TEI Encoding of a Classical Mixtec Dictionary Using GROBID-Dictionaries

Jack Bowers[1,2,3], Mohamed Khemakhem[1,4,5,6], Laurent Romary[1]

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

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

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

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

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

[6] BBAW – Berlin-Brandenburg Academy of Sciences and Humanities, Berlin, Germany

E-mail: iljackb@gmail.com, mohamed.khemakhem@inria.fr, laurent.romary@inria.fr

Abstract

This paper presents the application of GROBID-Dictionaries (Khemakhem et al., 2017; Khemakhem et al., 2018a; Khemakhem et al., 2018b; Khemakhem et al., 2018c), an open source machine learning system for automatically structuring print dictionaries in digital format into TEI (Text Encoding Initiative) to a historical lexical resource of Colonial Mixtec 'Voces del Dzaha Dzahui' published by the Dominican Fray Francisco Alvarado in the year 1593. The GROBID-Dictionaries application was applied to a re-organized and modernized version of the historical resource published by Jansen and Perez Jiménez (2009). The TEI dictionary thus produced will be integrated into a language documentation project dealing with Mixtepec-Mixtec (ISO 639-3: mix) (Bowers & Romary, 2017, 2018a, 2018b), an under-resourced indigenous language native to the Juxtlahuaca district of Oaxaca Mexico.

Keywords: Mixtec; TEI; GROBID-Dictionaries

1. Introduction to the resource

This paper presents the creation of a TEI dictionary of the earliest lexical resource[1] of a Mixtec language: the Vocabulario published by the Dominican Fray Francisco Alvarado in the year 1593[2]. This resource was automatically converted from PDF format to a structured TEI dictionary using the application GROBID-Dictionaries (Khemakhem et al., 2017; Khemakhem et al., 2018a; Khemakhem et al., 2018b; Khemakhem et al. 2018c), an open source machine learning system for automatically

[1] Not including the codices which were pictographic and not specific to any local variety of Mixtec. Though the author did not represent all the features of the language such as tone, nasalization among other features is resource is thus the first with any representation of the phonetic characteristics of a Mixtecan Language (Jansen & Perez Jiménez, 2009).

[2] This document was likely compiled from the few existing resources at the time, namely the Doctrina en Lengua Mixteca by Fray Benito Hernández published in 1567 and 1568 respectively from sources compiled in Teposcolula Mexico (Jansen & Perez Jiménez, 2009).

structuring print dictionaries in digital format into TEI (Text Encoding Initiative). The PDF source used in the transformation is from a re-organized and modernized version of the historical dictionary published by Jansen and Perez Jiménez (2009). The TEI dictionary produced contains roughly 26,600 entries and related entries.

The Mixtec variety sampled by Alvarado to create this vocabulary was that of Yucu Ndaa (Teposcolula) *dzaha dzavui*[3], which according to the sources is thought to have been used as a *lingua franca* of the Mixteca region at the time and the language is presently in the field of Mixtecan commonly referred to as "Classical Mixtec" or "Colonial Mixtec" (Jansen & Perez Jiménez, 2009).

The vocabulary was produced by the Orden de los Predicadores (O.P.) aka. the Dominican Order, who wanted to learn the language as part of the evangelization efforts in order to be able to communicate with Mixtecs in their own language for the purposes of conversion. In this same year a grammar was published by Fray Antonio de los Reyes (also of the Teposcolula - Yucu Ndaa variety)[4].

There are several inter-related potential uses of the output of this endeavour[5] for philological, linguistic, anthropological purposes including: 1) the creation of a machine searchable data set for the study of the Yucu Ndaa variety itself, and/or the historiographical and philological issues related to the collection and specifics of the vocabulary collected; 2) creating an open, highly structured resource for other Mixtecan lexical projects; 3) combining the first two to potentially create a more cohesive body of pan-Mixtecan resources and a set of vocabulary for cross Mixtecan comparison; 4) the TEI format can easily be exported into other formats (e.g. tab separated plain text, etc.) for non-TEI users, i.e. the format is fully extensible.

And in line with the above, this endeavour was undertaken in order to integrate the contents of this historical resource into a TEI-based language documentation project dealing with Mixtepec-Mixtec (ISO 639-3: mix) (Bowers & Romary, 2017, 2018a, 2018b), an under-resourced indigenous language native to the Juxtlahuaca district of

[3] In the present day, there are dozens of Mixtec varieties with different levels of mutual intelligibility, estimates range from 52 (Simons & Fennig, 2018) to 85 distinct varieties (Instituto Nacional de Lenguas Indígenas, 2015).

[4] Both the source of the document (Jansen & Perez Jiménez, 2009) and Mesolore provide excellent overviews of issues relevant to the study and understanding of the contents of the vocabulary and thus those seeking a more extensive description thereof, should consult these studies.

[5] Note these benefits discussed are on top of the essential work done by Jansen and Perez Jiménez (2009) who made the resource much more user-friendly in implementing a number of normalizations, altering the entries to Mixtec -> Spanish, provided an indepth discussion of the source and its context, and provided a vision of the resource as a potential basis for pan-Mixtecan etymological and philological comparison. We share this vision and assert that the application of TEI enables the use of the resource as a machine and human readable database.

Oaxaca Mexico[6]. Mixtepec-Mixtec (spoken in the Juxtlahuaca district of Oaxaca) like Teposcolula is in the "Mixteco Alto" region, and the linguistic relation between modern Mixtepec-Mixtec and the historical variety Yucu Ndaa is quite clear in a significant portion of the vocabulary.

