TEI Encoding of a Classical Mixtec Dictionary Using GROBID-Dictionaries

Jack Bowers[1], Mohamed Khemakhem[2], Laurent Romary[3]

[1,2,3] Inria-ALMAnaCH - Automatic Language Modelling and ANAlysis & Computational Humanities, Paris, France

[1] EPHE - École Pratique des Hautes Études, Paris, France

[1] ACDH - Austrian Center for Digital Humanities, Vienna, Austria

[2] UPD7 - Université Paris Diderot - Paris 7, Paris, France

[2] CMB – Centre Marc Bloch, Berlin, Germany

[2] BBAW – Berlin-Brandenburg Academy of Sciences and Humanities, Berlin, Germany
E-mail: iljackb@gmail.com, mohamed.khemakhem@inria.fr, laurent.romary@inria.fr

Inria
INVENTEURS DU MONDE NUMÉRIQUE

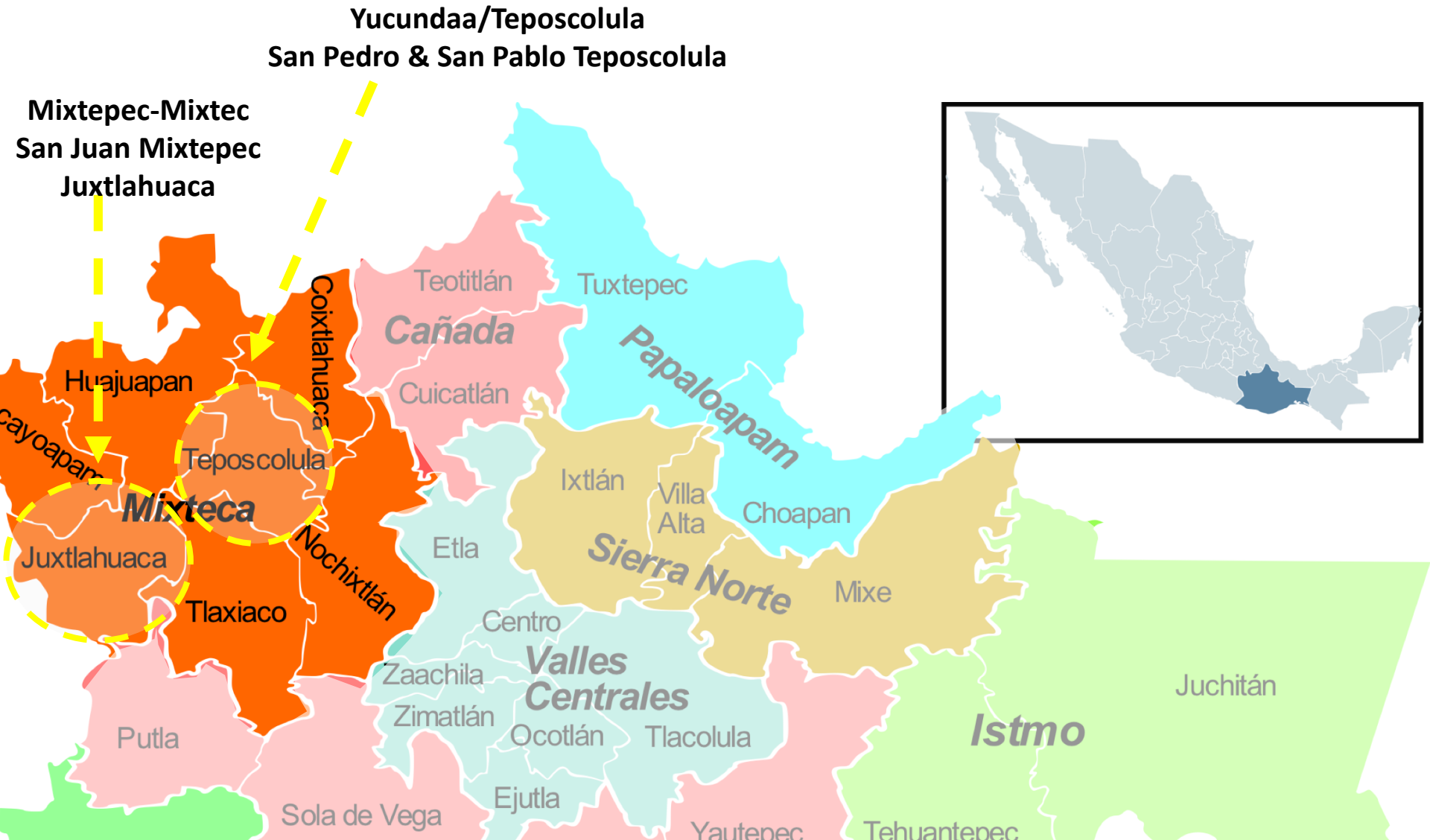
Overview of the Source & Output

- ‘Vocabulario en lengva misteca’ published by the Dominican Francisco de Alvarado (1593)
- Variety from Teposcolula Mexico (Mixteca Alta)
  - Classical Mixtec/Colonial Mixtec/Yucu Ndaa
