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

Jack Bowers

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'École Pratique des Hautes Études

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

Soutenue par
Jack BOWERS
le 8 octobre 2020

École doctorale n° 472
École doctorale de l'École Pratique des Hautes Études

Spécialité
Linguistique

Composition du jury :

Guillaume, JACQUES Directeur de Recherche, CNRS	*Président*
Alexis, MICHAUD Chargé de Recherche, CNRS	*Rapporteur*
Tomaž, ERJAVEC Senior Researcher, Jožef Stefan Institute	*Rapporteur*
Enrique, PALANCAR Directeur de Recherche, CNRS	*Examinateur*
Karlheinz, MOERTH Senior Researcher, Austrian Center for Digital Humanities and Cultural Heritage	*Examinateur*
Emmanuel, SCHANG Maître de Conférence, Université D'Orléans	*Examinateur*
Benoit, SAGOT Chargé de Recherche, Inria	*Examinateur*
Laurent, ROMARY Directeur de recherche, Inria	*Directeur de thèse*

École Pratique des Hautes Études | PSL

1. Introduction to Project

This dissertation describes the documentation project of the Mixtepec-Mixtec language[1] (MIX) *sa'an savi* 'rain language' using the Text Encoding Initiative, or TEI (www.tei-c.org) as the encoding format. The benefits of the outcomes of this work are to: present an account of how the TEI and related XML technologies can be used as the primary encoding, metadata, and annotation format for multi-dimensional linguistic projects, including under-resourced languages; evaluate the current tools, standards and practices used in LD; as well as to create a body of linguistic resources (LR) for the MIX language and community. Due to the array of different data and resources produced, this project has components that equally fall within the fields of: digital humanities (DH), language documentation (LD), language description and corpus linguistics. Because of this overlapping relevance, over the processes of attempting to carry out this work in line with best practices in each sub-field, this work has brought to light the potential, and the need to more concretely identify, discuss, and further bring together the overlapping interests, technologies, practices and standards relevant to, and used in each.

The primary output of the project is an open source body of reusable and extensible multimedia language resources including: a multilingual TEI Dictionary, a collection of audio recordings published and archived on Harvard Dataverse (Bowers, Salazar, and Salazar 2019)[2], and a corpus of texts derived from a combination of spoken language transcriptions and written language encoded and annotated in TEI, as well as linguistic and lexicographic descriptions and

[1] Mixtepec-Mixtec Iso 639-3 [mix]; Glottolog [mixt1425]
[2] https://doi.org/10.7910/DVN/BF2VNK

analyses of the Mixtepec-Mixtec language[3]. As MIX is an under-resourced language, the aim has been to integrate as many of the available resources in the language as possible into the TEI corpus with a common encoding and annotation scheme, which depending on the source, requires different degrees of manual work, scripting and the use of digital tools to achieve. The LR created are in turn being used to further knowledge of all aspects of the language itself within the fields of linguistics and lexicography allowing for empirical corpus-based grammatical descriptions and analyses of aspects of the language's features. However, as will be discussed, while linguistic analyses and description (section 2) have been produced as a result of this work, particularly in the form of an analysis of the semantics of body-part terms (Bowers, in press), the main output, and focus of this dissertation is to describe the structure, sources and contents of the corpus, archive and dictionary.

In the process of data collection, annotation, and encoding, I have sought to capture content relevant to every linguistic level from phonetic to semantic and etymological, as well as potential sub-dialectal and even idiolectal variation. In conjunction with the complexity of the data, given the maximally broad scope of linguistic and lexicographic research being pursued, both at present in my own work, as well as in anticipation of future re-use, it is essential to have a means of organizing all the various components of the languages resources within a dynamic, flexible and non-software dependent system. Also, given the lack of dictionary resources for the language[4], it is especially important that what is created is reusable and extensible so that it may continue to evolve, with the possibility of being easily exported or converted to other formats and made accessible in a user friendly format, with the Mixtec community members in mind.

As the scope of this work is multi-faceted and spans multiple academic fields, over the course of this work I have encountered important issues from a number of different disciplines, and have had to continuously find ways to address them in a way that does justice to the language, the goal of providing a quality output for the Mixtec community, adhering to ethical

[3] The GitHub repository (https://github.com/iljackb/Mixtepec_Mixtec) contains the annotated files making up the corpus and the TEI dictionary.

[4] While at the time of submission there is no other dictionary resource for Mixtepec-Mixtec proper, there is a small dictionary (Galindo Sánchez, 2009) for the Abasolo del Valle variant of Mixtec spoken in the Playa Vicente in state of Veracruz by a community who migrated in several stages from the 1930's to the 1950's from the San Juan Mixtepec area. This variety is generally accepted to be the same as Mixtepec-Mixtec.

best practices and finally creating an output that conforms to best practices in digital humanities, TEI and language documentation.

In pursuit of these goals, TEI was chosen as the format for encoding and annotating the corpus, born-digital dictionary, and metadata that would best accommodate all of the aforementioned research goals and desired output. Notably, as will be discussed, in contrast to the patchwork array of tools and in some cases, tool-dependent data formats for each of the main components used in language documentation and computational linguistics, using TEI allows for the entirety of the data to be encoded and annotated in the same format. TEI is widely accepted in the digital lexicographic community as the de facto standard for the encoding of both retro-digitized and born-digital dictionaries and is being increasingly used for annotated lexical text corpora. Additionally, it has extensive metadata related features embedded in each file which allow for creation of features structures for the linguistic fields, people and places, as well as linking between linguistic content and related media without having to produce and edit metadata and content separately.

While TEI is well established and increasingly more widely adopted for projects and resources dealing with major world languages, particularly those of Europe and North America, it is far less adopted in projects dealing with indigenous languages. Aside from publications related to the current project (Bowers, 2015; Bowers and Romary, 2017; 2018a,b; 2019), Czaykowska-Higgins and Holmes (2013) Czaykowska-Higgins et al. (2014) describe creation of a TEI dictionary and an interface application from legacy resources for the indigenous language Moses-Columbia Salish “Nxaʔamxcín”. Additionally, of note is the recent Mesolex project (DEL Grant #HAA-266482-19)[5] for which a primary output is to collect lexical resources from a number of Indigenous Meso-American languages (including varieties of Mixtec) and convert them into a commonly searchable TEI format. A major benefit of the use of TEI in dealing with an under-resourced language is that it allows for the encoding of documents that can be used both as an annotated linguistic corpus resource, which (along with simple schemas), can be simultaneously presentable for human consumption, as well as for researchers in other fields.

[5] https://securegrants.neh.gov/publicquery/main.aspx?f=1&gn=HAA-266482-19

While, as will be discussed, the creation of such flexible multi-purpose resources is at the core of the mission of digital humanities, it has not traditionally been a major priority for most fields of linguists.

In some cases, the use of TEI for documentation work has required the use of the markup vocabulary for new, or less common applications in order to accommodate the particular nuances of the data. Additionally, it requires the use of different combinations of TEI components and features which are less often used together, and thus for which there is little to no examples in the guidelines, nor are there precedented use cases in the literature (one particularly glaring such omission is interlinear glossed text (IGT). It cannot be denied that at times adopting this approach, as opposed to other major toolkits such as SIL's FLEx[6], ELAN[7], Toolbox[8], etc.[9]., has been cumbersome, both in the time required to manually annotate, organize contents, to write conversion scripts and the fact that I am not able to take advantage of many of the user-oriented output features of the aforementioned tools. However, having taken the time to work out the various issues benefits, not only this project in mapping out how to accommodate new unique combinations of features for a non-Indo-European indigenous language, it also has served as a comprehensive survey of gaps both in the TEI, as well as in the field of data standardization, interoperability and interchange.

Furthermore, it is hoped that the adoption of TEI for this work, in combination with the survey of commonly used tools and data formats in LD will contribute to the implementation of new measures to: increase the usability of TEI for potential future users and projects seeking to do similar things, both in terms of the development of new tools for non-experts as well as in setting a precedent that can be emulated; establish a body of scripts and stylesheets to convert between different data formats, and finally to further the cause of data standards and interchange.

[6] https://software.sil.org/fieldworks/

[7] https://tla.mpi.nl/tools/tla-tools/elan/

[8] https://software.sil.org/toolbox/

[9] Though there are certain components of the more commonly adopted toolkits that may seem more user friendly, there are numerous reasons that these programs were not a good fit for this work. These issues will be discussed in this dissertation.

In working with under-resourced languages, it is imperative to be able to integrate *any* potential contemporary, or historical data source, which can come from wide array of different digital or analog formats. In order to build the necessary capacity to integrate and processes such data, the development of toolkits such as GROBID Dictionaries (Khemakhem et al., 2017) is essential. GROBID Dictionaries scans and processes PDF lexical resources and outputs into a TEI dictionary. This innovative technology represents a major component of the development of tools that enable researchers to digitize and create structured dictionary corpora from existing resources (where existing) (Khemakhem et al., 2017). Moreover, as the tasks and approaches become more widely adopted, it will hopefully give rise to a demand for the development of ever more user-friendly software options for carrying out such tasks, and/or the adaptation of existing software toolkits to enable them.

While I present positive components, outcomes and prospects of this work, I also present issues in which some aspects of the work, in which my methodological or technological approach, or the output itself remains to be improved, and about which questions remain to be addressed moving forward. Finally, this dissertation presents only the groundwork of the methodological issues and of course the linguistic output. It is my intention that all dimensions of this work be continued moving forwards, thus herein I present the preliminary results of the technical and some linguistic components of this project.

2. Introduction to Language

Mixtepec-Mixtec is spoken in the 72 communities, neighborhoods, and colonies 'colonias' of the San Juan Mixtepec municipality[10]. In Mexican government data[11], the language is referred to as Western-Central Mixtec (*mixteco de oeste central*); Josserand (1983) classifies the variety as falling within the Southern Mixteca Baja dialect region[12], bordering on the Mixteca

[10] Though not available in any public government source, an unofficial document containing a list of places in the San Juan Mixtepec municipality and their known inhabitants compiled by SIL researcher Gisela Beckmann can be found here: https://github.com/iljackb/Mixtepec_Mixtec/blob/master/misc-sources/Pueblosy%20su%20estatus%20alfabetico.doc (source: Gisela Beckmann, personal communication July, 2020)

[11] https://www.inali.gob.mx/clin-inali/html/v_mixteco.html#47

[12] The term "dialect region" is used in accordance with the classifications referenced from Josserand (1983). As a side note, the term "dialect" has traditionally been used to dismissively refer to indigenous languages in Mexico, and

Alta region[13] and as a separate dialect branch[14], though it is likely that this classification needs revision as more varieties (particularly those in the Juxtlahuaca area are documented). Within Mexico, MIX is also spoken by several thousand speakers living in Baja California, Tlaxiaco, Santiago Juxtlahuaca, and within the United States by significant populations in California, particularly around Santa Maria (where the two project collaborators were raised and one still resides) and Oxnard, Oregon, Florida, and Arkansas.

The number of estimated Mixtec varieties ranges from 52 Ethnologue[15] (Simons and Fennig, 2018) to 81 INALI (2008). As the sources of Ethnologue have traditionally been census from the Mexican government, INALI is likely the most reliable source[16]. Statistics for the speaker demographics and status of Mixtepec-Mixtec have not been collected since 2000 (with a census in 2010 that collected information only by language family) which put the number of speakers at 9,166[17]. An up-to-date evaluation of its speakers is needed as in there is conflicting information regarding its endangerment status. According to the ELDP[18] the status is 'Threatened' whereas according to Ethnologue[19] its status is 'Stable'[20].

Based on first hand observations and in discussing the issue with MIX speakers, the status of 'Threatened' is certainly the more accurate, as the combination of the: more widespread use of Spanish in entertainment, internet, school, as well as the large numbers of MIX speakers who live outside of the speech area whose children are not exposed to the language outside the home, particularly those whose parents speak Spanish or English is observably lowering the number of new speakers. In addition to the pragmatic/demographic issues, as is the case in many

is considered derogatory. Thus, the term "variety" is generally used when referring to different Mixtec (or other indigenous) languages.

[13] Despite these classifications, I have heard native MIX speakers describe their variety as belonging to Mixteco Alto grouping.

[14] https://glottolog.org/resource/languoid/id/mixt1425 (accessed 2019/12/29)

[15] https://www.ethnologue.com/subgroups/mixtec (accessed 2019-08-20)

[16] It should be noted that as of October 2019, Ethnologue is now a paid service to "high-income countries" and thus access is restricted thus without subscription access, the sources can no longer be checked as to where the numbers are based on.

[17] https://www.ethnologue.com/subgroups/mixtec (accessed 2019-08-20)

[18] http://www.endangeredlanguages.com/lang/10531 (accessed 2019-08-20)

[19] https://www.ethnologue.com/language/mix (accessed 2019-08-20)

[20] This discrepancy is particularly curious due to the fact that the ELP page (which gives the status as 'Threatened') cites Ethnologue as the source which gives the status as 'Vigorous'.

indigenous, post-colonial societies, historically and into the present day, speakers of indigenous languages have been victims of racism and discrimination in Mexico as well as abroad in diaspora communities. This, in combination with an attitude that speaking indigenous languages doesn't have any benefits, has undoubtedly played a role in influencing some parents to neglect to pass on the language to their children (Basurto, Hernández Martínez, and Campbell, in press).

Additionally, children of MIX speakers who live in urban areas are increasingly likely to only have receptive knowledge of Mixtec as in their everyday lives they interact with people who may not speak Mixtec, including other indigenous people and thus Spanish becomes the only practical language of communication. Furthermore, among even those who do speak Mixtec, there is a situation of diglossia in which their usage of Mixtec is restricted to certain contextual situations and importantly, topics of discussion. This situation has the effect of limiting the extent of daily life for which Mixtec has vocabulary; the domains in which Mixtec is not used then speakers either will use Spanish loanwords or (at least for bilingual speakers) will switch to Spanish.

2.1 Brief Overview of Mixtepec-Mixtec Language Typology and Features

As the main focus of this dissertation is the language documentation and the particular approach taken with regard to the technological approach, it is not a major goal herein to provide a comprehensive linguistic description of the Mixtepec-Mixtec language. The idea is that the priority has been given to collecting and annotating the materials for both the purpose of ensuring the resources will be preserved and well documented. However, in this section I provide a rudimentary description of some of the major features of MIX language, which will provide a reference for some of the linguistic examples shown herein, both in the corpus and dictionary, and which will form the basis of a more comprehensive grammar to be elaborated on in the near future with quantitative evidence from an expanded corpus as well as acoustic evidence from additional transcribed speech contents.

Note also that Salazar et al. (2020) as part of a field methods course taught by Eric Campbell at University of California, Santa Barbara (UCSB) is in the process of writing a

grammar of the language[21] with Jeremías Salazar. With this stated, in order to provide some linguistic context for many of the linguistic features discussed throughout the examples discussed in this dissertation below I give a concise overview of the MIX language structure and its most notable features. As the data from fieldwork is transcribed and integrated in the the corpus, future work will focus on providing corpus-based quantitative analyses of the language features, including the phonetics and phonology.

2.1.1 Phonological System

Aspects of the phonology of Mixtepec-Mixtec have previously been described by Paster and Beam de Azcona (2004, 2005); Paster (2005, 2010)[22]; as well as Pike and Ibach (1978). This section gives an overview of some of the basic components of the phonology as described by the previous authors with some minor differences and additions according to the data observed thus far in our project[23].

Past literature in Mixtecan (Josserand, 1983) as well as MIX (Paster and Beam de Azcona, 2005) refer to the concept of the "couplet", which defines the structure of Mixtec lexical roots. The root shape of Mixtepec-Mixtec according to Paster and Beam de Azcona (2005) must contain two vowel slots and can comprise of sequences from the following template: (C)(C)V(C)V.

2.1.1.1 Consonants

Below is a chart of the MIX inventory of simple phones. In the following sub-sections, these, as well as the set of complex phones (affricates, pre-nasalized and labialized) will be discussed along with examples, an overview of phonetic variants (where applicable), and their phonotactic distributions as they occur within the lexical roots.

[21] The title of the grammar of Salazar et al. (2020) refers to 'Yucunani Mixtepec Mixtec', as one of my primary colleagues in this project, Jeremías Salazar (who is from Yucunani), has also been the primary consultant and collaborator in the UCSB course, and is the main author of that grammar in progress.
[22] Note that the consultant for the Paster and Beam de Azcona papers at UC Berkeley is one of the two primary collaborators, and sources in this project as well.
[23] Given that the majority of the spoken language data collected in this project is yet to be processed and transcribed, future studies of this data will be made possible both from corpus, and acoustic phonetic perspectives, which will add a much more scientific basis to the understanding of the language's phonology, and provide more evidence for some of the areas which still need more study or more concrete evidence.

	Bilabial	Labio-dental	Alveolar	Post-Alveolar	Palatal	Velar	Labio-velar	Glottal
Stop	p		t			k		ʔ
Nasal	m		n		ɲ			
Trill			r					
Tap or Flap			ɾ					
Liquid			l					
Fricative		v	s	ʃ				
Glide					j		w	

Table 1: Mixtepec-Mixtec simple consonant inventory

2.1.1.1.1 Stops

Mix has four phonologically distinct stops: /p/, /t/, /k/ and /ʔ/. The voiced bilabial stop /p/ is relatively rare and is only found in loanwords e.g. *pain* 'shall', from Spanish *paño* (Paster and Beam de Azcona, 2005), and *paa* (from Spanish *padre*) 'father'. The alveolar /t/ and velar /k/ stops can occur as syllabic onsets in word-initial, or word-medial context, and the alveolar stop is always articulated with a dental quality. The glottal stop never occurs word-initially, and most commonly occurs as an onset word-medially. Additionally, the glottal stop can also occur in word internal coda position, which is the only consonant that can occur outside of a syllabic onset. Voicing in stops is in non-contrastive[24], with the exception of /p/ ~ /b/, the former of which is rare ,and occurs in the context of loanwords, and the later which is a co-variant of /v/.

phone	orthography	phonetic forms	examples
/p/	p	[p]	[páĩ̂] pain *'skirt'* (loanword from Spanish *paño* (Paster and Beam de Azcona 2005) [páâ] paa *'father'* (loanword from Spanish *padre)*
/t/	t	[t̪]	[tẫã̄] taan *'earthquake'* [t̪ìt͡sī] titsi *'belly'*

[24] The lack of phonological contrast of voice in MIX is reflected in the Spanish spoken by native Mixtec speakers, in which it is common to ambiguate the pronunciation of the words *cuando* 'when' and *cuanto* 'how much'.

/k/	k	[k] ~ [ɣ]	[kàā] kaa *'metal'* [ká] ~ [ķá] ~ [ɣá] ka (*demonstrative particle*) [ʃʧàkī] xchaki *'brain'*
/ʔ/	'	[ʔ]	[káʔà̰] ka'an *'speak'* [tóʔlō] to'lo *'rooster'* [jàʔvī] ya'vi *'plaza'*

Table 2: Mixtepec-Mixtec stop inventory and examples

Note that both Pike and Ibach (1978), and Paster and Beam de Azcona (2005) also include the voiced velar stop /g/ as a separate phone from the voiceless /k/. I do not share this view, as the only context in which this phonetic form appears is in the context of its prenasalized form, thus the conditioning environment is parallel to the appearance of the voiced alveolar stop [n̤d]; specifically, it is the result of Post-nasal Voicing Assimilation as described by Paster and Beam de Azcona (2005).

Another variant of the /k/ is sometimes pronounced as [ɣ], though this is largely limited to the context of the demonstrative particle *ka* (most commonly pronounced as [ķá]), and the marker of first person plural inclusive *ko* (most commonly pronounced as [ķó]). These particles are exclusively placed following lexical items and phrases they modify and thus, given that lexical items in MIX almost exclusively end in vowels, this variation is likely due to a combination of a processes of intervocalic voicing > [ķ], then lenition and spirantization > [ɣ].

2.1.1.1.2 Nasals

MIX has three phonologically distinct nasals: /m/, /n/, /ɲ/. Each nasal can occur in word-initial or word-medial onset context.

phone	orthography	phonetic forms	examples
/m/	m	[m]	[máʔà] ma'a *'racoon'* [kùmĭ] kumi '*four'*
/n/	n	[n]	[nànĭ] nani *'name'* [n̤ĭnà] tina *'dog'*
/ɲ/	ñ	[ɲ]	[ɲánī] ñani '*brother', 'kindsman'*

			[ɪ̀ŋʊ̀] iñu *'six'*

Table 3: Mixtepec-Mixtec nasal consonant inventory and examples

Note that the velar nasal [ŋ] is present in the language only as a conditioned variant in the context of the prenasalized velar stop /nk/, which is a result of Nasal Place Assimilation (discussed below in section 2.1.1.1.7).

2.1.1.1.3 Liquids, Trills, Taps and Flaps

The liquid /l/ occurs as a syllabic onset in word-initial and word-medial contexts. Likewise, the flap /ɾ/ is found in some native words, as well as in Spanish loanwords; it can occur as a syllabic onset in word-initial and word-medial contexts. The trill /r/ primarily appears in loanwords, but can also be found in some native onomonapeia words as well. Both the tap and the flap are relatively rare in MIX.

phone	orthography	phonetic forms	examples
/l/	l	[l]	[lūū] luu *'small'* [súlú] sulu *'child'* [tóʔlō] to'lo *'rooster'*
/ɾ/	r	[ɾ]	[ɾà] PRON.3SG.MASC.FORM [sāɾà] sara *'then'*
/r/	rr	[r]	[káru] karru *'car'* (loanword from Spanish *carro)* [t̪íríˈ] tirri *'bumble bee'* (onomatopoeia based on buzzing sound)

Table 4: Mixtepec-Mixtec liquid, tap and flap inventory and examples

2.1.1.1.4 Fricatives

There are three fricatives: /v/, /s/, /ʃ/, all of them can occur word-initially and as the onset of a word-medial syllable. In line with Paster and Beam de Azcona (2005), the labio-dental fricative /v/ is freely variable in each of these contexts with the voiced-bilabial stop [b] and the bilabial fricative /β/. In Spanish loanwords ending with alveolar fricatives [s], the post-alveolar fricative /ʃ/ also appears in offset position.

Paster and Beam de Azcona (2005) suggest that the variation of the labial /v/ between [v] ~ [β] ~ [b] is most common word medially, in particular following glottal stops, and that word-initially the labio-dental fricative form [v] is maintained. Based on observation in our data, this does not seem to be the case as there are numerous examples of this variation in word-onsets (see Table 5).

phone	orthography	phonetic forms	examples
/v/	v	[v] ~ [β] ~ [b]	[vílú] ~ [βílú] vilu *'cat'* [víkŏ] ~ [βìkŏ] ~ [vìkŏ] viko *'cloud'* [sàvĭ] savi *'rain'*
/s/	s	[s]	[sa̰ʔá̰] sa'an *'language'* [kŏsò] koso *'azde'*
/ʃ/	x	[ʃ]	[ʃìnĭ] xini *'head'* [nd̪ŭʃì] ntuxi *'honey'* [lónīʃ] lonix *'monday'* (from Spanish *lunes)*

Table 5: Mixtepec-Mixtec fricative inventory and examples

2.1.1.1.5 Glides

The palatal glide /j/ mostly occurs in word-initial contexts, however it can be found in a few lexical items in medial position, mostly in items which are clearly products of derivation or compounding. The alveo-velar glide /w/ is most commonly present independently from its typical labial offset usage in certain variant pronunciations of the root *kue* [kʷē] (plural marking particle) in which the initial stop is deleted, and only the /w/ is left as the onset. Additionally, the /w/ is also present in loanwords from Spanish. There is one item identified so far *yeua* 'female horse' that has an alveo-velar glide in a contexts other than the two aforementioned.

phone	orthography	phonetic forms	examples
/j/	y	[j]	[jâá] yaa *'tongue'* [kùjāʧĭ] kuyachi *'to approach'* (derived from inchoative prefix *ku-* and adpostition *yachi* 'near')
/w/	-u-	~[w]	[wê] (plural marker), PRON.1PL.EXCL (variant of [kʷê])

			[jéwâ] yeua ‘*female horse’* [hwǻâ] ~ [wǻâ] ‘Juan’ (Spanish name)

Table 6: Mixtepec-Mixtec glide inventory and examples

2.1.1.1.6 Affricates

MIX has four basic affricates: /st/, /ts/, /ʧ/, /sk/. In each of these voicing is also non-contrastive. Whereas /st/ and /sk/ only appear in word-initial position (with the exception that /sk/ can appear word-medially in Spanish loanwords); /ts/ and /ʧ/ occur in both word-initial, and word-medial syllabic onsets. The alveolar stop-fricative affricate /ts/ is often voiced in word-medial (intervocalic) positions and less regularly voiced in word-initial context.

phone	orthography	phonetic forms	examples
/st/	st	[st̲]	[st͡īkǐ] stiki *‘bull’* [mɛ́st͡rù] mestru *‘teacher’* (loanword from Spanish ‘maestro’)
/ts/	ts	[t͡z] ~ [t͡s] ~ [t̲z̲]	[t͡sāʔă] tsa’a *‘foot’* [nt͡zìt͡sì] ~ [nt͡zìt͡zì] ntsitsi *‘wing’*
/ʧ/	ch	[ʧ] ~ [ʤ]	[ʧĩkʷíî] chikuii *‘water* [kàʧǐ] kachi *‘cotton’*
/sk/	sk	[s͡k]	[s͡kɛ́ɨ̱ă] sketa *‘I run’* [s͡kǻʔà] skaka *‘interpret’*

Table 7: Mixtepec-Mixtec affricate inventory and examples

2.1.1.1.7 Prenasalized phones

MIX has five distinct prenasalized phones: /mp/, /nt/, /nk/, /nts/, /nʧ/. These clusters primarily occur in word-initial position, but /nt/, /nts/ and /nʧ/ less commonly occur medially, most often where they have undergone a process of derivation in which a derivational prefix assumes word-initial position (for discussion, see section 2.1.8), or in lexical items which have undergone historic compounding processes. The bilabial pre-nasal /mp/ is rare, and thus far, has only been observed in the single lexical item *mpaa* ‘compadre’.

As discussed by Paster and Beam de Azcona (2005), there are two Assimilation processes visible in MIX prenasals. First, the prenasalized stops and affricates are voiced as a result of

Post-Nasal Voicing Assimilation, e.g. /nt/ is realized as [nd̪], and /nʧ/ is realized as [ndʒ], etc. Note however that /mp/ is the exception to this (possibly because the voiced bilabial stop [b] is part of the phonological space of the phone /v/). The second Assimilation process is Nasal Place Assimilation, specifically, the place of articulation of the pre-nasals are non-contrastive, and they assimilate to that of the following consonant, e.g.: /nk/ is realized as [ŋk̟], /nts/ is often realized as [nt͡z], etc. The prenasalized velars /nk/ vary in their pronunciation between a full velar nasal [ŋ] and a nasalized close vowel [ĩ], some of this variation is reflected in the orthography with some lexical items containing the *in* and others *nk* (see Table 8 below).

phone	orthography	phonetic forms	examples
/nt/	nt	[nd̪] ~ [nt̪]	[nd̪āʔá] nta'a *'hand'* [kònd̪ò] konto *'knee'*
/nk/	ink (*or*) nk	[ĩŋk̟] ~ [ŋk̟]	[ĩŋk̟āà] ~ [ŋk̟āà] inkaa *'to be located'* [ŋk̟ójò] Nkoyo *'Mexico'*
/np/	mp	[mp]	[mpáà] mpaa *'compadre'*
/nts/	nts	[nt͡z] ~ [nt͡s] ~ [nt̻͡z̻]	[nt͡zìt͡sì] ntsitsi *'wing'* [kũt͡sáʔnũ̄][25] kuntsa'nu *'governor' 'king', queen'*
/nʧ/	nch	[ndʒ] ~ [nʧ]	[ndʒíí] nchíí *'where'* [nìkàndʒīĩ] nikanchii *'sun'*

Table 8: Mixtepec-Mixtec prenasalized consonant inventory and examples

The sequence /nts/ is most frequently observed with /a/ and /i/, e.g. *ntsi-* and *ntsa-*, and primarily only in word-onset contexts, with certain exception being where an inflectional prefix is added to a verb, or where a derivational prefix is added to a lexical item e.g. *kuntsa'nu* is either a compound or a derivation and seems likely to be comprised of: the stem *tsa'nu* 'elder' and a segment *kun-*, which may potentially be either a reduced form of the potential copula *kuu,* or the second inchoative prefix *ku-*[26].

[25] I am not sure of the first tone on the noun *kuntsa'nu* because I have only observed it in written texts, thus I have included no tone diacritic on the first vowel. I can be highly confident of the rest of the tones because of the extensive number of observations of the lexical root *tsa'nu* [t͡sáʔnũ̄] 'elder'.

[26] It is not clear however where the nasal *ku**n**-* in *ku**n**tsa'nu* may have come from given that neither of these prefixes are nasalized. The potential prefix *kù-* may be nasalized as *kùn-* [kṹ] when preceeding an onset nasal but this doesn't apply here.This could be an indication this is neither the potential copula, or the inchoative prefix. This merits

2.1.1.1.8 Prespirantized phones

There are two regularly occurring pre-spirantized phones in MIX: /sn/ and /sʧ/, which often occur in causative verbs (see section 2.1.8), and only appear in word-initial position[27]. In each of these clusters, there is a tendency for the [s] to vary with the post-alveolar [ʃ]. Additionally, in the case of the prespirantized nasal /sn/, the nasal place of articulation may vary between the alveolar [n] and palatal [ɲ].

phone	orthography	phonetic forms	examples
/sn/	xn	[s͡n] ~ [ʃ͡n] ~ [ʃ͡ɲ]	[ʃ͡nú̱bikǒ] ~ [s͡nú̱bìkǒ] ~ [s͡ɲú̱bìkǒ] Xnubiko *'San Juan Mixtepec'*
/sʧ/	xch	[s͡ʧ] ~ [ʃ͡ʧ]	[s͡ʧóʔō] ~ [ʃ͡ʧóʔō] xcho'o *'chop'*

Table 9: Mixtepec-Mixtec prespirantized consonant inventory and examples

There is one lexical item in which there is a prespirantized velar stop /ʃk/, *xkama* [ʃ͡kamà][28] which is a loanword from either Spanish *jicama*, or possibly Nahuatl[29] *xīcamatl*; in whichever case, the vowel in the initial syllable was reduced and deleted, leaving just the prespirantized velar stop [ʃ͡k]. Thus, the historical and phonological processes that lead to /sn/ and /sʧ/, and that lead to /ʃk/ are completely unrelated.

2.1.1.1.9 Labialized phones

There are three labialized phones in MIX: /kʷ/, /nkʷ/, /skʷ/. All appear in word-initial onset positions, and only /kʷ/ appears in word medial contents.

phone	orthography	phonetic forms	examples
/kʷ/	ku	[kʷ] ~ [v] ~ [w]	[kʷàʔá] kua'a *'sister'* [ʧīkʷíi̯] chikuii *'water'*
/nkʷ/	nku	[ngʷ]	[ŋk̟ʷīī] nkuii *'fox'*

further investigation. However, the important point of emphasis here is the fact that the only instances of /nts/ in word-internal contexts are as a result of compounding, derivational or other inflectional processes.

[27] As will other complex phones, the only way that pre-spirantized phones may occur word-internally is where there is a process of inflection or derivation, none have been observed in word-internal context as a result of compounding.

[28] I am not sure of the first tone of *xkama* [ʃ͡kamà], thus I have left it without a tonal diacritic.

[29] Whether the item was borrowed directly from Nahuatl into Mixtec, or was borrowed via Spanish, the origin of the item is Nahuatl (see: https://nahuatl.uoregon.edu/content/xicamatl).

/skʷ/	sku	[skʷ]	[sk͡wáʔā] skua'a *'to study', 'learn'*

Table 10: Mixtepec-Mixtec labialized consonants and examples

The sequence [kʷ] has two main variants. The first is in the word *kue* [kʷē], the plural marker, as well as in compounds containing this particle; herein it is sometimes reduced through lenition and deletion to [wē]. In the second, sequences of [kʷ] may be realized as [v][30]: e.g. *kui*, can be observed as [vi]; *takua* [t̪akʷa] 'because' is sometimes observed as *tava* [t̪ava][31].

2.1.1.1.10 Intervocalic sonorant gemmination

Intervocalic sonorants are lengthened in certain (though not all) lexical roots (Paster and Beam de Azcona, 2005). In contrast to vowels however, consonant length in MIX is non-contrastive. Table 11 shows a list of several examples from both Paster and Beam de Azcona (2005) and that also have been observed in our transcribed data. For this, the aforementioned authors posit a rule of Sonorant Gemmination which states that a mora is linked to the medial sonorant in intervocalic contexts.

orthography	**IPA transcription**	**gloss**
ana	[án:à]	*'heart'*
kuñu	[kūɲ:ů̀]	*'body'*
iñu	[ĭ̀ɲ:ů̀]	*'six'*
kumi	[kùm:ĭ]	*'four'*
kolo	[kól:ó]	*'male turkey'*
uni	[ùn:ì]	*'three'*

Table 11: Examples of intervocalic sonorant lengthening

While this is certainly an observed phenomenon, it occurs irregularly and there are many observed instances of these same lexical items, as well as other intervocalic sonorants that are

[30] There are two different lexical items that are spelled *kui* and both dislay this variation; one is the potential copula inflected for third person *kuu* + *-i,* and the other is the third person general pronoun (see section 2.1.3).
[31] Note that the tones are not yet determined for either of these items and thus the IPA has no tonal diacritics. Also, the variant 'tava' has only been found in booklets published by SIL and has not been observed in speech from Yucunany speakers, or in any of the (as of yet) transcribed speech from speakers from other towns.

not lengthened. Thus, it may be better to refer to this as a tendency rather than a rigid, formal rule.

