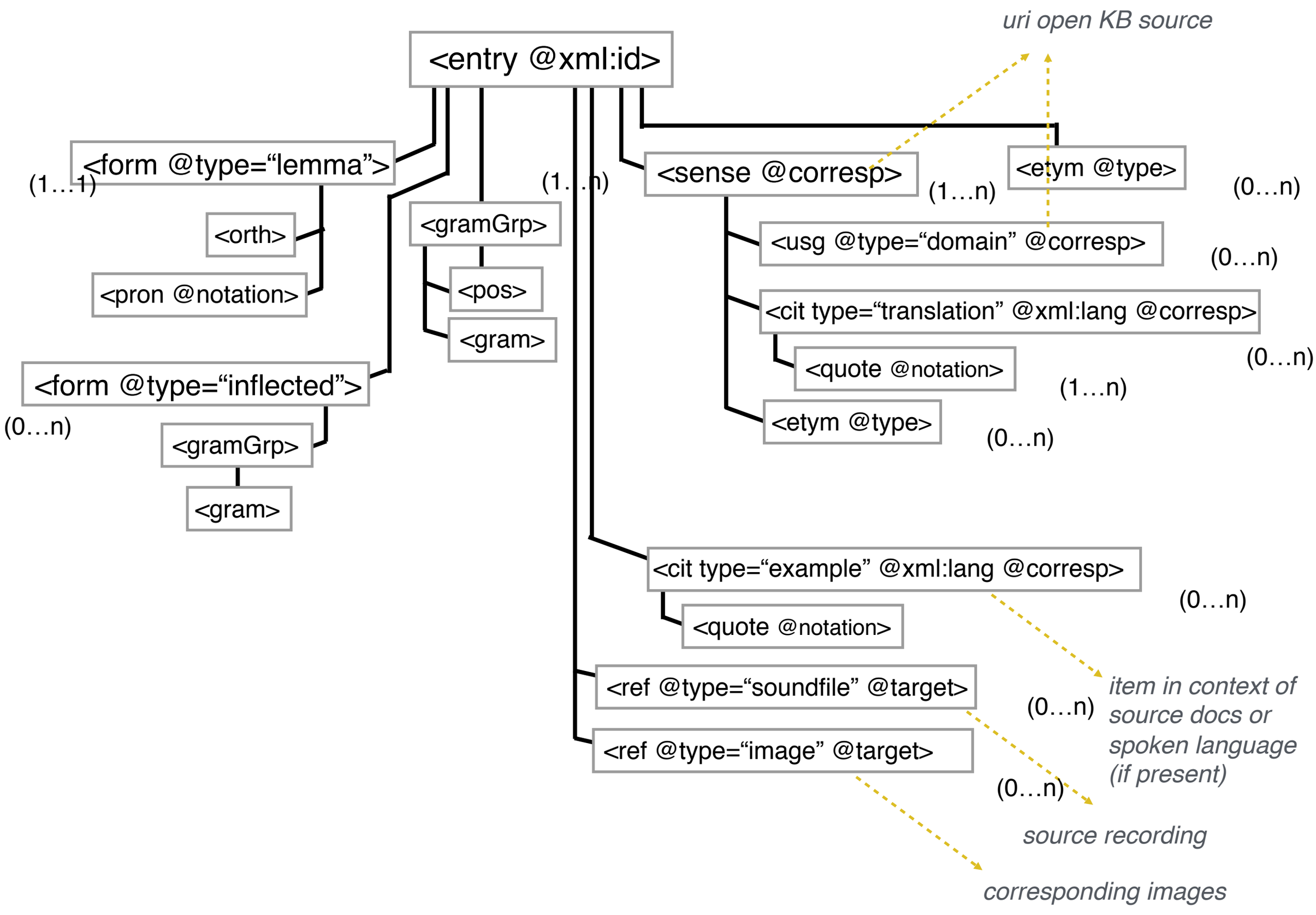
TEI Encoding

```
<div>
    <label>1. PHONOLOGICAL WORD</label>
    <p>
      In Mixtepec Mixtec1 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 /<c notation="ipa">ʔ</c>/:

      <seg xml:id="d1e24" xml:lang="mix" notation="ipa">
        <w xml:id="d1e25" orig="k[o.]¹lo¹">koˀloˀ</w> <w xml:id="d1e27" orig="ko¹">koˀ</w>
      </seg>
      '<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>',
      <linkGrp type="translation">
        <link target="#d1e24 #d1e28"/>
      </linkGrp>
          ……..

    </p>
```