- Entries based on three earlier dictionary sources:
  - Castilian-Nahuatl (Valley of Mexico, 1571)
  - Castilian-Zapotec (Valley of Oaxaca, 1578)
  - Castilian-Latin (1492)
- PDF re-organized, modernized version ‘Voces de Dzaha Dzahui’ (Jansen & Pérez Jiménez, 2009)
- TEI dictionary produced contains roughly 26,600 entries and related entries.
- Structure is TEI Lex-0 compatible

Utility/Purpose of Endeavour

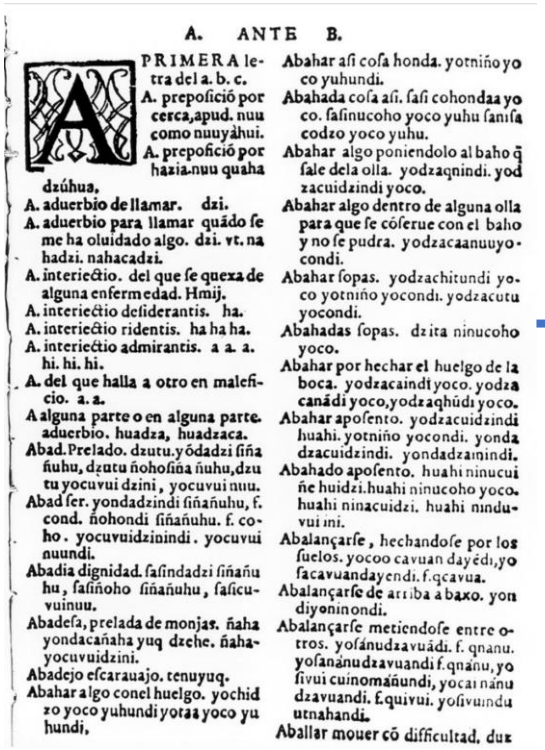
- Increase coverage of relevant lexical material in Mixtepec-Mixtec documentation (ISO 693-3 [mix])
  - Link and cross-reference in Mixtepec TEI dictionary
- Machine searchable data set for:
  - study of the Yucu Ndaa variety
  - historiographical and philological research
- Create a more cohesive body of pan-Mixtecan resources
  - Vocabulary for cross Mixtecan comparison; (81 Varieties of Mixtec)
- TEI format can easily be exported into other formats for non-TEI users

La Mixteca (Mixtec Region)