2.1.1.2 Vowels

As described by Paster and Beam de Azcona (2005) and Pike and Ibach (1978), the MIX systems has five vowel places: /i/, /e/, /a/, /u/, /o/. The high close front, and open central vowels: /i/, /u/ and /a/ are the most frequent, while the close-mid vowels /e/ and /o/ occur much less frequently; all vowels have contrastive simple and long oral, and nasalized forms. Tables 12 and 13 show the inventories of MIX oral and nasalized vowels respectively.

	Front	Central	Back
Close	**i** **i**:		**u** **u**:
Close-Mid	**e** **e**:		**o** **o**:
Open		**a** **a**:	

Table 12: Mixtepec-Mixtec inventory of oral vowels

	Front	Central	Back
Close	**ĩ** **ĩ**:		**ũ** **ũ**:
Close-Mid	**ẽ** **ẽ**:		**õ** **õ**:
Open		**ã** **ã**:	

Table 13: Mixtepec-Mixtec inventory of nasalized vowels

Long vowels most commonly only occur in syllabic/word-initial context in which they make up the entire lexical item or in which they are preceded by an onset consonant[32], e.g. VV or CVV. Exceptions to this can be found in items which are the result of compounding: e.g. [nìkàndʒīĩ] *nikanchii* 'sun' (/nì/ + /kaa/ 'to get up', 'climb' + /ndʒīĩ/ 'to shine')[33]; however, the components of some apparent compounds such as [ʧĩkʷíî] *chikuii* 'water' (/ʧĩ/ + /kʷíî/) do not have any obvious semantic meaning that would be relevant to the whole meaning.

[32] I include complex consonants in this CVV classification, e.g. /nt/, /kʷ/, /nʧ/, etc.
[33] The component of *nikanchii* 'sun' /nì/ seems to be the completive prefix, however it is not clear how this would contribute to the meaning. It is possible this portion could come from another, yet unrecognized historical lexical source.

According to Paster and Beam de Azcona (2005), MIX has no phonological diphthongs with the exception of loanwords from Spanish (e.g. [pa̋î̋] *pain* ‘skirt’), and where there are adjacent non-identical vowels, they belong to different syllables. There are a few instances of /ai/ and /io/ in lexical roots that do not seem to be loanwords (at least from Spanish) e.g. [ʧái] *chai* ‘chair’; [s͡kʷĭā] *Skuia* ‘Santiago Juxtlahuaca’; [tsīò] *tsio* ‘side’; [ʃîô] *xio* ‘dress’, ‘skirt’; [kʷáî] *kuai* ‘male horse’. These can vary in pronunciation however, sometimes there is a partial or full epenthetic palatal glide [j] ~[ʲ] that unsystematically occurs between the two vowels, e.g. it is [s͡kʷĭʲā] *Skuia* ‘Santiago Juxtlahuaca’; [tsīʲò] *tsio* ‘side’.

2.1.1.2.1 Close front vowels

MIX has a large number of lexical items that are comprised of long nasal and/or oral close-front vowels that are minimal pairs based on tone.

phone	orthography	phonetic forms	examples
/i/	i	[i]	[ī̀nì] ini *‘inside’*
/iː/	ii	[iː]	[îĭ] ii *‘husband’* [ì́] ii *‘sacred’*
/ĩ/	in	[ĩ]	[ʧ́ʔĩ̄] ti’in *‘rat’* [ĭ̃ĭ̃] iin *‘salt’*
/ĩː/	iin	[ĩː]	[ĩ̋ĭ̃] íin *‘hail’* [ĩ̄ĭ̄] in *‘one’* [ĩ̀ĭ̀] iin *‘nine’* [ĩ̌ĭ̃] iin *‘skin’, ‘leather’*

Table 14: Mixtepec-Mixtec close front vowels

2.1.1.2.2 Close back rounded vowels

The short, oral close back vowel /u/ can occur in onset or offset position, whereas the long oral /uː/ is only observed as offsets. The short nasal /ũ/ is only phonologically contrastive in offsets but may appear in other positions as a result of passive nasalization spreading (see section 2.1.1.2.7). The long nasal /ũː/ also overwhelmingly occurs as an offset, but there is (at least) one exception in which it makes up the entire lexical item, e.g. *uun* ‘yes’.

phone	orthography	phonetic forms	examples
/u/	u	[u]	[ùnà] una *'eight'* [jújú] yuyu *'dew'*
/uː/	uu	[uː]	[kúūˀ] kuu *'be'* (potential copula)
/ũ/	un	[ũ]	[tz͡ɑ̀ˀù] tsa'un *'fifteen'*
/ũː/	uun	[ũː]	[ũ̄ũ̄] uun *'yes'* [kúũ̄] kuun *'to fall'*

Table 15: Mixtepec-Mixtec close back rounded vowels

2.1.1.2.3 Close-mid front vowels

The close-mid front vowel forms /e/, /eː/, /ẽ/ and /ẽː/ are the least frequent of all vowel places in MIX. There are no observed instances of a lexical item (other than Spanish loanwords) beginning with close-mid front vowels in MIX, and they only occur following a consonant in syllabic offsets. The lexical item *ke'en* 'several' is thus far the only known instance of a short nasalized /ẽ/.

phone	orthography	phonetic forms	examples
/e/	e	[e] ~ [ɛ]	[sèˀē] ~ [sèˀɛ̄] se'e *'offspring', 'child'*
/eː/	ee	[eː] ~ [ɛː]	[mēé] mee *'very'*
/ẽ/	en	[ẽ]	[kéˀẽ̄] ke'en *'several'*
/ẽː/	een	[ẽː]	[xe̋ẽ̋] xeen *'sharp', 'dangerous'*

Table 16: Mixtepec-Mixtec close-mid front vowels

2.1.1.2.4 Close-mid back vowels

As mentioned, the set of close-mid back vowels in MIX comprises of /o/, /oː/, /õ/, and the long nasalized phone /õː/. The short oral vowel /o/ is the only form to appear in word-initial position. The long nasalized form /õː/ has only been observed a small number of items in which a lexical root with long nasalized close back rounded root vowel /ũː/ is inflected for first person plural inclusive[34].

[34] Note that this process of replacing the root vowels with long close-mid rounded vowels inflect for 1st person plural inclusive is not prototypical of that inflection, as it predominantly marked with either a pronoun/enclitic (*-kó, yóó*), or as a single moraic close-mid back rounded vowel (*-o* [ó], *-on* [ó] or [ǒ]) which assimilates to root nasalization, e.g. *nti'i* 'all' > *nti'o* 'all of us' (see section 2.1.3 below for more information on person marking). This phenomena will be further investigated and discussed in future studies when more data is available.

phone	orthography	phonetic forms	examples
/o/	o	[o]	[òkò] oko *'twenty'* [sòʔō] soko *'ear'*
/oː/	oo	[oː]	[kòǒ] koo *'snake'*
/õ/	on	[õ]	[nákòʔǒ] nako'on *'let's (incl) go'*
/õː/	oon	[õː]	[ɲȭǒ] ñoo '*our (incl) town, village*' (possessive of [ɲū̃ǔ] '*town, village*') [sáʧȭǒ] sachoon *'we (incl) work'* (1pl.incl inflection of [sáʧū̃ū̃] *'work'*)

Table 17: Mixtepec-Mixtec close-mid back vowels

2.1.1.2.5 Open Central vowels

The MIX system has long and short, as well as nasal and oral open central vowels. Thus far, there are only two lexical items observed that are made up of just a single short vowel, both are grammatical in function, and are comprised of the open central oral vowel /a/: the particle *a* [ā][35], which occurs in sentence-initial position indicating a yes-no question, and the conjunction *a* [á] 'or'.

phone	orthography	phonetic forms	examples
/a/	a	[a]	[ā] a (sentence initial yes-no question particle) [á] a *'or'* [máʔà] ma'a *'racoon'*
/aː/	aa	[aː]	[kàā] kaa *'metal'*
/ã/	an	[ã]	[áʔâ] a'an *'no'*
/ãː/	aan	[ãː]	[ãã̄] aan *'yes'*

Table 18: Mixtepec-Mixtec open central vowels

2.1.1.2.6 Vowel Harmony

As mentioned briefly by Paster and Beam de Azcona (2005), a majority of monomorphemic lexical roots are comprised of multiple instances of the same vowel place,

[35] I have posited a mid tone on the question particle [ā] as this is seemingly the most common realization, but in the tokens in this collection, it seems to vary, and could potentially be low.

which is a result of historical processes and is not a synchronic phonological function. In the vast majority of these items, the harmonized vowels are separated by stops and nasal consonants.

Vowel Combinations	examples
Close Front	[ṯ́ʔī] ti'in *'rat'* [īnì] ini *'inside'* [kìṯ̱ǐ] kiti *'animal', 'horse'* [nd̪íʔì] nti'i *'everything','everyone'*
Close-Mid Front	[sèʔē] se'e *'offspring'* [vēʔē] ve'e *'house'* [kɛ́ʔɛ̄] ke'en *'several'*
Open Central	[máʔà] ma'a *'racoon'* [áʔâ] a'an *'no'* [ndāʔá] nta'a *'hand'*
Close Back Rounded	[kùʔù] ku'u *'woman's sister'* [chṵ̄ʔú] chu'un *'spider'* [jūʔú] yu'u *'mouth'*
Close-Mid Back Rounded	[òkò] oko *'twenty'* [sòʔō] so'o *'ear'* [jōʔó] to'o *'rope'*

Table 19: Lexical roots displaying vowel harmony

2.1.1.2.7 Passive Nasalization

As discussed by Pike and Ibach (1978) and Paster and Beam de Azcona (2005), it is common to see non-contrastive nasalization on certain vowels, most often following a nasal consonant (*progressive nasalization*), but in some cases preceding a nasal (*regressive nasalization*). In the context of nasal consonants, there is no phonological contrast between nasal and oral vowels. Additionally, in couplets (e.g. words with CVCV or VCV), passive nasalization usually occurs in both syllables or neither, and only rarely in one. In the small number of cases where only one syllable is nasalized, it is the second syllable (Paster and Beam de Azcona, 2005).

Table 20 shows examples of lexical items that typically display such non-contrastive nasalization, and Table 21 shows examples in which items with similar or identical sequences of vowels and nasals that do not regularly undergo passive nasalization[36]. Though it remains to be further systematically studied, it appears that passive nasalization may be more common with post-nasal back rounded vowels, and with palatal nasal consonants.

orthography	IPA	gloss
nuu	[nù̘ŭ̘]	*'face'*
iñu	[ì̘ɲːù̘]	*'six'*
tsanu	[t͡sà̘nū]	*'brother's wife'*
ñuma	[ɲŭmá]	*'wax'*
kuñu	[kŭɲù̘]	*'meat', 'muscle'*

Table 20: Examples of items with non-contrastive nasalization

orthography	IPA	gloss
naa	[nāá]	*'carry'*
uni	[ùnːì]	*'three'*
tina	[t̪ĭnà]	*'dog'*
nama	[nàmá]	*'soap'*
koni	[kóní]	*'female turkey'*

Table 21: Examples of items not displaying passive nasalization

2.1.1.2.8 Passive Glottalization

Also attested by Paster and Beam de Azcona (2005) is the fact that in the context of intervocalic glottals, vowels may be realized as creaky voiced variants, e.g. a (generic) VʔV sequence may be realized as V̰V̰. This process also may occur in combination with nasalized vowels, e.g. ṼʔṼ may be realized as Ṽ̰ʔṼ̰ or Ṽ̰Ṽ̰.

[36] I use the orthography as a reference to compare with the commonly realized phonological forms as it was developed by native speakers, and their spelling conventions should be considered an indication of their judgements of the given word forms.

2.1.1.3 Tones

MIX is a tonal language with three tone levels (*low, mid, high*), as well as a *rising* and *falling* tones which can occur on a single mora and can combine in the context of bimoraic long vowels to create different sequences of global tone patterns[37]. Included in Table 22 below are examples of low, mid and high tones, as well as rising and falling tones in lexical items.

Tones	examples
Low	[sùt̪ù] sutu *'priest'* [òkò] oko *'twenty'*
Mid	[vēʔē] ve'e *'house'* [jātʃī] yachi *'near'*
High	[kóní] koni *'female turkey'* [lótʃǐ] lochi *'vulture'*
Rising	[jǒsō] yóso *'metate'* [jōsǒ] yosó *'(grassy) plain'* [t̪ǐnà] tina *'dog'* [jǔt̪ī] yuti *'sand'* [t̪ínānă] tinana *'tomato'*
Falling	[súkû] suku *'high'* [kōt̪ô] koto *'sarape'* [āʔâ̂] a'an *'no'* [sâʔvà] sa'va *'frog'* [sâʔmă] sa'ma *'clothes'*

Table 22: Mixtepec-Mixtec basic tones with examples

Rising tones are much more commonly observed in the single moraic context than falling tones. It should be noted that Paster and Beam de Azcona (2005), Pastor (2004), and Pike and Ibach (1978) describe both single moraic and bimoraic contours as a series of level tones, and do

[37] A full inventory of the possible tone level combinations over VV spans is still being studied at present. Thus, it is possible that instances of additional contour combinations may be found, or that some of those described herein may require revision. Further descriptions based on observations of transcribed speech will be published in future stages of this project. Note also that recordings and notes created in the Salazar et al. (2020) project at University of California at Santa Barbara were also consulted for determining certain tones in lexical items for which there was previously no, or few quality recordings.

not distinguish simple *rising* and *falling* tones occurring on a single mora as distinct phonological units. In certain conditioning contexts, *high* and *rising* as well as *low* and *falling* tones are interchangeable and non-contrastive.

A primary reason *falling* and *rising* are treated herein as distinct phonological tones (as opposed to a sequence of specific tone levels as in previous studies), is that there are no known instances of two lexical roots whose only distinction is the difference between the onset and offset tone level in a rising or falling contour occurring in a single mora (e.g. *CV̌CV and *CV̋CV are both assumed to be phonologically equal to CV̌CV). Thus, the specific tonal onset or offset level on a single mora does not seem to be minimally contrastive, and the basis for these phonological tones is simply their upward or downward F0 contour.

Over the course of bimoraic (long) vowels (CVV or VV syllables), nearly every sequential combination of the three level tones has been observed, however there does not seem to be any contrast between *Low High* VV patterns[38] and *Low Rising* (see Table 24)[39]. These combined sequences result in long level tones and various global falling and rising tone contours where the onset and offset tones differ[40]. Examples of each combination of level tones are shown below in Table 23.

[38] There is an acoustic difference between a *low rising* and what would be a *low high*, which is the degree of the upward slope (F0 pitch increase) is much steeper in combinations involving a rising tone rather than a simple upward slope between two level tones (such as that which occurs on a *low mid*, or *mid high* VV sequence).
[39] It has been shown by Ohala (1978) and Ohala and Ewen (1973) that it takes longer to produce a pitch increase than decrease (e.g. to produce the contours required for *low high* or *low rising* tones). Additionally, citing these studies, Silverman (2003) has shown that there can be diachronic effects to a language's tonological inventory resulting from function interactions of such phonetic factors, and notably, that there are unique patterns and physiological pressures observable in rising tones. While the pattern in MIX is not specifically mentioned in the Silverman (2003) study, the phonetic bases for these works may offer an avenue for understanding how diachronic and phonetic factors may be relevant to the idiosynchracy in this gap in the tone distribution patterning, i.e. given the physiological requirements to produce a *low high* and *low rising* tone contour, the signals produced may not have been salient enough to remain distinct phonological tone patterns, which could have lead them to merge into one single pattern, e.g. *low rising.*
[40] The distinction between a falling or rising tone and a sequence of two distinct level tones is determined by the degree with which a given tone contour ascends or descends within the space of a single mora. Future studies will present an extensive acoustic and quantitative basis for this classification.

Tones	examples
Low Low	[ʧṵ̀ṵ̀] chuun *'star'* [nt̰à̰à] ntaa *'flat', 'truth'* [ḭ̀ḭ̀] iin *'nine'*
Low Mid	[vèē] vee *'heavy'* [ʧà̰ā] chaa *'man'* [kàā] kaa *'metal'*
Mid Low	[ɲṵ̄ṵ̀] ñuu *'town', 'village'* [yōò] yoo *'cup, drinking vessel'* [sāà] saa *'bird'*
Mid Mid	[ḭ̄ḭ̄] in *'one'* or (indefinite determiner) [lūū] luu *'small'*
Mid High	[mēé] mee *'very'* [kʷēé] kuee *'not'*[41]
High Low	[ʧáì] chai *'chair'* [mpáà] mpaa *'god-father (of son)', 'compadre'*
High Mid	[ɪ́ḭ̄ḭ̄] iin *'to exist', 'there is'* [kʷíī] kuii *clear'*
High High	[ɪ́ɪ́ḭ̄ḭ̄] íin *'hail'* [nʤáá] nchaa *'blue'*

Table 23: Combinations of level tones on CVV couplets

Table 24 below shows examples of combinations of level tones with falling and rising tones observed thus far.

[41] The lexical item *kue* [kʷēé] 'not' is often reduced in length in fast, or casual speech and in these cases the tone is often realized simply as *high* [kʷé] or *rising* [kʷě].

Tones	examples
Low Rising	[xe̋ĕ̋] xeen *'sharp', 'dangerous'* [ĭĭ̋] iin *'salt'* [ĭ̋ĭ] ii *'sacred'* [nűŭ̋] nuu '*face'* [nàă] naá *'to end'*
Mid Rising	[vīī] vii *'pretty', 'healthy looking'* [nāā] naa *'dark'*
High Rising	[kʷíĭ] kuii *'green'* [kʷĭ̃ĭ̃] kuiin *'narrow'* [íĭ] ii *'husband'* [ĭ̃ĭ̃] iin 'skin'
Low Falling	[ʃîô] xio *'dress', 'skirt'* [kʷàâ] kua *'about', 'approximately'*
High Falling	[páâ] paa *'father'* (loanword from Spanish *padre* [ˈpa.dɾe]) [hwáân̂] 'Juan' (loanword from Spanish *Juan* [ˈhwan]) [kwáâ] kuaa *'blind'* [kʷáân̂] kuaan *'yellow'* [náâ] náa *'to carry'*
Rising Mid	[t͡sz̆āā] tsaa *'new'* [ŋǩʷīī] nkuii *'fox'*
Falling Mid	[tân̄ān̄] taan *'earthquake'*

Table 24: Global multi-level tone patterns on CVV couplets

Of the two CVV items identified as the pattern *high mid low* by Paster and Beam de Azcona (2005) and Pastor (2004) ([páâ] paa *'father'* and [hwáân̂] 'Juan') both are Spanish loanwords and the tone pattern adopted in the Mixtec forms reflects the Spanish stress pattern. In these cases, the original stress on the first vowel, in MIX becomes a long bimoraic vowel with the stress (high tone) on the first mora, and a falling tone on the final mora. In Table 24, these are

represented as *High Falling.* Note that the non-stressed portion of Spanish loanwords shows a tendency for deletion in MIX, and that word-final nasal consonants are deleted and the preceding vowels are nasalized.

In this section, I have discussed only tones that occur on a single vowel, and those sequences that occur on long vowels in a single syllable (e.g. VV or CVV), and have not sought to provide a full inventory of tone melodies that occur over the course of multisyllabic lexical roots (e.g. CVCV, VCVV, CVCVCV, etc.). Additionally, issues of tone sandhi, and a full examination of the role of lexical tone in MIX morphology will also be further examined in the a more comprehensive presentation of the MIX linguistic system.

2.1.2 Basics of Information Structure

Syntactically, like other Mixtecan languages, MIX is an VSO language examples (1)-(3), though this can be changed in the context of pragmatic focus shifts such as in interrogatives (ex. 4), responses to WH questions (ex. 5), emphatic statements (ex. 6). Also, like other Mixtecan varieties, there is no case and word ordering plays a major role in syntactic and pragmatic function. Note that the language content in this section is presented in the working MIX orthography as used by SIL Mexico[42].

(1) INTRANSITIVE

tsátsi chaa
IPFV\eat man
'the man is eating'

(2) TRANSITIVE

tsátsi chaa kuñu
IPFV\eat man meat

[42] Glossed examples are given in orthography due to the fact that a significant number of them are from text sources for which no audio is available. Thus, in order to be consistent in the transcription method, the orthography is used. In cases where the tone is both known, and functionally relevant to the vocabulary, and lexical phenomena presented, IPA examples are also given in the tables. Future iterations of the description of the language will be presented with full tone data.

‘the man is eating meat’

(3) DITRANSITIVE

kun-kua’a xu’un nuu Jack

POT-give\1SG money ADPOS[face] Jack

‘I will give money to Jack’

(4) WH-NARROW FOCUS SHIFT

nchíí yee =ni

where live =2SG.FORM

‘Where do you live?’

(5) REPLY TO WH-NARROW FOCUS

nuu chuun inkaa =yu

ADPOS[face] work COP.LOC =1SG

‘I’m at work’

(6) DEMONSTRATIVE EMPHASIS

sutu =ka ni-kani =yu

priest =PTCL.DEM PFV-hit =1SG

‘that priest hit me’

2.1.3 Marking Person and Pronouns

Verbs, predicative adjectives, nouns, adverbs, adpositions and in some cases conjunctions (for comitative functions) are marked for person either with: a morphological inflection (which can be a vowel and/or tone change), an enclitic or pronoun. Note however that verbs are only marked for person when the nominal subject is not explicitly specified. Where there are two consecutive verbs, such as in volitive modal contexts, e.g. (ex. 7), both the first and second verb are inflected for person, however the second uses the irrealis stem whereas the first the realis (see section 2.1.7 for description of verb stems and mood in MIX):

(7) tsátsi chaa

IPFV\eat man

'the man is eating'

(8) kúni =yu katsi

IPFV\want =1SG eat[IRREAL]\1SG

'I want to eat'

(literally) 'I want I eat'

The usage of morphemes vs the enclitics shown above for marking the primary argument of a verb are conditioned by the phonological properties of the stem, particularly the tone and vowel environments. Additionally, in some cases pragmatics may also play a role. For a more detailed description of the phonological factors which condition the use of a morpheme, a tone change, or an enclitic see: Paster and Beam de Azcona (2004, 2005); Paster (2005). MIX has at least three sets of pronouns: the dependent enclitic pronouns; the independent emphatic pronouns; and demonstrative pronouns. Table 25 shows the inventory of the clitic/pronouns, morphemes and emphatic pronouns.

The emphatic pronouns are used in reflexives, for emphasis, contrast, and topic shifting and are a combination of *mee* [mèě][43] the basis emphatic form with an enclitic pronoun or the corresponding morpheme. These pronouns in Table 25 can be used as subjects (examples (4), (5), (8) above), or objects (ex. (6) above) in transitive and intransitive phrases, and can be used in marking possession as well (see section 2.1.5).

[43] In the tokens collected, in isolation, this is most commonly articulated as *low high*, though in speech contexts (and depending on the tone context on the offset) a significant minority of token have a *mid high* pattern which makes it homophonic with the adverb *mee* 'very', though, given their different semantic and discourse contexts of usage, this is likely not a problem or a point of confusion. In the context of the emphatic pronouns which combine with the clitics, the tone pattern of this first portion varies between the most common pattern of *mid high* [ēé] and *mid mid* [ēē], this latter seems to occur where the following tone is *high* (e.g. [mēēní] 2sg.form, [mēēŋá] 3sg.form.f); in Table 25 I have transcribed the most common realization of these tones for each pronoun.

Person	Gender/Entity	Clitic/Pronoun	Morphemes	Emphatic
1.	(sg)	yu [jù]	*low* or *falling* tone (on final V)	mee [mèě]
	Exclusive (pl)	kue [kʷê]		meekue [mēékʷê]
	Inclusive (pl)	ko [ḵó] yóó [jóó]	-o [ó] ~ -on [ő]	meeko [mēéḵó]
2.	Familiar (sg)	ku [ḵǔ]	-u [ú] ~ -un [ű]	meu [mēű]
	Familiar (pl)	kueyu [kʷējú] koyu [kōyú]		meekueyu [mēékʷējú]
	Formal (sg)	ni [ní]		meeni [mēēní]
	Formal (pl)	kueni [kʷēní]		meekueni [mēékʷēní]
3.	General (sg, pl)	ña [ɲà] kui [kʷi] ~ vi [vi][44]	-i [ì] ~ -in [ǐ] -a [à] ~ -an [ȁ]	mii [mīí] meeña [mēéɲà]
	Informal (pl)	kueyi [kʷējì] koyi [kōyí]		meekueyi [mēékʷējì]
	Formal: Masculine (sg)	ra [ɾà]		meera [mēéɾà]
	Formal: Masculine (pl)	kuera [kʷērà]		meekuera [mēékʷērà]
	Formal: Feminine (sg)	ñá [ɲá] ná [ná]	-í [í] ~ -ín [ǐ] -á [á] ~ -án [ȁ]	meeñá [mēēɲá] meená [mēēná]
	Formal: Feminine (pl)	kueñá [kʷēɲá] kuená [kʷēná]		meekueñá [mēékʷēɲá] meekuená [mēéwēná]
	Formal: Human (sg)	na [nà]		meena [mēénà]
	Formal: Human (pl)	na [nà]		meekuena [mēénà]
	Animal	ti [t̪í]		meeti [mēēt̪í]
	Deity/Holy	ya [jà]		meeya [mēéjà]
	Wood	tu [t̪ū]		meera [mēét̪ū]
	Spherical	ti [t̪î]		meeti [mēét̪̂]
	Child	tsi [t͡sī]		meetsi [mēét͡sī]
	Liquid	ra [ɾá]		meera [mēēɾá]

[44] I am unsure of the tone of the 3rd pers sg general pronoun variants *kui* and *vi* as I've only observed them in orthographic form literature.

Table 25: MIX enclitic and emphatic pronouns in working MIX orthography[45]

Some of the pronouns in Table 25 are derived from the nouns they stand for as shown in Table 26:

Full Form Noun	Meaning	Enclitic/Pronoun
ña'a [ɲàʔā]	'thing'	ña [ɲà]
ña'á [ɲāʔá]	'woman'	ñá [ɲá]
kiti [kǐʈ̲ǐ]	'animal'	ti [ʈ̲í]
tutu [ʈ̲ûʈ̲ǔ]	'wood'	tu [ʈ̲ū]

Table 26: Full form source nouns and their corresponding enclitic pronouns

2.1.3.1 Demonstrative Pronouns and Components

Demonstrative pronouns are comprised of certain enclitic pronouns with the demonstrative particle *-ka*; e.g.: *ñaká* [ɲàɻ̣á], which can mean 'that', 'there', 'these', 'those'; *ñáká* [ɲáɻ̣á], meaning 'that woman' (from the formal female pronoun ñá)[46]; *naka* [nàká] 'those people' (same *na* as in the third person general formal pronoun/enclitic). There is also the distal pronoun *ika* [īkā] meaning 'there'. These also function emphatically and can be used to disambiguate co-referenced participants in a discourse.

(8) **ñaká** n-tsatsi cha n-tsi'i chikuii
those PFV-eat\1SG and PFV-drink\1SG water

[45] Note for the: animal, wood, spherical, child and liquid forms, there are also plural versions of each the enclitic and emphatic pronouns following the same patterns (e.g. for enclitics: kue+PRON and for emphatic: meekue+PRON) but were not included for reasons of space.

[46] Other Mixtec varieties, e.g.: Chalcatongo Mixtec (Macaulay, 1996); Diuxi-Tilantongo (Kuiper and Oram 1991); Jamiltepec Mixtec (Johnson, 1988); Ayutla Mixtec (Hills, 1990) amongst numerous others have attested "free form" independent pronouns which include 1st, 2nd, and other persons. It may be possible that the MIX pronouns *yo* (2sg.inf) and *yóó* (1pl.incl) shown in Table 2 may in fact be instances of this, as they have clear cognates in numerous other varieties, e.g.: *yòò'* (inclusive) Ayutla (Hills, 1990); *yò'ó* (inclusive) Jamiltepec (Johnson, 1988), *yo̲'ó/yò* (2sg.inf) Diuxi-Tilatongo (Kuiper and Oram, 1991). In all observations in the MIX data, these only occur as objects of a transitive verb. Thus, it is possible that there is another set of 1st and 2nd person independent pronouns that would be counterparts to the full nouns of the 3rd person forms from which enclitic pronouns such as *ñá, tu, ti,* (e.g.: *ña'á* 'woman', *tutú* 'wood', *kiti* 'animal' respectively) though more research is needed.

‘I ate those and drank water’

The particle *ka* [ká̰] seen in these forms is primarily used to carry out demonstrative emphasis, mostly following nominal subjects, objects and even obliques, and it is also an active component in the pragmatic and information structure changes which license certain grammaticalized extensions of BPT (see Bowers (in press) for discussion). Note also there is another particle *ka* [kà̰] which as seen in other varieties, including Chalcatongo Mixtec (Macaulay, 1996), in which it is described as the additive particle (see examples (10), (42)).

(9) DEMONSTRATIVE

chaa **=ka**

man **=PTCL.DEM**

‘that man’

(10) ADDITIVE

ma= kua’a **=ka** staa katsi-a

NEG=give\1SG **=PTCL.ADD** tortilla eat-3SG.INF

‘I will not give him anything more to eat’

Additionally, there is another demonstrative proximal pronoun *ño’o* [ɲóʔō], ‘this’ or ‘here’ (ex. 11), which appears to be the pronominal counterpart of *yo’o* [jóʔō] (see example (12), also (19), (24)), which can function as a proximal demonstrative determiner, e.g. ‘this (X)’, or a proximal locative pronoun meaning ‘here’.

(11) nchii kuu **ño’o**

what COP **PRON.DEM.PROX**

‘what is this?’

(12) staa **yo’o**

tortilla **DET.DEM.PROX**

‘this tortilla’

2.1.4 Copular and Related Expressions

MIX has several copular verbs which follow the same inflection patterns as regular verbs, and certain adjectives may occur as predicates[47]. The primary two copula in MIX are *kaa* [káā], and *kuu* [kúū], in numerous other varieties of Mixtec, e.g.: Chalcatongo: (Macaulay, 1996); Diuxi-Tilatongo: (Kuiper and Oram, 1991); Ayutla (Hills, 1990), the cognates of these forms are classified as the *realis* and *potential*. Though, as shown in examples (16) and (17), there are certain complimentary usages of the two copula, their distribution is not in line with such a distinct classification along the lines of *realis* and *potential*[48].

(12) ka’nu ta **ku-**i
big very **COP-3**
‘it is very big’

(13) nchii **kuu** ño’o
what **COP** PRON.DEM.PROX
‘what is this?’

(14) che’e **kaa** xini patsa’nu
beautiful **COP** hat grandfather
‘Grampa’s hat is nice’

(15) nixi **ka-**u
how **COP-**2SG.INF
‘How are you?’

[47] Note that it hasn’t yet been determined what are the precise factors for which adjectives may function as predicates.
[48] Further evidence that *kuu* is not itself potential is the fact that it can inflect for potential aspect: *kun-kuu* and perfective aspect *ni-kuu.* Additionally, *kaa* can also inflect for potential *kun-kaa.*

An interesting dichotomy between the two can be found in comparing the following question and answer pair (ex. 16) and (ex. 17) where in the former, *kuu* is used and in the latter *kaa* is used:

(16) nchii hora **ku-i**
what time **COP-3S**
'what time is it?' (Nieves and Beckmann, 2007b)

(17) **kaa** iñu ntaa
COP six o'clock
'It's three o'clock' (Nieves and Beckmann, 2007b)

In the corpus, the copula *'kaa'* is also observed often in the context of phrases meaning to 'look like':

(18) tono **kaa** ti'in+ita
look.like **COP** skunk[rat+flower]
'It looks like a skunk' (Rojas Santiago et al., 2014)

However, in a phrase meaning 'to be similar to', the order is reversed:

(19) yutu yo'o tsá'-i kui'i ña **kaa** tono limu
tree this IPFV/give-3 fruit that **COP** like lime
'This tree produces fruit that is similar to limes' (Rojas Santiago et al., 2014)

There is also another copula-like verb *iin* [ĩ̃ĩ], which can function in a number of different sense, including as an existential copula 'there is'; 'to be'.