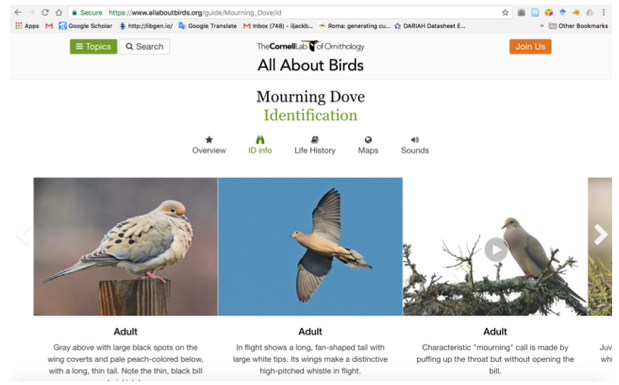
(III) TEI Dictionary

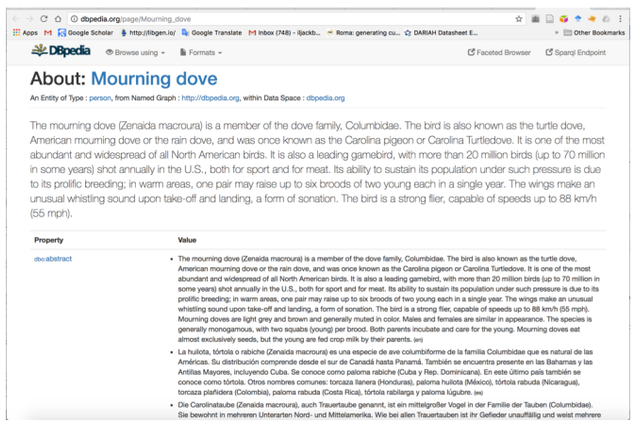
Mixtec Codex Nuttal- British Museum

TEI Dictionary Structure



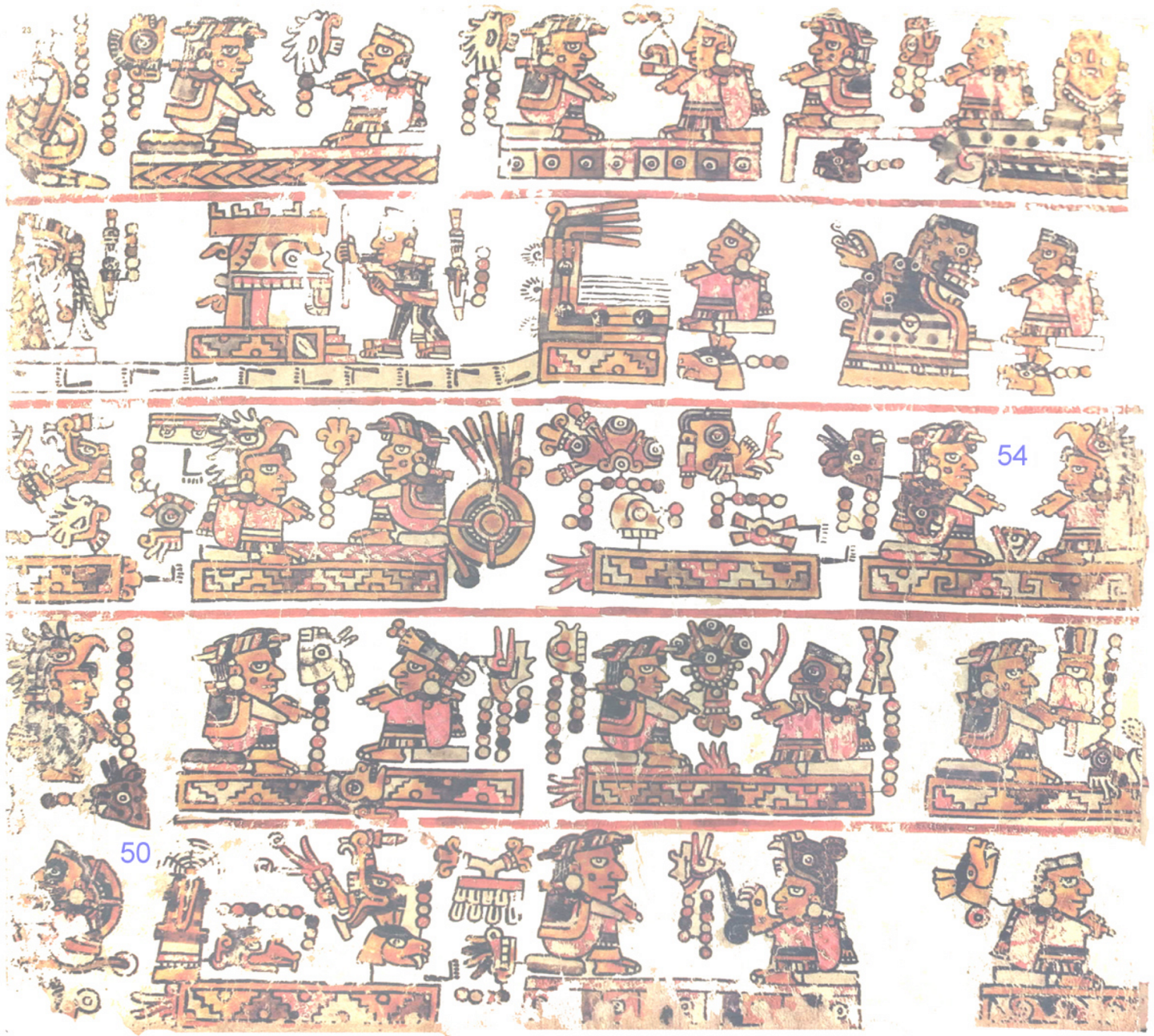
TEI Dictionary Entry:
Basic example: TEI






```
<entry xml:id="bird-mourning_dove">
  <form type="lemma">
    <orth>lakuku</orth>
    <pron notation="ipa">laɻku˥kuˀ</pron>
  </form>
    <gramGrp>
      <pos>noun</pos>
    </gramGrp>
  <sense corresp="http://dbpedia.org/resource/Mourning_dove">
    <usg type="domain" corresp="http://dbpedia.org/resource/Bird">Bird</usg>
     <cit type="translation" xml:lang="en" corresp="https://en.wiktionary.org/wiki/mourning_dove">
    <!— hypernymOf(Bird) —>
      <quote>mourning dove</quote>
     </cit>
     <cit type="translation" xml:lang="es" corresp="https://es.wiktionary.org/wiki/tortolita">
      <quote>tortolita</quote>
     </cit>
  </sense>
  <cit type="example" corresp="/SIL_docs/L152/L152-tok.xml#L152-01-01">
    <quote>In kii ra iin <oRef>lakuku</oRef> kunia tanta'i tsi in ncho'o, cha koo xu'in sa'i viko.</quote>
  </cit>
  <ref type="soundfile" target="N_mourning_dove_01_TS.wav"/>
  <!-- could also include references to images (where available) -->
</entry>
```

(III) TEI Dictionary
ii. Etymology



Mixtec Codex Bodley - Bodleian Library Oxford

Overview of Issues & Sources in Mixtec Etymology

- Sense based changes (*metaphor, metonomy, grammaticalization*) identified in related Mixtec languages important for cognitive linguistics (Brugman, 1983; Beckman, 1995;…)
  - *especially body-part terms*

Sources

- Reconstructed Proto-Mixtecan from Longacker
- Loanwords easily identifiable (mostly from Spanish)
  - *phonological changes evident in these*
- Oldest sources of documentation of Mixtec languages from 1592

Issues, Challenges

- *Need to balance personal/academic/intelectual interests with the practicalities and needs of the language community*
- *Need Mixtec editors to provide Mixtec language versions of content (requires new use of language vocabulary!)*
- Other issues: *one man job*

Overview of Components of Etymological Markup in TEI

Bowers & Romary (2016) propose expansion and refinement of etymology section of the TEI dictionary

TEI Lex0 Etym (Bowers & al., 2018) builds off this…..