2. OCR technology and indigenous language dictionaries

In recent years there have been growing efforts to apply OCR to digitize indigenous language resources, which is increasingly necessary as language communities are seeking to make the limited materials they have more widely available and to avoid situations where paper copies of content are not only inaccessible but at risk of complete loss if physical copies fall victim to any number of potential man-made or natural disasters.

Maxwell and Bills (2017) discuss the application of OCR methods in creating a structured, machine readable XML lexicon for indigenous language resources, including Tzeltal-English, Muinane-Spanish and Cubeo-Spanish dictionaries. Additionally, Ranaivo-Malançon et al. (2017) discuss the conversion of Melanau-Mukah-Malay and Iban-Malay indigenous language dictionaries from PDF sources into HTML files, which were then parsed using a Python HTMLParser to extract the dictionary content to be saved as comma-separated plain text files.

More advanced approaches using machine learning techniques have been seen since in recent years. The most successful one that showed enough potential for scalability and adaptation is the cascading parsing of print dictionaries implemented in GROBID-Dictionaries (Khemakhem et al., 2017; Khemakhem et al., 2018a; Khemakhem et al., 2018b; Khemakhem et al., 2018c). The technique is based on Conditional Random Fields (CRF) (Lavergne et al., 2010) which allow, along with dedicated libraries for manipulating PDF documents, the end-to-end extraction of lexical structures into TEI compliant resources. The extensibility of GROBID-Dictionaries, along with being language agnostic, have motivated our present work to speed up the process of building a structured resource for the historical Mixtec language resource.

2.1 Different versions of the resource

In both the automatic structuring process used to create the TEI dictionary and in the specifics of the content, the history of the organization of the lexical resource plays a significant role. While the original dictionary created by Alvarado was Castilian - Mixtec, the version by Jansen and Perez Jiménez (2009) was transformed to be Mixtec - Castilian. Below is an example of the original Castilian - Mixtec entry structure taken from the PDF version with the original structure created by Mesolore. Not only is this

[6] Mixtepec-Mixtec is an Otomonguean language spoken by roughly 9,000 – 10,000 people, and in addition to the native communities in Mexico, it is also spoken by communities of several people living in California, Oregon, Washington, Florida and Arkansas in the United States.

lexicon Castilian based, but it is organized in such a way that an entry often contains multiple Mixtec forms, has unclear indicators of grammatical information, the components of the Mixtec items are not appropriately delimited, in some cases they were not consistently spelled, and finally in many cases the Mixtec forms had other senses that were placed in separate entries. The original content was thus not a user-friendly resource.

> Aceptar persona. Yodzacainuundi yositoninondita, f. coto, yotniño nuundita, yonaquai nuundita, yonaquaicahandisita.

Figure 1: Dictionary structure prior to the restructuring of Jansen and Perez Jiménez (2009).

Jansen and Perez Jiménez (2009) split up the contents of this into five separate entries and applied several normalizations to the orthographic representation to produce a more uniform convention. These changes both improve the organization of the Mixtec content and more clearly reflect the linguistic structure. The results of which are shown below[7]:

> **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
> **yona quay cahandi sita**: aceptar persona

Figure 2: Revised version of the entry shown above, separated into four separate entries in Jansen and Perez Jiménez (2009)

The changes made in the aforementioned source, particularly the use of bold type for the Mixtec forms, the addition of a colon ":", semi-colon ";", and comma "," delimiters between form and sense, different senses, and separate glosses in a single sense all rendered the contents much more amenable to the application of GROBID, as the contents of entries are much more clearly demarcated. Issues specific to GROBID will be discussed in more depth in the following section.

[7] Note in Figure 2, the Spanish gloss of the entry **yodza cay noondi** contains content that was not in the original shown in Figure 1; this was apparently taken from elsewhere in the in the original dictionary as *yodza cay noondi* has been given as the gloss for multiple Spanish terms. One of the major achievements of Jansen and Perez Jiménez (2009) was the consolidation of this information.

3. GROBID-Dictionaries

3.1 System overview

GROBID-Dictionaries (Khemakhem et al., 2017; Khemakhem et al., 2018a) is a machine learning infrastructure for parsing and structuring digital dictionaries based on CRF models (Lavergne et al., 2010). The infrastructure has been tested with digitized and born digital dictionaries in several languages, and is still under development. In the following section, we present the part of the tool's up-to-date architecture reflecting the logical (lexicographic) structure of the present dictionary.

3.2 Cascading CRF models for lexical information parsing

The lexical information extraction in GROBID-Dictionaries relies on a cascade parsing of the structures in an input dictionary. At each parsing level a CRF model, being trained on samples of the target dictionary, has the goal of predicting a set of labels representing TEI structures. In Figure 3 we present the architecture of different models and labels recognized by the system in the case of the present dictionary.