(20) EXISTENTIAL COPULA: *iin*
iin ve'e na'nu
exist building very.big

‘there is a very big building’

Though it is not yet clear what, if any semantic or other lexical criteria determine whether an adjective can be predicative, when they can, they inflect identically to verbs with the same pronoun/enclitics, or morphemes:

(21) NOUN-ADJECTIVE

yutu suku

tree tall

‘tall tree’

(22) PREDICATING ADJECTIVE

suku =yu

tall =1SG

‘I am tall’

2.1.5 Noun Phrases, Possession and Related Expressions

In MIX, like other varieties of Mixtec, noun phrases precede modifying adjectives (ex. 23), and demonstrative determiners (ex. 24); in possessive (ex. 26) and (ex. 27) and part-whole constructions (ex. 25), nouns are expressed in the same syntactic order as are possessive phrases, with the first noun (the part) preceding the head of the phrase (the whole), e.g. N(part/possessed)-N(whole/possessor). The indefinite article *in* (and numbers in general)[49], as well as the plural marker *kue* however, both precede the noun they modify.

(23) NOUN-ADJECTIVE

yutu **suku**

tree **tall**

‘tall tree’

[49] The indefinite article *in* [ĩ̃ĩ] is the number ‘one’, the orthography represents it distinctly because the number nine *iin* is also a long, high front nasal vowel, with a low, [ḭ̃ḭ̃].

(24) NOUN-DEMONSTRATIVE DETERMINER

yutu **yo'o**

tree **DET.DEM.PROX**

'this tree'

(25) NOUN-GENITIVE/PART-WHOLE

xiní **chaa**

hat **man**

'the man's hat'

(26) POSSESSIVE

maa **=yu**

mother **=1SG**

'my mother'

(27) POSSESSIVE BPT

nu**u̲**

face**1SG**

'my face'

(28) INDEFINITE ARTICLE

in chaa

ART.INDEF.SG man

'a man'

(29) PLURAL MARKER

kue= chaa

PL= man

'the men'

Additionally, oblique phrases with adpositions also mirror this same structure, which as shown by Brugman (1983), Brugman and Macaulay (1986) and Bowers (in press), this is not coincidental as many of the prepositions are metaphorical extensions of relational nouns, most notably body part terms, which are in their most primitive sense, part-whole noun phrases, e.g.:

(30) **nuu** + ve'e
face + house
'front of the house'

(31) BPT IN STATIC ADPOS PHRASES
ntú'u saa =ka **nuu** ve'e
IPFV\sit bird =PTCL.DEM **face** house
'that bird is sitting in front of the house'

(32) inká-i **tsa'a** yutu
IPFV\COP.LOC-3 **foot** tree
'It is under the tree'

But the semantics of the particular body part is evident in the usage of a given extended adpositional sense depending on the term being related to, as shown in (ex. 33), in relating to objects that are physically akin to four legged animals, the BPT *titsi* [t̪ìt͡sī] is used instead of 'foot'. In the expression translation to 'under the table', the configuration of an object under a table is more akin to being under a four legged animal, whereas when something sitting at the base of a tree is more akin to being at the feet of a human:

(33) ntú'-i **titsi** mesa
IPFV\sit-3 **stomach** table
'It is sitting under the table'

(34) BPT IN DYNAMIC ADPOS PHRASES

ntsaa =kue **nuu** chuun

PFV\arrive =1PL.EXCL **face** work

‘We arrived at work’

These extended BPT are extended in adposition phrases beyond the domain of space and motion as shown in examples (35) and (36) show *nuu* [nùŭ] ‘face’, and (37) shows *tsa’a* [t͡sàʔă] ‘foot’ in oblique ditransitive phrases with indirect objects:

(35) FACE IN TRANSFER OF POSSESSION

kun-kua’a xu’un **nuu** Jack

POT-give\1SG money **face** Jack

‘I will give money to Jack’

(36) FACE IN TRANSFER OF INFORMATION

ntakani =na **nu̱** ña ntivi karru =ku

PFV\tell =3PL.FORM.GEN **face**\1sg REL PFV\break car =2SG.INF

‘Someone told me your car broke down’

(37) FOOT IN EXCHANGE FOR

kun-cha’vi =yu **tsa’**-i

POT-pay =1SG **foot** -3

‘I’m going to pay for it’

Note from the examples above, that even in the extended sense (ex. 35-37) in which the meaning has grammaticalized well beyond the original nominal sense, the BPT-N information structure remains. The extensions of the BPT, particularly in the context of spatial and motions phrases, can be best analyzed using the concepts of trajector and landmark from Cognitive Grammar (Langacker, 1986, 1987), see Bowers (in press) for such an analysis.

2.1.6 Conjunctions and Adverbs

When marked for person, the structure of predicative adjectives, adverbs, and conjunctions also mirrors that of V-PERS$_{(SUBJ)}$, e.g.: ADJ-PERS, ADV-PERS, CONJ-PERS. The conjunction *tsi* [t͡sī] ‘with’, ‘and’ (which occasionally is observed as an adposition ‘to’), is inflected as: *tsi-an* ‘with him/her/it (informal)’:

(38) ntuu **tsi** tsikuaa
day **and** night
‘day and night’

(39) ni-kitsaa =kuera **tsi-an** ñuu yo’o
PFV-arrive =3PL.M.FORM **with-3SG.INF** town this
‘they arrived in this town with it’ (Mendoza Santiago, 2008)

When inflected, certain adverbs come between the base and the inflection or clitic, note example (40) shows the use of the BPT *sata* [sàt̪ă] (*inflected for first person singular as* [sàt̪â]) in an extended adverbial sense meaning ‘backwards’ (see Bowers (in press) for in-depth analysis and discussion). Additionally, example (41) shows both an inflected conjunction and the presence of the adverbial *ta* [t̪à] ‘very’, which comes between the verb and the enclitic *yu* (1sg).

(40) tsíka **sat̪a**
IPFV\walk **back\1SG**
‘I’m walking backwards’

(41) kúni **=ta =yu** káka+nuu tsi-an
IPFV\want **=very =1SG** stroll [walk+face] with-3SG.INF
‘I really want to take a stroll with him’ (Gómez Hernández, 2008a)

In the following example, the additive particle *ka* follows the adverbial *so* and precedes the enclitic pronoun of the subject *ti*, this also represents an example of the comparative:

(42) luu **so =ka =ti**

small **very =PTCL.ADD =3SG.ANML**

'It is so much smaller' (Rojas Santiago et al., 2014)

Note however that in standard VSO information structure, most adverbs are not marked and occur in sentence final position:

(43) ni-kuun savi **takuni**

PFV-fall rain **yesterday**

'it rained yesterday'

2.1.7 Verbal Inflections: Aspect and Mood

According to Bickford and Marlett (1988), verbs in Mixtec languages inflect for aspect, and mood rather than pure tense, and although the various aspects can refer to events in the present, past and future, they refer to the internal temporal structure of a situation as opposed to a specific location in time. Bickford and Marlett (1988), Macaulay (1996), and numerous others have shown that there is a primary distinction between Realis and Irrealis mood, which is reflected in a dichotomy between verb stems in Mixtec languages. Accordingly, many, (though not all) MIX verbs have a realis and irrealis form[50]:

Verb	Realis	Irrealis
'walk'	tsika	kaka
'sing'	tsita	kata
'cry'	tsaku	kuaku

[50] Note that in Mixtec lexicography, the gloss form of the verb is the irrealis form, according to Mille Nieves of SIL Mexico, this is the equivalent form (both phones and tones) to the stem on an inflected verb in the potential aspect. In other varieties of Mixtec such as Chalcatongo Mixtec (Macaulay, 1996) the tones are not the same on the cognate forms of realis and irrealis stems, with the exception that it is possible to identify where the offset base tone is low or falling (due to the behavior of the 1st sg inflection), I am not yet sure of how to identify the underlying tones on the realis forms. For this reason, I have left these forms in Table 27 without tones in their SIL orthographic forms.

'give'	tsa'a	kua'a
'sleep'	kixi	kusu

Table 27: Realis and Irrealis verb forms in MIX

As described by Macaulay for Chalcatongo Mixtec, some verbs whose realis and irrealis stems differ display various types of alternations between the given forms, the most common of which is an alternation between the realis *ts* and irrealis *k*, though there are others including: x- and k- alternation (MIX *ts* and *k*); x- and k- alternation plus tone alternation; x- and k- alternation plus vowel alternation; x- and k^w- alternation; tone alternation (only); and several others[51].

The realis forms are used with: the Perfective (also referred to as *Completive*[52]), Imperfective (also referred to as *Incompletive*, or *Continuative*), Habitual, and the Progressive aspects[53]. Irrealis forms are used for the Potential aspect, imperatives, as well as the Modal[54]. MIX verbs are thus marked for aspect and mood with a combination of the verbal stems (where applicable) in addition to prefixes, and/or tone.

[51] In MIX the due to a lack of processed speech data, specifically with regards to the irrealis base forms, most particularly with respect to the tones, the specific details and extent of the alternations is still under investigation and I have refrained from attributing tones to the realis and irrealis base forms to avoid incorrect assertions.
[52] Amongst the other studies of Mixtecan varieties that use the term *Completive* and *Incompletive* are: Paster and Beam de Azcona (2005) for (Yucunani Mixtepec Mixtec); Macaulay (1996) for Chalcatongo Mixtec; Kuiper and Oram (1991) for Diuxi-Tilatongo Mixtec; Hills (1991) for Ayutla Mixtec (though the latter two use *Continuative* rather than *Incompletive*);
[53] Kuiper and Merrifield (1975), Macaulay (1996), Bickford and Marlett (1988), amongst others have discussed the issue of the Progressive aspect in other Mixtec varieties, amongst the characteristics of which are additional verb stems in addition to the standard Realis – Irrealis contrast, though only in the context of motion verb phrases. This issue is related to the semantics of motion and arrival; however, the specific behavior of the progressive aspect verb stems in MIX in comparison to cognate varieties requires a more in-depth analysis and will be addressed in further works.
[54] The term *Modal* is used in line with Macaulay (1996) in describing the cognate function for Chalcatongo Mixtec.

2.1.7.1 Imperfective

The imperfective aspect is used to express present situations, and is not marked with a prefix, but with a high tone on the initial vowel[55] of the realis verb form[56]. It should be noted that Paster and Beam de Azcona (2005) describe an exception to this rule of marking imperfective with a high tone in which, when the first vowel on irrealis verb root has a mid tone, this tone remains unchanged in marking the imperfective, (see example for *sketa* 'run' in Table 28).

Verb (irrealis)	Imperfective
katsi *'eat'*	**tsátsi [t͡zátsî]** *'I am eating'*
ko'o *'drink'*	**tsí'i [t͡zíʔî]** *'I am drinking'*
ka'an *'speak'*	**ká'an yu [kǻʔà̰ jù]** *'I am speaking'*
kuaku *'cry'*	**tsákuia [t͡zákʷīà]** *'he/she is crying'*
kusu *'sleep'*	**kíxi yu [kíʃì jù]** *'I am sleeping'*
sketa *'run'*	**skéta [sk͡ētâ]** *'I am running'*

Table 28: Verbs in their irrealis (gloss) and imperfective forms

(44) tsí'i ntixi michuni
IPFV\drink\1SG pulque right.now

[55] Whereas in the working orthography, the low tone marking the perfective aspect is not represented, the high tone marking the imperfective is represented with a high tone diacritic above the first vowel in the verb stem. This is also true in cases where the first vowel maintains a *mid* tone level.

[56] Note, as investigation of the tone patterns of the verb lemmas is still in progress, as in many cases, the only observation of certain verbs has been in written sources in which tone is only represented in the imperfective and in certain minimal pairs. Thus, in showing these forms, I use the working orthography in which tone in only marked in the imperfective aspect and in certain minimally distinctive lexical items.

‘I’m drinking pulque right now’

(45) ká’an =kuená sa’an savi
IPFV\speak =3PL.FEM.FORM Mixtepec-Mixtec
‘They (elder women) are speaking Mixtepec-Mixtec’

(46) tsáku vari kúni =ta =yu tanta’a cha koo xu’un
IPFV\cry\1sg because IPFV\want =very =1SG get.married\1SG and NEG.exist money
‘I’m crying because I really want to get married but there’s no money’

(47) tsátsi =na tikoo tsi ntuchi
IPFV\eat =3PL.FORM tamale and bean
‘they’re eating tamales and beans’

2.1.7.2 Perfective

The perfective aspect is typically used for isolated past events. As described by Paster and Beam de Azcona (2004) and Paster (2005), it is usually marked by the verbal prefix ni- (IPA: [nì]) (48), and on verbs with onset pre-nasalized stops and affricates (*nt-, nts-)*, it is marked with low tone on the first vowel of the stem (50). Additionally, though in certain tonal and phonological conditions, can be marked with either: a combination of a pre-nasal n- along with a tone change (low-tone) on the first vowel (49), or where a verb has a root initial mid-tone, the perfective is marked simply by a (low-) rising tone change on the first vowel (51)[57].

Verb (irrealis)	**Imperfective**	**Perfective**
ya’a *‘cross, pass’*	**yá’i [já?ì]** *‘he/she’s crossing’*	**ni-ya’i [nìjà?ì]** *‘he/she crossed’*

[57] I have indeed observed this phenomena of the perfective being marked by only a *low rising* tone on the first vowel of the verb *sketa* ‘run’, which is one of the six verbs presented as evidence of this phenomena. While Paster and Beam de Azcona transcribe it as *low mid*, in my observations it is simply a rise from a *low* starting point, thus I transcribe it simply as *rising.*

ko'o *'drink'*	**tsí'i [t͡zíʔì]** *'I'm drinking'*	**ntsi'i [nt͡zìʔì]** *'I drank'*
ntava *'fly'*	**ntava [nd̲ávà]** *'it is flying, it flies'*	**nta̲va ti [nd̲àvà]** *'it flew'*
sketa *'run'*	**skéta [sk̟ētâ]** *'I am running'*	**ske̲ta [sk̟ětâ]** *'I ran'*

Table 29: Contrasting between verbs in Irrealis, Imperfective and Perfective

(48) ni-ya'a uvi hora
PFV-pass two hour
'two hours passed'

(49) n-tsi'i chikuii tsi luluu kafé
PFV-drink\1SG water and little.little coffee
'I drank water and a very small coffee'

(50) nta̲va taka =ka xini =yu
PFV\fly woodpecker =PTCL.DEM head =1SG
'the woodpecker flew over my head'

(51) ske̲ta nuu chuun takuni
PFV\run\1SG face work yesterday
'I ran to work yesterday'

2.1.7.3 Potential

The potential is generally used for non-actual, and relative future situations, and is marked by the prefix *ku-* [kú] ~ *kun-* [kŭ][58], the nasalized co-variant *kun-* appears where the

[58] There are two variants of form of the future prefix: [ŭ], and [ŋ́]; both of these are usually represented in the orthography as *kun-*. It is noteworthy that the potential prefix is likely derived from what is referred to in other

onset of an irrealis verb stem is a velar stop /k/, or begins with a nasal (both full nasal phones and prenasalized phones). In all instances in the observed data, the use of the prefix with the nasal and it's nasalized vowel variant occurs where verb stems begin with *k.*

Imperfective	Potential
skéta [s͡kētâ] *'I am running'*	**ku-sketa [kús͡kētâ]** *'I will run'*
tsí'i na [tz͡íʔī nà] *'they are drinking'*	**kun-ko'o na [kűkòʔō nà]** *'they will drink'*
skuáchi [sk͡ʷáʧî] *'I am chopping'*	**ku-skuachi [kúsk͡ʷàʧî]** *'I will chop'*
tsá'i [tz͡áʔì] *'he/she is giving'*	**kun-kua'i [kűk̟ʷàʔì]** *'he/she will give'*

Table 30: Contrasting forms between Imperfective and Perfective verbs

(52) ku-sketa xchaan
POT-run\1SG tomorrow
'I will run tomorrow'

(53) i'iin ñachaa ku-ntuta'an =ra kumi chika
each the.men POT-recieve =3SG.MASC four plantain
'.. the men will each receive four plantains' (Beckman and Nieves, 2008b)

(54) kun-ku'u =yu ntuku iki katsi
POT-go =1SG look.for\1SG calabaza eat\1SG
'I will go look for calabaza to eat' (Gómez Hernández, 2007a)

(55) kun-ko'o =kuera ntixi tsini vichi

Mixtec varieties as the potential copula *kúu;* Macaulay (1996) notes that in Chalcatongo Mixtec, the cognate of the aforementioned potential copula (also *kúu*) also has a common variant comprised of just the vowel ú.

POT-drink =3PL.MASC.FORM pulque tonight

'they (elder men) will drink pulque tonight'

2.1.7.4 Imperatives

Imperatives[59] use the irrealis verb form, while informal commands take only the irrealis stem, when giving a command to an elder or otherwise respected person, the formal *=ni* is used as well[60].

Irrealis	Imperfective	Imperative
sketa *'run'*	**sketa ku** *'you are running' (informal)*	**sketa** *'run!' (2sg.inf)*
katsi *'eat'*	**tsátsi ni** *'you (formal) are eating'*	**katsi ni** *'eat!' (formal)*
ka'an *'speak'*	**ka'un** *'you (informal) are speaking'*	**ka'an** *'speak!' (2sg.inf)*

Table 31: Comparison of Irrealis, Imperfective and Imperative verb forms

(56) kaka chinu inkaa =yu

walk[IMP] over.to COP.LOC =1SG

'walk over to me'

(57) Kuntu'u nuu̱

sit[IMP] face\1SG

'sit down in front of me'

(58) katsi =ni

eat[IMP] =2SG.FORM

[59] I am still looking into negative imperatives and thus they will not be discussed here.

[60] While the imperative forms take the structure of lemmas (the irrealis form), at the time of publishing, I am still investigating whether there is a predicatable tone pattern in the imperative verb forms as in some observations the tones appear to be the same but in others it does not. Thus, tones are not included in the examples in Table 31.

‘eat!’ (polite)

(59) kua’a =ni ntaku
give[IMP] =2SG.FORM broom
‘give me the broom’ (polite)

2.1.7.5 Habitual

The habitual aspect is marked by the prefix *ntsi-* (IPA: [nt͡sì]) on the realis stem, and can express past habitual behavior, or past ongoing actions:

(60) che’e ta ntsi-kana =ti
beautiful so HAB-sing =3SG.ANML
‘it was so beautiful when it sang’ (Ramos Hernández, 2007)

(61) ntsi-kuntu’un =ti nta’a in yutu
HAB-sit =3sg.anml branch[hand/arm] ART.DEF.SG tree
‘it was sitting on the branch of a tree’ (Gómez Hernández, 2008b)

(62) tsini =na tu’un yutu ña ntsi-kaa ñuu yo’o
know =3PL.FORM story tree REL HAB-stand town this
‘they know the story of the tree that used to stand in this town’ (Mendoza Santiago, 2009)

(63) ntsi-tsatsi staa
HAB-eat\1SG tortilla
‘I was eating tortillas’

2.1.7.6 Modals

The modal, marked with the prefix *na-* (IPA: [ná]) on the realis stem (where distinct), and can express numerous functions, including: hortatives, intentions, necessity, hypotheticals, possibilities and subjunctive-like moods.

(64) na-ko'on
MOD-go[1PL.INCL]
'let's go!'

(65) kua'a sa'mu na-kiku na-chinchee yo
give[IMP] clothes MOD-sew\1SG MOD-help\1SG 2SG.INF
'Give (me) the clothes, I can help you sew' (Gómez Hernández, 2007b)

(66) na-tsinu sa'mu ra na-ko'on viko
MOD-be.finished clothes CONJ MOD-go[1PL.INCL] party
'when the clothes are done, let's go to the party' (Gómez Hernández, 2007b)

(67) ta ni-ne'e xu'un na-ntakuaan ntivi
when PFV-get\1SG money MOD-buy\1SG egg
'when I get money, I'll buy eggs' (Beckmann and Nieves, 2007)

(68) ntsi-ntu'un nchatu nuu avión =ka na-kitsa-i
HAB-sit\1SG wait\1SG face airplane =PTCL.DEM MOD-arrive-3SG
'I was sitting down, waiting for the airplane to arrive'

(69) takua na-kuu ki'in avión
so.that MOD-be.able catch\1SG plane
'..so that I could catch the plane'

(70) ku-yakua nta'a tatu na- ke'e nuu sta-u
POT-get.dirty hand\1SG if MOD-touch\1SG face tortilla-2SG.INF
'I'll get my hands dirty if I touch your tortilla' (Gómex Hernández, 2007a)

2.1.7.7 Negation

Negation in MIX is primarily expressed with the verbal prefix *ma-* [mà], or the adverbial *kue* [kʷēé] (*or* [kʷě]), which can modify adjectives and verbs. In Chalcatongo Mixtec, Macaulay describes the cognate of *ma-* (which takes the same form) as a *negative mood marker*, whose meaning is the opposite of *na-* (also cognate of the same form).

(71) ma- sana + in-o sa'an =ko
NEG- forget -1PL.INCL language =1PL.INCL
'we must not forget our language' (Beckmann and Nieves, 2008c)

(72) ma- tsíni =na tu'un + yata ñ-oo
NEG- IPFV\know =3PL.GEN legend town-1PL.INCL
'they don't know the legend of our town' (López Santiago, 2008)

(73) A ma- kuu chinche-u yu
Q NEG- be.able help -2SG.INF PRON.1SG
'Can you not help me?' (Gómez Hernández, 2007a)

(74) Kue va'a kíku =ku
NEG well IPFV\sew =2SG.INF
'You're not sewing well' (Gómez Hernández, 2007b)

(75) Kue kúni =yu sachuun
NEG IPFV\want =1SG IPFV\work\1SG
'I don't want to work' (Gómez Hernández, 2007a)

(76) Kue tsitsini =yu michu'ni in libru ka'vi =yu
NEG eat.breakfast =1SG right.now ART.INDEF.SG book read =1SG
'Right now, I'm not eating breakfast, I'm reading a book'

(77) kue nchichi

NEG difficult

'easy'

In Chalcatongo Mixtec (Macaulay, 1996), the prefix *ma-* only occurs with verbs in the potential mood, and in the small number of instances observed in the corpus, it does appear that it mostly occurs with the *irrealis* verb stems[61]. However, in MIX, it can also occur with a verbs in the perfective, which recall take the realis verb stem (for verbs in which they are distinct):

(77) ma- ni- kuu sketa =ti

NEG- PFV-be.able run =3SG.ANML

'it could not run'

(78) ma- ni-ntakuaan =kue nchii nchai

NEG- PFV-buy =1PL.EXCL any food

'We did not buy any food'

(79) ma- n-tsini lochi =ka

NEG- PFV-know vulcher =PTCL.DEM

'the vulture didn't know' (Gómez Hernández, 2008c)

(80) ma- n-tsa' -i mii katsi ña'a =ka

NEG- PFV-allow -3SG PRON.EMPH.3SG [IRREALIS]eat woman =PTCL.DEM

'He didn't allow himself to be eaten by that woman' (Gómez Hernández, 2008d)

Additionally, there is only one observed instance of negation being marked solely by a tone change, which occurs with the potential of the verb 'give', with the first vowel of the stem changing to a low-rising tone contour. However, the standard negation *ma-* can also be used without the tone change. The tone change as a means of negation has been documented in Ayutla Mixtec (Hills, 1990) in which it is the primary means of marking negation in that variety:

[61] Note that some verbs are inherently potential and have only irrealis forms such as *kuu* 'to be able to' and *kuni* 'to want'

Potential (affirmative)	Potential (negative)
kun-kua'a [kűķʷàʔà] *'I will give'*	**kua'a [ķʷǎʔà]** *(or)* **ma-kun-kua'a [mà kűķʷàʔà]** *'I will not give'*

Table 32: Negation of verb kua'a 'to give'

2.1.8 Derivation

MIX, like numerous other Mixtec varieties has a series of derivational prefixes which can be combined with verbs or nouns to create new lexical items, they are described below[62]:

2.1.8.1 Causative

The causative prefix *sa-* is clearly derived from *sa'a* [sáʔā] 'to do, make', and combines to express concepts related to causation or certain kinds of activities, there are also variants which can appear as simply: *s-* or *x-* [ʃ]:

Source	Causative
va'a *'good'*	**sava'a** *'to construct, 'build'*
chuun *'work'*	**sachuun** *'to work'*
na'a *'appear'*	**sna'a** *'to show, teach'*
núu *'come down'*	**xnuu** *'to bring down'*
tutsi *'hurt'*	**stutsi** *'to hurt'*
tsio *'side'*	**satsio** *'to separate'*

[62] Note that at the time of publishing I do not have sufficient evidence for the tones of many of the derivational lexical items. In order to avoid publishing inaccurate transcriptions, and to keep the contents consistent, I do not include IPA transcriptions for these lexical items.

Table 33: Causative verbs and their lexical sources

Note, this causative form can be observed in the name of the primary Mixtepec-Mixtec town (San Juan Mixtepec) *Xnubiko*, also *Snubiko* which can be parsed as: *xnuu* 'bring down from' + *biko 'clouds'*[63].

2.1.8.2 Iterative

The iterative prefix *nta-* combines to express repetition or recommencement; in other varieties of Mixtec, the iterative has been referred to as the *repetitive* (see: Macaulay, 1996: Chalcatongo Mixtec), and takes the form of *na-*:

Source	Iterative
kaka *'walk'*	**ntakaka** *'to walk again*
kana *'to yell, call'*	**ntakana** *'to tell'*
tu'u *'word'*	**ntatu'u** *'to discuss, talk over'*
kuni *'know'*	**ntakuni** *'to recognize'*

Table 34: Iterative verbs and their lexical sources

2.1.8.3 Inchoative

The inchoative has two different prefix forms *ntu-* (from *ntu'u* 'to become') and *ku-* (from *kuu* potential copula) and express some kind of transition[64]:

Source	Iterative
tsaa *'new'*	**ntutsaa** *'to renew'*
va'a *'good'*	**ntuva'a** *'feel better'*
vii *'clean, beautiful'*	**ntuvii** *'to become clean'*

[63] In discussions with several speakers, this componential meaning is still understood in the placename.
[64] Source of information about inchoatives is Mille Nieves (personal communication: July 26, 2017)

yachi *'close'*	**kuyachi** *'to approach'*
kuaa *'blind'*	**kukuaa** *'to go blind'*

Table 35: Inchoative verbs and their lexical sources

2.1.8.4 Combinations of Derivatives

Note, there is at least one observed example of a lexical item which combines the causative and iterative prefixes, note also that the order in which they are attached is: causative *sa-* attaches directly to the lexical base and the iterative *nta-* attaches to the causative. The basis of this is likely that the act of sharpening entails a repeated motion and the end result is that the sharpened object is made dangerous.

Source	Iterative + Causative
xeen *'dangerous'*	**ntasaxeen** *'to sharpen*

Table 36: Causative and iterative combined derivation

Final Notes on Linguistic Description

Once again, the very limited linguistic description presented herein is far from complete and is not at the core of the purpose of this dissertation (which is to present the MIX language resources, corpus, dictionary and annotation methods in the context of the interface between fields of language documentation and digital humanities). The topics and linguistic features presented above, as well as numerous others not included, will be discussed in further detail in future publications with comparative analyses of cognate phenomena as presented in Mixtecan literature. Also, as the encoding of the corpus and unannotated audio materials collected so far are processed, this will enable quantitative corpus analyses. See also Bowers (in press) for an in depth discussion of the semantics of body-part terms in MIX, as well as overviews of the basics of the relativizer and nominalizer *ña* (see Hollenbach, 1995b; for discussion of parallel functions in several cognate Mixtecan languages), and an introduction to the semantics of spatial language.

3. Mixtepec-Mixtec Documentation Project Origins and Methods

As alluded to already, this dissertation presents a project that has made significant contributions to both a LD outcome for the MIX language, as well as to digital humanities/digital lexicography in the way that the TEI has been taken beyond the confines of its traditional usage. However, due to the manner in which this work began (as an informal pursuit of mutual interest), issues pertaining to the availability of data for the language, and logistics in working with collaborators, until the last few years, it was not necessarily conducted in the way a prototypical LD project would be, as it was not originally conceived of as a LD project. Additionally, the technological aspect was developed out of both analytical (linguistic), and practical needs (corpus annotation method, metadata management, etc.) and particularly early on, was conducted in an ad-hoc manner. In this section, I give a brief overview of the origins of the project, its development and then in the following sections I discuss issues stemming from literature on the relevant topics, notably those pertaining to language documentation and digital humanities and how this work approaches key issues.

The project documenting MIX came into being incrementally beginning in a graduate field methods course at San José State University (San José, California) in 2010, while I was pursuing my M.A. Linguistics. The consultant for the semester was Jeremías Salazar, who is from the town of Yucunani[65] in the San Juan Mixtepec district[66] and who moved with his family to Santa Maria, California, which is now a major population center for Mixtepec Mixtecs as well as numerous other Mixtec people (see Reyes Basurto et al., in press). During the field methods course much of the work was focused on issues such as phonetics, phonology and basic information structure. For this work, I with some colleagues took it upon ourselves to manage and collect recordings made in consultation sessions, most of which was recorded using a Sony PCM-D50 Linear PCM Recorder at a rate of 96kHz/24-bit. For annotation, the Praat software system (Boersma and Weenik, 2020) was used. On our own initiative, myself, two colleagues and Jeremías continued consultation work after the course was over[67]. Within the next year

[65] https://www.geonames.org/8880392/yucunani.html
[66] http://www.geonames. org/3518634/san-juan-mixtepec.html
[67] The speaker collaborators have not been paid and have participated in this work on a voluntary basis. The only 'formal' arrangements to participate have been in the form of traveling with the express purpose of working together, both are described below.

Jeremías moved out of state, but we[68] continued to work with his brother Tisu'ma Salazar, who also lived in the San Francisco Bay Area, he became the primary consultant and collaborator for this work since that point. Tisu'ma had previously worked as a language consultant while he was a student at UC Berkeley, which produced several descriptions of phonological and morphological aspects of the language (Paster, 2005, 2010; Paster and Beam de Azcona, 2004, 2005). Upon graduation in 2012, myself and Tisu'ma continued to work together.

Until roughly three years into the work (which was being pursued as a part-time, unofficial endeavor), the main goal and scope of the research was to learn about the linguistic features of the language, particularly: phonetics, phonology, information structure as well as issues related to semantics, mainly metaphor, metonymy and grammaticalization. As I started to become more deeply interested in these issues, it became necessary to try to implement a system in which I could store, annotate and retrieve the all level of linguistic information along with their interfaces. Around this same time, having discussed the goals for our mutual collaboration with my Mixtec colleagues, it became clear that their goals for their role of in our work together were that the output should also be something of use to the community. And this is when the work began to be consciously pursued as a corpus creation and language documentation project, however this was challenging in a number of ways.

Because at the time I had no real training in language documentation, my early approach was to find methods in computational and corpus linguistics to manage, store and process the data. However, as practically every linguistic subfield had their own separate practices for storing, annotating and searching data (though seemingly none were uniformly adopted and none of which were particularly user friendly), there was not any established practices for representation of linguistic interface data structure, ambiguity or sufficient representation of important metadata. Furthermore, of the mostly Python-based approaches such as NLTK (Loper and Bird, 2002) that did exist were not oriented towards producing the kind of user friendly data needed in a language documentation project.

[68] The voluntary consultation sessions were attended by myself and two colleagues from the M.A. Linguistics program at San Jose State until 2012 when we all graduated. After this point only I, along with the speaker/collaborator continued the work. See (Corpuz, 2012) for an output from the collaborative work by my colleague Larry Corpuz Jr.

Additionally, as is common in dealing with indigenous and under-resourced languages, variation (phonetic, orthographic and other) was ubiquitous in the dataset. While it was important for me to keep variation that may be relevant, given that most computational linguistic toolkits and practices were developed using major (western) world languages as the basis (namely English, German, French and Spanish), thus there was not proper support for languages with certain features such as tone, or nascent orthographic systems. Moreover, at the time, there was not even proper Unicode support for characters with diacritics (which is needed in Mixtec). Thus, there was an extreme gap in the ability to manage and use the data within the given systems.

Around the same time, it was becoming increasingly necessary to go beyond plain text/tab separated corpus that I was using to store the vocabulary output and that a more dynamic data structure was needed, which lead me to TEI which had established modules and guidelines for structured encoding of both text corpora and dictionaries. In 2013 I began compiling a TEI dictionary for storage of the vocabulary as well as etymological information (see section 7.5)[69]. While it was clear that the TEI and XML technology was the best choice for my particular needs, as I got deeper into the work in creating a dictionary, it became clear that there were numerous areas in which it was not sufficiently developed to accommodate the kinds of details and features I wanted to include, particularly in the areas of applying true linguistic analysis to etymology[70], and other features that are particularly pertinent to working with an indigenous under-resourced language (see chapter 7 for details). These gaps are attributable to the facts that: the TEI, particularly the Dictionary module has mostly been designed for, and by lexicographers as opposed to linguists, and that the vast majority of projects adopting it were for European languages (Bowers and Romary, 2018a).

[69] https://github.com/iljackb/Mixtepec_Mixtec/blob/master/MIX-Lexicon-TEI-Dict.xml
[70] As a major focus of the linguistic investigation into MIX was centered around cognitive factors involved in the etymology of body-part terms, such as metaphor and metonymy, amongst other key processes, the need to establish a more stable and expressive means to encode this information in TEI was the motivation for (Bowers and Romary, 2016).