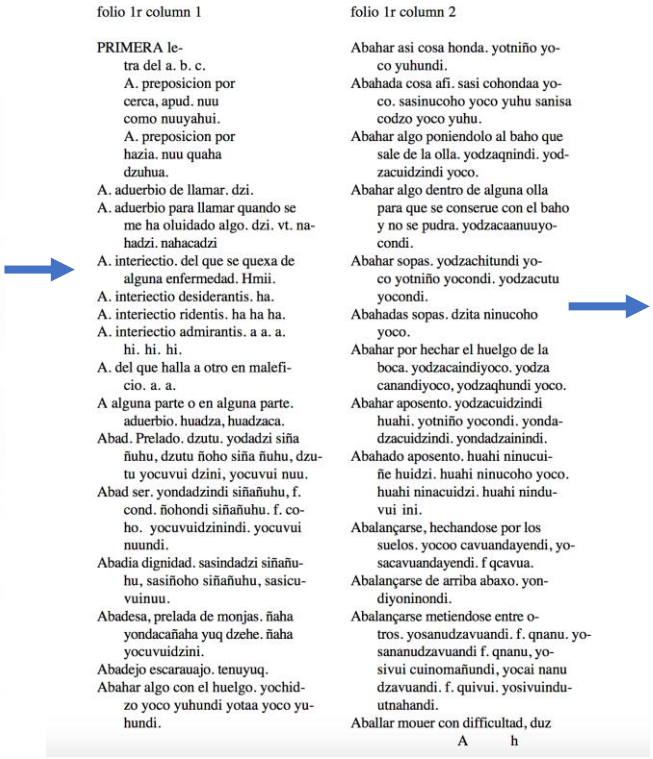
Versions of Resource

- Original (Printed: 1593) > (facsimile edition 1965)
- Mesolore (Bakewell & Hamman, 2001)
  - Digitized from scanned copy
- Jansen and Pérez Jimenez 2009



A. ANTE B.

PRIMERA letra del a. b. c.
A. prepoſició por cerca,apud. nuu como nuuyàhui.
A. prepoſició por hazia.nuu quaha dzúhua.
A. aduerbio de llamar. dzi.
A. aduerbio para llamar quãdo ſe me ha oluidado algo. dzi. vt. na hadzi. nahacadzi.
A. interiectio. del que ſe quexa de alguna enfermedad. Hmij.
A. interiectio deſiderantis. ha.
A. interiectio ridentis. ha ha ha.
A. interiectio admirantis. a a. a. hi. hi. hi.
A. del que halla a otro en maleficio. a. a.
A alguna parte o en alguna parte. aduerbio. huadza, huadzaca.
Abad. Prelado. dzutu.yõdadzi ſiña ñuhu, dzutu ñohoñia ñuhu,dzu tu yocuvui dzini, yocuvui nuu.
Abad ſer. yondadzindi ſiñañuhu, f. cond. ñohondi ſiñañuhu. f. coho. yocuvuidzinindi. yocuvui nuundi.
Abadia dignidad. ſaſindadzi ſiñañu hu, ſaſiñoho ſiñañuhu, ſaſicuvuinuu.
Abadeſa, prelada de monjas. ñaha yondacañaha yuq dzehe. ñahayocuvuidzini.
Abadejo eſcarauajo. tenuyuq.
Abahar algo conel huelgo. yochid zo yoco yuhundi yotaa yoco yu hundi,
Abahar aſi coſa honda. yotniño yo co yuhundi.
Abahada coſa aſi. ſaſi cohondaa yo co. ſaſinucoho yoco yuhu ſaniſa codzo yoco yuhu.
Abahar algo poniendolo al baho q̃ ſale dela olla. yodzaqnindi. yod zacuidzindi yoco.
Abahar algo dentro de alguna olla para que ſe cõſerue con el baho y no ſe pudra. yodzacaanuuyocondi.
Abahar ſopas. yodzachitundi yoco yotniño yocondi. yodzacutu yocondi.
Abahadas ſopas. dzita ninucoho yoco.
Abahar por hechar el huelgo de la boca. yodzacaindi yoco. yodza canãdi yoco,yodzaqhũdi yoco.
Abahar apoſento. yodzacuidzindi huahi. yotniño yocondi. yonda dzacuidzindi. yondadzaindi.
Abahado apoſento. huahi ninucui ñe huidzi.huahi ninucoho yoco. huahi ninacuidzi. huahi ninduvui ini.
Abalançarſe, hechandoſe por los ſuelos. yocoo cavuan dayẽdi,yo ſacavuandayendi. f.qcavua.
Abalançarſe de arriba a baxo. yon diyoninondi.
Abalançarſe metiendoſe entre otros. yoſánudzavuãdi. f. qnanu. yoſanánudzavuandi f.qnánu, yo ſivui cuinomãñundi, yocai nánu dzavuandi. f.quivui. yoſivuindu utnahandi.
Aballar mouer cõ difficultad. duz