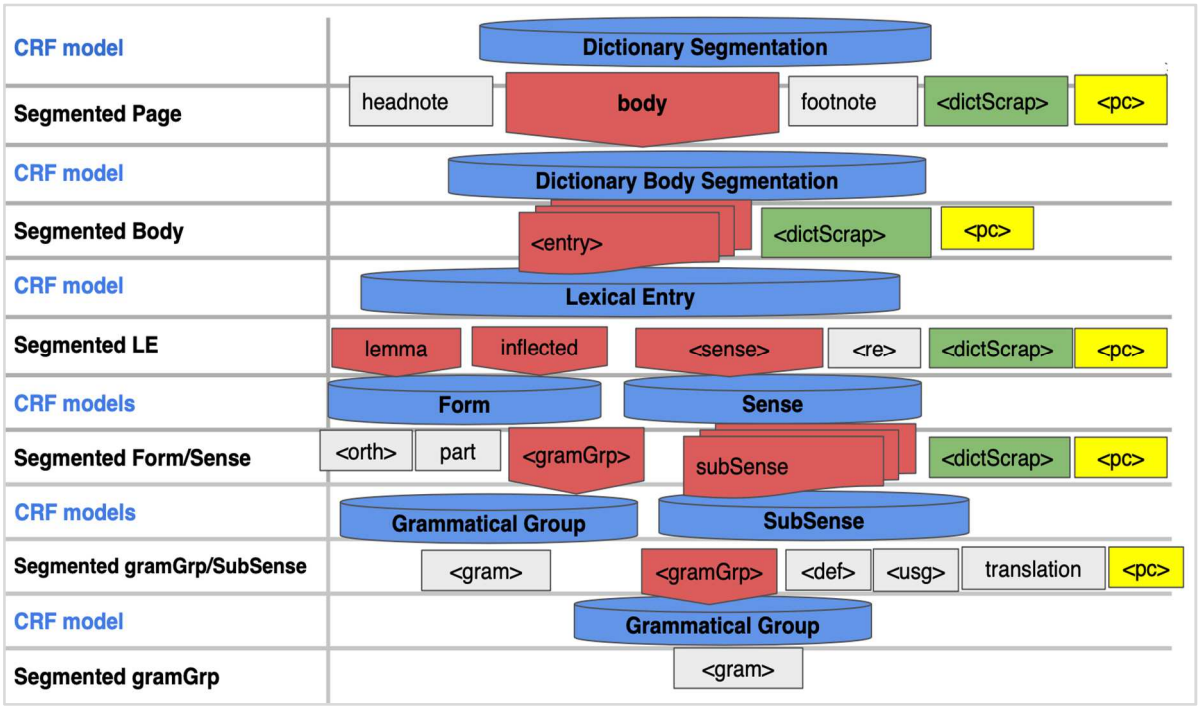


Figure 3: Parts of GROBID-Dictionaries' architecture activated for parsing the Mixtec dictionary

As a reminder, a text cluster recognized by a CRF model could be either directly wrapped into a valid TEI structure – represented in Figure 3 with angle brackets – or into a pivot XML element – represented in Figure 3 without brackets. Pivot elements are implemented when a TEI construct is typed, such as the <form> element that could be typed with either "lemma" or "inflected", or as for definitions which are serialized in TEI using <def> construct. We have used pivot elements just for the training stage, which are then rendered in the final output as a valid TEI construct. <pc> is present at almost all segmentation levels, as marking up such information is useful for the machine learning model to learn field limits in a continuous sequence of tokens. A simple find/replace post-processing can remove such valid TEI tags if needed.

Compared to what has been already achieved in Khemakhem et al. (2018a), several improvements have been carried out to cover more lexical features encountered both in this dictionary as well as in other samples of similar lexical description depth:

1. Forms in lexical entries are differentiated into lemma and inflected.

2. The form model parses morphological and grammatical information of different forms, replacing the old model which was designed to extract the main information related to the lemma .

3. After being extracted and segmented into sub-senses, if semantic nesting needs to be reflected then senses can be parsed by a SubSense model to recognize definitions, usage, grammatical information and translation equivalents.

3.3 Experiment

In training the different models required in the architecture we have encountered several challenges related to the logical and physical (typographic) structure of the dictionary. We detail in this section the major obstacles, the implemented solutions, the impact on the annotation process and the results of the experiment.

3.4 Automatic parsing: features and challenges

The logical structure of the dictionary has been affected by the fact that dictionary had been re-compiled from an earlier version, which, as mentioned above, greatly improved the quality and organization of the resource in many ways. While the dictionary looks fairly simple in structure, due to a mixture of issues related to the original vocabulary collection in combination with some conventions in the updated formatting which are not clearly specified by Jansen and Jiménez Perez (2009), it contains some complexities which pose some serious obstacles to parsing, and most of these features are described below.

3.4.1 Forms and related entries

While thanks to the revisions by Jansen and Jiménez Perez (2009) the form section is nicely delimited from the sense by the use of the bold type, there are nonetheless quite a few different features present in that section with unique conventions for demarcation.

The most common supplemental feature in the forms is the inclusion of an inflected form (which is only a single part of the verb phrase), which is delimited by the combination of a comma, followed by the grammatical feature in italic type.

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

Figure 4: Entry with inflected (future) form *cuico*

There are several different conventions used in related entries and variant forms, none of which have enough instances sufficient for automatic recognition and structuring:

huau ndaha / saha: artejo

Figure 5: Entry for "knuckle" specific to "hand" ndaha or "foot"

In the entry *huau ndaha/ saha* the form *ndaha* is "hand" while *saha* is "foot", thus the content is a related entry and is only part of the full form of the second lexical entry (as the lexical items for knuckles in Classical Mixtec are equivalent to "hand knuckles" and "foot knuckles"). In the entry below, it appears that there are alternate terms which translate into Spanish as *en buen tiempo* ("in good time") and these alternative phrases are separated by the first comma with the grammatical feature (either verbal tense or mood) preceding the inflected form.

quevui iñe huaha, quevui iñe huii, *futuro* cuiñe: en buen tiempo

Figure 6: Entry with variant term for part of phrase huaha ~ huii.