Additionally, due to the fact that I both wanted to create as large of a collection as possible in the output, and that I needed to increase my own knowledge of the language in order to carry out unsupervised translation, annotation and glossing, there was a need to accumulate more linguistic data. Thus, with permission of the publisher, TEI encoded versions of the SIL booklets (originally in PDF file form) were created and added to the annotated corpus[71]. Along with the transcriptions for original recordings, these documents from SIL form the majority of the text sources in this project's corpus, and at present, they actually make up the vast majority of published content written in the language.

The fact that MIX is an under-resourced language and has no prior linguistic analysis beyond the phonological system (c.f. Pike and Ibach, 1978; Paster, 2005, 2010; Paster, and Beam de Azcona, 2004, 2005), corpora, or even a firmly established orthographic system meant that there could be no means of translating or annotating the corpus other than manually. As is common in dealing with such languages in which there is an extremely limited number of potential participants (especially given that this work was not funded), there were very few options for approaches to annotating the corpus (see Thieberger et al., 2016). Thus, the approach taken with the text corpus has been to first create the translations, then, pending the availability of one of the two collaborators, go through and correct and complete the translations for each document as needed. More in-depth annotations are then added afterwards.

As a result of mostly working with only one speaker at a given time outside of the speech community, there was little opportunity to collect much spoken language in natural contexts. Thus, throughout the first several years of the project, I would often focus on collecting vocabulary content mostly using translation elicitation[72]. While this was of course not best practice in LD (see Himmelmann, 1998; Woodbury, 2003) it allowed for the collection of much of the most essential vocabulary, and me to both study the particular phenomena I was interested

[71]Original source materials are from: http://mexico.sil.org/resources/search/code/mix?sort_order=DESC&sort_by=field_reap_sortdate and the TEI encoded and annotated contents are available from: https://github.com/iljackb/Mixtepec_Mixtec/tree/master/SIL_docs

[72] Though most of the vocabulary were obtained through elicitation, in the study of spatial configurations, several sets of images were created for the purpose.

in. A few exceptions to this were in cases where speaker collaborators would occasionally make recordings of conversations that took place in their daily life or went on trips to the region[73].

The project continued when I moved to Paris (2014-2015) and then Vienna (2015-present) for professional reasons. Over the course of this time, the issues that I have encountered in continuing this work with my Mixtec colleagues living in the USA created a whole array of unique factors and constraints to the manner in which this work has been carried out until the present, though thanks to mobile messaging service, social media and teleconferencing such as Skype, Google Hangouts, etc. semi-regular communication has been possible.

In 2017 with funding from DARIAH Tisu'ma was able to come to Vienna for two weeks to assist with various aspects of the work. Additionally, in summer 2019, with funding from the EPHE and Inria I was able to finally spend three weeks the region[74] accompanied by both long-term project collaborators Jeremías and Tisu'ma Salazar where we stayed with their parents in the city of Santiago Juxtlahuaca. All audio contents obtained in this latter trip were recorded with a Tascam DR-05X Linear PCM Recorder at a rate of 96kHz/24-bit[75]. All of the recordings created and full (TEI) metadata for the contents created from these trips and the rest of the project are available on our Dataverse repository under the name "Mixtepec Mixtec Lexical Resources"[76] (Bowers, Salazar, and Salazar, 2019), this will be discussed in more depth in later sections.

In order to build a basis for a maximally comprehensive lexicographic dataset, the work being done is not limited to simple documentation and treatment of MIX and resources from related, and historical Mixtec varieties are being integrated into the project, particularly in the dictionary component (see section 7 discussing the TEI Dictionary). Additionally, as described in Bowers, Khemakhem, and Romary (2019), using the OCR toolkit GROBID dictionaries

[73] In content collected from recordings made by speakers, informed consent to record various conversations was obtained for most (though unfortunately not all) recordings made, though due to the low quality of the recording device used most of these recordings have not been usable.
[74] First in the San Francisco Bay Area in California (USA) and then since 2015 in Vienna (Austria).
[75] As will be discussed in following sections, metadata records for all media files created specifies the specific recording equipment, elicitation methodology and several other key factors.
[76] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/BF2VNK

(Khemakhem et al., 2017) a TEI dictionary from a historical Mixtec dataset of Classical Mixtec[77] originally published by the Dominican fray Francisco Alvarado in the year 1593 has been created and added to the project output. Integrating such resources provides a rich resource for comparative historical data that not only enhance the quality of the Mixtepec-Mixtec dictionary, but it can also be re-used by those working with any other Mixtec variety.

4. On the Intersections and Divergences of Language Documentation, Description, Digital Humanities and Corpus Linguistics

As this work exists at the interface of multiple subfields: digital humanities/digital lexicography, language documentation, corpus linguistics, amongst others, there is a wide variety of literature from these various fields relevant to different aspects of this work, but very little that covers every key aspect. A fundamental necessity of any given LD project is to provide a documented collection of primary language data, along with lexical information from potentially any level of language (i.e. phonetics/phonology, morpho-syntax, semantics, lexicon or dictionary information, etc.), often with transcriptions and annotations (e.g. interlinear glossed texts). Additionally, imperative, is the need to: organize the data, provide access, publish, and analyze information, i.e. to ensure maximum re-usability as well as potentially empirical verification via best practices and ideally, the use of data standards (Bird and Simons, 2003b; Thieberger, 2010, 2012, 2014; Gawne and Berez-Kroeker, 2018). These issues are of course also equally relevant to any multi-faceted linguistic, lexicographic and/or corpus linguistics projects (between which the distinction may, in some cases be somewhat arbitrary) (see Cox (2011) for in-depth discussion of the overlap and divergences between corpus linguistics and LD). This broad scope presents highly complex technological and logistic challenges in terms of: software, data format, markup and workflow.

In this section I discuss key issues, principles and theoretical foundations from key literature pertaining to these various fields which are at the core of this work, namely: DH, LD, language description, the intersection of DH and LD, data design and management, best practices and ethical issues in LD, as well as issues in working with under-resourced languages.

[77] Classical Mixtec is also referred to as “Colonial Mixtec”.

4.1 On Language Documentation and Digital Humanities

Digital Humanities is peculiar in that it is not actually one single field, rather it is a means of working with, encoding, annotating and presenting work done in any number of topics in humanities (e.g. History, Literature, Linguistics, Lexicography, etc.) which, traditionally carry out their work in separate academic departments, buildings, conferences and journals. DH has evolved into a distinct, yet multi-disciplinary field largely because within the traditional confines of these separate academic cultures and practices in the various fields of Humanities, the use of technological tools was not institutionally prioritized, either academically or in their respective institutional programs and departments.

The digitization of legacy data, and the creation of new *born-digital* data is crucial for preservation and re-use and facilitates exponentially faster search, retrieval and analysis of source material which benefits researchers and their various potential audiences alike. As is common in Humanities, content from one source can be relevant to multiple fields, e.g. historical literature, epigraphy and numismatics, while all specialized studies on their own, also are major primary sources of historical language data. Thus, their contents and analyses, as well as their provenance, etc. are all potentially relevant to historical linguists, as well as potentially historians, anthropologists, amongst others. Given these facts, there has been a need for those in these fields to seek to develop and exchange methods and knowledge from the various technologies outside of their own fields and for the development of data standards which permit the exchange of digital dataset and analyses.

Likewise, LD is fundamentally cross disciplinary, as according to Himmelmann (1998) guidelines for LD are necessarily much broader than that of possible sub-disciplines of linguistic description/analysis, as they can concern any of the following:

- sociological and anthropological approaches to language (variationist sociolinguistics, conversation analysis, linguistic and cognitive anthropology, language contact, etc.);
- "hardcore" linguistics (theoretical, comparative, descriptive);
- discourse analysis, spoken language research, rhetoric;
- language acquisition;

- phonetics;
- ethics, language rights, and language planning;
- field methods;
- oral literature and oral history;
- corpus linguistics;
- educational linguistics;

To this, Austin (2013) adds:

- ethnography
- psychology
- library science
- archiving
- media- and recording arts
- pedagogy

Furthermore, Himmelmann (1998) states that the major theoretical challenge for linguists in LD is synthesizing a coherent framework from all of the disciplines listed above, which is also at core of Digital Humanities (see Penfield (2014) for an in-depth overview of the unique issues facing inter-disciplinary studies in academia). Though lexicography and linguistics projects[78] are not rare, it is quite rare to hear about language documentation[79] in the context of DH. Moreover, while those working on language documentation rarely consider their work to be in the digital humanities field, this is indeed changing with current trend towards LD aligning linguistic methods with aims and approaches central to DH, namely in the focus on re-usability, compatibility and extensibility, as well as in producing replicable research and research data (Bird and Simons, 2003b; Thieberger, 2010, 2012, 2014; Gawne and Berez-Kroeker, 2018). Bird and Simons (2003b) is a seminal paper cited in both DH and LD contexts discussing key issues to language documentation and description pertaining to: content, format, discovery, access,

[78] it is actually rarer to actually hear a DH project described as "linguistics" as that field generally describes itself as "computational linguistics" or "corpus linguistics" when involving digital methods.
[79] Though, the work carried out in many European dialectology projects is very similar in many way to language documentation (see Bowers and Stöckle (2018) for an example of work done in the DH domain on Bavarian varieties in Austria as part of the long-term cultural legacy project Datenbank der bairischen Mundarten in Österreich)

citation, preservation and rights; while the target audience for this work and subject matter was those doing language documentation and description, many of the principles and issues in this work are also cannon in the practice of DH in general.

There have been two particularly influential projects in the domain of technology assisted language documentation which clearly reflect the intrinsic connection between DH and LD: DOBES (Dokumentation bedrohter Sprachen) project (2000-2011)[80], which also created an archive of endangered languages, and E-MELD (Electronic Metastructure for Endangered Languages Documentation) (Boynton et al., 2006)[81]. These projects sought to identify key issues, and make recommendations towards the establishment of best practices on key issues relevant to both LD, DH, as well as corpus linguistics in order to both ease the process and increase the sustainability and interoperability of the output. Topics covered include: data and archival formats, metadata, annotation, analysis, standards, tools, workflow and management[82].

4.2 On Language Description vs Language Documentation

A key point to clarify is the separation between collection, description and analysis of *primary data* in which the goal of documentation is the recording and production of records of natural spoken language and linguistic description is simply a byproduct (Himmelmann, 1998, 2006; Austin, 2006; Woodbury, 2003; Mous, 2007; Good, 2011). Most fundamentally, the main goal of language documentation is data collection, with representation and diffusion with the production of grammars, dictionaries, new material creation, as well as annotation and analyses being secondary.

Given that the target audience of a language documentation project is potentially much more diverse including (in particular) community members, researchers from other fields as well, anthropologists, ethnologists, etc., a major challenge to those working in a LD context is to

[80] http://dobes.mpi.nl/ (accessed 2019/12/31)
[81] http://emeld.org/ (accessed 2019/12/31)
[82] Other influential projects in the development of best practices and of language documentation as a distinct field were the Endangered Languages Documentation Programme (ELDP) (2002-present) (https://www.eldp.net/); and Documenting Endangered Languages (DEL) interagency initiative of the United States National Science Foundation and the National Endowment of the Humanities (https://www.nsf.gov/pubs/2005/nsf05590/nsf05590.htm) (2005-2020)

develop a coherent framework or set of principles for capturing and representing the content relevant to this variety of disciplines, but providing it in a way that does not preclude one field or purpose over the others.

> *"A clear separation between documentation and description will ensure that the collection and presentation of primary data receive the theoretical and practical attention they deserve." (Himmelmann, 1998, p.164)*

Perhaps the most key distinction between language documentation and description is the role of data in combination with the goals and motivations for the work: while as described above, the goal for the former is the creation of well documented media and other primary language resources for preservation and re-use, for the latter the main goal is the production of grammars analysis and (in some cases) dictionaries with the primary target audience being linguists with the purpose of supporting some linguistic analysis (Himmelmann, 1998, 2006; Woodbury, 2003; Austin, 2006; Austin and Grenoble, 2007).

This harsh distinction between the two was challenged by Nathan and Austin (2004); Austin and Grenoble (2007) who argue that the creation of maximally usable, quality comprehensive documentation (in the form of multiple 'entry points' such as transcription, translation and annotation) is necessarily reliant upon linguistic analysis and that linguistic analysis is inherently needed to discover and evaluate the lexical contents of a documentation collection. Himmelmann (2006, 2012) himself states that despite the fact that language documentation and description can be separated fairly clearly on the grounds of methodology, and epistemology doesn't necessarily mean they can, or must actually be separated in practice. For instance, linguistic analysis is necessary in identifying and determining where crucial speech genres, lexical forms, paradigms, sentence constructions, etc. is already contained and where it is missing. Where analysis is necessary for such tasks, it is necessary to document the features and the basis for their identification and treatment, such as in segmenting linguistic spans and features that may affect basic meaning, etc.; the documentation of these issues amounts to linguistic description and it has implications to both discovery, and potential re-usage

While the distinction of language documentation and description is a difference in focus between primary data (e.g. audio/video recordings, transcriptions, etc.), and analytical output and resources (e.g. dictionaries, grammars and analyses), in most cases it is likely that a project doing documentation will also perform some form of description (Good, 2011). As discussed in section 3, this is indeed true in this work, which started as a linguistic endeavor in which the desire was to learn about the language and the production of a dictionary, and a digital corpus was originally designed as a means to that end, and it was only later that it was consciously pursued as a language documentation, though with the goal of producing resources that can be used by the speaker community.

4.3 Language Resources and Data

LD resources are de-facto corpora of under- resourced and/or researched languages; this necessarily means that they differ from corpora of major languages in terms of their purpose, production, content, sources and size (Mosel, in press). The specifics of each aspect of these differences will of course differ according to the given situation and history of the language in question, but whereas a major language likely has a full array of pre-existing spoken and written resources to choose from in any number of domains and registers, a much larger pool of speakers, and a naturally increasing body of sources, a LD project may literally have no pre-existing resources of any mode or genre. Thus, the sources of data may be sporadic and from an irregular diverse pool of sources, which could potentially even encompass the entirety of existing LR for a given language.

In a typical LD project, the main source of content will likely be audio or video files recorded of native speakers. These files are then transcribed in a time aligned format using some software such as Praat, ELAN (Brugman and Russel, 2004), or EXMARALDA (Schmidt and Wörner, 2009) (see section 4.4.2 for further discussion). In addition to audio or video, there may be texts integrated into a corpus, either original writings from speakers, or pre-existing sources of any genre available (see section 6 for examples and discussion in this project).

Another major difference in purpose is that while major language corpora are ubiquitously used for linguistic and possible other levels of research (and/or perhaps training of

technological software), LD corpora are likely needed for a wide variety of purposes, including cultural and linguistic heritage, education materials, as well as research. Additionally, in cases where the project is based on creating collections (corpus creation) of LR, especially in the case of indigenous, or threatened status, major challenges are: a) the creation of original content (consultation sessions, etc.); b) accumulation of resources from external sources in pursuit of corpus creation; c) the integration of these resources into a common data formats so that they can be searched from a common query interface and eventually output in a presentation format for community oriented output. Key to meeting these challenges from the data perspective are the issues of interoperability, interchange, standards and tools.

Finally, of the highest importance, is the issue of creating and managing metadata, both in the near and long-term view for archival and preservation, as well as for issues related to research, reuse, analysis, etc. The following subsection presents an overview of the role, recommendations and practice in metadata.

4.3.2 Metadata

Metadata, or 'data about data' (Nathan and Austin, 2004) is of course a central aspect of any language documentation output and is particularly important in work with resources for endangered or under-resourced languages. The need for the creation of records of this information in language documentation is a key, and (to various degrees of specificity) required component in language archival, discovery as well as data management. Quality metadata records is essential in enabling resource discovery for a diverse potential audience, as well as to validate the quality of the data and record important demographic and methodological, bibliographic details (Aristar-Dry and Simons, 2006; Himmelmann, 2006). Metadata records should minimally include: date, place of occasion, type of speech event, participants, language(s) used, access rights, as well as the properties of the data files described (Aristar-Dry and Simons, 2006; Himmelmann, 2006). Additionally, recording factors pertaining to the creation of the given linguistic content (such as circumstances and methods of elicitation), is important for potential evaluation of the quality of the content (Nathan and Austin, 2004; Austin, 2006; Himmelmann, 2006; Nathan, 2010).

The metadata records created for language documentation resources are fundamentally connected to, and their adoption has been driven by, the archives in which they are deposited which enables organized search and discovery, particularly in a digital, online environment (Simons and Bird, 2003a,b).

On the macro level, Good (2011) summarizes the most basic documentary contents that require metadata as the categories: *project*[83], *corpus*, *session*, *resource* and *people*. A *resource* (audio, video or transcription) is created during a *session* (which is of course, an event), in many if not most cases, more than one resource may be created in a single session. *People* (speakers, researchers, etc.) can be declared on the level of the *project,* but will of course be referenced throughout the documentation of the individual sessions. Collections of sessions may be the main components of a *corpus,* and a collection of corpora may be joined as the components of a *project*. However, as noted by Good, the concept of *corpus* and *project* are subjective and may be employed differently by different teams.

Nathan and Austin (2004) make the distinction between "thick" and "thin" metadata. *Thin metadata* according to the authors is metadata that is focused on resource discovery, and is akin to the type and depth of information used in library cataloguing practice, in which basic information provided by publishers is used in such as: title, provenance, author, publisher, date, ISBN. *Thick* metadata is the core language and linguistic content such as transcriptions, annotations and analysis which are necessitated by the nature of audio and video data, for which thick metadata, such as time-aligned annotations are needed in order to provide a more significant basis via which the content can be discovered as without text annotations, the core content resources are inaccessible through any other means than a secondary user having to listen to it themselves.

Austin (2013) extends proposals to Woodbury (2011) and calls for a theory of language meta-documentation or "Meta-documentary Linguistics" the focus of which would be to expand documentary models, processes and practices by drawing upon practices common in other

[83] However Good also states that 'project' and 'corpus' is more of a subjective notion and may be more likely to be varied in how they are referenced in the context and practices of specific individuals and teams carrying out the given documentary work.

disciplines such as social and cultural anthropology, archaeology, archival and museum studies, as well as issues relevant to interpreting legacy documentation and materials. With regard to the meta-documentation of researchers, Austin argues for the documentation of the following:

- identification of project stakeholders and their roles;
- attitudes and ideologies of consultants and their community with regard to their language, the documenter and project;
- the relationships between researchers, project participants and the wider community;
- goals and methodology of the project, including research methods, tools;
- corpus theorization (see Woodbury, 2011);
- theoretical assumptions underlying annotation and translation (glossing and annotation practices);
- issues related to potential for the project and output to contribute to revitalization;
- background knowledge and experience, training of the researcher and main consultants;
- the conditions under which the project was carried out;

Thus, the concept of thick *metadata* as advocated for by Nathan and Austin (2004) and Austin (2013) is a proposal to reconsider the scope of metadata beyond simply superficial details akin to what might find in a library catalogue to a more comprehensive account of potentially any factor that might be relevant to the resources created by the project.

4.3.3 On Data Formats: Files and Markup

The second essential component to discuss is that of file formatting. According to best practices as per Aristar-Dry and Simons (2006), any digital language resource should be: preservable, intelligible, and interoperable (the specifics of which of course depend on the data type), and these are addressed in this section.

In terms of preservation, file formats that are ‘lossless’ are essential, i.e., the file format should not lose any contents through compression. Additionally, file formats should be ‘open’, i.e. they should not be proprietary, which means that access to their contents are dependent on a particular vendor's software. Examples of open and lossless file formats are as follows:

- Audio: .wav, .aiff
- Images: .tiff
- Video: .avi
- Text: .txt, .xml, .html

Additionally, the format should be transparent (Aristar-Dry and Simons, 2006), which means that the format doesn't require any special knowledge or algorithm to read or interpret and that there is a one to one correspondence between numerical values and information represented. For example, plain text (.txt) documents have a 1-to-1 correspondence between numbers and characters and audio files using Pulse-code modulation (PCM) (e.g. .wav and .aiff) has a 1-to-1 correspondence between the numbers and amplitude of the sound wav (Aristar-Dry and Simons, 2006). While each for can be read by any program that handles text or audio respectively, .zip and .mp3 files require implementation of complex algorithms to restore the original correspondences of the files and thus to read and access (ibid).

For annotated text documents and lexica in LD, the two fundamental recommendations that have been widely accepted and established are the use of Unicode[84] and XML[85] (Bird and Simmons, 2003b; Austin, 2006; Good, 2011). According to Good (2011), XML has several attributes that make it particularly well suited for LD. First of all, it can be expressed in plain text (i.e. the element values, or even the attribute values if extracted), and doesn't use any special characters beyond that found in plain text files and can make use of widely-adopted open format which facilitates archiving; while designed as a machine-readable markup, (depending on the implemented vocabulary or tag schema) the tags have semantic value and are themselves human readable. This means that even in the case that a dataset is not documented (as long as the specific implementation is not arbitrary or inconsistent), the structural logic still will likely be comprehensible to humans using simple text editors. The fact that XML is by itself (somewhat) self-documenting makes it particularly conducive to long-term preservation since even in the absence of documentation and/or metadata, its intrinsic structure ensures a certain degree of

[84] https://home.unicode.org/
[85] https://www.w3.org/XML/

interpretability. Finally, the flexibility of XML enables the expression and encoding of a wide array of data types; XML has been widely adopted in commercial and academic contexts, as a result of which there are many tools for processing and manipulating XML, which he asserts makes the format and its existing infrastructure ideal for the creation of resources specific to LD.

While a data markup format such as XML is perhaps on the surface intimidating to non-experts, and would not elicit much positive feedback if shown directly to many community members of a LD project, in combination with numerous different open software, stylesheets and schemas (e.g. CSS, XSLT and XQuery) it can be fairly easily be rendered for human consumption (i.e. in a *presentation format*), the data can be extracted, and it can be transformed into other data formats such as HTML, PDF and more. Note however, that while XML has the aforementioned benefits to working in LD contexts, it is not a standard on its own and its optimal implementation to a lexicographical or corpus project depends on the establishment of a schema or the adoption of existing markup vocabularies in the form of standards, which is described in the following section.

4.4 Standards

In addition to choosing the right file formatting, data standards are an increasingly essential component each with regard to: metadata, corpus markup, descriptive resources such as dictionaries, corpus annotation, and grammatical descriptions and inventories. According to Romary (2011), data standardization should serve to stabilize knowledge contained in the data as well as be structured in such a way as to prevent future roadblocks from arising in working with the standardized data in the future. Furthermore, the use of data standards facilitates data interchange between users and tools, and allows users to take advantage of the fact they are already documented, thus saving the user time in having to design and describe their own markup system (Romary, 2011). This discussion begins with metadata standards and moves on to corpus markup, descriptive resources, corpus annotation, and grammatical descriptions.

4.4.1 Metadata Standards

Metadata standards in LD as in corpus linguistics serve several purposes, the primary of which is that they promote discovery and systematic access to resources within archives,

comparison of resources across languages and collections, as well as relationships between resources. There are two primary competing standards for metadata in language documentation which are widely adopted and recommended, these are: OLAC[86] (Open Language Archive Community) (Simons and Bird, 2001, 2003a,b, 2008; Bird and Simons, 2003a) and IMDI (ISLE Meta Data Initiative)[87]. Each of these standards primary serialization is XML, but they can be extended to RDF as well.

A metadata record can be either embedded within a resource described (particularly if the resource is XML) or it can be stored separately, which of course will always be the case with metadata describing the contents of media files. Additionally, though not a commonly established or demanded standard in the LD community, the TEI (also XML-based) shares the full array of capacities of the two leading standards, with the main difference being that whereas the IMDI and the OLAC are strictly metadata standards, the TEI is a standard that covers any and every aspect of corpus linguistics, lexicography, etc. and thus it will be discussed in the contexts of the standards for each subtopic described above.

4.4.1.1 OLAC

The Open Language Archives Community (OLAC), which originated from Open Archive Initiative (OAI) is an extension of the Dublin Core Metadata Initiative (DCMI)[88] (see also Bird and Simons, 2003a) and is a major metadata scheme used in language archives. The adoption of this scheme is required for any language resources to be registered within the OLAC infrastructure which when searched, links users to the external repositories using its metadata standard. The OLAC system has flat (nonhierarchical XML) structure comprised of all fifteen elements from the Dublin Core vocabulary: *Title, Subject and Keywords, Description, Resource Type, Source, Relation, Coverage, Creator, Publisher, Contributor, Rights Management, Date, Format, Resource Identifier, Language*; plus, several qualifier categories: *Subject.language,* (Resource) *Type.functionality, Type.linguistic, Format.cpu, Format.encoding, Format.markup,*

[86] http://www.language-archives.org/
[87] https://tla.mpi.nl/wp-content/uploads/2012/06/IMDI_Catalogue_3.0.0.pdf
[88] https://dublincore.org/

Format.os, Format.sourcecode. OLAC identifies language in the various elements using ISO 639 standards (parts 1,2,3).

Additionally, the OLAC schema can be extended according to the following categories[89]: Discourse Types, Linguistic Field, Linguistic Data Type and Participant Roles. Note that each extended category has a set of potential values. Finally, OLAC can be extended with external vocabularies according to different project needs or subcommunity.

4.4.1.2 IMDI

The ISLE Metadata Initiative (IMDI), developed by the Max Planck Institute (MPI) is designed to provide interoperability for browsable and searchable corpora and descriptions of language resources. IMDI is the required metadata schema for the DOBES archive. IMDI distinguishes between two categories of metadata: *session* and *catalogue.* In contrast to OLAC, the IMDI (as will be shown with TEI below) has hierarchical structures which support a more complex encoding of the information and is thus is more in-depth than OLAC vocabulary, allowing for a wide array of categories of information that are integral to creating a well-documented LD resource.

The session data describes the primary data from a 'session', or specifically the occasion in which the language resource was created, including written texts which could include the circumstances and conditions of the utterance event, administrative information pertaining to, and the content of the event (IMDI, 2003). While given that the IMDI is a fairly deeply structured standard, this description does not cover every detail about it herein, thus for more information, see (IMDI, 2003), nonetheless, the major element groups for sessions are: *Session, Project, Content, Actors, Resources, Source, References.*

IMDI *Catalogue* data is used to catalogue the resources which are made to describe such content as "published corpora" which are not appropriately described on the level of *Session* metadata (IMDI, 2009). While again, this description only covers the upper nodes of the schema,

[89] http://www.language-archives.org/REC/olac-extensions.html

the major elements of the *Catalogue* data schema are as follows: *Name, Title, Id, Description, Document Languages, Subject Languages, Location, Content type, Format, Quality, Smallest Annotation Unit, Applications, Data, Project, Publisher, Author, Size, Distribution Form, Access, Pricing, Contact Person, Reference Link, Metadata Link, Publications, Keys.*

Though IMDI can be managed with any XML editor such as Oxygen XML editor, the Arbil[90] tool from MPI TLA (The Language Archive) is an application that is specifically dedicated to creating and organizing metadata records for archiving (Withers, 2012). As of Fall 2020, the SOAS has made significant progress in a toolkit that shows promise in filling the gaps left by ARBIL with the program *lameta*[91], which allows users to prepare data for archiving by creating IMDI records and associates them with project files.

4.4.1.3 TEI

TEI capacity for metadata documentation has not previously discussed in the context of LD, and while the "Text" in the Text Encoding Initiative would indicate that the standard is limited to texts, the TEI vocabulary is very much able to encode the same pieces of information describing any kind of language resource (media or text) as the OLAC and IMDI. In the TEI, the metadata is specified in the various sub-elements of the <teiHeader>, while the actual content can be encoded within the <body> section. A major difference between the TEI, OLAC, and IMDI is that the TEI is both a highly specialized markup vocabulary for metadata as well as for the content of a text document itself.

The TEI header, as is the entire XML-based structure, is organized hierarchically with five principal elements within which sub-elements encode the key areas of language resource metadata discussed above. These elements are described below as per (TEI Guidelines Ch2 Organization of the TEI Header)[92]. Additionally, the contents described below are also part of the ISO 24624:2016 standard: *Transcription of Spoken Language* developed as part of TC 37 for language resource management.

[90] https://tla.mpi.nl/tools/tla-tools/arbil/ (Note that ARBIL is now classified as a 'legacy software' and there is no longer any active support or further development)
[91] https://github.com/onset/lameta
[92] https://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html

<fileDesc> (file description) "contains a full bibliographic description of an electronic file"[93]: within the file description, there are three mandatory sub-element blocks <titleDesc>, <publicationStmt> and <sourceDesc>. Within these sections, several key areas of information from the other metadata standards described above is encoded, particularly:

- <titleStmt> (title statement) within which the title <title> and information pertaining to the person(s) involved are specified, e.g. <respStmt> (responsibility statement) which assigns a specific role in a file along with a person and/or organization. Herein other information about sponsor(s) or funder(s) can also be included in <sponsor> and <funder>;
- <publicationStmt> (publication statement) within which information concerning publication and distribution of the files or texts is placed. Within the publication statement bibliographic information can be specified with (<biblFull> or <fileDesc>). The individuals or group(s) who hold the rights to publish and distribute the materials can be specified within <authority> and the license can be stated with <availability>;
- <sourceDesc> (source description) defines the source of the text or file contents, and can be used to specify the source file as well as declare whether the encoded text is born digital. Within <sourceDesc> a link to an external media or text file can be declared with <media> or <ptr> (in the case of a text file). <recordingStmt> (recording statement) can be used to record key details relevant to the creation of the material, particularly for audio and video files; some of the content in this element is akin (though not exactly parallel) to that in IMDI *Session.* Key elements in <recordingStmt> are as follows:

    - <respStmt> used for specifying the participants in the recording (speakers, researchers, etc.);
    - <equipment> specifies the mechanisms used for the recording;
    - <location> specifies the location of the recording event, within which several sub-elements are used for more granular geographic details, e.g. <placeName>,

[93] Note however that despite this definition, the file described does not actually need to be electronic as it could be describing an analogue resource as well.

<country>, <region>, amongst others[94] all of these are akin to IMDI elements by similar names, e.g. (*Continent, Country, Region*);

- <date> records a date of the recording event[95] (has identically named parallels in both IMDI and OLAC);
- It is also within this section that the means or context of the communicative event (likely akin to categories in OLAC *Content* sub-categories: *CommunicationContext.Interactivity and CommunicationContext.Planning Type*) recorded can be specified though there is not an exclusively designed element for this. See section 6.3.4.3 for explanation of how this was done in this project;

<encodingDesc> (encoding description) describes the relationship between a digital text and the potential source(s). It also can be used to document editorial content pertaining to a number of other details, of particular relevance are the following sub-elements:

- <projectDesc> (project description), used to define the aim, purpose, and/or methodology of the project itself. This element is an exact parallel to IMDI *Project.Description;*
- <tagsDecl>[96] (tags declaration) describes detailed information about the tagging. This element is related to the DCMI *Smallest Annotation Unit* (though <tagsDecl> is broader in scope than the latter);
- <classDecl> (class declaration) contains taxonomies defining concepts used in the text;

<profileDesc> (text-profile description) provides descriptions about non-bibliographic aspects of a text or other type of file where applicable, the language(s) used, and the situation in which it was produced[97]. Elements relevant to the LD context include:

[94] <location> is not limited to the context of <recordingStmt>, and for the full list of child elements see: https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-location.html

[95] <date> is not limited to the context of <recordingStmt>, and for the full list of contexts see: https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-date.html

[96] As an alternative to the use of <tagsDecl>, it is possible, and possibly more common to specify this information using the TEI ODD (One Document Does it All) to declare the specific usage of tags in a document: https://wiki.tei-c.org/index.php/ODD

[97] It would be possible to include some of the information pertaining to the actual recording event which are also contained by <recordingStmt> within <sourceDesc> in <profileDesc>. For more specifics about <profileDesc> see https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-profileDesc.html

- <langUsage> with the attribute @xml:lang encodes the language, <language> specifies the language in the one or more of the project's working languages (these elements have parallels in both OLAC *Language*, and in the various sub-categories of IMDI *Content.Languages.Language)*;
- <textDesc> (text description) describes the text with regard to its parameters. This element block contains information used to classify key areas of a text such as its factuality (encoded with element <factuality>), domain (encoded with element <domain>), whether it's interactive (<interaction>), the purpose of the text (<purpose>) amongst others[98];
- <settingDesc> (setting description) defines the context in which the language interaction takes place, this could be used to specify where a linguistic research event took place or the setting of a fictional work as well;
- <particDesc> (participant description) documents the people who have participated in the creation of a given file or project in general along with a number of other important details relevant to LD metadata. The various elements in this section cover areas associated with the sub-categories in IMDI *Collector* and *Participants*. The <person> element[99] can specify the name(s) (including variants thereof) of participants, their role(s) in the project, residence (<residence>), which can include specific or ranges of dates (e.g. <residence notAfter="1994">), languages spoken (<langKnowledge>), education (<education>), and a large number of other elements for other important details;

<revisionDesc> (revision description) contains a summary or record of the revisions (versioning) made to the file and <xenoData> (non-TEI metadata) is a generic container element where non-TEI metadata or other content can be placed.