- Etymology element (<etym>): structuring etymology processes through typing and recursivity
- Typology of etymological processes
- Etymons and their forms
- Related forms (cognates, and others)
- Temporality of etymological processes
- Bibliographical references in etymologies
- Prose description of etymological process and content

Etymological Processes: Borrowing

There is a significant amount of loanwords in the language (the vast majority of which are from Spanish).

```
<entry xml:id="mother">
  <form type="lemma">
    <orth xml:lang="mix">maa</orth>
    <pron xml:lang="mix" notation="ipa">māā</pron>
    <gramGrp>
      <pos>noun</pos>
    </gramGrp>
  </form>
  <sense>
   <!— other info here -->
    <cit type="translation">
      <form>
        <orth xml:lang="en">mother</orth>
      </form>
    </cit>
  </sense>
   <etym type="borrowing">
    <cit type=“etymon">
      <lang xml:lang="en">Sp.</lang>
       <form>
         <orth xml:lang=“es">madre</orth>
         <pron xml:lang="es">'madre</pron>
       </form>
    </cit>
  </etym>
</entry>
```

Etymological Markup: Inheritance & Cognates

Generally inheritance from an ancestor form can be inferred via comparative with cognates in related Mixtec (and Mixtecan) varieties;

Mixtepec Mixtec

xini

[ʃìnĭ]

'head'

Proto-Mixtecan

*ʃīnī

'head'

Mixtec Variety	form	source
Ayutla	shīhih	(Hills, 1990)
San Martín Duraznos	ʃīɲī	(Padgett, 2017)
Guadalupe Nundaca	ʃīɲī	(Padgett, 2017)
Santa Rosa Caxtlahuaca	ʃīɲī	(Padgett, 2017)
Santa Catarina Noltepec	ʃīɲī	(Padgett, 2017)
San Miguel Cuevas	ʃīɲī	(Padgett, 2017)
Yucunicoco	ʃīɲī	(Padgett, 2017)
Coicoyán de las Flores	ʃìnĭ	(Padgett, 2017)
Chalcatongo Mixtec (San Miguel El Grande)	šinì	(Macaulay, 1996)

Etymological Features: Cognates

Related forms extracted (manually) from academic papers on other Mixtec varieties

….

(Chalcatongo Mixtec: San Miguel El Grande) *šinì (Macaulay, 1996)*;

(Ayutla Mixtec) *shīhih (Hills, 1990)*;

(Coatzospan Mixtec) *ʃīɲī (Padget, 2017)*;

(Guadalupe Nundaca) *ʃīɲī (Macaulay, 1996)*;

```
<cit type="cognate">
  <lang>Chalcatongo Mixtec</lang>
  <usg type="geo">
    <placeName>San Miguel El Grande</placeName>
  </usg>
  <form>
    <pron notation="trans-macaulay-mig" xml:lang="mig">šinì</pron>
  </form>
  <ref type="source" target="#Macaulay-ChalcatongoMixtec-1996">(Macaulay, 1996)</ref>
</cit>

<cit type="cognate">
  <lang>Ayutla Mixtec</lang>
  <form>
    <pron notation="trans-hill-1990-miy" xml:lang="miy">shīhih</pron>
  </form>
  <ref target="#Hills-AyutlaMixtec-1990">(Hills, 1990)</ref>
</cit>

<cit type="cognate">
  <lang>San Martín Duraznos</lang>
  <form>
    <pron notation="ipa" xml:lang="smd">ʃɪɲī</pron>
  </form>
  <ref target="#Padgett-2017">(Padget, 2017)</ref>
</cit>

<cit type="cognate">
  <lang>Guadalupe Nundaca</lang>
  <form>
    <pron notation="ipa" xml:lang="gna">ʃɪɲī</pron>
  </form>
  <ref target="#Padgett-2017">(Padget, 2017)</ref>
</cit>
```

Etymological Features: Cognates & Bibliographic Sources

<bibl xml:id="“Macaulay-ChalcatongoMixtec-1996">
Macaulay, M. (1996). A grammar of Chalcatongo Mixtec (University of California publications in linguistics ed.). University of California Press.</bibl>

<bibl xml:id="“Hills-AyutlaMixtec-1990">
Hills, Robert A. 1990. A syntactic sketch of Ayutla Mixtec. Studies in the Syntax of Mix-tecan Languages, vol. 2, Summer Institute of Linguistics and the University of Texas at Arlington Publications in Linguistics, no. 90, ed. C. Henry Bradley and Barbara E. Hollenbach, pp. 1-260. Dallas: Summer Institute of Linguistics and the University of Texas at Arlington. </bibl>