In only a few instances, where the entry itself is an inflected form of another entry this information is stated in square brackets within the form (bold) portion of the entry. However, this content is mixed between the feature (below *imperativo*), Spanish translation, and then the Mixtec form (below *yosa cahindi*). These instances are actually duplicates of existing entries.

qua cahi [*imperativo* de yosa cahindi]: ir por algo generalmente

Figure 7: Entry whose form is an inflected form of a separate (related) entry.

3.4.2 Sense information

In entries with multiple gloss-like definitions but which are to be considered a single sense, commas separate the contents[8]:

quevui yahui: feria, mercado

Figure 8: Entry with two glosses of a single sense.

In some cases, the entry is divided into multiple senses (which themselves have one or more gloss), these separate senses are delimited by semicolons.

ñuhu nisitu: cavada tierra; labrada tierra

Figure 9: Entry with two distinct senses.

There are cases of exceptions to these, for instance, while a comma usually delimits different glosses, in a few examples one is used in a normal grammatical way, delimiting clauses. In the following example, the definition is *fofa cosa* "soft thing" and the content after the comma states "such as dirt".

ñuhu tisaha: fofa cosa, como tierra

Figure 10: Entry showing a comma delimiting separate glosses of same sense.

3.4.3 Usage, etymology grammatical information

In many entries there is supplemental information about the sense given by the original author which generally specifies some aspect of the usage. This is represented in the Jansen and Jiménez Perez (2009) version in round brackets.

ama: bien está (otorgando); sí

Figure 11: Example of usage information in sense.

Likewise, there are some entries with supplemental information which is grammatical in nature, and this is also placed in round brackets but is distinguished from the usage information with italics.

amana: ¿cuándo? (*adverbio interrogativo*), ¿en qué tiempo?

Figure 12: Example of grammatical information in sense

[8] Despite the structuring, there are many cases in which the use of the sense delimiter ";" does not seem to delimit strictly distinct senses.

However, there are certain cases in which there is grammatical information as well as a translation in the round brackets. Though the structure is distinct, in that within the brackets the grammar information is in italics delimited by a colon and the Spanish translation is to the right of the colon, there are not enough instances to train the system to automatically recognize this.

ca nayndo saha qhundo: llevar alguna cosa (*imperativo*: llevarás esto)

Figure 13: Example of grammatical information and translation of inflected form in sense

In some entries Jansen and Jiménez Perez (2009) added notes of where the sense is metaphorical in nature, these also are represented in round brackets within the sense section. The number of these instances is also not sufficient for the system to recognize and structure this content.

ña tuvui nini dzavua yuqua iyondi: vivir pobre (por metáfora); pobre estar

Figure 14: Example of metaphor specified in sense information

3.4.4 Other issues

Hyphenated content which is present in the source due to line breaks and additional varied use of brackets for various lexical content are also present in the data source, and contain too few instances to provide enough training data annotations for ML to create the desired output. Such content (as well as that mentioned above which lacks sufficient quantities for successful training) are structured in the TEI output either manually, semi-manually, or automatically using XSLT, much of which will be described in a later section.

3.5 Annotation

Covering instances of all the aforementioned observations in a few pages is a very hard task for the annotation. And given the multi-stage annotations, where the annotation is focused at each level on marking up all possible variations of specific structures, the number of pages required to be annotated can grow exponentially.

3.6 Page sampling process

As a random sampling was not an option in the case of this dictionary, we tried to cover the variation of logical and physical structures by selecting pages that represent the maximum number of challenges. We selected just a few pages containing related entries as they are sparsely distributed, and we also had to give up the annotation of

some structures, such as "morphological variants", given their low number and inconsistent typographic representation. More useful information about language, comparison, transcription, etc. can be found in the prose section, which we decided to ignore in the scope of this experiment.

We have selected and annotating 14 pages from different parts of the dictionary: 10 for training and four for evaluation. We detail the annotated instances for each model, except for the first one dealing with the prediction of the main regions of a page, which has less lexical importance with regard to the scope of this work, in Table 1.

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>

Table 1: Page Sampling Statistics.

3.7 Results and discussion

The results of the first two models, **Dictionary Segmentation** and **Dictionary Body Segmentation** were almost perfect, with an above 98 F1 score. In the following table we detail the performance of the rest of the models on the field level, in which the evaluation takes into consideration the prediction of all the tokens of field and not only single tokens. We do not show the evaluation of the **Grammatical Group** model as it has only one label.

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

Table 2: Field Level Evaluation of the Lexical Models.

The evaluation shows the high performance of the models in predicting lexical structures with the exception of related entries, grammatical information, sense usages within sense and orthography of inflected forms.

In the case of related entries, the training and evaluation datasets combined contain just 32 instances representing two logical representations (collocates and non-collocates) and four physical variations (with/without brackets and with/without commas). We consider the quantity of instances used for training is not sufficient for the **Lexical Entry** model to learn the distribution of such a structure.