[98] For a full array of child elements and content model of <textDesc> see: https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-textDesc.html

[99] In <particDesc>, a <person> can occur as a direct child or within a <listPerson> element. <listPerson> can be typed to specify the particular type of participants. For a full array of child elements and content model of <person> see: https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-person.html

4.4.1.4 AILLA

It should be noted that not all archives use IMDI or OLAC, AILLA (The Archive of the Indigenous Languages of Latin America)[100] uses its own metadata inventory (though the categories are quite analogous to those in OLAC in particular[101]. Additionally, unlike the previous three metadata vocabularies, the AILLA metadata is not based on XML and is a simple Excel spreadsheet which contains separate collections of controlled vocabulary data fields for the categories of:

- Collection (22 subcategories)
- Languages (12 subcategories)
- Resources (50 subcategories)
- Media Files (30 subcategories)
- Contributors (24 categories)
- Terms (15 subcategories with multiple potential values)

4.4.1.5 CMDI

Another metadata initiative that should be mentioned but will not be discussed in detail is the Component MetaData Infrastructure (CMDI)[102] developed by and used by CLARIN[103]. The purpose of the CMDI is to provide a framework to describe the metadata contents of any metadata blueprint on the basis of "components" which are grouped into description formats called "profiles". These facets of a metadata record are stored and shared in Clarin's Component Registry[104] and expressed as an XML file to promote reuse. There is a purpose-specific editor for CMDI called COEMDI[105]. Additionally, there is the Virtual Language Observatory (VLO)[106] which is an online service that provides an interface for uniform search and discovery of language resources and tools based on CMDI metadata records (see Van Uytvanck et al., 2012). Because this schema is intended as a meta-schema for conversion and depositing resources in the

[100] https://www.ailla.utexas.org/
[101] https://ailla.utexas.org/site/depositors/metadata
[102] https://www.clarin.eu/content/component-metadata
[103] https://www.clarin.eu/
[104] https://catalog.clarin.eu/ds/ComponentRegistry/#/
[105] http://clarino.uib.no/comedi/page
[106] https://www.clarin.eu/content/virtual-language-observatory-vlo

CLARIN infrastructure, and there are tools and services to carry out the conversions, and editing I will not get into the details of the metadata components in CMDI.

4.4.1.6 Issues in Metadata Compatibility and Interchange

A preliminary mapping of IMDI *Session Descriptions* (version 2.5) and OLAC (version 0.3) was attempted in 2001 (IMDI, 2001). A major task for the near future will be to formally map the parallels between TEI and each of the two more prominent LD metadata schemas and to develop an XSLT stylesheet to carry out such a conversion. Creating such a mapping is an important step in: a) showing that TEI can express the same information, and thus should be an acceptable schema for metadata in LD, and b) making it possible for those who use TEI (like in this project) to easily convert, or at least extract the required data areas in order to produce OLAC, IMDI or CMDI records required for depositing LD data in so many archives.

However, as is the case with the attempt at mapping IMDI and OLAC, in which there are numerous areas in which OLAC cannot express the details and types of information expressed in IMDI, it will also be the case between TEI and each of these other two systems that TEI can express a much wider variety of information, including the areas of "thick metadata" and meta-documentary called for by Nathan and Austin (2004) and Austin (2013) (see section: 4.3.3). While this is a good thing in terms of an individual project being able to record the full array of information they want, it also poses a problem, which is common to the TEI in that the vast amount of encoding options, and a lack of established practice in this specific sub-domain means that there are in many cases, multiple ways to markup a single phenomenon.

There are conversion schemas[107] between CMDI and several different metadata standards including OLAC, DCMI, IMDI, and several others, however not TEI. The reason given for this is that the TEI has too many variants, which in certain areas this is true, however in the area of metadata, the TEI header is actually one that should provide the easiest point from which to create a conversion schema. The eventual development of conversion schemas between TEI, OLAC and IMDI should provide a clear and key step in demonstrating the feasibility of creating a schema between TEI and CMDI.

[107] https://www.clarin.eu/faq-page/274#t274n3483

4.4.2 Standards and Formats for Corpora and Time-aligned Speech

The process of spoken language annotation is of course a major source of data in a LD project and given the nature of how much labor goes into transcription of audio files (35 to 1 time ratio); typical translation to a major language (25 to 1 ratio) and even more when grammatical or other transcriptions are involved (potentially above 100 to 1 ratio). Given this, the use of highly efficient tools for such tasks is the norm, and it is unheard of to undertake such work without task specific software. As a result, the data formats and standards used and produced in LD, as well as corpus projects for major languages with spoken language data are driven by these tools. In this section I discuss the data formats, interchange and interoperability, and will further discuss the tools and their core functions in section 4.4.4.

A significant problem exists in that different languages, research interests, linguistic subfields and methodological traditions have developed different transcription conventions, using different tools which themselves come with their own data model and formats (Schmidt, 2011). This has led to the situation that there are still no widely adopted standards for spoken language transcription (Schmidt, 2011). Time-aligned transcription can be divided into two levels:

- *macrostructure* which consists in temporal information, classes of transcription and/or annotation features. Macrostructure is generally implemented within the data models of the tools;
- *microstructure,* which is the way that relations between linguistic units (e.g. words, pauses, morphological and other semi-lexical units, etc.) and transcriptions are represented;

4.4.2.1 Time-aligned Transcriptions: Macrostructure

With regard to *macrostructure,* a basic fact of speech transcription tools is that despite reading and writing different file formats, the fundamental concepts and models are all necessarily based on the representation and organization of the same features, or variants of the same base model containing: a time-aligned annotation ‘triple’ comprised of a starting point, an

end point and the transcription and potentially annotation (Schmidt, 2011; Schmidt, et al. 2008). These components can be further divided into tiers and sub-tiers of a given linguistic or conceptual type which can be assigned to a specific speaker. The models implemented in the various tools can differ in several key ways which are briefly described.

- Timelines can be *implicit* or *explicit*; specifically, the timeline of the recorded content can exist on its own, and pointed to in transcription start- and end-times point, or the time points are only stated in the transcriptions or annotations.

```
<tier name="speaker_1_words">
  <annotation start="0.12" end="0.27">
    you
  </annotation>
  <annotation start="0.27" end="0.36">
    lie
  </annotation>
</tier>
<tier name="speaker_1_POS">
  <annotation start="0.12" end="0.27">
    PRO
  </annotation>
  <annotation start="0.27" end="0.36">
    V
  </annotation>
</tier>
```

```
<timeline>
  <anchor id="T0" time="0.12"/>
  <anchor id="T1" time="0.27"/>
  <anchor id="T2" time="0.36"/>
</timeline>

<tier name="speaker_1_words">
  <annotation start="T0" end="T1">you</annotation>
  <annotation start="T1" end="T2">lie</annotation>
</tier>
<tier name="speaker_1_POS">
  <annotation start="T0" end="T1">PRO</annotation>
  <annotation start="T1" end="T2">V</annotation>
</tier>
```

Figure 1: XML structure of implicit (left) vs explicit (right) time alignments from Schmidt et al. (2008)

```
0       Tokens  1       1.29
0.38    Spanish cabeza  1.11
0.38    IPA     ʃini↗   1.11
0.38    Mixtec  xiní    1.11
0.38    English head    1.11
```

Figure 2: TSV example of implicit transcription timeline exported from Praat TextGrid annotation

- Tiers can be layered (*single* vs *multi-layer*): whereas some models are as simple as having a *single* tier per speaker, others have a *multi-layer* model in which multiple tiers can be used to encode different levels of information for a single speaker.

```
<track name="speaker_1">
  <el start="0.12" end="0.27">
    <attribute name="words">you</attribute>
    <attribute name="POS">PRO</attribute>
  </el>
  <el start="0.27" end="0.36">
    <attribute name="words">lie</attribute>
    <attribute name="POS">V</attribute>
  </el>
</track>
```

```
<tier name="speaker_1_words">
  <annotation start="0.12" end="0.27">you</annotation>
  <annotation start="0.27" end="0.36">lie</annotation>
</tier>
<tier name="speaker_1_POS">
  <annotation start="0.12" end="0.27">PRO</annotation>
  <annotation start="0.27" end="0.36">V</annotation>
</tier>
```

Figure 3: XML structure of single (left) vs multi-layer (right) alignments from Schmidt et al. (2008)

- *Multi-layer* tool models allow for the hierarchical classification of the various tiers in terms of their structure and semantics. In such a model there can be multiple tiers that are considered subordinate to another.
- Tiers can contain *simple* or *structured annotations*: in some tools/models, a tier can have internal structure, whereas in others, the smallest unit of information is text strings.
- Speaker assignment to tiers: whereas in some models speakers can be explicitly assigned to a given tier, others such as Praat have no officially built in way to do this in their data model[108] (see @name in Figures 1 and 3).

The primary tools that are used in LD are ELAN and Praat with the former being ever increasingly the most widely adopted tool. Also, of interest is EXMARaLDA, while less used in the domain (it has been mainly used for pragmatics and discourse analysis, and dialectology) has many features, and in particular an XML-based data model that is highly conducive to important issues such as: extensible format, interoperability, recording of key metadata.

Briefly, several other tools that have been discussed in earlier LD and language technology literature and used for speech and/or video transcription and annotation but which are no longer widely adopted, and thus will not be discussed in further detail; these tools are:

[108] It is possible to keep track of speakers in Praat indirectly however, through naming tiers (e.g. Orth_JS) for the orthographic transcription of speaker "JS". This method is not ideal as if there are multiple tiers, and multiple speakers, this would mean creating duplicate tiers for each speaker. Another work around method is possible if there is only one speaker, which is to include the speaker's initials or name in the file name (e.g. "20190612-tamales-JS.wav" which could be annotated with a file named "20190612-tamales-JS.TextGrid").

- CLAN/CHAT (MacWhinney, 2000)[109]
- ANVIL (Kipp, 2001);
- EMU (http://ips-lmu.github.io/EMU.html)
- XTrans (https://www.ldc.upenn.edu/language-resources/tools/xtrans)
- TranscriberAG[110] (Barras et al., 2001)

In the following section the three aforementioned tools will be briefly discussed with an emphasis on the key factors of: their capacity for capturing and encoding phonetic information, their underlying data models (i.e. *macrostructure*) as well as, their interoperability with formats from other tools and standards.

4.4.2.1.1 ELAN

ELAN is a tool developed by the MPI in Nijmegen and is the most common tool used for spoken language and video transcription and annotation for language documentation. ELAN uses its own EAF (Elan Annotation Format) which is serialized in XML and is based on the Abstract Corpus Model (Brugman and Wittenburg, 2001; Sloetjes et al., 2011) and was influenced by the Annotation Graph Model (Bird and Liberman, 2001). The annotation graph model has similarities with the models used by other annotation systems, including EXMARaLDA and Praat among others (Cassidy and Schmidt, 2017). The annotations can be multi-tiered, and are organized as points on an external timeline, which can be assigned to a given speaker as well as annotator. ELAN tiers have internal structure, they can be typed and can be related hierarchically to a parent (or subordinate) tier.

[109] See Meankins (2007) for a review of CLAN and its difficulties in the context of a language documentation project.
[110] *TranscriberAG* is formerly known as *Transcriber: see* http://transag.sourceforge.net/

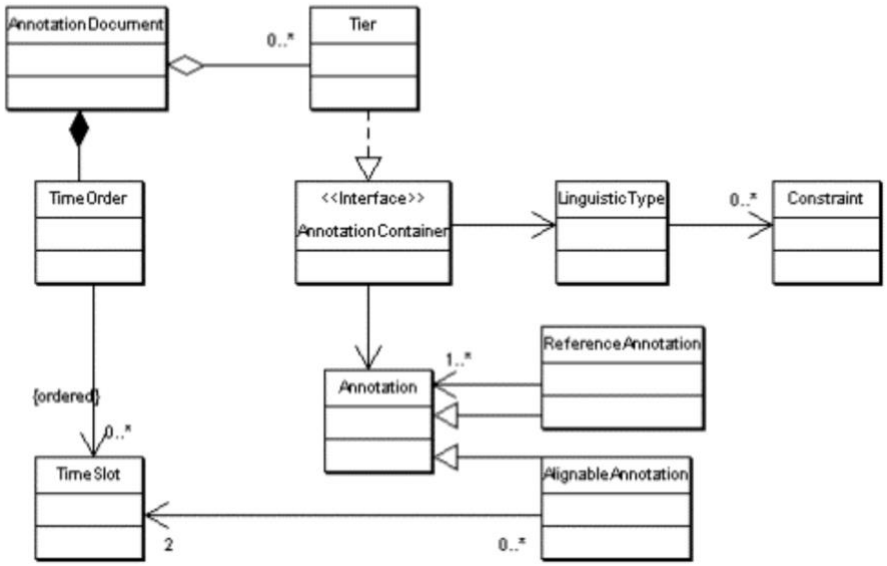


Figure 4: UML Diagram of the core part of the Abstract Corpus Model from Brugman and Russel (2004)

ELAN offers a high level of interoperability and interchange with other transcription tool by both being able to read files from other transcription tools such as Praat, Transcriber, CHAT, and Audacity, as well as files for lexical database and annotation software such as Toolbox, FLEx, Shoebox as well as CSV/Tab-delimited files. It can also export files to be read as Toolbox, FLEx, Shoebox, CSV/Tab-delimited files, Tiger XML, Interlinear text file, HTML, Praat among others. Thus, given the underlying XML data model and the extensive compatibility both in terms of importing and exporting, ELAN is a tool that is highly conducive to the mission of creating interoperable resources in LD.

4.4.2.1.2 Praat

Praat is another of the foremost time-aligned speech annotation tools and its strength is in the ease of the annotation process, as well as the vast amount of options for visualization and quantitative analysis of the acoustic signal, i.e. spectrogram, formats, waveform, pitch contour (F0), intensity, pulses, etc. Additionally, unlike ELAN or EXMARaLDA, Praat has the ability to automate a large number of tasks via the Praat scripting language. Files are saved in a plain text format called .TextGrid which can be extracted into binary, tab-separated or CSV file formats. Note that, in contrast to ELAN and EXMARaLDA, Praat annotation tiers can be either: ‘point

tiers' which annotate a single point in the timeline, or 'interval tiers' which annotate a span of time.

However Praat lacks many, or even most of the key features that are needed in a typical LD project, notably: it entirely lacks the capacity to explicitly include metadata (such as *speaker, date, session, place, etc.*), there are not parent/child tier relations, it doesn't allow for any association with controlled vocabularies, and possibly most troublesome is that it lacks an option for any XML export[111], and it lacks interoperability as it doesn't allow for import of other standard annotation formats or files from other related software (however as discussed, both ELAN and EXMARaLDA do allow for the importing and exporting of Praat files to and from their given systems). As discussed, other programs such as ELAN and EXMARaLDA use XML which is much more difficult to deal with if integrating into a plain text-based format than is the case in the reverse direction. Thus, the fact that Praat cannot import or export is largely due to the fact that it has no XML capabilities which greatly increases the ease of data transformation needed for interchange.

4.4.2.1.3 EXMARaLDA

EXMARALDA is an example of a toolkit that enables both annotation of time-aligned speech (audio or video) as well as various corpora and lexicon functions. An explicit aim in the development of EXMARaLDA is to facilitate the exchange of corpus data and long-term archival, to which end UNICODE and XML are key basic components of the system (Schmidt and Wörner, 2009). EXMARaLDA is described as a *data-centric* system in that the key contents and properties of the data itself are the driving force of the tools for processing it, which make it quite unique in that aspect (Schmidt and Wörner, 2009). EXMARaLDA's data model is a variety of annotation graph based on Bird and Liberman (2001) and is similar to ELAN, FOLKER and Praat to name a few. Within this system, there are three types of file formats: *Basic-Transcriptions, Segmented-Transcriptions, List-Transcriptions* (Cassidy and Schmidt, 2017).

[111] It is however possible, as will be discussed herein (see section 6.3.3.1) to convert Praat plain text (tsv) transcriptions to TEI.

As EXMARaLDA is a tool created for the study of issues in the domain pragmatics, it lacks tools for acoustic phonetic analysis such as those featured in Praat, whose core purpose is acoustic phonetics. As with Praat, EXMARaLDA can have layers of tiers and the annotations are *simple* and tiers cannot be assigned to parents tiers. Like ELAN, the annotations point to the full timeline in the data, speakers can be explicitly assigned to tiers, and tiers can be typed with one of the values of: *'transcription', 'description'* or *'annotation'*.

4.4.2.2 Time-aligned Transcriptions: Microstructure

Microstructure pertains to the particular transcription conventions used to denote mostly nonlinguistic content, and their usage is mostly specific to a particular corpus or project and are generally not published for a general audience (Schmidt, 2011). The applications of most of these transcription systems has generally been conversation and discourse analysis. Notable transcription conventions are as follows:

Transcription Convention	Transcription Example
HIAT - Halbinterpretative Arbeitstranskriptionen: (Ehlich and Rehbein, 1976; Ehlich, 2003)	((coughs)) You must/ you (should) let • it be. ((laughs)) Pleease!
CHAT - Codes for Human Analysis of Transcripts (MacWhinney, 2000)	&=coughs you must... you should let # it be. &=laughs please!
DT1 - Discourse Transcription (DuBois et al., 2003)	(COUGH) you must-- you <X should X> let .. it be. @@ please?
GAT - Gesprächsanalytisches Transkriptionssystem: (Selting et al., 2009)	((coughs)) you must- you (should/could) let (-) it be; ((laughs)) plea:se-
cGAT - (Selting et. al., 2009)	((coughs)) you must you (should/could) let (-) it be ((laughs)) please

Table 37: Notable transcription conventions with examples

In all such formats, lexical items are generally transcribed using orthography, but some depart from standard orthographies in certain cases, and most express the key features of: standard words, unfilled pauses, audible non-speech events (breathing, laughing, coughing),

uncertainty (providing alternatives to uncertain portion), incomprehensibility. Below these are listed as per their given convention:

Audible non-speech events (coughing, laughing, breathing):

- ((coughs)), ((laughs)): HIAT, GAT, cGAT
- (COUGH), @@: DT1[112]
- &=coughs, &=laughs: CHAT

Uncertainty/incomprehensibility is represented in four of the five systems (*with CHAT being the exception*):

- round brackets (e.g. you (should) let): HIAT
- <X X> (e.g. you <X should X> let): DT1
- possible words separated by forward slash (you (should/could) let): GAT, cGAT

Pauses:

- bullet (e.g. let • it be): HIAT
- hash (e.g. let # it be): CHAT
- "(-)" (e.g. let (-) it be): GAT, cGAT
- two periods (e.g. let .. it be): DT1

Self-repair or interruption:

- (e.g. You must/ you) HIAT
- (e.g. you must… you) CHAT
- (e.g. you must-- you) DT1
- (e.g. you must- you) GAT

Pronunciation variation, such as vowel lengthening is only represented in two of the five systems:

- vowel duplication in orthography (e.g. Pleease!): HIAT

[112] Note, the DT1 convention distinguished between laughing and coughing.

- colon (e.g. plea:se): GAT

Other key types of differences in these conventions has to do with the unit or span of speech, with the exception of cGAT, the rest of the systems divide speech according to either *utterances* (HIAT, GAT) or *intonation phases* (GAT) or *intonation units* (DT1). For *utterances*, HIAT distinguishes between declarative or exclamative *mood* with the first indicated with the period (e.g. let • it be.) and the second by the exclamation point (e.g. Pleease!). CHAT distinguishes between three different utterance types: *interrupted,* indicated by an ellipsis (e.g. you must... you); *declarative*, marked by a period (e.g. let # it be.) and emphatic, marked by an exclamation point (e.g. please!). Note that these criteria are based on pragmatics.

GAT distinguishes between the given portions of speech according to the pitch levels (i.e. *intonation phrases*). Level pitches are indicated by a hyphen (e.g. you must- you (should/could) let (-) it be;) whereas falling pitch is indicated by a semicolon (e.g. ((laughs)) plea:se-). DT1 shows three types of *intonation units*, the first being an interrupted unit (e.g. you must-- you); the second a terminative unit marked by a period (e.g. let .. it be.); and the third a so called 'appeal', represented by a question mark (e.g. please?).

4.4.2.3 ISO 24624:2016 and TEI Representation of Spoken Language Transcription

Schmidt (2011) *TEI-based approach to standardizing spoken language transcription* laid out a blueprint through which time aligned transcriptions can be formatted in TEI from EXMARaLDA. This work mapped out key factors necessary to express the structures of the aforementioned data model with other key formats such as ELAN. The TEI based format articulated therein was the basis for the ISO 24624:2016 guidelines, which was developed in joint agreement between ISO and the TEI consortium, and it is dually part of the TEI guidelines and an ISO standard. This standard uses the TEI elements to encode both the *micro*- and *macrostructure*, as well as metadata of spoken language transcription documents. Below I give a brief overview of the most essential components of each, with examples from the standard. For a full account of the encoding mechanisms, consult ISO 24624:2016 directly.

On the level of macrostructure, the major components deal with: *the timeline; utterances; free dependent annotations; grouping of utterances and dependent annotations; independent (generally non-linguistic) contents outside of utterances; inline paralinguistic annotation* and *global divisions of transcriptions.*

The timeline (<timeline>) is used to define points in the recorded speech content, each timeline is represented by a <when> element with an @xml:id which is referred to in the given events recorded in the transcription using @start, @end and @synch. The absolute time values from the beginning of the recording are specified in @interval. The attribute @since denotes the point in time from which the given <when> is measured.

```
<timeline unit="s" origin="#T0">
  <when xml:id="T0" absolute="2009-02-04T20:42:00"/>
  <when xml:id="T1" interval="2.13" since="#T0"/>
  <when xml:id="T2" interval="3.74" since="#T0"/>
  <when xml:id="T3" interval="4.71" since="#T0"/>
  <when xml:id="T4" interval="unknown" since="#T0"/>
  <when xml:id="T5" interval="8.53" since="#T0"/>
  <when xml:id="T6" interval="11.36" since="#T0"/>
  <when xml:id="T7" interval="13.91" since="#T0"/>
  <when xml:id="T8" interval="15.47" since="#T0"/>
  <!-- [...] more when elements -->
</timeline>
```

Figure 5: <timeline> element from ISO 24624:2016

Utterances are represented by the <u> (utterance) element which is the most important unit for the representation of transcriptions. In terms of standard practice with tools such as ELAN, etc., content represented in <u> corresponds to a contiguous span of speech by a given speaker and must be assigned to a speaker using the @who attribute, which refers to a person whose initials are placed in the value of @xml:id on a <person> element, which is declared in the header (see section 4.4.1.3 for description of metadata in the header). A <u> block may optionally be embedded in an <annotationBlock> element. The temporal information of an utterance can be stated in the attributes @start and @end, which point to the given @xml:id attributes in <when>. In a transcription in which the contents of <u> are not tokenized (see description of tokenization below), the <anchor> element can be used to delimit specific temporal points in the contents.

```
<!-- u with start and end attributes only (minimal temporal structure) -->
<u who="#SPK1" start="#T0" end="#T1" xml:id="u2">Good morning! </u>

<!-- u with embedded anchor elements (additional temporal structure) -->
<u who="#SPK0" start="#T1" end="#T4">Okay. <anchor synch="#T2"/>Très bien, <anchor
synch="#T3"/>très bien.
</u>

<!-- u with an attribute for language -->
<u who="#SPK1" start="#T0" end="#T1" xml:id="u2" xml:lang="en">Good morning! </u>

<!-- two <u>s with partial overlap -->
<u who="#SPK0" start="#T0" end="#T2">Do not <anchor synch="#T1"/>interrupt me!</u>
<u who="#SPK1" start="#T1" end="#T1">Sorry, <anchor synch="#T2"/>mate!</u>
```

Figure 6: Examples of various encodings and features of utterance mechanisms from ISO 24624

Free dependent annotations using <spanGrp> and embedded [113] elements are a method of standoff annotation in which the contents of a transcription are annotated separately from the <u> using pointers. The @type attribute can be used on <spanGrp> to specify what is the specific annotation and the @to and @from attributes[114] should be used on each to synchronize to the timeline (these times should be identical to the contents being annotated within the <u> being annotated).

```
<!-- annotations from an en (=English translation) tier -->
<!-- using a reference to the timeline -->
<spanGrp type="en">
  <span from="#T1" to="#T2">Okay. </span>
  <span from="#T2" to="#T4">Very good, very good.</span>
</spanGrp>

<!-- part-of-speech annotations -->
<!-- using a reference to ids of <w> elements →
 <spanGrp type="pos">
  <span from="#w148" to="#w148">PersPron</span>
</spanGrp>

<!-- 1:n relation between tokens and annotations -->
<u><w xml:id="w1">I</w><w xml:id="w2">dunno</w></u>

<spanGrp type="lemma">
  <span from="#w1" to="#w1">I</span>
  <span from="#w2" to="#w2">
```

[113] Note that the use of to annotate linguistic tokens is in line with the mechanism described in the ISO 24611 standard *Morphological Annotation Framework* (MAF).
[114] Note that @to and @from are not entirely necessary as one may decide to just specify the start time of each token using @start or @synch, which points to the point on the timeline.

```
      <span>do</span>
      <span>not</span>
      <span>know</span>
    </span>
  </spanGrp>

  <!-- hierarchically organized annotation -->
  <u>
    <w xml:id="w3">John</w><w xml:id="w4">loves</w><w xml:id="w5">Mary</w>
  </u>
  <spanGrp type="phraseStructure">
    <span from="#w3" to="#w5">
      <span>S</span>
      <span from="#w3" to="#w3">
        <span>NP</span>
        <span from="#w3" to="#w3">N</span>
      </span>
      <span from="#w4" to="#w5">
        <span>VP</span>
        <span from="#w4" to="#w4">V</span>
        <span from="#w5" to="#w5">
          <span>NP</span>
          <span from="#w5" to="#w5">N</span>
        </span>
      </span>
    </span>
  </spanGrp>
```

Figure 7: Examples of various encodings and features of standoff annotation of utterances using <spanGrp> from ISO 24624

Grouping of utterances and dependent annotations can be done using the element <annotationBlock> which serves to group the utterance content with its annotations in <spanGrp>. In this case the temporal points associated with a given utterance (@start, @end), or even a non-utterance event, as well as the speaker information (@who) can be stated on the level of the <annotationBlock> element.

```
<!-- an utterance grouped with corresponding annotations -->
<annotationBlock who="#SPK0" start="#T0" end="#T1">
  <!-- the transcribed text from the primary tier -->
  <u>
    <!-- [...] (see above) -->
  </u>
  <!-- additional annotations from a sup (=suprasegmentals) tier -->
  <spanGrp type="sup">
    <!-- [...] (see above) -->
  </spanGrp>
  <!-- additional annotations from a translation tier -->
  <!-- with an xml:lang attribute capturing the language of the translation —>
  <spanGrp type="translation" xml:lang="en">
```

```
    <!-- [...] (see above) -->
  </spanGrp>
</annotationBlock>
<!-- an annotationBlock without subordinate <u> element —>
<annotationBlock who="#SPK0" start="#T0" end="#T1">
  <vocal>
    <desc>laughter</desc>
  </vocal>
</annotationBlock>
```

Figure 8: Examples of various encodings and features of <annotationBlock> from ISO 24624

Global divisions of transcriptions can optionally be encoded with the element, which can either be encoded as a direct parent element to the <u>, or the <annotationBlock> and could potentially contain multiple instances of the latter two element blocks. If desired, the @type and @subtype attributes may be used to classify the given contents.

```
<div type="greeting">
  <annotationBlock who="#SPK0" start="#T0" end="#T1">
    <!-- [...] u and spanGrp elements, see above -->
  </annotationBlock>
  <annotationBlock who="#SPK1" start="#T1" end="#T2">
    <!-- [...] u and spanGrp elements, see above -->
  </annotationBlock>
</div>

<!-- [...] -->
<!-- final section of the interaction -->
<div type="farewell">
  <annotationBlock who="#SPK1" start="#T112" end="#T113">
    <!-- [...] u and spanGrp elements, see above -->
  </annotationBlock>
  <annotationBlock who="#SPK0" start="#T113" end="#T114">
    <!-- [...] u and spanGrp elements, see above -->
  </annotationBlock>
</div>
```

Figure 9: Examples of use of element as per ISO:24624

Because of the complicated and language-specific nature of the concept of "word", the <w> element is simply defined as a *token* in ISO:24624. Id's should be given to each <w> in order to allow pointing. Tokens can be typed (@type) with possible values of "assimilated", "truncated", or "repetition". The @ana attribute can be used to annotate grammatical (e.g. part of speech) or other linguistic or other features can be encoded directly (as opposed to using the standoff annotation method with <spanGrp> described above). The @lemma attribute can be

used to associate a given token with a lemma in a lexicon, and @lemmaRef can be used to point to a definition of a lemma in an external (potentially online) lexicon. The @xml:lang attribute can also be used on <w>.

```
<!-- an utterance divided into tokens -->
<u who="#SPK0" start="#T0" end="#T2">
  <w xml:id="w148">I</w>
  <w xml:id="w149">am</w>
  <w xml:id="w150">very</w>
  <w xml:id="w151">much</w>
  <w xml:id="w152">aware</w>
  <w xml:id="w153">of</w>
  <w xml:id="w154">that</w>
</u>

<!-- token marked as assimilated via a type attribute -->
<u who="#SPK0" start="#T0" end="#T1">
  <w xml:id="w1">what</w>
  <w xml:id="w2" type="assimilated">cha</w>
  <w xml:id="w3">got</w>
  <w xml:id="w4">cookin</w>
</u>


<u who="#SPK0" start="#T0" end="#T2">
  <w xml:id="w148" lemma="I" ana="PRO">I</w>
  <w xml:id="w149" lemma="be" ana="V">am</w>
  <w xml:id="w150" lemma="very" ana="ADV">very</w>
  <w xml:id="w151" lemma="much" ana="ADV">much</w>
  <w xml:id="w152" lemma="aware" ana="ADJ">aware</w>
  <w xml:id="w153" lemma="of" ana="PREP">of</w>
  <w xml:id="w154" lemma="that" ana="PRO">that</w>
</u>

<!-- language encoded as attribute on the token: with code-switching →
<u who="#SPK0" start="#T0" end="#T2">
  <w xml:id="w148" xml:lang="en">I</w>
  <w xml:id="w149" xml:lang="en">am</w>
  <w xml:id="w150" xml:lang="fr">enchanté</w>
  <w xml:id="w151" xml:lang="fr">mon</w>
  <w xml:id="w152" xml:lang="fr">cher</w>
  <w xml:id="w153" xml:lang="fr">ami</w>
</u>
<!-- a token with an accentuated syllable -->
<!-- the accentuation being represented in a separate span element -->
<annotationBlock who="#SPK0" start="#T0" end="#T2">
  <u>
    <!-- [...] -->
    <w xml:id="w152"><seg xml:id="seg152a">awe</seg>some</w>
    <!-- [...] -->
  </u>
  <!-- [...] -->
  <spanGrp type="prosody">
```

```
    <span from="#seg152a" to="#seg152a">accentuated</span>
  </spanGrp>
</annotationBlock>

<!-- the same phenomenon encoded inline -->
<w xml:id="w152"><seg type="accentuated">awe</seg>some</w>

<!-- a token with a short pause inside -->
<w xml:id="w152">abso<pause type="short"/>lutely</w>

<!-- a token with a time anchor inside -->
<w xml:id="w152">a<anchor synch="#T3"/>ware</w>
```

Figure 10: Examples of use of element as per ISO:24624:2016

On the level of microstructure, the major components deal with: tokens; pauses, audible and visible non-speech events; punctuation; uncertainty, alternatives, incomprehensible and omitted passages; units above and below the <u> level. Note that all these (mostly) non-linguistic features are those for which the various microstructure transcriptions systems (i.e. HIAT, CHAT, DT1, GAT, cGAT) described in the previous section were developed, however the use of the XML elements of the TEI/ISO 24624:2016 guidelines effectively negates the need to keep these transcriptions in the final data, as these features using elements and attributes which produces a much less obtrusive annotation. However, as tools are used to transcribe spoken language, microstructure transcriptions would never be done directly into the TEI/ISO 24624:2016 format, thus there needs to be (XSLT) conversion scripts developed to convert between the various systems. These are each outlined below along with other issues such as: uncertainty, alternatives, incomprehensible or omitted passages and punctuation.