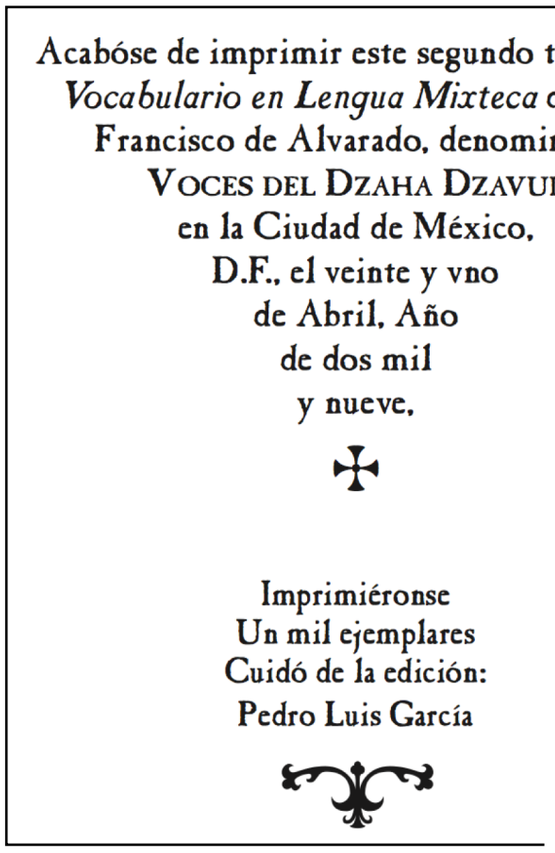
<bibl type="thesis" xml:id="“Padgett2017">
Padgett, E. (2017). Tools for assessing relatedness in understudied language varieties: a survey of Mixtec varieties in Western Oaxaca, Mexico. University of North Dakota. </bibl>

```
<cit type="cognate">
  <lang>Chalcatongo Mixtec</lang>
  <usg type="geo">
    <placeName>San Miguel El Grande</placeName>
  </usg>
  <form>
    <pron notation="trans-macaulay-mig" xml:lang="mig">šinì</pron>
  </form>
  <ref type="bibl" target="“#Macaulay-ChalcatongoMixtec-1996">
      (Macaulay, 1996)</ref>
</cit>
<cit type="cognate">
  <lang>Ayutla Mixtec</lang>
  <form>
    <pron notation="trans-hill-1990-miy" xml:lang="miy">shīhih</pron>
  </form>
  <ref type="bibl" target="#Hills-AyutlaMixtec-1990">(Hills, 1990)</ref>
</cit>
<cit type="cognate">
  <lang>San Martín Duraznos</lang>
  <form>
    <pron notation="ipa" xml:lang="smd">ʃīɲī</pron>
  </form>
  <ref type="bibl" target="#Padgett-2017">(Padget, 2017)</ref>
</cit>
<cit type="cognate">
  <lang>Guadalupe Nundaca</lang>
  <form>
    <pron notation="ipa" xml:lang="gna">ʃīɲī</pron>
  </form>
  <ref type="bibl" target="#Padgett-2017">(Padget, 2017)</ref>
</cit>
```

Integrating Sources: Colonial Mixtec Dictionary

- *Using GROBID Dictionaries (Khemakhem et al., 2017) to automatically create structured TEI dictionary from PDF*

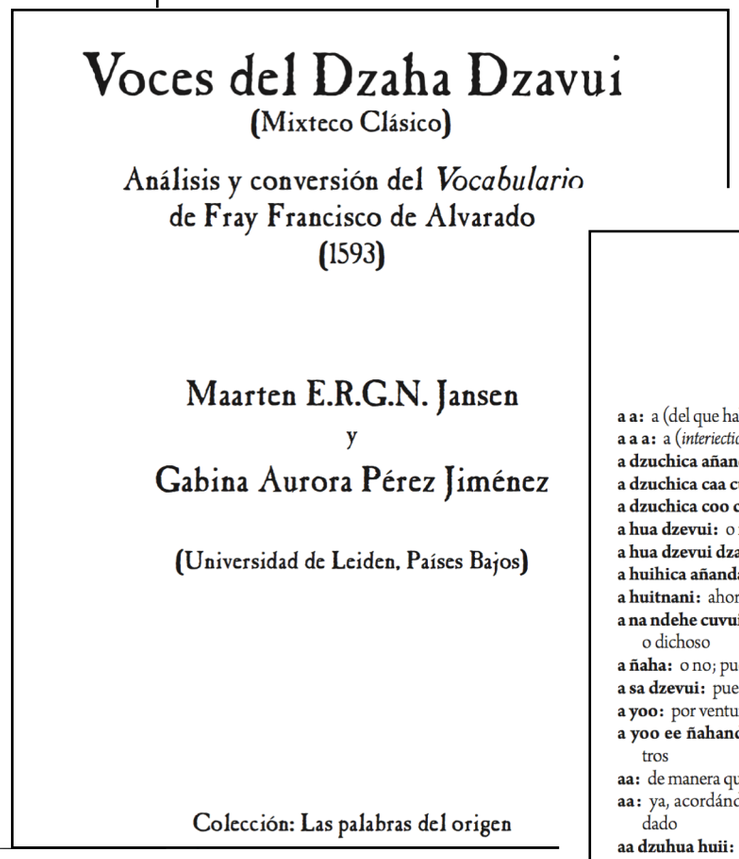
Provides new source of etymons!!



Acabóse de imprimir este segundo tomo del *Vocabulario en Lengua Mixteca* de Fray Francisco de Alvarado, denominado VOCES DEL DZAHA DZAVUI, en la Ciudad de México, D.F., el veinte y vno de Abril, Año de dos mil y nueve.

✠

Imprimiéronse
Un mil ejemplares
Cuidó de la edición:
Pedro Luis García



Voces del Dzaha Dzavui
(Mixteco Clásico)

Análisis y conversión del *Vocabulario* de Fray Francisco de Alvarado (1593)

Maarten E.R.G.N. Jansen
y
Gabina Aurora Pérez Jiménez

(Universidad de Leiden, Países Bajos)