folio 1r column 1

PRIMERA letra del a. b. c.
A. preposicion por cerca, apud. nuu como nuuyahui.
A. preposicion por hazia. nuu quaha dzuhua.
A. aduerbio de llamar. dzi.
A. aduerbio para llamar quando se me ha oluidado algo. dzi. vt. nahadzi. nahacadzi
A. interiectio. del que se quexa de alguna enfermedad. Hmii.
A. interiectio desiderantis. ha.
A. interiectio ridentis. ha ha ha.
A. interiectio admirantis. a a. a. hi. hi. hi.
A. del que halla a otro en maleficio. a. a.
A alguna parte o en alguna parte. aduerbio. huadza, huadzaca.
Abad. Prelado. dzutu. yodadzi siña ñuhu, dzutu ñoho siña ñuhu, dzutu yocuvui dzini, yocuvui nuu.
Abad ser. yondadzindi siñañuhu, f. cond. ñohondi siñañuhu. f. coho. yocuvuidzinindi. yocuvui nuundi.
Abadia dignidad. sasindadzi siñañuhu, sasiñoho siñañuhu, sasicuvuinuu.
Abadesa, prelada de monjas. ñaha yondacañaha yuq dzehe. ñaha yocuvuidzini.
Abadejo escarauajo. tenuyuq.
Abahar algo con el huelgo. yochidzo yoco yuhundi yotaa yoco yuhundi.

folio 1r column 2

Abahar asi cosa honda. yotniño yoco yuhundi.
Abahada cosa afi. sasi cohondaa yoco. sasinucoho yoco yuhu sanisa codzo yoco yuhu.
Abahar algo poniendolo al baho que sale de la olla. yodzaqnindi. yodzacuidzindi yoco.
Abahar algo dentro de alguna olla para que se conserue con el baho y no se pudra. yodzacaanuuyocondi.
Abahar sopas. yodzachitundi yoco yotniño yocondi. yodzacutu yocondi.
Abahadas sopas. dzita ninucoho yoco.
Abahar por hechar el huelgo de la boca. yodzacaindiyoco. yodza canandiyoco, yodzaqhundi yoco.
Abahar aposento. yodzacuidzindi huahi. yotniño yocondi. yondadzacuidzindi. yondadzainindi.
Abahado aposento. huahi ninucuiñe huidzi. huahi ninucoho yoco. huahi ninacuidzi. huahi ninduvui ini.
Abalançarse, hechandose por los suelos. yocoo cavuandayendi, yosacavuandayendi. f qcavua.
Abalançarse de arriba abaxo. yondiyoninondi.
Abalançarse metiendose entre otros. yosanudzavuandi. f. qnanu. yosananudzavuandi f. qnanu, yosivui cuinomañundi, yocai nanu dzavuandi. f. quivui. yosivuindutnahandi.
Aballar mouer con difficultad, duz

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

Comparisson of Resource Versions

(Mesolore 2001 Version)

Aceptar persona. Yodzacainuundi yositoninondita, f. coto, yotniño nuundita, yonaquai nuund

(Jansen & Pérez Jiménez 2009 Version)

yodza cay noondi: deshollejar; abajar la cabeza para mirar algo profundo; aceptar persona; anillo poner en el dedo; echar los ojos en algo; inclinarse bajando la cabeza para mirar hacia abajo; poner los ojos en algo para hurtarlo; poner los ojos en algo que parece bien