Independent (generally non-linguistic) contents outside of utterances, such as pauses and incidents can be encoded with the <pause> and <incident> elements respectively. These elements must also have temporal information recorded using @start and @end. Audible non-speech events such as: breathing, laughing, coughing, etc. can be expressed with the <incident> element in an embedded <desc>.

```
<incident start="#T1" end="#T2">
  <desc>roar of thunder outside</desc>
</incident>
```

Figure 11: TEI markup mechanism for expression of non-speech events as per Schmidt (2011)

Unfilled pauses can be expressed by the <pause> element with the attribute @dur (duration) and if necessary @start and @end as well.

```
<!-- pause inside an utterance -->
<u who="#SPK0" start="#T0" end="#T2">
  <w>I</w>
  <w>am</w>
  <pause dur="PT1.2S"/>
  <w>aware</w>
  <w>of</w>
  <w>that</w>
</u>

<!-- measured pause outside <u>, with its own start and end attributes —>
<pause dur="PT0.61S" start="#T10" end="#T11"/>
```

Figure 12: TEI markup mechanisms for expression of unfilled pauses as per ISO:24624

Annotations to mark the occurrence of some paralinguistic component of an utterance such as a change in pace, pitch, tempo or rhythm can be done using the <shift> element. The @synch value assigns the feature to a point in the timeline and the attribute @new specifies the particular change in the speech quality.

```
<!-- a change of tempo encoded as a <shift> milestone -->
<u start="#T1" end="#T4" who="#SPK1">
  And he was <shift feature="tempo" new="faster" synch="#T2"/>up and away
  <shift feature="tempo" new="normal" synch="#T4"/>
</u>
<!-- the same phenomenon encoded as an annotation in a <span> -->
<annotationBlock start="#T1" end="#T4" who="#SPK1">
  <u>
    And he was <anchor synch="#T2"/>up and away
  </u>
  <spanGrp type="sup">
    <span from="#T2" to="#T4">faster</span>
  </spanGrp>
</annotationBlock>
```

Figure 13: TEI markup mechanism for denotation of shift in speech quality as per ISO:24624

Audible and visible non-speech events such as: non-verbal communicative functions (e.g. laughter, shaking of the head); secondary modes of communication (e.g. body language, hand gestures, facial expressions), as well as events such as background noise (e.g. telephone ringing), and activities (e.g. shifting, rummaging through one's pocket) can be represented using one of the <vocal>, <kinesic>, or <incident> elements in combination with <desc>, which is used to

specify the given non-linguistic event. These events can be given timestamps in combination with <anchor> or temporal durations using @start and @end. Additionally, they can be attributed to a particular individual using @who.

```
<!-- coughing encoded as vocal element between tokens and anchors of a u -->
    <u who="#SPK0" start="#T4" end="#T6">
        <anchor synch="#T4"/>
        <w>dépend</w>
        <vocal>
          <desc>cough</desc>
        </vocal>
        <anchor synch="#T5"/>
        <w>un</w>
        <w>peu</w>
        <anchor synch="#T6"/>
      </u>

      <!-- simultaneous laughter by the same speaker -->
      <!-- encoded as vocal element within the same annotationBlock --> <!-- with start and end points -->
      <annotationBlock who="#SPK0" start="#T4" end="#T6">
        <u>
          <anchor synch="#T4"/>
          <w>dépend</w>
          <anchor synch="#T5"/>
          <w>un</w>
          <w>peu</w>
          <anchor synch="#T6"/>
        </u>
        <vocal start="#T4" end="#T6">
          <desc>laughing</desc>
        </vocal>
      </annotationBlock>

      <!-- (backchannel) nodding as kinesic element on the level of annotationBlock --> <!-- with speaker
assignment and start and end points -->
      <annotationBlock who="#SPK0" start="#T6" end="#T9">
        <!-- [...] -->
      </annotationBlock>
      <kinesic who="#SPK1" start="#T7" end="#T8">
        <desc>nods</desc>
      </kinesic>
```

Figure 14: Examples encoding visible and audible non-speech events as per ISO:24624

As described above, in the traditional micro-structure conventions (e.g. GAT, HIAT, CHAT, etc.), punctuation characters are used to denote nonlinguistic content, thus when using such systems, it is not possible to use full orthographic punctuation in spoken language transcription. However, in the TEI-based representations of spoken language content, these features are represented using purpose-specific elements and attributes. Thus, in the case that

orthographic punctuation is included, it can be encoded using the <pc> (punctuation character) element. The attributes @type and @unit can also be used to specify additional functional information if desired.

```
<!-- punctuation represented as pc elements -->
<u who="#SPK0" start="#T4" end="#T6">
  <w xml:id="w330">No</w>
  <pc>,</pc>
  <w xml:id="w331">I</w>
  <w xml:id="w332">mean</w>
  <w xml:id="w333">I</w>
  <w xml:id="w334">knew</w>
  <pc type="declarative">.</pc>
</u>
```

Figure 15: Example encoding orthographic punctuation as per ISO:24624

Uncertainty, alternatives, incomprehensible and omitted passages are dealt with using the <unclear>, <choice> and <gap> elements. Unclarity can be expressed in a number of ways: in this system, <unclear> can be used, and in the case of multiple alternatives (such as in the example above with “(should/could)”) and the attribute @reason can be used to specify the cause of the uncertainty (e.g. background noise, unclear speech, recording quality, etc.). Where there are possible alternate interpretations, the element <choice> can be used as a parent element to the given alternatives.

```
<!-- uncertain passage -->
<u who="#SPK0" start="#T4" end="#T6">
  <w>you</w>
  <unclear reason="background noise">
    <w>should</w>
  </unclear>
  <w>let</w>
  <!-- [...] -->
</u>
<!-- uncertain passage with alternatives for a single word-->
<u who="#SPK0" start="#T4" end="#T6">
  <w>you</w>
  <unclear>
    <choice>
      <w>should</w>
      <w>could</w>
    </choice>
  </unclear>
  <w>let</w>
  <!-- [...] —>
</u>
```

Figure 16: Representation of unclear content in ISO:24624

Where a passage of speech is completely incomprehensible, the <gap> element should be used also in combination with the @reason attribute with the value of "incomprehensible". The @dur attribute can be used to specify the duration of the gap. Additionally, <gap> can be used for portions that were left untranscribed for other reasons as well.

```
<!-- incomprehensible passage within an utterance -->
<u who="#SPK0" start="#T4" end="#T6">
  <w>good</w>
  <w>morning</w>
  <gap reason="incomprehensible" unit="syllables" quantity="2"/>
</u>
<!-- incomprehensible passage between utterances -->
<!-- with start and end attributes -->
<u who="#SPK0" start="#T4" end="#T6">
  <w>good</w>
  <w>morning</w>
</u>
<gap reason="incomprehensible" dur="PT8.9S" start="#T6" end="#T7"/>
```

Figure 17: Representation of gaps in content in ISO:24624

Divisions of an utterance element (<u>) are represented as a typed <seg> (e.g. *utterance, intonation units, intonation phrases*), and a @subtype can be used for additional classification (e.g. *declarative, interrogative*, for the mode of an utterance or potentially *falling, rising*, etc. for a tone or intonation). A typical use would be for the equivalence of a distinct "sentence" in spoken language. ID's (@xml:id) can be used for pointing when using standoff annotation.

```
<!-- u divided into two seg elements (utterances according to HIAT/CHAT) -->
<u who="#SPK0" start="#T40" end="#T43">
  <seg type="utterance" subtype="declarative" xml:id="seg23">
    <w xml:id="w319">And</w>
    <gap reason="incomprehensible"/>
    <w xml:id="w320">disappointed</w>
    <w xml:id="w321">when</w>
    <w xml:id="w322">you</w>
    <w xml:id="w323">got</w>
    <w xml:id="w324">to<anchor synch="#T41"/>gether</w>
  </seg>
  <anchor synch="#T42"/>
  <seg type="utterance" subtype="interrogative" xml:id="seg24">
    <gap reason="incomprehensible"/>
    <w xml:id="w325">you</w>
    <pc>,</pc>
    <w xml:id="w326">Victoria</w>
  </seg>
</u>
```

Figure 18: Representation of division of an utterance (<seg>) in content in ISO:24624

Finally, as mentioned previously, the TEI elements for metadata described in the previous section are also in accordance with the ISO 24624:2016 standard, thus this portion of the standard need not be covered in this section.

Schmidt (2011) laid the groundwork for the TEI-based serialization of ISO:24624:2016 while demonstrating concretely how to the TEI can be used to represent all the important macro- and micro-structural features needed and represented in the common transcription tools (e.g. ELAN, EXMARaLDA, Praat, etc.) and transcription conventions (e.g. CHAT, HIAT, GAT, etc.). As a result, it is clear that in many cases, it should be possible to convert between the given formats, and that TEI/ISO 24624:2016 would be fully able to either be an exchange and/or an underlying format for such tools. Further support for this assertion is also made as a part of this dissertation, particularly in the form of the use of TEI to represent transcription files created in Praat.

A major issue in working with and creating a corpus of both transcribed speech and text resources is the fact that the common formats used to search and work with these different data types are often completely separate. This problem is exacerbated by the fact that LD projects generally use different tools (with different data models) for the tasks of annotating text corpora and transcribed speech. This problem could be resolved by defining a common mapping or ontology of the features that are in each linguistic resource type on both an abstract and technical level for non-spoken linguistic resources in the same manner as was done by Schmidt (2011) for TEI/ISO:24624. In fact, ISO: 24624 states that a secondary goal of the standard is to relate transcribed data with standards for annotated corpora. The feasibility of this idea will in fact be shown, both in the following section concerning corpus and annotation standards, as well as throughout the second half of this dissertation in which the TEI encoding of the MIX resources are described.

4.4.2.4 Corpora and Annotation

Gries and Berez (2017) list the following characteristics as prototypically defining a language corpus:

- it should consist of one or more *machine-readable* Unicode text files;
- it should be *representative* for a particular kind of speaker demographic, register, variety, or a language in general;
- it should be *balanced*: the sampling of the given speakers, registers, varieties should be proportional to the overall population of those that speak the language;
- it should contain data from *natural communicative settings* or contexts: ideally the language data should be as untainted as possible from the process of data collection and should not have just been created for the purpose of creating the corpus;

However, when dealing with an under-resourced language, there are unique issues and those compiling a corpus do not always have the luxury of adhering to every one of the aforementioned conditions. As discussed in section 4.3, in accumulating a collection of materials in a language for which there are very few existing (such as is with this MIX project), the corpora will likely be much smaller, and less topically diverse than those of major languages (Ostler, 2008; Gries and Berez, 2017; Mosel, in press). In LD projects, it is usually the case that most of the sources will be from *primary data* (e.g. recordings) and what Himmelmann (2006) refers to as the *apparatus* of the corpus (e.g. transcription, annotations and metadata). In building a corpus of an endangered or under-resourced language, it may be necessary to integrate resources from a wide array of different formats into a single corpus along with the primary data which requires both effort in terms of the workflow (including programming and/or tools), as well as proper data formats within which all the sources can be integrated and accessed. Furthermore, in most cases an LD corpus will also be bilingual and have interlinear glossed text (IGT) (Mosel, in press) which adds another dimension to the process of annotation and markup.

Whereas in a corpus project for a major language, or even an extinct or ancient language, the target audience tends to be much more focused, be that for academic, various specialized purposes, or a popular general audience. In such cases, the types of interfaces used to access the data are more easily determined and catered towards the given target users. In creating corpus materials for under-documented languages, even if the project is created with an academic focus, there is also an ethical responsibility of making the resources available to the speech community

so they can be reused. To do this requires extra effort in making user-friendly interfaces and data formats that serve both the technical and social purposes.

In integrating the array of resources and producing both a: (maximally) seamless, accessible body of resources, and one that is extensible and conducive to long-term durability, the use of XML and Unicode are now well entrenched in best practices and are the basic format of key metadata standards, and standards for tools in spoken language transcription. The use of TEI for the encoding of text corpora is widely adopted and has been the basis for the encoding format for such large scale projects as the British National Corpus[115] and the Polish National Corpus (Przepiórkowski and Bański, 2009) as is highly intertwined with international standards for various types of linguistic contents particularly those managed by ISO/TC 37 SC 4 *Language Resource Management* (see Stührenberg, 2012; Romary, 2015a). The encoding of many corpora projects which are not done according to a specific standard, are nonetheless either based on, or designed to be compatible with the TEI, notably early pioneering initiatives such as: EAGLES (Expert Advisory Group on Language Engineering Standards)[116]; Corpus Encoding Standard (CES)[117], XCES[118], as well as MATE (Multilevel Annotation, Tools Engineering) (Lehmberg and Kai, 2008). Another standard EpiDoc[119] (Elliot et al., 2006-2017), which is based on a subset of the TEI and is this fully compatible. Though EpiDoc is for the encoding of ancient sources such as monuments, inscriptions such as epitaphs, papyri, etc., and generally not relevant to LD, it could very well be relevant for project documenting languages with ancient ancestral sources.

Although in a monolingual context, it is feasible that a corpus may have no annotation, the reality is that, especially in LD, a corpus will likely be annotated with at least translations, so standards for corpus encoding are discussed along with annotation. Given the central role of tools such as ELAN, Praat, EXMARaLDA, in the speech transcription domain and FLEx and Toolbox in the domain of annotation of texts and lexicons, unless the project has members who can create conversion schemas, the compatibility of the different resources (e.g. spoken language

[115] https://www.english-corpora.org/bnc/ Note that Gries and Berez (2017) state that the BNC represents an example of a prototypical corpus.
[116] http://www.ilc.cnr.it/EAGLES/home.html
[117] https://www.cs.vassar.edu/CES/
[118] http://www.xces.org/
[119] https://sourceforge.net/p/epidoc/wiki/Home/

and text-based) in an LD project is often dependent on these tools. Below in this section, the specific of the data models and standards for corpora are discussed and analyzed for structural and conceptual compatibility, and later in section 4.4.4, these issues will also be further discussed with respect to the capabilities of the specific software for importing and exporting various data formats.

Corpora can be annotated for lemma information, part-of-speech and/or morpho-syntax, syntactic parse trees (based on the part-of-speech tagging), various types of semantic annotation. Corpus annotation content may differ based on the purpose, as we have seen, spoken language corpora can be annotated for phonetics[120], prosody, sign-language, gesture, conversational interaction. Parallel corpora can contain translations of text or transcriptions into one or more languages, and/or interlinear glossed texts, which is a mixture of translational glosses and morpho-syntactic annotation. Other types of corpora may be annotated for discourse and pragmatics, and learners corpora can be annotated with any of the features mentioned above in combination with information about errors that language learners have made in a given context[121].

There are several manners of corpus annotation: *inline* or *embedded annotation; multi-tiered or interlinear annotation; standoff/standalone annotation, relational databases* (Gries and Berez, 2017). Below I present a brief discussion of each, with examples and point out which tools and/or standards use them.

In *inline* or *embedded annotation,* the annotation is included in the same line in the file as the annotated content and is used in lemmatization and part-of-speech tagging (Gries and Berez, 2017). In the example below from the XML version of the BNCwe corpus, the @hw attribute stands for 'headword' and the @c5 stands for the CLAWS5 tagset[122]

[120] Until relatively recently, phonetics were transcribed with ASCII formats like (X-)SAMPA which were highly limited in terms of the features it could represent, but now with Unicode, the standard IPA alphabet can be used without problems of machine readability or rendering.
[121] See Gries and Berez (2017) for a more detailed overview of each of these types of corpus annotations with discussion of specific corpora.
[122] http://ucrel.lancs.ac.uk/claws5tags.html

```
<s n="1">
   <w c5="VVB" hw="introduce" pos="VERB">Introduce</w>
   <w c5="NP0" hw="brenda" pos="SUBST">Brenda</w>
   <w c5="PNQ" hw="who" pos="PRON">who</w>
   <w c5="VBZ" hw="be" pos="VERB">'s</w>
   <w c5="VVG" hw="go" pos="VERB">going</w>
   <w c5="TO0" hw="to" pos="PREP">to</w>
   <w c5="VVI" hw="speak" pos="VERB">speak</w>
   <w c5="PRP" hw="to" pos="PREP">to</w>
   <w c5="PNP" hw="we" pos="PRON">us</w>
   ....
</s>
```

Figure 19: Partial extract of part-of-speech and lemma tagged sentence from BNCwe XML corpus

TEI has several different mechanisms for encoding these kinds of annotations using attributes on the <w> element: @pos, @lemma. Note also that the @ana can be used to annotated other content (non-pos where @pos is used) or it can be used in the place of @pos and @lemmaRef can be used to place a pointer to the definition of a lemma in an online lexicon where needed.

```
<s n="1">
   <w pos="VVB" lemma="introduce">Introduce</w>
   <w pos="NP0" lemma="brenda">Brenda</w>
   <w pos="PNQ" lemma="who">who</w>
   <w pos="VBZ" lemma="be">'s</w>
   <w pos="VVG" lemma="go">going</w>
   <w pos="TO0" lemma="to">to</w>
   <w pos="VVI" lemma="speak">speak</w>
   <w pos="PRP" lemma="to">to</w>
   <w pos="PNP" lemma="we">us</w>
   ....
</s>
```

Figure 20: TEI version of extract from BNCwe XML corpus

In *multi-tiered or interlinear annotation*, the annotations are placed on different lines of the same file as the annotated content. As the name would indicate, this is a prototypical annotation format in interlinear glosses, which is of course the most common annotation performed in LD projects. These types of annotations are produced in FLEx, Toolbox, ELAN, Praat, EXMARaLDA, etc.

The examples below show first a plain IGT glossing of a phrase in MIX, and then an ELAN annotation thereof in order to concretely demonstrate how the multi-tier/interlinear structure is encoded.

nakatsi lochi
na-katsi lochi
HORT-eat[3SG] vulcher
'I hope a vulcher eats you'

Figure 21: IGT of Mixtepec-Mixtec phrase

```
<TIER ANNOTATOR="JB" LINGUISTIC_TYPE_REF="default-lt" PARTICIPANT="TS" TIER_ID="Orth">
  <ANNOTATION><ALIGNABLE_ANNOTATION ANNOTATION_ID="a1" TIME_SLOT_REF1="ts3" TIME_SLOT_REF2="ts15">
      <ANNOTATION_VALUE>nakatsi lochi</ANNOTATION_VALUE>
    </ALIGNABLE_ANNOTATION>
  </ANNOTATION></TIER>
<TIER ANNOTATOR="JB" LINGUISTIC_TYPE_REF="default-lt" PARTICIPANT="TS" TIER_ID="tokens">
  <ANNOTATION><ALIGNABLE_ANNOTATION ANNOTATION_ID="a2" TIME_SLOT_REF1="ts1" TIME_SLOT_REF2="ts2">
      <ANNOTATION_VALUE>na-</ANNOTATION_VALUE>
    </ALIGNABLE_ANNOTATION>
  </ANNOTATION>
  <ANNOTATION><ALIGNABLE_ANNOTATION ANNOTATION_ID="a3" TIME_SLOT_REF1="ts9" TIME_SLOT_REF2="ts12">
      <ANNOTATION_VALUE>katsi</ANNOTATION_VALUE>
    </ALIGNABLE_ANNOTATION>
  </ANNOTATION>
  <ANNOTATION>
    <ALIGNABLE_ANNOTATION ANNOTATION_ID="a4" TIME_SLOT_REF1="ts13" TIME_SLOT_REF2="ts18">
      <ANNOTATION_VALUE>lochi</ANNOTATION_VALUE>
    </ALIGNABLE_ANNOTATION>
  </ANNOTATION></TIER>
<TIER ANNOTATOR="JB" LINGUISTIC_TYPE_REF="default-lt" PARTICIPANT="TS" TIER_ID="gloss">
  <ANNOTATION><ALIGNABLE_ANNOTATION ANNOTATION_ID="a9" TIME_SLOT_REF1="ts4" TIME_SLOT_REF2="ts8">
      <ANNOTATION_VALUE>HORT-</ANNOTATION_VALUE>
    </ALIGNABLE_ANNOTATION>
  </ANNOTATION>
  <ANNOTATION><ALIGNABLE_ANNOTATION ANNOTATION_ID="a10" TIME_SLOT_REF1="ts10" TIME_SLOT_REF2="ts11">
      <ANNOTATION_VALUE>eat[3s.inf]</ANNOTATION_VALUE>
    </ALIGNABLE_ANNOTATION>
  </ANNOTATION>
  <ANNOTATION><ALIGNABLE_ANNOTATION ANNOTATION_ID="a11" TIME_SLOT_REF1="ts14" TIME_SLOT_REF2="ts17">
      <ANNOTATION_VALUE>vulcher</ANNOTATION_VALUE>
    </ALIGNABLE_ANNOTATION>
  </ANNOTATION></TIER>
<TIER ANNOTATOR="JB" LINGUISTIC_TYPE_REF="default-lt" PARTICIPANT="TS" TIER_ID="en">
  <ANNOTATION><ALIGNABLE_ANNOTATION ANNOTATION_ID="a12" TIME_SLOT_REF1="ts5" TIME_SLOT_REF2="ts16">
      <ANNOTATION_VALUE>I hope a vulcher eats you</ANNOTATION_VALUE>
    </ALIGNABLE_ANNOTATION>
  </ANNOTATION></TIER>
```

Figure 22: Sample interlinear annotation from ELAN (not showing timeline)

It should be kept in mind from the discussion in the previous section that ELAN allows for the association of tiers to controlled vocabularies, and thus it is possible to associate a given tier with more information than is shown in the example above.

The following example is an export of a FLEx '.flextext' file from an interlinearized annotation of the inflected MIX verb phrase meaning *you wash*, the gloss form of which is *ntakacha* and the inflection for second person singular informal is marked with the affix '-u'.

```
<paragraph guid="092b0b5f-349d-42c5-0a94a82">
  <phrases>
    <phrase guid="77925b0a-0b9f-462b-c6545771">
      <item type="segnum" lang="en">1</item>
      <words>
        <word guid="ddcdce5e-e52d-44a9-208b325">
          <item type="txt" lang="mix">ntakachu</item>
          <morphemes>
            <morph type="stem" guid="d7f713e8-e8cf-c04f186933">
              <item type="txt" lang="mix">ntakach</item>
              <item type="cf" lang="mix">ntakacha</item>
              <item type="gls" lang="en">wash</item>
              <item type="msa" lang="en">v</item>
            </morph>
            <morph type="suffix" guid="d7f713dd-e8cf-04f186933">
              <item type="txt" lang="mix">-u</item>
              <item type="cf" lang="mix">-u</item>
              <item type="gls" lang="en">2sg.inf</item>
              <item type="msa" lang="en">v:Any</item>
            </morph>
          </morphemes>
          <item type="gls" lang="en">wash</item>
          <item type="pos" lang="en">v</item>
        </word>
      </words>
      <item type="gls" lang="en">you wash</item>
    </phrase>
  </phrases>
</paragraph>
```

Figure 23: Sample interlinear annotation from FLEx

The XLingPaper format (Black, 2009; Simons and Black, 2009; Black and Black, 2012) is an XML-based format for writing linguistic articles and grammatical descriptions and is used by the FLEx program; SIL's Parser and Writer for Syntax[123] (PAWS) (Black and Black, 2012)

[123] https://software.sil.org/paws/

and the XMLMind editor[124]. An example showing the way IGT are formatted in this data model below:

```
<example num="xPluralIZ">
   <listInterlinear letter="xPluralIZ.a">
      <lineGroup>
         <line>
            <langData lang="lZap">ca yoo</langData>
         </line>
         <line>
            <gloss lang="lGloss">PL house</gloss>
         </line>
      </lineGroup>
      <free>
         <gloss lang="lFree">‘houses’</gloss>
      </free>
   </listInterlinear>
```

Figure 24: Sample interlinear annotation from XLingPaper

The data structure used in SIL’s Toolbox is plain text based and backslashes are used in defining data fields, this is known as the SIL ‘standard format’ (.sfm) and is also an output option of FLEx. Note Toolbox can now also export to XML as well.

```
\t Anong oras?
\m ano -ang oras
\g what.is the hour
\p pron art n
\f What time is it?
```

Figure 25: Sample interlinear annotation in Toolbox in .sfm format

A significant problem with this format is that the fields can be completely user defined (though there are suggested fields users can choose) leading to extreme variation between projects, and as the files are plain text instead of XML-based, the corresponding words and annotations are only aligned with whitespace characters (Arkhipov and Thieberger, 2018).

In the TEI, though parallel XML element structures to those used in FLEx for IGT annotation exist, there has been a remarkable absence of discussion or direction on the issue; other than publications related to this project (Bowers and Romary, 2017, 2019) whose methods will be discussed in section 6.4.5 the only discussion of the use of TEI mechanisms for IGT

[124] https://software.sil.org/xlingpaper/

annotation seems to be Langendoen and Simons (1995), in which feature structures (see section 4.4.3.2.3 below) from the earlier P3 version of the standard (serialized in SGML). The following example shows the IGT parsing of just the first word in Northwest Alaska Inupiatun (which is notoriously an agglutinative language).

```
<fs type=word>
  <f name=form><str>akutchilighmik-uvva</str></f>
  <f name=gloss><str>about making Eskimo ice cream.</str></f>
  <f name=analysis>
    <fs type=morpheme>
      <f name=type> <sym value=root> </f>
      <f name=form> <str>akut</str></f>
      <f name=lexForm> <str>akutuq</str></f>
      <f name=gloss><str>ice cream.</str></f>
    </fs>
    <fs type=morpheme>
      <f name=type>  <sym value=suffix></f>
      <f name=form> <str>chi</str></f>
      <f name=lexForm> <str>si</str></f>
      <f name=gloss><str>RSL</str></f>
    </fs>
    <fs type=morpheme>
      <f name=type> <sym value=suffix></f>
      <f name=form> <str>ligh</str></f>
      <f name=lexForm> <str>liq</str></f>
      <f name=gloss> <str>GER</str></f>
    </fs>
    <fs type=morpheme>
      <f name=type>  <sym value=suffix></f>
      <f name=form><str>mik</str></f>
      <f name=lexForm> <str>mik</str></f>
      <f name=gloss>  <str>s.MOD</str></f>
    </fs>
    <fs type=morpheme>
      <f name=type>  <sym value=enclitic></f>
      <f name-form> <str>uvva</str></f>
      <f name=lexForm> <str>uvva</str></f>
      <f name=gloss> <str>now</str></f>
    </fs>
  </f>
</fs>
```

Figure 26: Sample interlinear annotation in early TEI (P3 SGML) feature structures from Langendoen and Simons (1995)

It should be noted however, that while encoding IGT using the present day TEI P5 version of feature structures[125], parallel to those shown above is indeed possible, one significant

[125] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/FS.html

drawback is that it uses entirely different elements than are typically used in text corpora as well as in the encoding of spoken language transcription described above (e.g. <w>, <seg>, etc.) (although as will be discussed in section 4.4.3.2.3, it is possible to encode language data in feature structures).

In the format GMT (*Generic Mapping Tool*) (Ide and Romary, 2001) the authors present a means of encoding a very similar IGT structure in XML to that shown from FLEx in Figure 26.

```
<struct type="W-level">
   <feat type="lemma">pomme_de_terre</feat>
   <feat type="pos">NOUN</feat>
   <struct type="W-level">
      <seg target="#w1"/>
         <feat type="lemma">pomme</feat>
         <feat type="pos">NOUN</feat>
   </struct>
   <struct type="W-level">
      <seg target="#w2"/>
      <feat type="lemma">de</feat>
      <feat type="pos">PREP</feat>
   </struct>
   <struct type="W-level">
      <seg target="#w3"/>
      <feat type="lemma">terre</feat>
      <feat type="pos">NOUN</feat>
   </struct>
</struct>
```

Figure 27: GMT Interlinear Glossed Text from Ide and Romary (2001)

Though no longer in use, GMT was a format that sought to provide generic XML representations of morphological and corpus linguistic contents and annotations and whose data models provided structural precedent for the ISO 24611:2012 *Morphological Annotation Framework* (MAF) and ISO 24612:2012 *Linguistic Annotation Framework* (LAF). Both LAF (by means of feature structures) and MAF (using TEI stand-off mechanisms) were initially designed to be serializable in TEI, thus these structures can surely be represented in TEI and make an ideal template for a parallel structure to the XML format used in FLEx. Thus, while despite there not being a fully established means of encoding IGT, an exact parallel structure to the FLEx example in Figure 27 can nevertheless be achieved in TEI. Note however that there are other potential ways of encoding this, this following example shows that the TEI has the full capacity to express precisely the same data contents and structure as FLEx.

```
<p>
  <seg type="phrases">
    <seg type="phrase" xml:id="p77925b0a-0b9f-462b">
      <num type="segnum" xml:lang="en">1</num>
    </seg>
    <seg type="words" xml:id="ddcdce5e-e52d-44a9">
      <w type="txt" xml:lang="mix">ntakachu</w>
      <seg type="morphemes">
        <seg type="morph" subtype="stem" xml:id="d7f713e8-e8cf-11d3">
          <m type="txt" xml:lang="mix">ntakach</m>
          <w type="cf" xml:lang="mix" lemma="ntakacha">ntakacha</w>
          <gloss type="gls" xml:lang="en">wash</gloss>
          <gloss type="msa" xml:lang="en">v</gloss>
        </seg>
        <seg type="morph" subtype="suffix" xml:id="d7f713dd-e8cf">
          <m type="txt" xml:lang="mix">-u</m>
          <w type="cf" xml:lang="mix" lemma="-u">-u</w>
          <gloss type="gls" xml:lang="en">2sg.inf</gloss>
          <gloss type="gls" xml:lang="en">v:Any</gloss>
        </seg>
      </seg>
      <gloss type="gls" xml:lang="en">wash</gloss>
      <gloss type="gls" xml:lang="en">v</gloss>
    </seg>
    <gloss type="gls" xml:lang="en">you wash</gloss>
  </seg>
</p>
```

Figure 28: TEI rendition of interlinear annotation from FLEx preserving full structure and attributes[126]

Note that, at least when dealing in XML, in order to be rendered in a human readable manner such as in the example Figure 28 from FLEx, the multi-tiered/interlinear annotation structure needs to be transformed or associated with a stylesheet, otherwise the various blocks of information will not be human readable. This means that either a software tool with this built-in capacity or a person with the capacity to create and edit such schemas will be needed in order to provide a maximally human readable output.

[126] The one minor modification that was made is that in the first @xml:id, in FLEx, it began with a number (e.g. 77925b0a-0b9f-462b-a799-43a4c6545771) which isn't allowed in TEI, thus the letter "p" was appended (e.g. p77925b0a-0b9f-462b-a799-43a4c6545771).

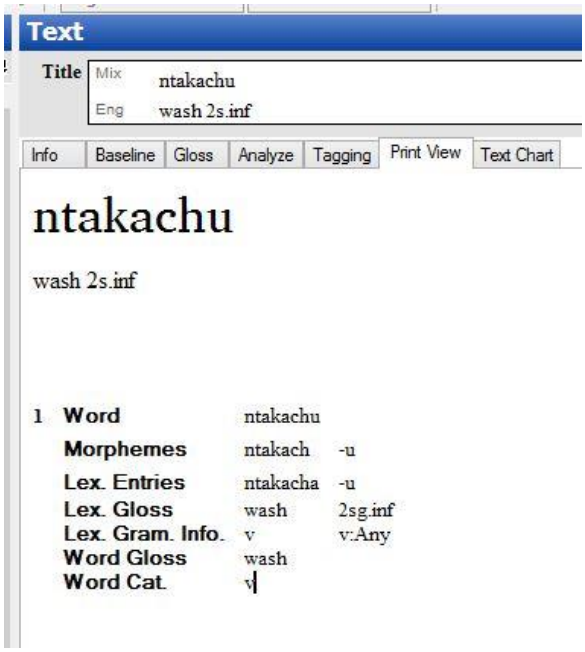


Figure 29: Print view of interlinear glossed text in FLEx

Standoff/standalone annotation is an annotation method in which the content and annotations are either: stored in a separate document (*remote standoff*), or separate from the content within the same document (*local standoff*). In XML, hyperlinks can be made between documents or within a single document, these links are defined using the XPath language to point to specific point (id values) which is the annotation target. A seminal work on standoff annotation was McKelvie et al. (1997) which addressed the issue of annotation of language corpora in the context of SGML (the predecessor of XML). Other notable works dealing with standoff annotation and TEI was Bański and Przepiórkowski (2009) which described the use of standoff annotation in the Polish National Corpus[127].

According to Gries and Berez (2017), that while unfortunately, standoff annotation remains rarely implemented in corpus annotation it is advantageous in several ways:

- where the base document is read only or very large;
- if the distribution of the source document is controlled, the annotations can still be made freely available;

[127] It should be noted that the use of <spanGrp> as the primary TEI mechanism for standoff annotation has become more common practice since Bański and Przepiórkowski (2009), see also Bański et al. (2016) and Ogrodniczuk (2011) for further discussion.

- the annotations may include overlapping hierarchies;
- it allows for alternative annotations (theoretical, general description, individual annotator variation);
- it avoids potentially highly complex documents;

Standoff annotation avoids problems which arise from the fact that XML does not allow for overlapping elements, which is important when there is a need annotation a series of elements which may, on one level be associated, but on another need to be annotated for two different features, and/or associated with different separate elements (Zinsmeister et al., 2008). Additionally, in the context of LD, keeping the annotations and analyses separate from the language content is a central aspect of best practice as per Himmelmann (1998).

Bański (2010) discusses several possible different types of semantics that can be expressed in standoff annotations: *inclusion* (to include certain components into an arrangement or grouping)*; replacement* (to replace content for corrections, normalization, etc.)*; multiple-point linking* (linking multiple elements, in TEI <link> is used)*; correspondence* (simply a mechanism to point to one or more targets and assign some kind of annotation value), *merger* (merges the attributes of a target document with those from an annotation)[128].