Colección: Las palabras del origen

A

a a: a (del que halla a otro en maleficio)
a a a: a (*interiectio admirantis*)
a dzuchica añandaa: poco más o menos
a dzuchica caa cuvui: poco más o menos
a dzuchica coo cuvui: poco más o menos
a hua dzevui: o no
a hua dzevui dzavua: pues no
a huihica añandaa: poco más o menos
a huitnani: ahora poco
a na ndehe cuvui ndatu nicay: o bienaventurado, o dichoso
a ñaha: o no; pues no
a sa dzevui: pues no
a yoo: por ventura alguno
a yoo ee ñahando: por ventura alguno de vosotros
aa: de manera que
aa: ya, acordándoseme lo que se me había olvidado
aa dzuhua huii: así, así (sonriéndose)
aa ndica huii: ya (acordándoseme lo que se me había olvidado)
adzi: o (*disyuntiva*); por ventura; quizás
adzi: suave cosa
adzi cuvui: o (*disyuntiva*); por ventura; quizás
adzi q cuvui: por ventura
adzi yoo: por ventura alguno
adzi yoo ee ñahando: por ventura alguno de vosotros
ahua: ay, quejándose la mujer
ama: así
ama: bien está (otorgando); sí
amana: ¿cuándo? (*adverbio interrogativo*), ¿en qué tiempo?
amana cuiya: en algún tiempo
amana cuvui huatu inindo: cuandoquiera que quisieres
amana na ndita ñumana nuundo: ¿cuándo has de despertar?
amana na ndotondo: ¿cuándo has de despertar?
amana na tahui inindo: ¿cuándo has de volver en ti?; ¿cuándo has de despertar?
amana quevui: en algún tiempo
amana quevui sa cuvui inindo: cuandoquiera que quisieres
amanaca: ¿cuándo, en qué tiempo?
amani: de tarde en tarde o raras veces
andaya: infierno, lugar de dañados
andevui: cielo
andevui isi ndaa tiñoo: cielo estrellado
angel nicoo coo ndaa ndita ña: ángel de mi guarda
angel yondaca ñaha: ángel de mi guarda
aniñe: palacio
anuhu: abismo; centro de la tierra; infierno, lugar de dañados
anuhu: profundo
anuhu maa: profundo
anuhu naa: infierno, lugar de dañados
anuhu ndahui: infierno, lugar de dañados
atana: ojalá
atu: amarga cosa; áspero al gusto
aya: amarga cosa; áspero al gusto

Jansen, M. E. R. G. N., & Perez, G. A. (2009). Voces del Dzaha Dzavui (mixteco clásico). Análisis y Conversión del Vocabulario de fray Francisco de Alvarado (1593).

Integrating Sources: Colonial Mixtec Dictionary

andevui: cielo

```
<entry xml:id="andevui">
  <form type="lemma">
    <orth xml:lang="nds-x-clmx">andevui</orth>
  </form>
  <pc>:</pc>
  <sense>
    <def xml:lang="es">cielo</def>
  </sense>
</entry>
```

```
<entry xml:id="sky">
  <form type="lemma">
    <orth xml:lang="mix">antivi</orth>
    <gramGrp>
      <pos>noun</pos>
    </gramGrp>
  </form>
  <sense corresp="http://dbpedia.org/resource/Sky">
    <usg type="domain">Meteorology</usg>
    <cit type="translation">
      <form>
        <orth xml:lang="en">sky</orth>
      </form>
    </cit>
    <cit type="translation">
      <form>
        <orth xml:lang="es">cielo</orth>
      </form>
    </cit>
  </sense>
  <etym type="inheritance">
    <cit type="etymon">
      <form>
        <orth xml:lang="nds-x-clmx">andevui</orth>
      </form>
      <gloss xml:lang="es">cielo</gloss>
      <ref type="bibl" target="#VOCESvocab-tei.xml">
        Francisco de Alvarado</ref>
    </cit>
  </etym>
</entry>
```

Jansen, M. E. R. G. N., & Perez, G. A. (2009). Voces del Dzaha Dzavui (mixteco clásico). Análisis y Conversión del Vocabulario de fray Francisco de Alvarado (1593).

Etymological Markup: Derrivation

MIX

ntasaxeen

‘to sharpen’

Derrivation

nta- + sa- + xeen

ITER + CAUS + dangerous

```
<entry xml:id="sharpen">
  <form type="lemma">
    <orth>ntasaxeen</orth>
  </form>
  <gramGrp>
    <pos>verb</pos>
    <gram type="transitivity">transitive</gram>
    <gram>causative</gram>
    <gram>iterative</gram>
  </gramGrp>
  <sense>
    <cit type="translation" xml:lang="en">
      <quote>sharpen</quote>
    </cit>
  </sense>
  <etym type="derivation">
    <cit type="etymon">
      <oRef>nta<pc>-</pc></oRef>

    </cit>
    <cit type="etymon">
      <oRef>sa<pc>-</pc></oRef>

    </cit>
    <cit type="etymon">
      <oRef>xeen</oRef>
      <gramGrp>
        <pos>adj</pos>
      </gramGrp>
      <gloss>dangerous</gloss>
    </cit>
  </etym>
</entry>
```

Etymological Processes: Derivation

```
<entry xml:id="sharpen">
  <form type="lemma">
    <orth xml:lang="mix">ntasaxeen</orth>
  </form>
  <gramGrp>
    <pos>verb</pos>
    <gram type="transitivity">trans</gram>
    <gram>causative</gram>
    <gram>iterative</gram>
  </gramGrp>
  <sense>
    <cit type="translation">
      <form>
        <orth xml:lang="en">sharpen</orth>
      </form>
    </cit>
    <cit type="translation">
      <form>
        <orth xml:lang="es">afilar</orth>
      </form>
    </cit>
  </sense>
  <etym type="derivation">
  ……
  </etym>
</entry>
```

nta-

sa-

xeen

MIX

ntasaxeen

‘to sharpen’

```
<etym type="derivation">
  ….
  <cit type="etymon">
    <form>
      <orth xml:lang="mix">nta-</orth>
    </form>
    <gramGrp>
      <gram>prefix</gram>
       <gram>iterative</gram>
    </gramGrp>
  </cit>
  <pc>+</pc>
  <cit type="etymon">
    <form>
      <orth xml:lang="mix">sa-</orth>
    </form>
    <gramGrp>
      <gram>prefix</gram>
       <gram>causative</gram>
    </gramGrp>
  </cit>
  <pc>+</pc>
  <cit type="etymon">
    <form>
      <orth xml:lang="mix">xeen</orth>
    </form>
    <gramGrp>
      <pos>adj</pos>
    </gramGrp>
    <gloss xml:lang="en">dangerous</gloss>
    <gloss xml:lang="es">peligroso</gloss>
  </cit>
</etym>
```