yosito ninondita, *futuro* coto: aceptar persona

yotniño nuundita: aceptar persona

yona quay nuundita: aceptar persona

OCR & Indigenous/Low-Resource Language Dictionaries

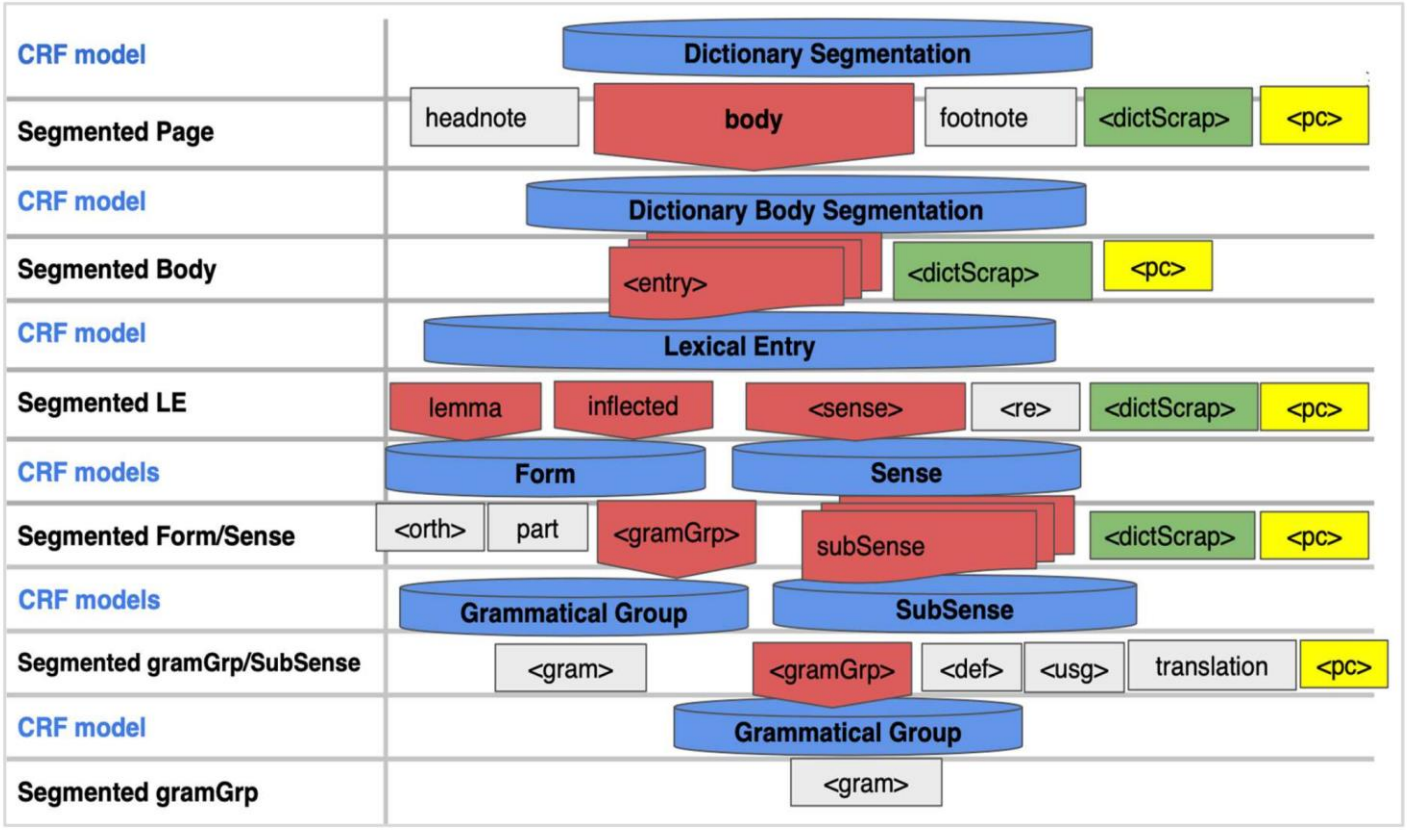
- Maxwell and Bills (2017) :
  - Tzeltal-English, Muinane-Spanish & Cubeo-Spanish
  - XML output

- Ranaivo-Malançon et al. (2017):
  - Melanau Mukah-Malay
  - PDF > HTML files > plain text
  - parsed using a Python

GROBID Dictionaries

- Cascading parsing of print dictionaries (see eLex paper: Khemakhem et al. 2017).
- Conditional Random Fields (CRF) (Lavergne et al., 2010) combined with dedicated libraries for manipulating PDF documents
  - end-to-end extraction of lexical structures into TEI compliant resources

GROBID Dictionaries



GROBID Sampling

Model	Training	Evaluation
Dictionary Body Segmentation	572 <entry>	270 <entry>
Lexical Entry	572 <sense> 572 <lemma> 28 <inflected> 10 <re>	269 <sense> 270 <lemma> 10 <inflected> 4 <re>
Sense	856 <subSense>	302 <subSense>
Form	787 <orth> 31 <part> 31 <gramGrp>	269 <orth> 11 <part> 11 <gramGrp>
SubSense	905 <def> 32 <usg> 7 <gramGrp> 9 <translation>	319 <def> 11 <usg> 8 <gramGrp> 2 <translation>

Evaluating GROBID Output

Model	Label	Precision	Recall	F1
Lexical Entry	<inflected>	90	90	90
	<lemma>	99.26	99.26	99.26
	<pc>	98.94	99.29	99.12
	<sense>	100	100	100
	<re>	0	0	0
Sense	<subSense>	100	100	100
	<pc>	100	100	100
Form	<gramGrp>	100	90.91	95.24
	<orth>	98.18	100	99.08
	<part>	70	63.64	66.67
SubSense	<def>	91.84	95.3	93.54
	<gramGrp>	100	25	40
	<pc>	76.81	88.33	82.17
	<translation>	100	100	100
	<usg>	60	90	72