For usage and grammatical information blocks, both structures are represented within senses as textual sequences wrapped in a round brackets. The only evident physical difference is the italics used to mark the grammatical information. An in-depth

investigation has shown that such a visual variation has not been translated consistently in the layout information associated with each token of the document and extracted by the PDF utilities libraries in GROBID-Dictionaries. Therefore, the **SubSense** model remains unable to differentiate these two physically similar structures. In the case of the <part> label, more annotated instances seem to be needed in the training dataset to strengthen the predictions of the **Form** model.

4. TEI structure of output

Because this resource is being converted to TEI in order to be integrated with the TEI-based project on the contemporary Mixtepec-Mixtec variety[9], the Classical Mixtec dictionary structure is designed to match the former as much as possible. The exception to this is that due to the inexact nature of the Spanish glossing the default element containing the Spanish is the definition element <def>, whereas is the Mixtepec-Mixtec TEI dictionary they are represented as pure translations.

4.1 Basic entry structure

```
<entry xml:id="fruit-plaintain">
  <form type="lemma">
    <orth xml:lang="mix">nchika</orth>
    <pron xml:lang="mix" notation="ipa">nʤiká</pron>
  </form>
 ....
  <sense corresp="http://dbpedia.org/resource/Plantain">
 ....
    <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>
```

```
<entry>
  <form type="lemma">
    <orth>chita</orth>
  </form>
  <sense>
    <def>plátano</def>
  </sense>
</entry>
```

Figure 15: Left, partial TEI dictionary entry for *nchika* 'plantain' in Mixtepec-Mixtec; right, view of (unenhanced) structure of the historically related form in Classical Mixtec.

Note that while the ISO 639-3 language code is applied to the Mixtepec-Mixtec entry and the Spanish 639-2 tag is applied to the translations of the Classical Mixtec entry, there are no ISO or other any other standardized language codes for 'Classical Mixtec', nor is there any documented modern Mixtec variety attributed to Teposcolula.

[9] For an in depth detailing of the structure and content in the Mixtepec-Mixtec TEI dictionary, see Bowers and Romary (2018a)

4.2 Inflected forms

The entries with inflected forms are shown below in the TEI output. Note that since these are mostly multi-word expressions verb phrases.

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

Figure 16: Entry with inflected form

```
<form type="lemma">
  <orth>yosico ini tnahandi</orth>
</form>
<pc>,</pc>
<form type="inflected">
  <gramGrp>
    <gram>futuro</gram>
  </gramGrp>
  <orth extent="part">cuico</orth>
</form>
```

Figure 17: TEI encoding of entry with inflected form

4.3 Senses

Entries with multiple senses and multiple glosses/definitions were generally handled well by GROBID, examples of the output of each case (unenhanced) are shown below with the source from the PDF entry above.

ñuhu nisitu: cavada tierra; labrada tierra

```
<entry>
  <form type="lemma">
    <orth>ñuhu nisitu</orth>
  </form>
  <pc>:</pc>

  <sense>
    <def>cavada tierra</def>
  </sense>
  <sense>
    <def>labrada tierra</def>
  </sense>
</entry>
```

ñuhu tisaha: fofa cosa, como tierra

```
<entry>
  <form type="lemma">
    <orth>ñuhu tisaha</orth>
  </form>
  <pc>:</pc>

  <sense>
    <def>fofa cosa</def>
    <def>como tierra</def>
  </sense>
</entry>
```

Figure 18: TEI encoding of entries with multiple senses and multiple definitions.

5. Post-editing: Modifications and enhancements

In terms of the source to target structure, the GROBID process was able to create a conversion of the PDF form of the resource into TEI which represented the majority of the features present in the dictionary. However, in the case of several features, further manual and semi-manual encoding enhancements were necessary in order to create a more dynamic and refined structure. These modifications were necessary due to either: a lack of sufficient tokens required for the machine learning process or to make it more compatible with the Mixtepec-Mixtec TEI corpus. These changes are described in this section.

Other key enhancements made to the output include the following:

- Spanish ISO 639-2 language tag added to all <def> elements
- Unique id's (@xml:id) are added to each <entry> and <re> which are based on the Spanish value (with underscores and token numbers added as needed)
- English translations are being added according to certain categories (at least for those which are sufficiently clear, as not all items can easily be translated)
- Domain tag (<usg type="domain">) is added in certain entries (some of the vocabulary in the source which are initially given <usg type="hint"> can be changed to domain)
- Records of normalizations and assumed phonetic equivalencies made by Jansen and Jiménez Perez (2009) are manually added in the header.

Below we discuss the formatting of content which was not sufficiently structured by GROBID and/or which needed additional structuring to bring it in line with best practice in TEI. In the examples we show both the output of GROBID and the revised TEI structure.

5.1 Related entries

In most cases related entries were correctly identified as such by GROBID, however because there are a number of different types of related entries most of which lack sufficient instances to train the system automatically recognize and encode then in detail, these items are manually refined in TEI.