Etymological Markup: Sense-based lexical innovation in Mixtec

Analysis of polysemy (*particularly body-part terms*) in Mixtecan languages (Brugman,1983), (Brugman and Macaulay,1986), (Hollenbach, 1995) provides evidence to support several key theoretical questions regarding patterns of lexical innovation; particularly, those involving:

(i) Lexical and cognitive strategies responsible for certain semantic changes (*i.e. Metaphor & Metonymy*);

(ii) Diachronic directionality, both on the semantic, and grammatical levels of the language (concrete > abstract)

for specific, indepth analysis of this phenomena see: (Bowers, 2016): http://bit.ly/2FnsKPU

Etymological Markup:
Examples of Mixtec Polysemy- **nuu** ‘*face’*

nuu

‘face’

nu-u̲

face\1sg

‘my face’

nuu ve’e

[face+house]

‘front (part) of the house’

intu’u saa-ka **nu-u̲**

sit[3SG.INF] bird-TPC **face\1SG**

‘the bird is sitting in front of me’

nuu tsa’a ña’a

[face] foot woman

‘..on the woman’s foot’

nuu yuku inkaa-yu

[face] forest cop.loc-1sg

‘I am in the forest’

ntava chumi-ka **nuu** yutu

CMPL/fly owl-TPC [face] tree

‘the owl flew into the tree’

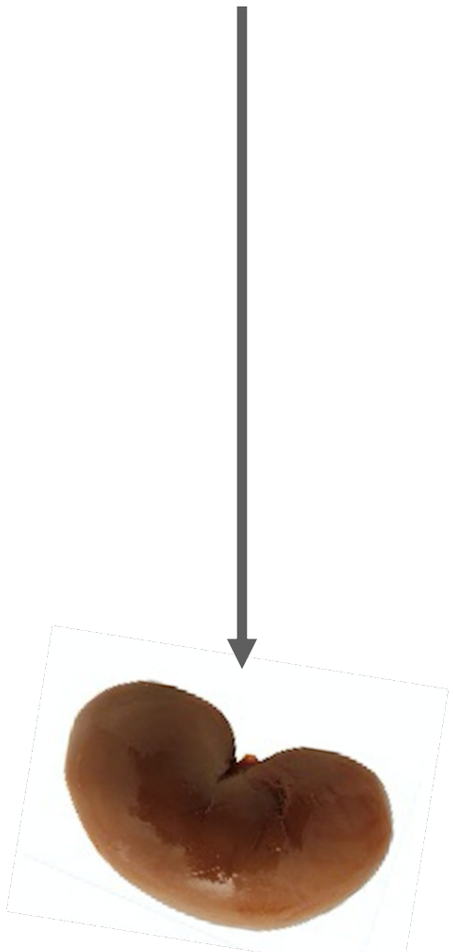
ntakoo chumi-ka **nuu** yutu

CMPL/get.up owl-TPC [face] tree

‘the owl got up, left the tree

Etymological Processes: Metaphor

source concept



target concept

```
<sense xml:id="bean" corresp="http://dbpedia.org/resource/Pinto_bean" n="1">
    <usg type="domain">Legume</usg>
    <usg type="domain">stapleFoods</usg>
    <cit type="translation">
      <form>
        <orth xml:lang="en">bean</orth>
      </form>
    </cit>
    <cit type="translation">
      <form>
        <orth xml:lang="es">frijol</orth>
      </form>
    </cit>
  </sense>
  <sense xml:id="kidney" corresp="http://dbpedia.org/resource/Kidney" n="2">
    <usg type="domain">InternalOrgan</usg>
    <cit type="translation">
      <form>
        <orth xml:lang="en">kidney</orth>
      </form>
    </cit>
    <cit type="translation">
      <form>
        <orth xml:lang="es">riñón</orth>
      </form>
    </cit>
    <etym type="metaphor">
     ……
    </etym>
  </sense>
```

Etymological Processes: Metonymy

kiti *(n.)*
1) (Es.) animal; (En.) animal
2) (Es.) caballo; (En.) horse
……

ANIMAL is a hypernymOf HORSE

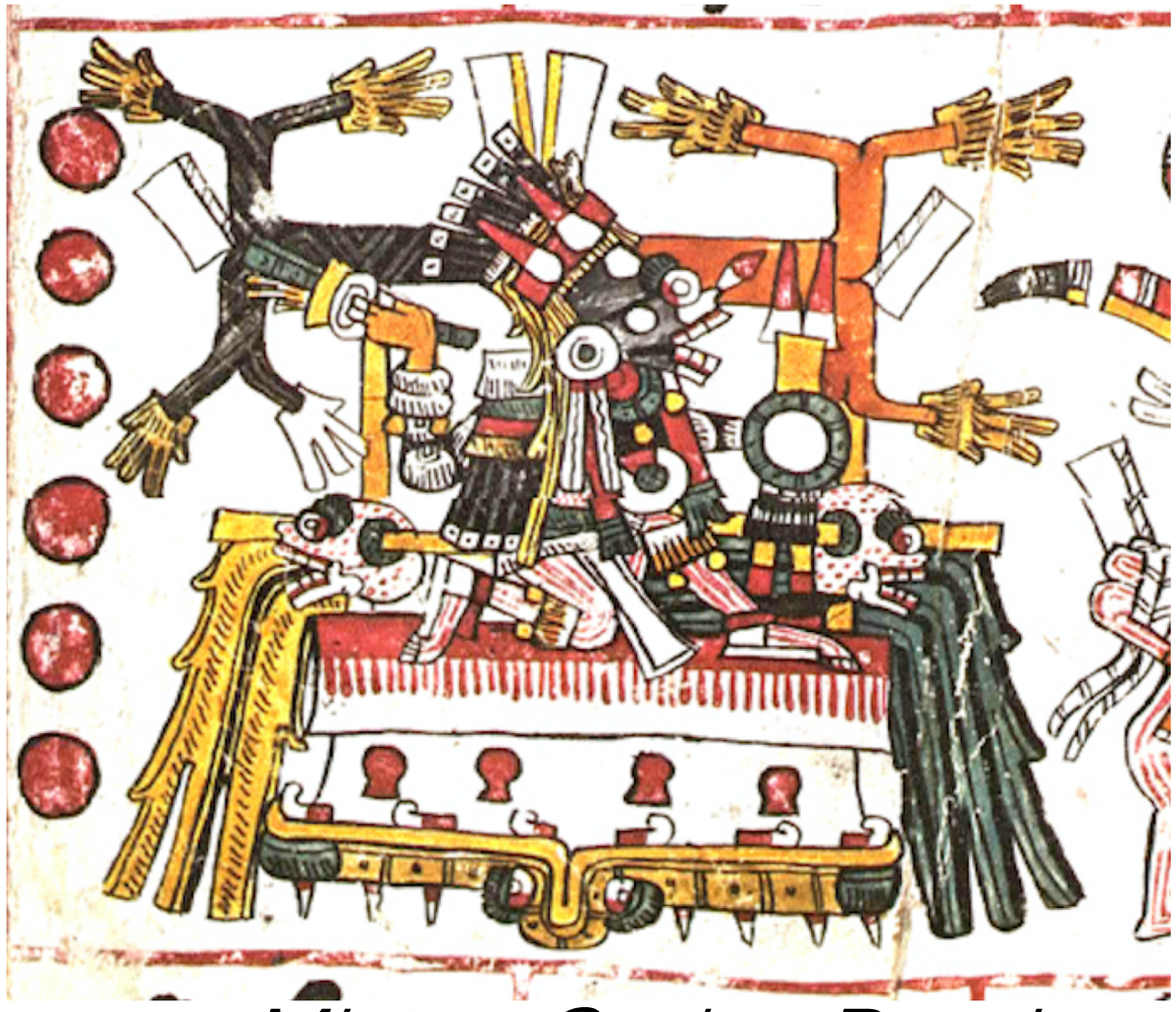
(implicit in data)