Integration into Mixtepec-Mixtec Project: TEI Structure of Output

- Goal to match the structure used in the Mixtepec-Mixtec TEI dictionary (Bowers & Romary 2018)

```
<entry xml:id="fruit-plaintain">
  <form type="lemma">
    <orth xml:lang="mix">nchika</orth>
    <pron xml:lang="mix" notation="ipa">nʤiká</pron>
  </form>
  <gramGrp>
    <pos>noun</pos>
  </gramGrp>
  <sense corresp="http://dbpedia.org/resource/Plantain">
   <usg type="domain">Fruit</usg>
    <cit type="translation">
      <form>
        <orth xml:lang="en">plantain</orth>
      </form>
    </cit>
    <cit type="translation">
      <form>
        <orth xml:lang="es">plátano</orth>
      </form>
    </cit>
  </sense>
...
</entry>
```

nchika [nʒiká] *(noun)*
[FRUIT] **plantain, plátano**

Mixtepec-Mixtec

```
<entry xml:id="plátano">
  <form type="lemma">
    <orth>chita</orth>
  </form>
  <gramGrp>
    <pos>noun</pos>
  </gramGrp>
  <sense corresp="http://dbpedia.org/resource/Plantain">
    <usg type="domain">Fruit</usg>
    <def xml:lang="es">plátano</def>
    <def xml:lang="en">plantain</def>
  </sense>
</entry>
```

chita *(noun)*
[FRUIT] **plantain, plátano**

Classical Mixtec

Integration into Mixtepec-Mixtec Project: Etymological/Historical Resource

antivi

[EARTH] **sky, cielo**

(Etymology)

Attested in: Yucu Ndaa **andevui** (Alvarado 1593)

```
<entry xml:id="sky">
  <form type="lemma">
    <orth xml:lang="mix">antivi</orth>
    ….
  </form>
  <sense corresp="http://dbpedia.org/resource/Sky">
    <usg type="domain">Earth</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">
    <seg type="desc" xml:lang="en">Attested in:</seg>
    <xr type="crossReference">
      <lang>Yucu Ndaa</lang>
      <ref type="entry">andevui</ref>
      <ref type="bibl" source="alvarado:andevui">Alvarado</ref>
      <date>1593</date>
    </xr>
  </etym>
</entry>
```

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

Integration into Mixtepec-Mixtec Project: Comparative Mixtecan Applications

Mixtepec-Mixtec
xini
[ʃiní]

Chalcatongo Mixtec
šinì
(Macaulay, 1996)

Ayutla Mixtec
shīhih
(Hill, 1990)

San Martín Duraznos Mixtec
ʃiɲī
(Padget, 2017)

Guadalupe Nundaca Mixtec
ʃiɲī
(Padget, 2017)

Coatzospan Mixtec
rkɨ
(Small, 1990)

Yucu Ndaa
dzini
(Alvarado, 1593)

TEI Structure of Output: Senses

- In 2009 source, separate senses were separated by semicolons and separate glosses (same sense) separated by commas

ñuhu nisitu: cavada tierra**;** labrada tierra

```
<entry xml:id="ñuhu_nisitu">
  <form type="lemma">
    <orth>ñuhu nisitu</orth>
  </form>
  <pc>:</pc>
  <sense>
    <def xml:lang="es">cavada tierra</def>
  </sense>
  <pc>;</pc>
  <sense>
    <def xml:lang="es">labrada tierra</def>
  </sense>
</entry>
```

ñuhu tisaha: fofa cosa**,** como tierra

```
<entry xml:id="ñuhu_tisaha">
  <form type="lemma">
    <orth>ñuhu tisaha</orth>
  </form>
  <pc>:</pc>
  <sense>
    <def xml:lang="es">fofa cosa</def>
    <pc>,</pc>
    <def xml:lang="es">como tierra</def>
  </sense>
</entry>
```