tay huasi cana / cay idzi yuhu: mozo que comien-za a barbar

Figure 19: Entry with related entry in source

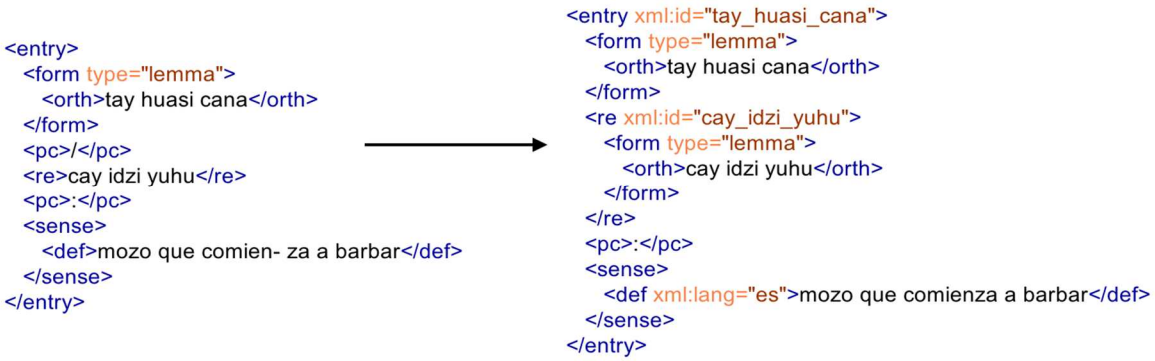


Figure 20: GROBID output (left) with revised TEI structure of form with related entry (right)

5.2 Collocate phrases in the form

In a small number of cases, there is collocate information included in the form. In TEI this is encoded using the <colloc> element.

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

Figure 21: Collocate of headword in source

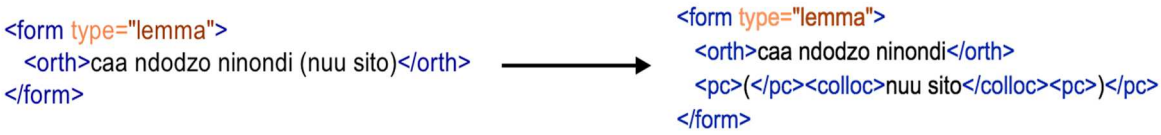


Figure 22: GROBID output (left) with revised TEI structure of form with collocate (right)

5.3 Modern Spanish Translations

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

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

Figure 23: Entry with modernized Spanish translation in source

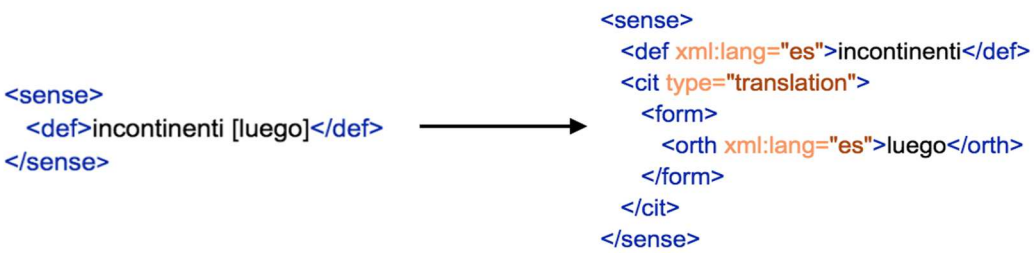


Figure 24: GROBID output (left) with revised TEI structure of modernized Spanish translation (right)

5.4 Inflected forms

In certain cases, even though the source did not have a given feature explicitly labelled, what they did include could be used to infer this, and then key features added in order to enhance the content and bring it in line with general lexicographic practice. One area where this was possible is where there were inflected forms.

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

Figure 25: Entry with inflected form in source

In these entries with the feature ***futuro***, there are two inferable features: first, that the entry is a phrase, (which mostly do not seem to have had simple lexicalized items in Mixtec), second, that given that the feature ***futuro*** is a feature of tense, that the part of speech verb can be inferred. The example below shows how these features are represented in the revised TEI structure by adding the @type="phrase" to <entry>, adding <pos>verb</pos> to the entry level, and by changing the generic <gram> to <tns> in the inflected future form[10]. This enhancement process is done using XSLT.

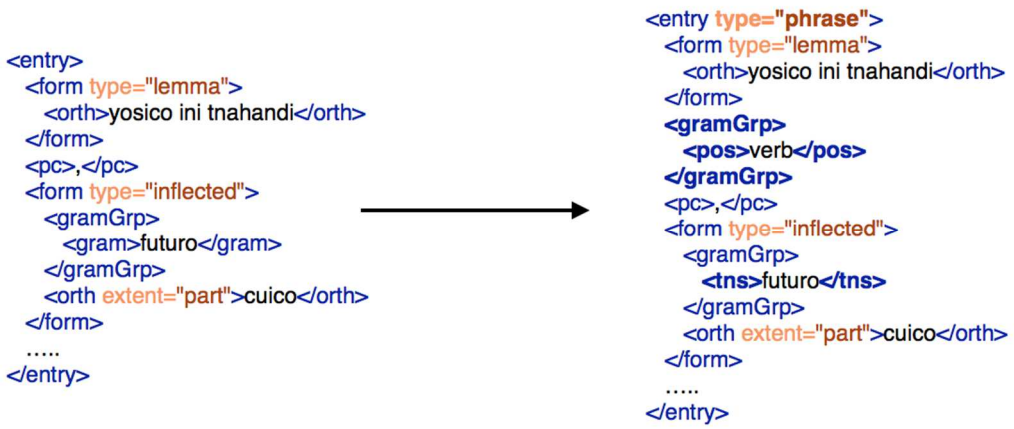
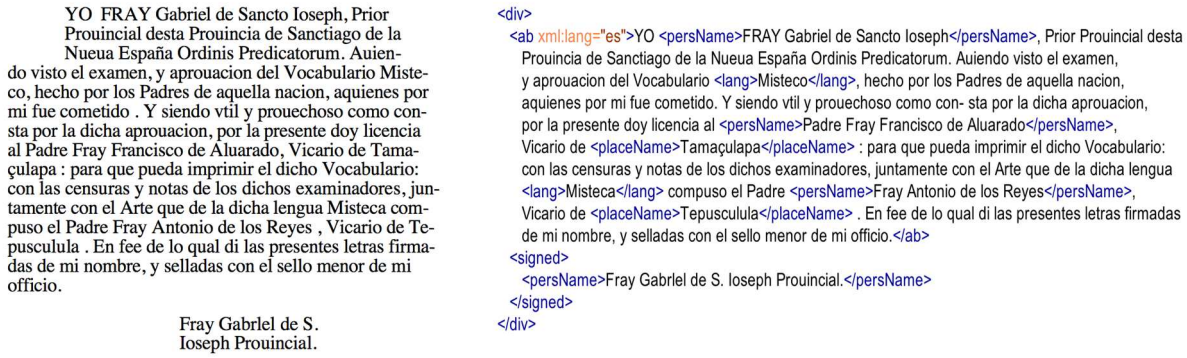