HORSE is a hypoymOf ANIMAL

```
<entry xml:id="animal-horse">
  <form type="lemma">
    <orth xml:lang="mix">kiti</orth>
    <pron xml:lang="mix" notation="ipa">kìt̪í</pron>
    <gramGrp>
      <pos>noun</pos>
    </gramGrp>
  </form>
  <sense xml:id="animal" corresp="http://dbpedia.org/resource/Animal" n="1">
    <usg type="domain">livingBeing</usg>
    <cit type="translation">
      <form><orth xml:lang="en">animal</orth></form>
    </cit>
    …..
  </sense>
```

```
  <sense xml:id="horse"  corresp="http://dbpedia.org/resource/Horse" n="2">
    <usg type="domain">Animal</usg>
    <xr type="hyponymOf">
      <ref xml:lang="mix">kiti</ref>
      <ref xml:lang="en">animal</ref>
      <ref type="sense" corresp="#horse"/>
    </xr>
    <cit type="translation">
      <form><orth xml:lang="en">horse</orth></form>
    </cit>
    <cit type="translation" xml:lang="es">
      <form><orth xml:lang="es">caballo</orth></form>
    </cit>
    <etym type="metonymy" subtype="categoryForMember">
     ….
    </etym>
  </sense>
</entry>
```

Etymological Processes: Multiple Processes



Mixtec Codex Borgia

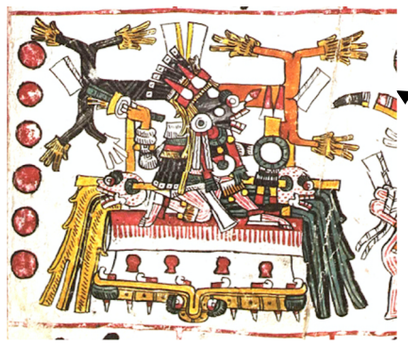
nta’a+ yutu

hand/arm + tree

‘tree branch’

‘rama’

Etymological Processes: Multiple Processes



nta’a yutu
hand/arm tree
‘tree branch’
‘rama’

```
<entry xml:id="tree-branch" type="compound">
  <form type="lemma">
    <orth>nta'a yutu</orth>
    <pron notation="ipa">ndà?á jùtú</pron>
    <gramGrp>
      <gram>noun</gram>
    </gramGrp>
  </form>
  <sense>
    <graphic url=".../imgs/apoala_tree.jpg"/>
    <usg type="domain">Botany</usg>
    <xr type="meronymOf">
      <ref xml:lang="mix">yutu</ref>
      <ref xml:lang="en">tree</ref>
    </xr>
    <cit type="translation">
      <form><orth xml:lang="en">branch</orth></form>
    </cit>
    <cit type="translation">
      <form><orth xml:lang="es">rana</orth></form>
    </cit>
  </sense>
  <etym type=“compounding">
    ……..
  </etym>
</entry>
```

```
<etym type="compounding">
  <etym type="metaphor">
    <cit type="etymon">
      <form>
        <orth>nta'a</orth>
      </form>
      <sense>
        <usg type="domain">HumanAntomy</usg>
        <xr type="meronymOf">
          <ref xml:lang="mix">kuñu</ref>
          <ref xml:lang="en">body</ref>
        </xr>
        <gloss xml:lang="en">hand</gloss>
        <gloss xml:lang="es">mano</gloss>
      </sense>
    </cit>
  <pc>+</pc>
  </etym>
  <cit type="etymon">
    <form>
      <orth>jutu</orth>
    </form>
    <gloss xml:lang="en">tree</gloss>
    <gloss xml:lang="es">arbol</gloss>
  </cit>
</etym>
```

Next Steps (near future)

- Make use of/ implement the @lemma in <w> to link all inflected word forms/phrases with their common lemma
- Implement Predicate Logic-Based linguistic structual descriptions
- Establish more refined translation typology
- Improve/standardize automatic processing, markup programming
- Publish in open repository (dataverse) & disseminate the corpus in CC-BY
- Register with OLAC
- Enhance translated content by linking to wiktionary
- Produce conversion scripts to make convertible to LOD (Lemon-ONTOLEX & LMF reserialzation)
- Produce corpus based studies of polysemy and etymological processes (particularly in Body-part terms):
  - *upcoming paper in publication for procedings of PUCP (2016)*
- Define concepts taxonomy for senses and domains
- Make data output compatible with FLeX toolkit used by SIL Mixtec researchers