TEI Structure of Output: Inflected forms

- Some entries have inflected forms, but they are usually only part of a MWE, these are indicated by italics following a comma on the left side of the colon delimiter

yosico ini tnahandi*, futuro* **cuico**: aficionados estar dos

```
<entry>
  <form type="lemma">
    <orth>yosico ini tnahandi</orth>
  </form>
  <pc>,</pc>
  <gramGrp>
    <pos>verb</pos>
  </gramGrp>
  <form type="inflected">
    <gramGrp>
      <gram>futuro</gram>
    </gramGrp>
    <orth extent="part">cuico</orth>
  </form>
  <pc>:</pc>
  <sense>
    <def xml:lang="es">aficionados estar dos</def>
  </sense>
</entry>
```

Added post conversion

Output: Collocate Phrases, Usage

caa ndodzo ninondi (nuu sito): echado estar (en la cama)

```
<entry xml:id="caa_ndodzo_ninondi">
   <form type="lemma">
      <orth>caa ndodzo ninondi</orth>
      <pc>(</pc><colloc>nuu sito</colloc><pc>)</pc>
   </form>
   <pc>:</pc>
   <sense>
      <def xml:lang="es">echado estar</def>
      <pc>(</pc><usg type="hint">en la cama</usg><pc>)</pc>
   </sense>
 </entry>
```

Output: Modern Spanish Translations

- There were a number of modernized Spanish translations added by Jansen and Pérez Jiménez (2009) which were placed in square brackets, responsibility added.

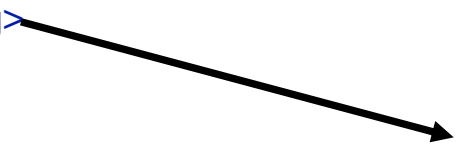
da queyeni: aprisa; incontinenti **[luego]**; y luego; luego a la hora; temprano

```
<sense>
   <def xml:lang="es">incontinenti</def>
   <pc>[</pc><cit type="translation" resp="#M.E.R.G.E.N.J #G.A.PJ">
      <form>
         <orth xml:lang="es">luego</orth>
      </form>
   </cit><pc>]</pc>
</sense>
```

Enhancing Output: Etymology

- Want to add linguistic analysis of etymology, in TEI do so according to Bowers & Romary (2016), Bowers & al. (2018)

```
<entry>
  <form type="lemma">
    <orth>yosa ndehe ichi</orth>
  </form>
  <pc>:</pc>
  <sense>
    <def xml:lang="es">fenecer</def><pc>,</pc>
    <def xml:lang="es">acabar el que muere</def>
    <pc>(</pc>
    <usg>por metáfora</usg>
    <pc>)</pc>
  </sense>
</entry>
```



```
<entry xml:id="yosa_ndehe_ichi">
  <form type="lemma">
    <orth>yosa ndehe ichi</orth>
  </form>
  <sense>
    <def xml:lang="es">fenecer</def><pc>,</pc>
    <def xml:lang="es">acabar el que muere</def>
  </sense>
  <etym type="metaphor">
    <seg type="desc">por metáfora</seg>
    <cit type="etymon" resp="#JB">
      <form>
        <orth>ichi</orth>
      </form>
      <def xml:lang="es">camino</def>
    </cit>
  </etym>
</entry>
```

Enhancing output:

GROBID process overall successful in converting contents of PDF into TEI. Howerver, further manual and semi-manual encoding enhancements were necessary These were necessary due to either:

- a lack of sufficient tokens required for the machine learning process
- make it more compatible with the Mixtepec-Mixtec TEI corpus

Potential Future Endeavours:

- Add English and Mixtepec-Mixtec translations
- Use as basis for a cross-Mixtec LR by adding translations from other Mixtec varieties as well
- Create TEI versions of the two related historical indigenous dictionaries published by the OP:
  - Molina 1571 (Nahuatl)
  - Cordova 1575 (Zapotec)
- Expand capacity of GROBID to be able to extract linguistic examples and content from linguistic papers

Conclusion

- GROBID can handle the vast majority of the work needed to create a highly structured TEI dictionary from PDF resources.
- Problem in achieving full standardization due to lack of language tag for the Mixtec variety as well as the early modern Spanish used in source
- Due to certain issues pertaining to the source document (structure and the sample size of certain structures) significant further manual and semi-manual work is required
- Enhancements by humans who understand certain details that are only accessible through detailed study are necessary.
- Sets further precedent for the use of TEI in digital lexicography in maximum integration of LR for indigenous/low-resource languages