The Figure below from Bański (2010) shows a diagram of the example originally presented in McKelvie et al. (1997), in which <w> tokens in a read only file are grouped into <s>[129] (sentence) blocks using standoff annotation. The resulting document or rendered content is shown on the right.

[128] Bański also mentions two more types of standoff semantics whose feasibility is questioned: *inverse replacement semantics* which is defined as: the inclusion of everything, but the element pointed at, and use the annotated value instead of it; *reverse inclusion* which is a literal interpretation of the semantics CES (Corpus Encoding Standard) standoff markup, it uses standoff annotations to virtually create a resulting annotation (assumedly rather than actually creating a new document).
[129] In TEI, an alternative to <seg> (or <seg type="S"> as is used in this project) to encode sentence blocks is <s>.

Source text (src.xml)

```
<source>
<w id="w1">word1</w>
<w id="w2">word2</w>
<w id="w3">word3</w>
<w id="w4">word4</w>
<w id="w5">word5</w>
<w id="w6">word6</w>
</source>
```

Annotation document

```
<result>
<s target="w1..w3" doc="src.xml"/>

<s target="w4..w6" doc="src.xml"/>
</result>
```

Result

```
<result>
<s target="w1..w3" doc="src.xml">
    <w id="w1">word1</w>
    <w id="w2">word2</w>
    <w id="w3">word3</w>
</s>
<s target="w4..w6" doc="src.xml">
    <w id="w4">word4</w>
    <w id="w5">word5</w>
    <w id="w6">word6</w>
</s>
</result>
```

Figure 30: From Bański (2010) showing remote standoff and inclusion semantics annotation of example from McKelvie et al. (1997)

A basic example of *local* or *embedded* standoff annotation was already shown in the previous section (Figure 7) from the ISO 24624:2016 standard, in which the TEI <spanGrp> mechanism is used to annotate the <w> tokens in an utterance by pointing to the respective @xml:id values.

```
<u>
  <w xml:id="w3">John</w>
  <w xml:id="w4">loves</w>
  <w xml:id="w5">Mary</w>
</u>
<spanGrp type="phraseStructure">
  <span from="#w3" to="#w5">
    <span>S</span>
    <span from="#w3" to="#w3">
      <span>NP</span>
      <span from="#w3" to="#w3">N</span>
    </span>
    <span from="#w4" to="#w5">
      <span>VP</span>
      <span from="#w4" to="#w4">V</span>
      <span from="#w5" to="#w5">
        <span>NP</span>
        <span from="#w5" to="#w5">N</span>
      </span>
    </span>
  </span>
</spanGrp>
```

Figure 31: Example of local standoff annotation as used in ISO 24624

Note that the method of standoff annotation applied to the MIX corpus and described herein (section 6.4) is *local/embedded* (within the same document) distinguished from a related

methodology of *remote standoff annotation* in which the annotations are stored in a completely different file from the original. This is not the only method for encoding standoff annotation in TEI, there is a new element <standOff>[130] in which the annotations are placed, which can be <spanGrp> or a wide array of other types of elements which are associated with a target for which they are designated as serving as an annotation.

It should be stated that while the method of using standoff annotations is of course advantageous in many ways, as: a) it is in line with the recommended practice of not mixing description and interpretation; and b) it allows for an infinite number of alternate or supplemental annotations by other editors, it also has its disadvantages which are not insignificant. Despite the flexibility and advantages, the use of a standoff annotation in corpora is unfortunately not widely adopted, due to the fact that it requires a dedicated tool to implement, search and retrieve data, without which is it not a realistic, or practical choice for an ordinary working linguist (OWL) (Bański, 2010; Gries and Berez, 2017)[131]. This is particularly true in the cases of OWL's working on endangered and indigenous languages in which usable community output is a central goal.

Furthermore, even using the tools of Oxygen XML editor and the ODD schema customization and templates which allows for keyboard shortcuts instant insertion of blocks of XML code and pop-up suggested values from a preset inventory, it is still time consuming to annotate using <spanGrp>. Additionally, in standoff annotations, the more annotations you have, the further away from the original content you are, which makes selecting the necessary xml:id's to point to incredibly cumbersome.

Finally, *relational databases* are similar to standoff/standalone formats in that the annotations are stored separately from the annotated content except that these databases require that all content be broken up, and the main content to be annotated does not keep its integrity and

[130] https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-standOff.html
[131] Wörner (2009) described the tool 'Sextant' (https://exmaralda.org/en/sextant-en/) which was designed as a tool for carrying out standoff annotations in spoken language transcription based on the Linguistic Annotation Framework (Ide and Romary, 2004) with component software within the EXMARaLDA system. However, this software is described as a "work in progress" on the EXMARaLDA and is not available for download.

requires a query function (and thus a competent programmer) to put it back together again (see Abiteboul et al, 2014 for discussion comparing XML vs relational database data structure). While SQL and other relational databases adhere to the principle of separation of source and annotation which potentially allow for overlapping hierarchies, and fast retrieval, there is a potential downside which is that the underlying model is more complex, and they require competent programmers to manage the data, maintain an extra software infrastructure and implement a user interface (see Chiarcos et al., 2008 for a discussion of an implementation of such as system using ANNIS software and the GraF model).

In annotating a LD corpus, due to a lack of data, most projects do not have the option to take advantage of automatic annotation tools available for major language (e.g. POS taggers, parsers, lemmatizers, etc.), thus it is usual for annotation to be carried out manually, which, the more features and deeply you annotate, may exponentially increase the labor needed. So, in the case of choosing a data model, it is often the case that OWL's will choose a tool that reduces as much of the labor as possible. There has been increasing progress made on the front of morphological parsing, particularly in the use of Finite State Transducer and neural analyzer systems for Crimean Tata, Yupik, and Arapaho respectively (see: Tyers et al., 2019; Schwartz et al., 2019; and Moeller et al., 2019).

4.4.3 Description Data Formats and Standards

In language documentation, apart from the general agreement on the values of XML and Unicode, there is not a widely adopted recommendation or practice for any particular annotation standard for encoding lexicons (e.g. dictionaries) or grammatical descriptions and inventories. As is the case with spoken language transcription and corpus annotation, for many LD projects, the choice of description format and standards are driven by the tools they use and possibly the archives they deposit with.

4.4.3.1 Lexicons and Dictionaries

Regardless of the tool and methods, developing lexicons and dictionaries is fundamentally a task of lexicography, which is of course on the language description side of the description vs documentation divide. Nonetheless, with maybe a few exceptions, the kind of data

collected for the creation of a dictionary or lexicon is generally the same for any standard semasiological dataset: namely, lexical entries, which will usually contain lemmas or headwords; orthographic and/or phonetic forms, possibly variants of each; grammatical and inflectional information; senses, definitions, translations and/or glosses (if multilingual, as most are in the LD context), usage examples (potentially for a project corpus), semantic domain, semantic relations, register, images, etymological information: forms, senses, dates, attestations, description, classification of etymological process (provenance, form changes, sense changes, etc.); bibliographic citations; cross references and much more. In the digital context, entries may also include speech files, videos, links to sources, and hyperlinked cross references.

For lexicons and/or dictionaries, there are two major standards that can guide modeling and encoding work on lexical data. On the one hand, the *Dictionary chapter* of the TEI guidelines (which is an open and community-based standard) and the ISO LMF (Lexical Markup Framework) standard (ISO 24613), which has been developed within ISO committee TC 37/SC 4 and was a basis for the OntoLex-Lemon model (McCrae et al., 2017) in the domain of linked open data (LOD) (for an in-depth discussion on the use of linked data for language resources see Gracia et al., 2014).

4.4.3.1.1 TEI Dictionaries

A core lexicographic component of the TEI is indeed the Dictionary module[132] which defines the components used for encoding lexica. TEI dictionaries are used to encode a wide variety of born-digital and retro-digitized dictionaries, for the purpose of human or machine oriented output. Due partially to the fact that the structural organization and thus, encoding needs of born digital and retro-digitized dictionaries are so diverse, there is a large degree of variability in the specifics of how any given TEI dictionary is structured. However, the basic components of a typical entry, and its primary child element blocks are shown in Figure 32 below.

[132] https://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html

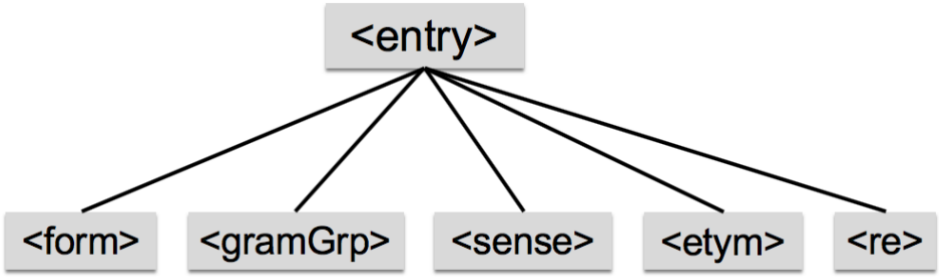


Figure 32: Most fundamental components of TEI entry

As shown in Figure 32 the primary components of an entry are <form> which generally will have a headword <form type="lemma">, inflected and variant forms can also be included using @type, as ”inflected” and “variant” respectively. Within <form> orthographic and phonetic forms are encoded distinctly using <orth> and <pron>. In a complex entry (or related entry) such as a compound or multi-word expression, the contents of <orth> or <pron> can be further segmented and linked to separate entries for each component part using <seg @corresp> (see Figure 33). Additionally, pointers to sources and/or media files such as audio can be embedded within form or elsewhere using the <media> element (see sections 6.3.4.1 and 7.4.4 for examples of how this is done in this project). Related entries (<re>) are embedded within an <entry> and can have all of the same structures and elements.

Grammatical information can be encoded for the main entry or within a specific sense of a word. An entry can have as many separate and/or embedded sub-senses as needed. Senses can have multilingual translations using <cit type="translation">, definitions <def>, examples <cit type="example">, domain and register using a typed <usg>. Cross-references can be added in a number of different places, typically in <sense> or <etym>. Within <sense> they typically denote sense relations e.g. synonyms, antonyms, meronymy, etc., all are expressed using the @type attribute.

Etymologies are contained within the <etym> element, they can be recursive, and can occur on the level of <entry> or optionally within a sense in the case that the sense has a specific etymology. Etymological processes can be typed, e.g. <etym type="borrowing">, or <etym

type="metaphor"> in the case of a sense etymology. Etymons, which can include forms, senses, and grammatical information are encoded using <cit type="etymon"> and cognates from related languages can either be encoded along the same lines as etymons (e.g. as <cit type="cognate">) or <xr> depending on the context. Prose can be encoded using <seg type="desc"> (for more details see section 7.5; Bowers and Romary, 2017; and Bowers and Romary, 2018b).

Figure 33 shows an example of a TEI encoded example of the same entry as the previous figure above modelled in LMF-UML also from Romary et al. (2019).

```
<entry>
  <form type="lemma" xml:id="center_form">
    <orth>center</orth>
    <pron>ˈsɛntɪ</pron>
    <gramGrp>
      <pos>noun</pos>
    </gramGrp>
    <form type="variant">
      <orth>centre</orth>
      <pron>ˈsɛntə˞</pron>
      <usg type="geo">U.K.</usg>
    </form>
  </form>
  <sense>
    <def>the point around which a circle or sphere is described</def>
    <cit type="example">
      <quote>earth's center</quote>
    </cit>
  </sense>
  <sense>
    <gramGrp>
      <pos>verb</pos>
    </gramGrp>
    <def>place in the middle</def>
    <cit type="example">
      <quote>center the picture on the wall</quote>
    </cit>
  </sense>
  <re type="multiWordExpression">
    <form>
      <seg corresp="#dead_form">dead</seg>
      <seg corresp="#center_form">center</seg>
    </form>
  </re>
</entry>
```

Figure 33: TEI encoded example of partial entry for *center* from Romary et al. (2019)

As mentioned, there have only been publications from only two different projects detailing the use of TEI for the creation of a digital dictionary for indigenous languages Czaykowska-Higgins and Holmes (2013), Czaykowska-Higgins et al. (2014); and Bowers and Romary (2017). Though the principles underlying the adoption of TEI as a markup format are supported and generally accepted within the LD community, particularly for those with a technological orientation, it has been slow to catch on, which likely has to do to several reasons.

A common criticism of TEI, which is often used as a counter argument to its adoption is that there are too many options for encoding the same features, and that given this variety and the number of projects that have already adopted TEI and have encoded a given feature in any number of ways, it is not an ideal format. In order to address this problem within the domain of lexicography (specifically the TEI Dictionary guidelines), the TEI Lex0 initiative (Bański et al., 2017; Romary and Tasovac, 2018) has been undertaken in order to provide a reduced array of encoding options for TEI digital dictionaries and a format to serve as baseline for interoperability of dictionaries in TEI and other formats as well, including OntoLex-Lemon (see McRae et al. 2019).

Additionally, ordinary working linguists (OWL's), especially OWL's in LD, who may not have extensive training technological issues are more likely to prefer to use lexicon development software with a GUI interface that provides the features and takes care of the technical aspects of data structure for them on the back-end[133] and that they can immediately use. Thus, because the tools they are using (particularly FLEx) do not use TEI, or provide an output option, it has not taken off in the domain of LD. Moreover, because TEI does not have a user-friendly software that can edit TEI dictionaries[134] in addition to the various other functions tools like FLEx have (see section 4.4.4.2), it has not, and is unlikely to become a widely adopted data standard in LD.

[133] It should be stated that despite the all-inclusive user interface and background data management functions, tools like FLEx are not entirely user-friendly as there is a learning curve and there are often many glitches. The lack of control and transparency of the system also makes the users highly dependent on the developers.
[134] This issue of a lack of native TEI editing software also applies to the areas of corpora and transcribed speech. In the domain of lexicography. However, Bowers, Stöckle, Breuer et al. (2019) describes the creation of a lexicographic editor tool which produces native TEI articles for the Dictionary of Bavarian Dialects in Austria project (WBÖ).

4.4.3.1.2 Lexical Markup Framework (LMF)

The LMF, originally published in 2008 and currently under revision in IS0 TC37, in its original state was, and remains a UML based model and is therefore decoupled from any specific serialization format (e.g. XML). However, parts 5 (ISO 24613-5) and 4 (ISO 24613-4) define two possible serializations in the LBX (Language Base Exchange) (George, 2013) and TEI respectively (see Romary et al., 2019 for an overview of the work in progress and Romary, 2015b for a preliminary mapping between TEI and LMF). The other components of the LMF reserialization are as follows:

- ISO 24613-1 - Core model: which defines basic classes required to model a baseline lexicon
- ISO 24613-2 - Machine Readable Dictionaries (MRD) model: contains components providing deeper specification of lexical description encapsulated within the core model. *Form* is for instance differentiated into *Related Form, Word Form, Stem* and *Word Part*
- ISO 24613-3 - Diachrony-Etymology[135]: categories related to word and meaning origin and change are defined
- ISO 24613-6 - Syntax and Semantics: semantic and syntactic components are gathered in this extension to be revised and integrated with the first three parts of the standard
- ISO 24613-7 - Morphology: morphology package will be defined in a separate part of the standard and will also be interconnected with the first three parts of the standard

The UML diagram in Figure 34 below shows the abstract modelling for the lexical entry *center* taken from Romary et al. (2019).

[135] I am a co-project leader with Fahad Khan of the Istituto di Linguistica Computazionale A. Zampolli– CNR on ISO 24613-3 Diachrony-Etymology

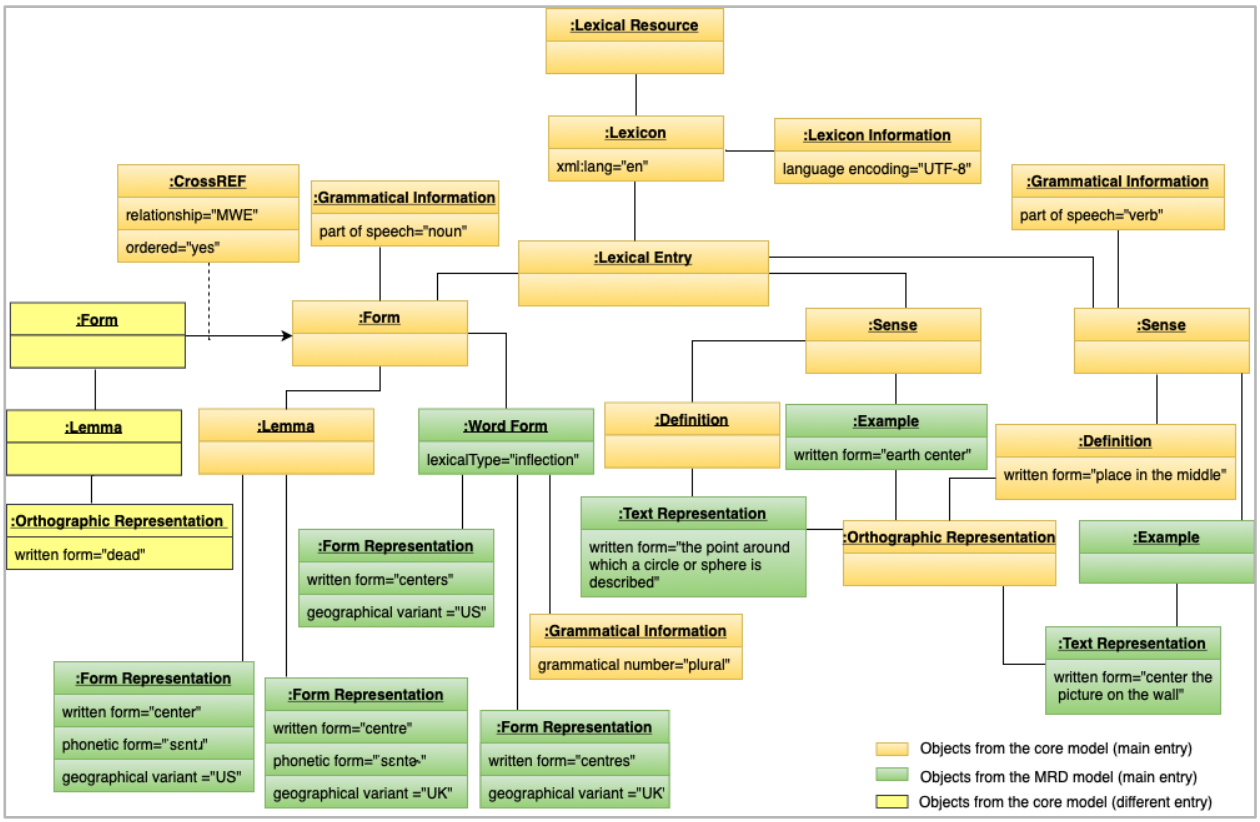


Figure 34: Example of the lexical entry "Center" encoded using the core (ISO 24613-1) and MRD (ISO 24613-2) metamodels from Romary et al. (2019)

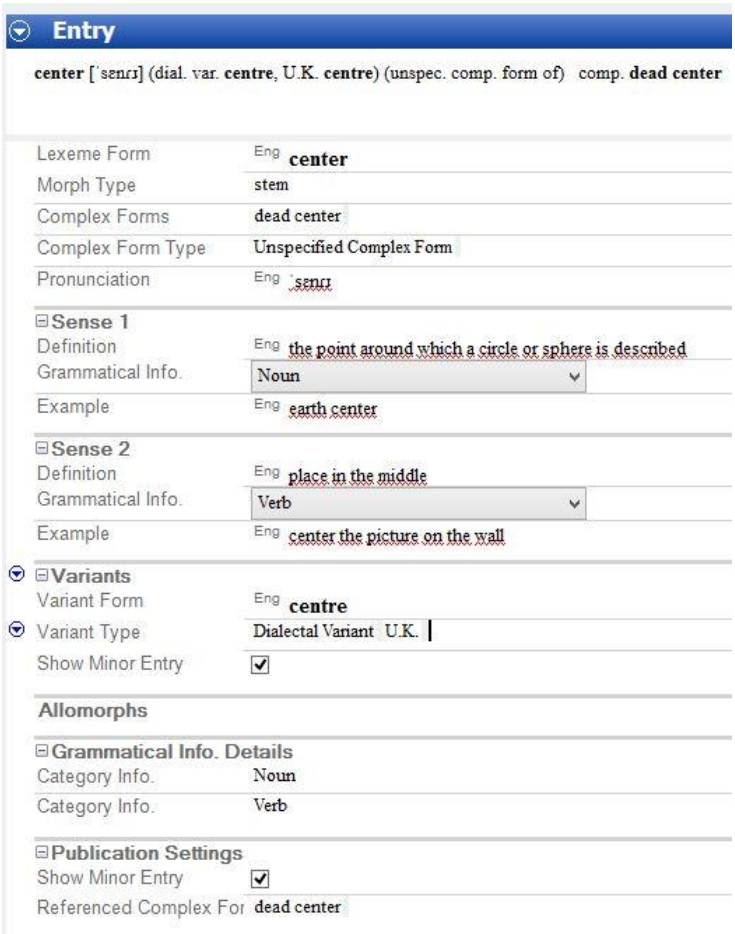
LMF is not widely adopted in lexicography much less in language documentation though some notable exceptions are: the DOBES archive, which uses LMF (as well as EAF - Elan Annotation Format for spoken language) as the underlying schemas[136], and the LEXUS software[137] from MPI uses LMF as the basis for its lexicon structuring (Ringersma and Kemps-Snijders, 2007). Finally, being a UML graph-based model, LMF is also a natural bridge between XML-based TEI and RDF based lexical data models such as Ontolex-Lemon, which was structurally designed based on the LMF data model (see McCrae et al., 2017).

[136] https://www.mpi.nl/corpus/a4guides/a4-guide-dobes-format-encoding.pdf
[137] https://tla.mpi.nl/tla-news/lexus-3-0-release-candidate/

4.4.3.1.3 LIFT

Another ‘standard’ that is necessary to mention is SIL’s LIFT (Lexicon Interchange FormaT)[138] which allows for the exchange of XML dictionary data between the SIL programs of WeSay, Lexique Pro and FLEx. While this format or standard is not in use outside of the SIL software ecosystem, the popularity of FLEx in particular amongst LD linguists, has meant that there are many projects whose XML dictionaries are exported into LIFT format. For the most part, the LIFT data produced by FLEx has the capacity to encode much of the same types of lexical information as in TEI, however it is not as transparent in the process of editing, as it is only through the tool that users can manipulate it (unless it is exported).

Although in the FLEx Fieldworks Language Explorer system the entry for ‘center’ looks as if it is all arranged in the same entry (e.g. Figure 35), examination of the actual entry in the



[138] https://github.com/sillsdev/lift-standard

LIFT export format shows that it is actually several entries, in which the (UK) variant *centre*, the multi-word expression *dead center* are separate entries.

Figure 35: Example of identical entry for *center* in Lexicon Edit view in FLEx tool

```
<entry dateCreated="2020-02-27T16:40:21Z" dateModified="2020-03-02T12:45:35Z"
  id="center_5a11" guid="5a11">
  <lexical-unit>
    <form lang="en"><text>center</text></form>
  </lexical-unit>
  <trait name="morph-type" value="stem"/>
  <relation type="_component-lexeme" ref="">
    <trait name="complex-form-type" value="Unspecified Complex Form"/>
  </relation>
  <pronunciation>
    <form lang="en"><text>ˈsɛnɹɪ</text></form>
  </pronunciation>
  <sense id="8440" order="0">
    <grammatical-info value="Noun"/>
    <definition>
      <form lang="en"><text>the point around which a circle or sphere is described</text>
      </form>
    </definition>
    <example>
      <form lang="en">
        <text>earth center</text>
      </form>
    </example>
  </sense>
  <sense id="0a60" order="1">
    <grammatical-info value="Verb"/>
    <definition>
      <form lang="en"><text>place in the middle</text></form>
    </definition>
    <example>
      <form lang="en"><text>center the picture on the wall</text></form>
    </example>
  </sense>
</entry>
```

Figure 36: Example of identical entry for “Center” LIFT export from FLEx tool

The significance of this is that the tool is presenting the data in a way that is not the same as the actual underlying (or at least the *exported*) data structure, thus in the case of LIFT and

FLEx, the tool design is the controlling factor in the way that the data is presented, which is different from the way that the standard format is structured[139].

4.4.3.2 Grammatical and Other Annotation Inventories and Features

In any corpus, lexicographic, as well as LD project, the grammatical and lexical (including semantic) features utilized in the annotation of corpus and/or dictionary should be explicitly declared somewhere so that both editors, and users of the resources can decipher annotations. Such inventories also serve to document the inventory of a language's features. Feature inventories, which may potentially include tags and definitions may be declared in the corpus, dictionary/lexicon documents themselves or in a separate location; the approach differs by project choice or by however the given software. Because in most linguistic projects at least a portion of the categories to be annotated are likely found across the world's languages, there are standardized inventories and ontologies designed to be reused, which both foster interoperability and saves editors from re-inventing the wheel. In this section I discuss these issues along with the specifics of how different tools and markup systems store and structure the inventory of categories used in annotation.

4.4.3.2.1 ISOcat

The ISO Data Category Registry (DCR) was created in 2008 in order to provide a database of standardized data concepts relevant to linguistic data, analysis and annotation, (Ide and Romary, 2004; Kemps-Snijders et al., 2008, 2009; Windhower et al., 2010; Windhower and Wright, 2012; Wright et al., 2013). Categories could be proposed and defined by the user community and referenced in data using each categories' URI which serves as its persistent identifier (PID), examples of such categories are: /*part of speech*/[140], /adjective/[141], etc. (ibid). As mentioned previously, ELAN has a feature in which annotation features can be associated with specific ISOcat data categories, this is also the case in both LMF and TEI data models.

However, as there were several key areas in which the previous system was seriously flawed as of 2014 (see Broeder et al., 2014), ISOcat has been undergoing a full data migration

[139] Views in FLEx can be created and the dialogue box can be customized.
[140] http://www.datcatinfo.net/datcat/DC-5660
[141] http://www.datcatinfo.net/datcat/DC-5748

and reorganization into a new registry[142] hosted by Interverbum Technology[143]. This change was necessary due to several flaws in the old system, including: duplicate categories, a lack of a clear definition of what a data category should entail, a flawed taxonomic macro-structure, unused features and functions, and a severe lack of systematic vetting of proposed new categories (which led to a proliferation of over 6,000 categories). At the time of submission, the reorganization process is still underway and such a long gap has undeniably been a setback for the cause of lexical standards.

4.4.3.2.2 Ontologies and Other Annotation Tagsets

Alternatively, there are additional, more structured annotation vocabularies in the form of ontologies designed for linguistic annotation in the context of linked open data, the most notable of which is General Ontology for Linguistic Description (GOLD)[144] (Farrar and Langendoen, 2003) and OLiA (Charcos and Sukhareva, 2015). OLiA, also a system primarily in use in the domain of LOD serves, though it differs from GOLD in that is designed as a means to integrate linguistic terms and concepts from multiple annotation vocabularies, rather than to serve as a single concept/tagset like GOLD[145].

GOLD was a product of the EMELD project and was created in order to provide a common interoperable lexical annotation vocabulary for all varieties of linguistic data. GOLD actually preceded ISOcat, and the data categories created, as well as their URI's and definitions therein were preserved and integrated into the ISOcat repository. The final version was in 2010 and it no longer remains actively developed. The ending of the project and the lack of subsequent maintenance was unfortunate as the ontology was not comprehensive enough to express all necessary linguistic concepts, which lead to problems in adoption, including in this project. GOLD is serialized both in OWL (Web Ontology Language) RDF and XML, for the sake of consistency, a sample feature is shown in the XML format below in Figure 37, note that the hierarchy is encoded with the value of the @parent attribute which defines 'MasculineGender' as a subclass of 'GenderProperty'.

[142] http://isocat.tbxinfo.net/
[143] At the time of publishing, this new system is not yet publicly available.
[144] http://linguistics-ontology.org/gold
[145] Another notable system is LexInfo (https://lexinfo.net/) which is a module of the OntoLex-Lemon system (see also Cimiano et al., 2011)

```
<concept uri="http://purl.org/linguistics/gold/MasculineGender"
parent="http://purl.org/linguistics/gold/GenderProperty">
   <label>MasculineGender</label>
   <definitions>
     <definition lang="eng">A gender property established on the basis of agreement, to which nouns may
be assigned based on semantic or formal criteria. In semantic gender systems, nouns belonging to the
masculine gender typically denote male humans as well as nouns meeting certain physical criteria. Some
gender systems differentiate masculine nouns from all other nouns (e.g. masculine/other or male
human/other), while others differentiate masculine, feminine and neuter nouns or several different gender
classes. [Corbett 1991: 30]</definition>
  </definitions>
</concept>
```

Figure 37: Concept 'MasculineGender' in GOLD linguistic ontology

OLiA (Charcos and Sukhareva, 2015) is a set of ontologies for linguistic annotations that is designed to mediate between different tag sets covering common linguistic phenomena (Chiarcos et al., 2008). An example of the type of issues OLiA addresses can be found in the variety of tag annotations applied to the English possessive determiner *her* which in various corpora cited by Chiarcos et al. (2008) as: PP$, TB, PRP$, DD, PRON(poss, sing), and APPGf. The OLiA system contains four types of ontologies, the first of which is the Reference Model, which defines the various linguistic features and categories which are: *MorphosyntacticCategory*, *SyntacticCategory*, *MorphophonologicalCategory*, *MorphophonologicalProcess*, *MorphosyntacticFeatures*, *SyntacticFeatures and SemanticFeatures*. The Annotation Model (which defines the various annotation schemes and tagsets); for each Annotation Model, there is a Linking Model, which serves to link and define the relationships between the properties and concepts in terms of the Reference Model. Finally, there is the External Reference Model which allows for the integration of external terminological repositories on the condition that they are encoded in OWL2/DL[146], which then can be linked via the Linking Models to the Reference Models.

Whereas GOLD, which emerged out of the EMELD project in which the central purpose was to provide technological recommendations and infrastructure to support the application of technology to the preservation of the world's languages, enjoyed a more general usage community, the target community and adoption of OLiA seems to be more limited to highly technically oriented to those carrying out highly automated and complex NLP and LOD tasks.

[146] https://www.w3.org/TR/owl2-overview/

Nonetheless, while the widespread adoption of tools like FLEx, which have their own built-in tagsets undoubtedly help in the creation of lexica that use a common terminology, the issue of bridging the gaps between different the annotation sets used across language projects is an important one and should receive more attention going forwards in the domain of language documentation. For more discussion on the use of OWL/RDF for corpus annotation interoperability, see Chiarcos (2012) with the POWLA system which is an OWL/DL implementation of the PAULA data model (Dipper, 2005; Chiarcos et al., 2008).

4.4.3.2.3 TEI and ISO 24610-1 Feature Structures

Components of the TEI were the basis for ISO 24610-1:2006 (*Language Resource Managemen*t — Feature structures — Part 1: Feature structure representation). Chapter 18 of the TEI guidelines[147] is dedicated just to feature structures which can be used in a number of ways, one is to either declare an inventory or directly annotate lexical or conceptual features for linguistic analysis, and the other is as an abstract means of grouping and relating structured information. With regard to the topic at hand, I will only discuss their usage as a mechanism for declaring an inventory of annotation.

The structure of the lexical features are just one part, the other is the actual features which are determined by the editor. Many different annotation tagsets exist from different linguistic domains which users can choose to apply to their projects. These inventories can be stored in a number of related, but distinct manners, notably in the form of: feature structures, feature declarations, or feature libraries, which depending on the particular needs of the project, can be structured in a number of ways[148] (see Figures 38, 39, 40 and 41 below).