Figure 26: GROBID output (left) with revised TEI structure of phrasal entry with inflected form (right)

[10] The reason that we did not train GROBID to automatically annotate the feature *futuro* as <tns> is that there are some instances of imperative forms listed in the same way. Thus, each of these features is further treated using XSLT specifically targeted, with the imperative forms being output in TEI as: <gram type="mood">.

5.5 Addition of original prologue

As there were different PDF versions created of this resource, some of them included content from the original that were not included in the others. Notably, in the version of Jansen and Perez Jiménez (2009) the original prologue content published in the original was not included; this content was thus added manually and is easily represented in the TEI output.



YO FRAY Gabriel de Sancto Ioseph, Prior Prouincial desta Prouincia de Sanctiago de la Nueua España Ordinis Predicatorum. Auiendo visto el examen, y aprouacion del Vocabulario Misteco, hecho por los Padres de aquella nacion, aquienes por mi fue cometido . Y siendo vtil y prouechoso como consta por la dicha aprouacion, por la presente doy licencia al Padre Fray Francisco de Aluarado, Vicario de Tamaçulapa : para que pueda imprimir el dicho Vocabulario: con las censuras y notas de los dichos examinadores, juntamente con el Arte que de la dicha lengua Misteca compuso el Padre Fray Antonio de los Reyes , Vicario de Tepusculula . En fee de lo qual di las presentes letras firmadas de mi nombre, y selladas con el sello menor de mi officio.

Fray Gabrlel de S.
Ioseph Prouincial.

```
<div>
  <ab xml:lang="es">YO <persName>FRAY Gabriel de Sancto Ioseph</persName>, Prior Prouincial desta
    Prouincia de Sanctiago de la Nueua España Ordinis Predicatorum. Auiendo visto el examen,
    y aprouacion del Vocabulario <lang>Misteco</lang>, hecho por los Padres de aquella nacion,
    aquienes por mi fue cometido. Y siendo vtil y prouechoso como con- sta por la dicha aprouacion,
    por la presente doy licencia al <persName>Padre Fray Francisco de Aluarado</persName>,
    Vicario de <placeName>Tamaçulapa</placeName> : para que pueda imprimir el dicho Vocabulario:
    con las censuras y notas de los dichos examinadores, juntamente con el Arte que de la dicha lengua
    <lang>Misteca</lang> compuso el Padre <persName>Fray Antonio de los Reyes</persName>,
    Vicario de <placeName>Tepusculula</placeName> . En fee de lo qual di las presentes letras firmadas
    de mi nombre, y selladas con el sello menor de mi officio.</ab>
  <signed>
    <persName>Fray Gabrlel de S. Ioseph Prouincial.</persName>
  </signed>
</div>
```

Figure 27: Left a PDF version of part of the original prologue and right its TEI encoding.

5.6 Etymology

In the Jansen and Perez Jiménez (2009) source there are roughly 70 instances which are labelled as being metaphorical in nature. These are labelled as follows:

yosa ndehe ichi: fenecer, acabar el que muere (por metáfora)

Figure 28: Entry for metaphorical term ‘yosa ndehe ichi’ as formatted in the source PDF.

Due to a lack of sufficient quantity for training, these items had to be manually identified and annotated as follows:

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

Figure 29: TEI (partially enhanced) encoded entry for metaphorical term ‘yosa ndehe ichi’

While at present we do not have enough of the Classical Mixtec language to provide full analyses of the majority of the instances of metaphor, this information is nonetheless encoded in the TEI structure as per the recommendations of Bowers and Romary (2016), and Bowers et al. (2018). A partial structured analysis is provided for the phrase "yosa ndehe ichi", of which only the portion *ichi* 'path' is discernible and which is represented as an etymon within the <etym> block in TEI. At a later stage, researchers who are more familiar with the language can enhance this content as needed.

6. Later steps

A logical and needed future aim would be to create a searchable TEI version of the grammar of the language published in 1593 *Arte en Lengva Mixteca* by Fray Antonio de los Reyes. Given that this resource is a grammar and not a dictionary-like text, this would not be a job for GROBID but another, general OCR tool. The text in the PDF available is of low quality, and it is likely that significant manual work would be necessary to carry out this task.