Inria
INVENTEURS DU MONDE NUMÉRIQUE

ÖAW
AUSTRIAN ACADEMY OF SCIENCES

Further Development (longer term)

- Create new materials in Mixtec
- Create permanant online hub for hosting, accessing, and adding data (*possibly including crowd sourcing*)
- Online website where users can read and view the original content and access the translations and annotations as well
- Enhance dictionary with MIX language definitions of entry content (to improve the usability for native speakers)
- Attempt automatic annotation of phonetics in Praat using build-in machine learning capacity
- Produce systematic studies comparing the speech of MIX speakers who reside only in Mexico and those who live in the US
- Integrate dataset into LD ontology for the concepts observed in etymology, integrate into model (e.g. *Framester*?)
- Extract & integrate available vocabulary for related Mixtec varieties
- Extract & integrate historical dictionary (from 1590's into dataset)
- Expand project to enable onomasiological-based collection and storage of all Mixtec varieties

Further work: Codex-based materials

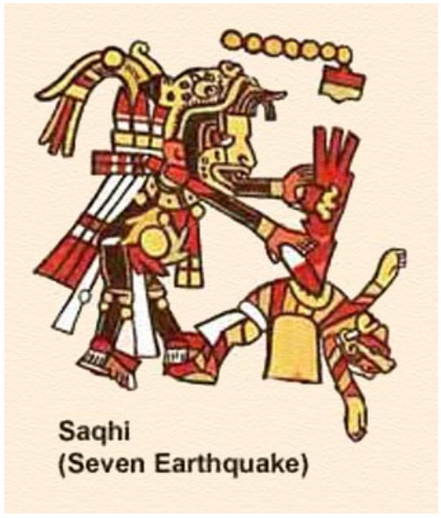
A major goal is to develop more materials in Mixtec, an ideal and un-tapped resource to use for this are the Mixtec codexes (….)

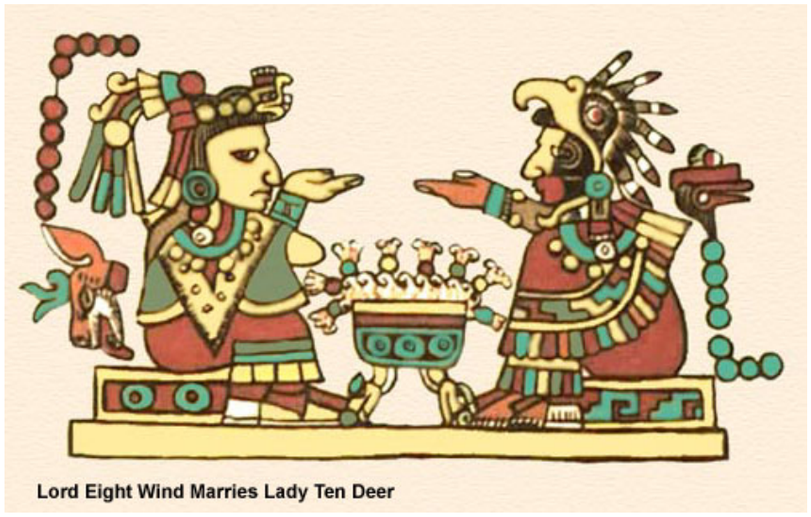
These can be used for both the creation of materials for adults, children and language learners;

img source: FAMSI

Image source: Pohl, J. (n.d.). FAMSI - John Pohl's - Ancient Books - Mixtec Group Codices. Retrieved March 6, 2018, from http://www.famsi.org/research/pohl/jpcodices/pohlmixtec8.html

Codex-based materials: translation of codex snippet





Utsa Taàn

Ntanta'a (lord)? Una Tachi ka tsi (Na'a) Utsi Isu

Yuku Saa Yuku Yoo ? ?

Image source: Pohl, J. (n.d.). FAMSI - John Pohl's - Ancient Books - Mixtec Group Codices. Retrieved March 6, 2018, from http://www.famsi.org/research/pohl/jpcodi

Academic Papers: Pike & Ibach (1978)

Paper is the 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)

The amount of data is too small to justify automation but it is important content both for comparing my results and in maximizing the quantity of data collected

Non-IPA

¢ = t͡s
š = ʃ
č = t͡ʃ
ʒ = t͡z
ǰ = d͡ʒ
ɡ = k̟

Tone levels are reversed from conventional description

^{3} = Low
^{2} = Mid
^{1} = High

Strings are often interrupted

š[i.]^{2}ši^{3}-¢i^{2} '

ko^{1}lo^{1}-ko^{1} 'our (excl.) male turkey', š[i.]^{2}ši^{3}-¢i^{2} 'his or her (child) aunt', s[o.]^{3}ko^{3}-yu^{3} 'my (polite) collarbone', z,[i.]^{3}¢a$^{3\text{-}1}$ 'sandal' ti^{3}k^{w}[a.]^{3}a^{2} 'butterfly'; but, laʔ^{2}la^{2}-¢i2 'his or her (child) mucus', ǰa^{1}ʔa^{1}-ni^{1} 'your (sing, polite) gravy'.

**I still need to decide the best way to represent tone… in my phonetic transcriptions*

Academic Papers: Pike & Ibach (1978)

ko[1]lo[1]-ko[1] 'our (excl.) male turkey', š[i.][2]ši[3]-¢i[2] 'his or her (child) aunt', s[o.][3]ko[3]-yu[3] 'my (polite) collarbone', z,[i.][3]¢a[3-1] ‘sandal' ti[3]k[W][a.][3]a[2] 'butterfly'; but, la[ʔ2]la[2]-¢i2 'his or her (child) mucus', ǰa[1ʔ]a[1]-ni[1] 'your (sing, polite) gravy'.