Furthermore, according to Mesolore, much of the Alvarado Classical Mixtec vocabulary was based on entries in the 'Molina Vocabulario' Castilian-Nahuatl dictionary (1571), a Castilian-Zapotec dictionary compiled by Juan de Cordova in the Valley of Oaxaca (1578), and Antoni de Nebrija's Castilian-Latin Dictionarium (1553). Thus, many of these resources have common content and it would be a natural and beneficial next step to create TEI versions of these to expand all of the benefits described in this work with regard to the current Classical Mixtec vocabulary to these other indigenous languages.

7. Conclusion

This project has shown that GROBID can handle the vast majority of the work needed to create a highly structured TEI dictionary from PDF resources. However due to certain issues pertaining to the source document used, its structure and the sample size of certain structures, significant further manual and semi-manual work is required in creating a maximally representative version of the content. Given the richness of the resource, in order to effectively achieve these enhancements it is essential that they are carried out by humans who understand certain details that are only accessible through detailed study.

Converting this resource into TEI brings the data into a highly structured extensible machine-readable format which can be systematically searched, extracted and exported into other data formats using simple XQuery and/or XSLT.

In creating this iteration of the historical resource, we have continued the work of previous scholars (specifically Jansen and Perez Jiménez, 2009) who worked to make this resource available to researchers and Mixtec communities. As this work was carried

out in order to integrate the important resource into an ongoing linguistic and lexicographic project dealing with the Mixtepec-Mixtec variety, we hope to demonstrate how the Alvarado resource can be used as both an etymological and comparative cross-reference between different varieties of Mixtec as well as how TEI is a highly beneficial data format.

8. References

Alvarado, F. de. (1593). *Vocabulario en Lengua Mixteca. Hecho por los Padres de la Orden de Predicadores.* En México: Con Licencia, en casa de Pedro Balli.

Bowers, J. & Romary, L. (2016). Deep Encoding of Etymological Information in TEI. *Journal of the Text Encoding Initiative*, (Issue 10). https://doi.org/10.4000/jtei.1643

Bowers, J. & Romary, L. (2017). Language Documentation and Standards in Digital Humanities: TEI and the Documentation of Mixtepec-Mixtec. In A. Kawase (ed.) *Proceedings of the 7th Conference of Japanese Association for Digital Humanities.* Kyoto, Japan: Doshisha University, pp. 21–23.

Bowers, J. & Romary, L. (2018a). Bridging the gaps between digital humanities, lexicography and linguistics: a TEI dictionary for the documentation of Mixtepec-Mixtec. *Dictionaries: Journal of the Dictionary Society of North America, 39(2).*

Bowers, J. & Romary, L. (2018b). Encoding Mixtepec-Mixtec Etymology in TEI. *Presented at the TEI Conference and Members Meeting*, Tokyo, Japan.

Jansen, M. E. R. G. N., & Pérez Jiménez, G. A. (2009). *Voces del Dzaha Dzavui (mixteco clásico). Análisis y Conversión del Vocabulario de fray Francisco de Alvarado (1593). Colegio Superior Para La Educatión Integral Intercultural de Oaxaca.*

Khemakhem, M., Foppiano, L. & Romary, L. (2017). Automatic extraction of TEI structures in digitized lexical resources using conditional random fields. In I. Kosem et al. (eds.) *Proceedings of eLex 2017 conference, Netherlands, Leiden.* Brno: Lexical Computing Ltd., pp. 598-613.

Khemakhem, M., Herold, A. & Romary, L. (2018a). Enhancing Usability for Automatically Structuring Digitised Dictionaries. *GLOBALEX Workshop at LREC 2018.* Presented at Miyazaki, Japan. Retrieved from ⟨hal-01708137v2⟩

Khemakhem, M., Romary, L., Gabay, S., Bohbot, H., Frontini, F. & Luxardo, G. (2018b). Automatically Encoding Encyclopedic-like Resources in TEI. In *The annual TEI Conference and Members Meeting.*

Khemakhem, M., Romary, L., Gabay, S., Bohbot, H., Frontini, F. & Luxardo, G. (2018c). Automatically Encoding Encyclopedic-like Resources in TEI. In *The annual TEI Conference and Members Meeting.*

Lavergne, T., Cappé, O. & Yvon, F. (2010). Practical very large scale CRFs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, pp. 504–513.

Maxwell, M. & Bills, A. (2017). Endangered data for endangered languages: Digitizing

print dictionaries. *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pp. 85–91.

Mesolore. Accessed at: http://www.mesolore.org/tutorials/learn/9/Introduction-to-the-Alvarado-Vocabulario (24 April 2019)

Proyecto de indicadores sociolingüísticos de las lenguas indígenas nacionales. (2015). Accessed at: http://site.inali.gob.mx/Micrositios/estadistica_basica/estadisticas2015/estadisticas2015.html (5 July 2017)

Ranaivo-Malançon, B., Saee, S., Othman, R. M. & Busu, J. F. W. (2017). Transforming Semi-Structured Indigenous Dictionary into Machine-Readable Dictionary. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, 9(3–11), pp. 7–11.

Reyes, A. de los. (1593). *ARTE EN LENGVA MIXTECA COMPUESTA*. Accessed at: https://books.google.at/books?hl=en&lr=&id=Vbh9oGk-YKwC&oi=fnd&pg=PA4&dq=%22Arte+en+Lengua+Mixteca%22+Antonio+de+los+Reyes&ots=yPfYu574TP&sig=S3ro_9u4ZAp1-7wGnPVoA4WPS4U (15 April 2019)