In this first example (Figure 38), which is in accordance to the approach taken in this projects' feature inventory (see section 6.4.1), the <fs> is simply an empty container for the feature (gender), which can take one of two values (FEM or MASC) which are expressed by the <symbol> element, all of which are contained in a <vAlt> (value alternation). The @value contains the full name of the feature and the @xml:id contains the value that is also the tag that

[147] https://www.tei-c.org/release/doc/tei-p5-doc/en/html/FS.html
[148] For a full description of the manners in which feature structures can be used and declared see TEI Guidelines chapter 18.

will be used in the annotation of a corpus. Note that <fs>, <f> as well as <symbol> can be directly assigned to a standardized data category (such as GOLD) using a persistent identifier (PID) in the @corresp, additionally there remains the @dcr:datacat and @dcr:ValueDatacat which could be used to attribute the feature with a category or value respectively from the ISOcat vocabulary[149].

```
<fs>
  <f name="gender">
    <vAlt>
      <symbol xml:id="FEM" value="feminine"/>
      <symbol xml:id="MASC" value="masculine"/>
    </vAlt>
  </f>
</fs>
```

Figure 38: ISO 24610-1 and TEI Feature structures for gender

Another manner of declaring the grammatical features in a TEI project is with <fsdDecl> (feature system declaration), in which each group of features which are grouped according to <fsDecl> (feature structure declaration) and <fDecl> (feature declaration) (not as <f> or <fs>), as in the previous example. These can be declared and defined in different layers and they can be grouped in a way that expressed different grammatical, or conceptual functions or relations in which the given sub features are involved, e.g. Figure 39 shows an example from the TEI guidelines with the declaration of the grammatical function of number agreement in English.

```
<fsdDecl>
  ....
  <fsDecl type="Agreement">
    <fsDescr>This type of feature structure encodes the features
      for subject-verb agreement in English</fsDescr>
    <fDecl name="PERS">
      <fDescr>person (first, second, or third)</fDescr>
      <vRange>
        <vAlt>
          <symbol value="1"/>
          <symbol value="2"/>
          <symbol value="3"/>
        </vAlt>
      </vRange>
    </fDecl>
    <fDecl name="NUM">
      <fDescr>number (singular or plural)</fDescr>
```

[149] These attributes belong to the TEI class att.datcat

```
        <vRange>
          <vAlt>
            <symbol value="sg"/>
            <symbol value="pl"/>
          </vAlt>
        </vRange>
      </fDecl>
    </fsDecl>
  </fsdDecl>
```

Figure 39: ISO 24610-1 and TEI Feature structures for number agreement in English

Figure 40below shows a third major way of grouping the features, within an <fvLib> (feature-value library) which can contain groups of features in <fs> and its various child elements.

```
<fvLib n="Major category definitions">
  <!-- ... -->
  <fs xml:id="N" type="noun">
    <!--  noun features defined here -->
  </fs>
  <fs xml:id="V" type="verb">
    <!-- verb features defined here -->
  </fs>
</fvLib>
```

Figure 40: ISO 24610-1 and TEI Feature Library for basic parts of speech

Additionally, there is <fLib> (feature library)[150] which can occur as a child of <fsDecl> which does nearly the exact same thing as <fsLib> with the exception that it uses <f> as the direct child rather than <fs>.

With regard to encoding semantic concepts such as a domain inventory, any one of these structures could be utilized, Figure 41 shows an example using a simple <fvLib>. The value to be used when tagging in a corpus is in the @xml:id of the given <f> or <symbol>, and if desired, the @corresp can be used to link to a URI of the concept from an external ontology or knowledge source (see section 6.4.7 and 7.4.1 for discussion in the context of this project).

```
<fvLib>
```

[150] https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-fLib.html

```
<fs type="domains">
  <f name="Universe" xml:id="Universe">
    <vAlt>
      <symbol value="Sun" xml:id="Sun"/>
      <symbol value="Wind" xml:id="Wind"/>
      <symbol value="Sky" xml:id="Sky"/>
      ...
    </vAlt>
    <!-- other domains here -->
  </f>
</fs>
</fvLib>
```

Figure 41: Example of the use of TEI feature structures to store/define semantic domain inventory

One downside is that there is no possibility of writing descriptive information such as definitions for each concept in any of the options[151]. Also, a problem is that whereas <string> would seem to be a perfect element for specifying multilingual versions of a term, it is only allowed to occur a single time within a <f>. These are each problems that need addressing in the TEI.

4.4.3.2.3.4 Grammatical and conceptual features in FLEx

The manner in which FLEx defines grammatical annotation features is in a format called LIFT Ranges, which are arranged and function similarly to TEI/ISO feature structures. One difference is in the fact that they are not taken from any standard set of data categories, nor is there any formal manner of associating them with such other than manually adding the new categories (though users can add custom categories).

The grammatical categories in the XML data structure, each larger feature set (e.g. etymology, grammatical information, semantic domain, person feature value, etc.) has a <range> block within which each member of the feature category is included in a <range-element>.

```
<range id="pers-feature-value" guid="55706aa1-2381-45a6-bba2-ea489bb4a636">
  <range-element id="1" guid="87395a09-b451-4311-b6df-7e14656dfd11">
    <label>
      <form lang="en"><text>first person</text></form>
    </label>
    <abbrev>
```

[151] In <fsdDecl> it is possible to include a <fsDescr> for each <fsDecl>, and a <fDescr> for each <fDecl> however this is insufficient as it only can be applied to a group of categories rather than to each individually.

```
        <form lang="en"><text>1</text></form>
      </abbrev>
      <description>
        <form lang="en">
          <text>First person deixis is deictic reference that refers to the speaker, or
            both the speaker and referents grouped with the speaker.</text>
        </form>
      </description>
    ...
    </range-element>
  ….
</range>
```

Figure 42: Example of annotation feature for 'first person' from FLEx LIFT Ranges

In the instance of a <range-element> for the category *1st person singular* shown in Figure 42, the category is given an @id value ("1"); a full value with the <label> element block, and the abbreviated value that is shown and used in interlinear glosses is encoded in the <abbrev> element block. A description/definition of the function of the feature is encoded in <description>.

On the level of semantics, the inventory of domains are arranged in a similar manner, with the difference that it is even more structurally shallow. Specifically, whereas with regard to grammar, each category gets a separate <range>, in the domain inventory, all lower levels, even those with member concepts, are simply encoded in <range-element>. The hierarchy is expressed by the combination of the value of the attribute @id, which for example on the range-element for 'Universe, creation' is '1', whereas for its member domain, 'Sky' is '1.1' and the use of @parent on a sub-ordinate domain to point to their direct parent. Note that this is the same method as seen in the XML serialization of the GOLD ontology in Figure 37.

```
<range id="semantic-domain-ddp4">
  <range-element id="1 Universe, creation" guid="63403">
    <label><form lang="en"><text>Universe, creation</text></form></label>
    <abbrev>
      <form lang="en"><text>1</text></form>
    </abbrev>
    <description>
      <form lang="en">
        <text>Use this domain for general words referring to the physical universe. ….'.</text>
      </form>
    </description>
  </range-element>
  <range-element id="1.1 Sky" guid="999581" parent="1 Universe, creation">
```

```
      <label><form lang="en"><text>Sky</text></form></label>
      <abbrev>...  </abbrev>
      <description>
        <form lang="en"><text>Use this domain for words related to the sky.</text></form>
        </form>
      </description>
    </range-element>
….
</range>
```

Figure 43: Example of domain features from FLEx's domain inventory in FLEx's LIFT-ranges format

One issue with the FLEx system of domains (Moe, 2003) is that there is a bit of equivocation between semantic domain (as a general topical and conceptual category) and concept, e.g. in the FLEx system, *Beautiful* and *Ugly* are domains whereas according to Cognitive Grammar a domain is a conceptual entity of varying complexity that provides a knowledge context or background information against which a lexical concepts are understood in language (Langacker, 1987; Evans and Green, 2006). Thus, one problem with the FLEx data model is that it can only ground domains in an external conceptual hierarchy and that hierarchy does not properly distinguish concepts (which should be associated with the sense) and domains, which should be a higher level semantic grouping a sense is associated with.

It is clear that the systems of feature structures used in the TEI and the lift-ranges data structures are compatible and should be mappable between each given system. The TEI features structures offer a wide variety of encoding grammatical feature, or conceptual inventories amongst other possible functions, however there remain several basic functions that somehow have not been clearly established, most notably: a) defining multiple iterations of a single feature such as full form, abbreviated and/or multiple languages; b) allowing for definitions of a single feature in-line, with the possibility to include examples.

4.4.3.2.3.5 Controlled Vocabularies in ELAN

ELAN can store grammatical and other controlled vocabulary features in XML files which are called ".ecv" files (External Controlled Vocabulary). The features in a controlled vocabulary file can be included within an ELAN template which is simply an empty EAF file (described above) used to store models and settings for ELAN transcriptions. The XML structure of the features is identical in both file types and is shown below.

```
<CONTROLLED_VOCABULARY CV_ID="Person">
    <DESCRIPTION LANG_REF="und">Grammatical category 'Person' as used in Mixtepec-Mixtec corpus annotation inventory and grammar</DESCRIPTION>
    <CV_ENTRY_ML CVE_ID="cveid_671cb">
      <CVE_VALUE DESCRIPTION="first person" LANG_REF="und">1PERS</CVE_VALUE>
    </CV_ENTRY_ML>
    <CV_ENTRY_ML CVE_ID="cveid_e6c">
      <CVE_VALUE DESCRIPTION="second person" LANG_REF="und">2PERS</CVE_VALUE>
    </CV_ENTRY_ML>
    <CV_ENTRY_ML CVE_ID="cveid_4f9aa">
      <CVE_VALUE DESCRIPTION="third person" LANG_REF="und">3PERS</CVE_VALUE>
    </CV_ENTRY_ML>
  </CONTROLLED_VOCABULARY>
  <CONTROLLED_VOCABULARY CV_ID="Number">
    <DESCRIPTION LANG_REF="und">Grammatical category 'Number' as used in Mixtepec-Mixtec corpus annotation inventory and grammar</DESCRIPTION>
    <CV_ENTRY_ML CVE_ID="cveid_a9603ec">
      <CVE_VALUE DESCRIPTION="singular" LANG_REF="und">SG</CVE_VALUE>
    </CV_ENTRY_ML>
    <CV_ENTRY_ML CVE_ID="cveid_aff5a8f3-5f3e-472b-b39a-5d66431b1d95">
      <CVE_VALUE DESCRIPTION="plural" LANG_REF="und">PL</CVE_VALUE>
    </CV_ENTRY_ML>
  </CONTROLLED_VOCABULARY>
  <CONTROLLED_VOCABULARY CV_ID="Inclusivity">
    <DESCRIPTION LANG_REF="und">Grammatical category 'Inclusivity' as used in Mixtepec-Mixtec corpus annotation inventory and grammar</DESCRIPTION>
    <CV_ENTRY_ML CVE_ID="cveid_aee9">
      <CVE_VALUE DESCRIPTION="inclusive" LANG_REF="und">INCL</CVE_VALUE>
    </CV_ENTRY_ML>
    <CV_ENTRY_ML CVE_ID="cveid_0f71">
      <CVE_VALUE DESCRIPTION="exclusive" LANG_REF="und">EXCL</CVE_VALUE>
    </CV_ENTRY_ML>
  </CONTROLLED_VOCABULARY>
</CV_RESOURCE>
```

Figure 44: Example of controlled vocabulary inventory from ELAN '.ecv' file

4.4.4 Tools, Formats, Standards and Interoperability

The growth of digital technology for recording, storage, and management of multimedia records has potentially been the most significant technical revolution in language sciences (Seifart et al., 2018). The innovations made in the various areas of this domain has enabled the creation and management of large-scale digital archives for all types of primary (e.g. audio and video files) as well as secondary (e.g. time-aligned transcriptions, annotations, etc.) linguistic data. Endeavors such as EMELD and DOBES sought to both survey the field and technological tools available as well stimulate the field into creating them, but given the pace of technological change, as well as the evolving needs of the field, and the complexities of the various tasks needed of tools for each aspect of language documentation and description content collection,

processing, management and presentation, it is not surprising that this issue is far from being settled.

A multi-dimensional LD project may comprise of any combination of: spoken language transcription, annotation of spoken and/or text-based corpus contents (e.g. POS, translations, semantics, etc.), lexicon development, linking any of the previous with media (either source or instantiation of vocabulary), annotation vocabularies (e.g. grammatical and/or domain inventories), search and retrieval, metadata and user-oriented presentation formatting. Additionally, another important possible factor in LD contexts, especially ones involving non-technical experts is usability. Furthermore, as discussed in the preceding subsections, the issue of data format and compatibility (i.e. standards) is of major importance in this domain and at present as they both facilitate interoperability (potentially within a given project and by others) as well as reusability.

As has been the case for much of the last decade (see: Nakhimovsky and Good, 2012), the most prominent tools used in LD are FLEx, Toolbox and ELAN, and while they each do many things well, and there has been considerable progress on the front of interoperability between these and other leading programs, none covers the full range of tasks needed. In the following section I briefly overview several of the major software tools used in LD and DH and give brief descriptions of: the key functions they carry out, and how they deal with standards and their respective capacities for interoperability and/or interchange.

4.4.4.1 Spoken language transcription tools

The primary function of speech transcription tools is of course time-aligned transcription of speech and/or video the output of which may be annotated for translations, and/or any number of lexical, pragmatic, semantic, contextual or other information as needed. The following table gives an overview of the tools and the types of media they can annotate:

ELAN	Praat	EXMARaLDA	CLAN/CHAT	ANVIL	Transcriber
audio, video	audio	audio, video	audio, video	video	audio, video

Table 38: Types of media for 6 major transcription software tools

As discussed in section 4.4.2.1, transcriptions can be *single-* or *multi-layered*, single layered tools only allow for a single transcription tier whereas multi-layered can take any number of tiers for a given speaker. Single-layered tools such as FOLKER and Transcriber are most fit to be utilized when only performing simple functions such as basic transcription of dialogues. In multi-layered tools such as ANVIL, ELAN, EXMARaLDA and Praat tiers can be freely defined by the user (e.g. *orthography, ipa, pos, gloss, English*, etc.) while in others, such as CLAN/CHAT they may be predefined by the software. In LD contexts, multi-layered tools are essential as there will be a need to at very least, transcribe the speech (in a working orthography and/or IPA), and likely some kind of annotation (e.g. interlinear glossing, translation, etc.).

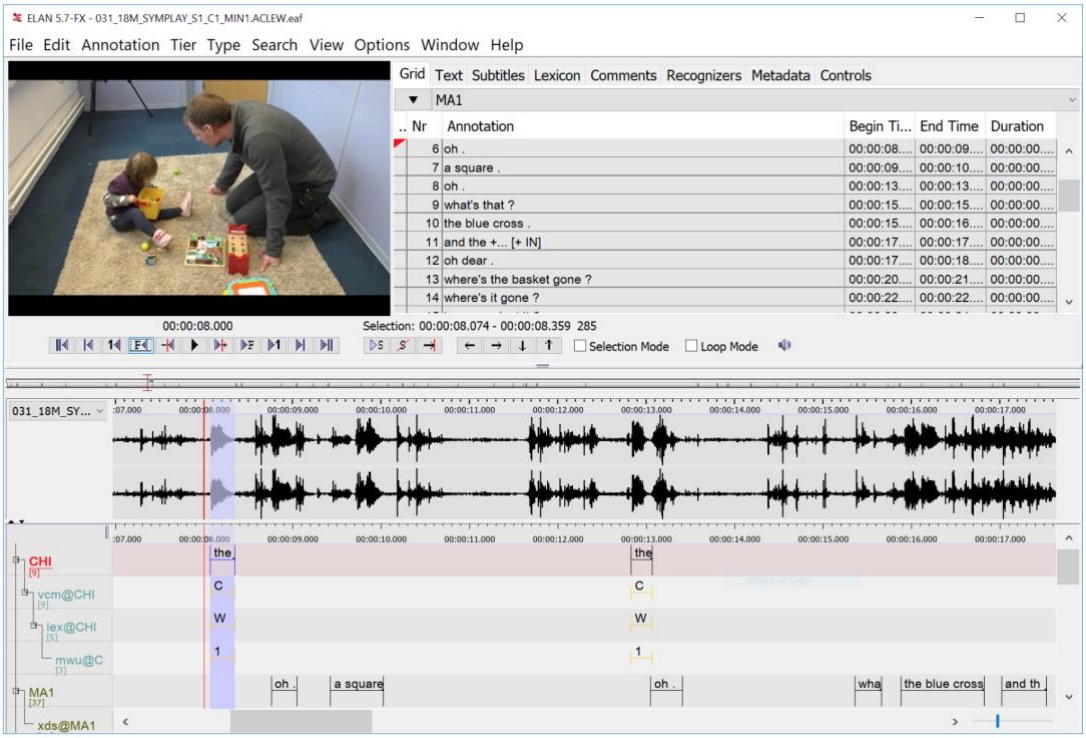


Figure 45: Example of multi-tiered annotation in ELAN[152]

As mentioned previously, ELAN and EXMARaLDA have the ability to assign specific speakers to tiers, though ELAN goes further, as has the ability to define the content of tiers beyond the strings on their labels. In the ELAN system, it is possible to define tiers as dependent tiers (with a parent), they can be assigned types (either from default values or self-defined), the

[152] Example retrieved from: https://tla.mpi.nl/tools/tla-tools/elan/

speaker annotated on a specific tier can be defined (as *participant*), the annotator can specify, and the content language can also be stated using ISO-639-3 language tags.

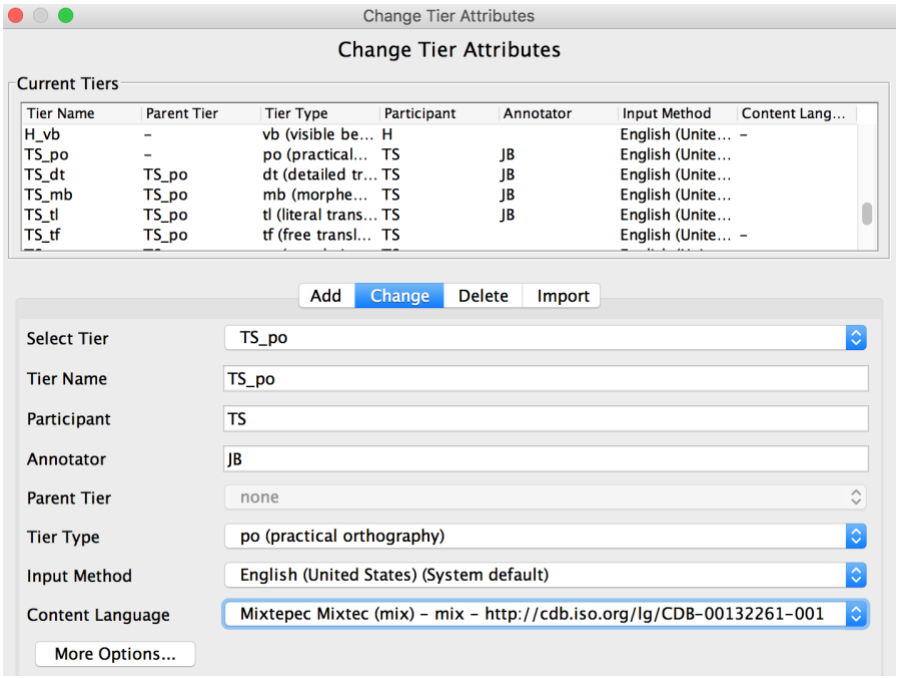


Figure 46: ELAN Tier Attributes function

With regard to full metadata records, ELAN can read and associate media files and annotations with IMDI or CMDI metadata files, though it cannot create or edit them, thus they need to be created elsewhere. EXMARaLDA (in the COMA application), can create and edit metadata records in the basic Dublin Core vocabulary (DCMI). Praat does not have the capacity for any of these tasks.

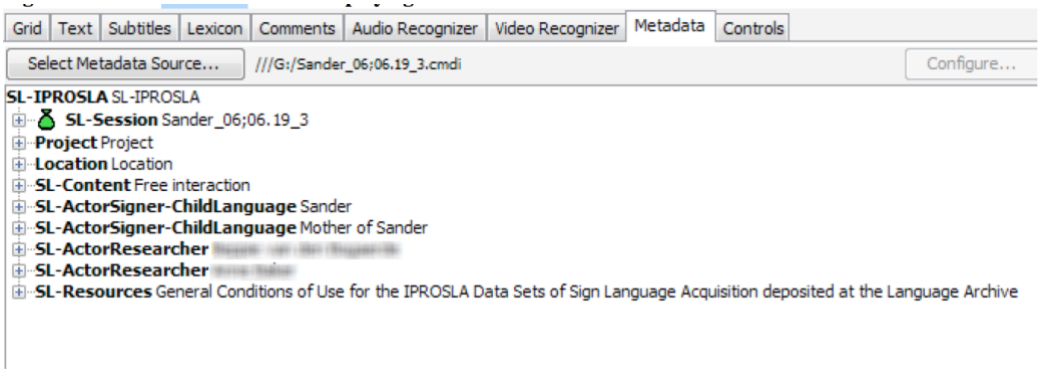


Figure 47: ELAN metadata display of CMDI metadata file(from version 5.9 guidelines)

Of all the tools for spoken language transcription, ELAN is the only tools which supports the explicit use of controlled vocabulary (CV's) inventories for annotation. CV can be from the ISOcat registry (currently lapsed, see section 4.4.3.2.1), other external sources, or they can be self-defined. Using a CV in annotation helps ensure that the annotations are less prone to individual variation that may otherwise arise. While it is of course possible to adhere to a CV in other annotation tools, in ELAN it is possible to associate specific annotation tiers with specific controlled vocabularies (e.g. POS, motion, gestures, etc.) and the tool will then allow for the possible values to appear in a suggested values drop down box, which saves time and reduces possible annotator error.

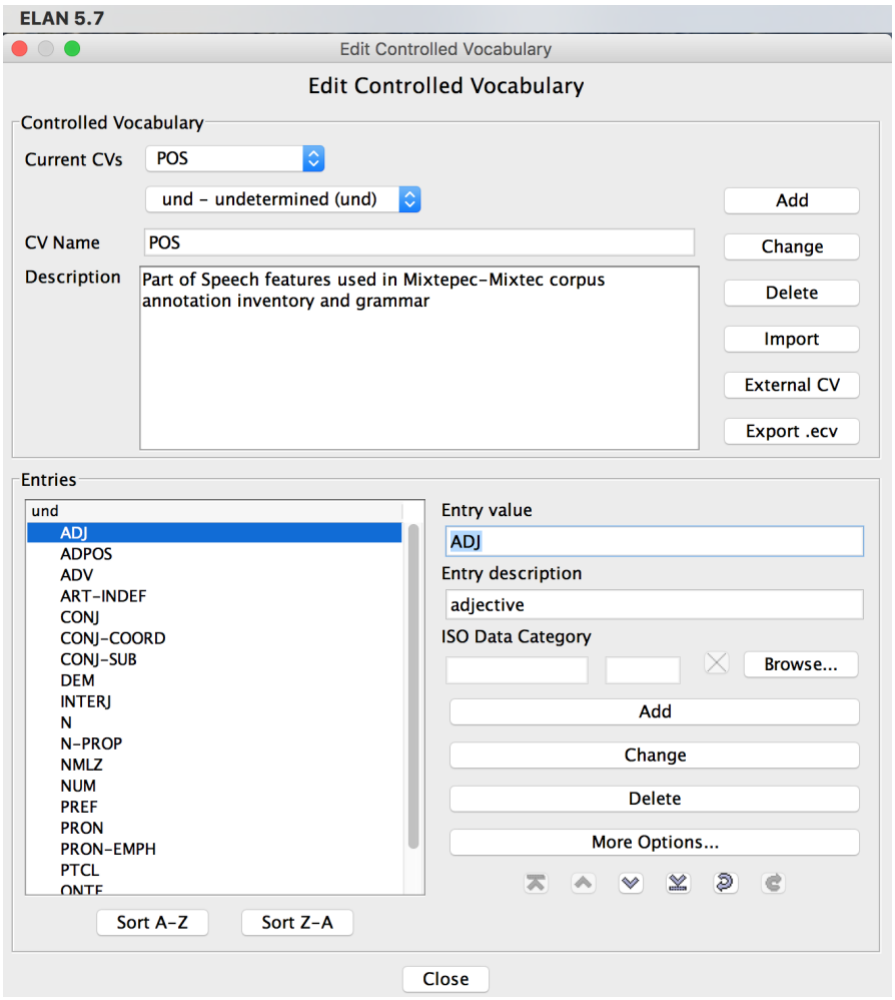


Figure 48: ELAN Controlled Vocabularies editor function

Annotated EAF files contain the CV inventory leaving the annotation system documented in the output, and CVs can be saved as separate XML files (.ecv), each of which is helpful in terms of reuse and portability. Finally, ELAN CVs can be defined in multiple languages, which is useful in the case of projects with multiple working and/or output languages.

Quantitative acoustic analysis may be a necessary component in language documentation projects, especially in determining phonological inventories, particularly in tonal languages. Of the tools commonly used in linguistics, lexicography and LD discussed herein most have only waveform signals, and most do not provide any way to extract quantitative information. The only tool that has the capacity for high level acoustic analysis (of those which are open source and which are already adopted within the linguistics and to some degree LD community), is Praat.

Praat can be used to measure the acoustic, articulatory and auditory readings of: resonance frequencies, pitch, duration, intensity. noisiness, place of articulation and glottal period which are visualized in the forms of: waveform (the direct visualization of a sound representing the air pressure fluctuations as function of time), pitch curve (frequency of periodicity), intensity curve (period averaged power of the speech signal), spectrum, spectrogram (the representation of high and low frequencies), and formant tracks (for an in depth overview of these features see Boersma, 2014; Ladefoged, 1996; Ladefoged and Maddieson, 1996).

The waveform is a basic aspect of speech analysis and shows where there is speech or silence (which is why even the programs that do not have the capacity to carry out or extract other acoustic data or functions all feature it). From a waveform it is possible to infer certain acoustic properties such as: spectral quality, periodicity and intensity (Boersma, 2014). Waveforms are however particularly useful in analysis of voice onset time (a key feature of stops), the example below from Boersma (2014) shows a comparative waveform of intervocalic fricative [aça] vs an intervocalic stop [aca], with the one on the right showing the voice onset time of the stop [c].

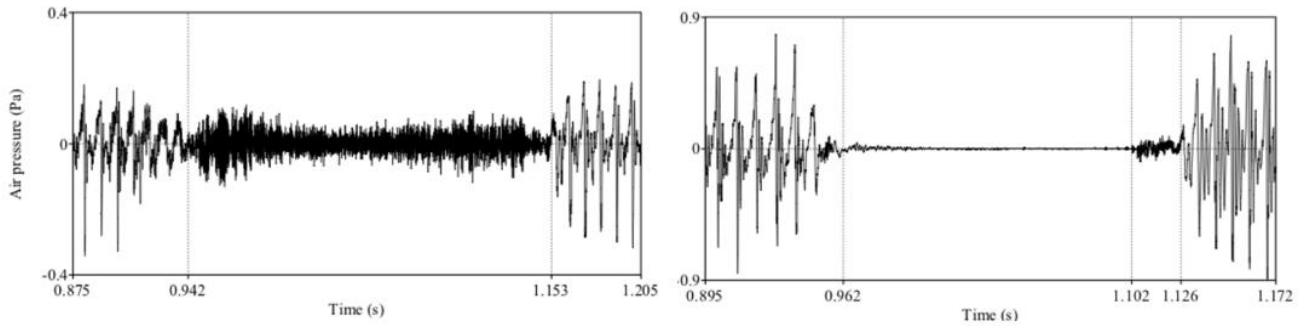


Figure 49: left waveform of intervocalic voiceless palatal fricative [aça]; right intervocalic voiceless palatal plosive [aca] from Boersma (2014)

Another key acoustic measurement is that of the spectrogram, which displays the frequency contents of a sound and which reflects the function of the basilar membrane in the inner ear, and divides the sound into the frequency components over the span of time of its duration (Boersma, 2014). Figure 50 shows the spectrogram and formants of the MIX lexical item *in* [ĩ:] meaning 'one'.

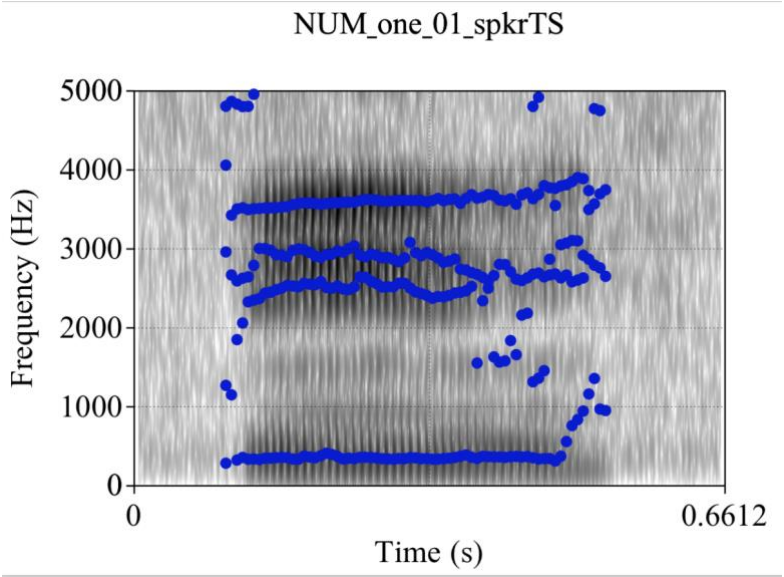


Figure 50: Spectrogram and transposed formants of MIX [ĩ:]

F0 is the basic acoustic correlate of lexical tone, as mentioned, Praat is the only major open source annotation tool that has the capacity to measure and plot this linguistic indicator, which is particularly important when documenting tonal languages. For annotated files, the full

pitch contour can be extracted and saved as a pitch tier file in Praat. While only Praat can generate this, it is possible to import and display this in ELAN by linking the files (.Pitch) in the Timeseries viewer. The following figure shows a plot of the F0 along with the TextGrid transcriptions from three minimal pairs in MIX.

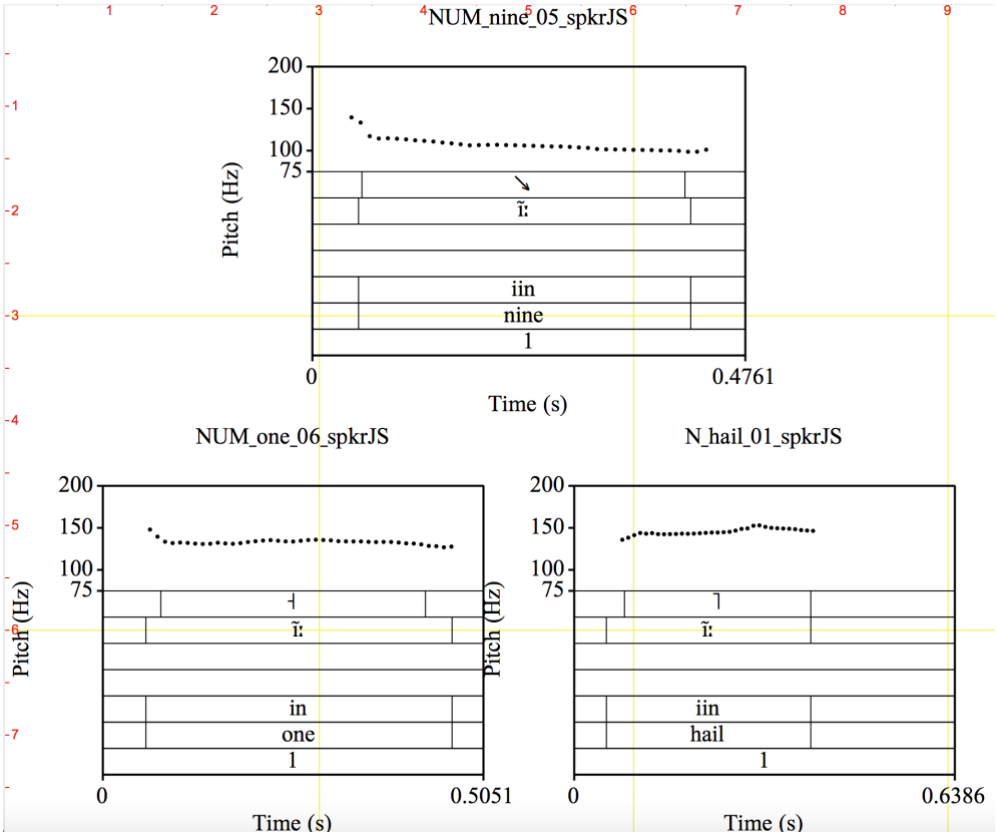


Figure 51: Plotting of F0 contour for tones of 3 transcribed MIX minimal pairs in Praat

Praat has the capacity to carry out a wide array of different analyses and functions from extracted acoustic data, and much of it can be done using the Praat scripting language, including: speech synthesis; listening experiments; speech manipulation; numerous statistical processes such as multidimensional scaling (MDS), principal component analysis, discriminant analysis; machine learning algorithms such as feedforward neural networks and discrete and stochastic Optimality Theory for automatic classification. ELAN also has several different automatic functions it can carry out using its audio Recognizers functions specific to either audio or visual contents, these include: phone-level segmentation, vowel tagging, silence recognition, speaker analysis and more.

Because in an LD context, transcriptions are going to be the basis for a lexicon and/or corpus, the output of the tools in which the spoken language transcriptions are made need to be compatible with the tools in the next stage of data management and/or processing workflow. Next, I discuss the most prominent tools used in lexicon management.

4.4.4.2 Lexicon and Dictionary Creation and Management Tools

Though an annotated lexicon or dictionary are technically descriptive rather than a simple documentary resource, it is likely to be a central component of an LD project. A lexical database need not be the dictionary directly, it can be used as a collection for the lexical as well as all of the encyclopedic knowledge about the concepts as well and then dictionaries can be derived therefrom (Arkhipov and Thieberger, 2018)[153].

In the development of a lexicon and/or dictionary in the context of LD, the reality will be that there will be a corpus of sources from spoken or written sources, which themselves can be either analogue, digital. Sources can come from transcribed spoken language such as any of the tools described above (e.g. ELAN, Praat, EXMARaLDA, etc.), published written sources such as PDF text documents (such as those integrated in this project from SIL booklets), scanned legacy resources (see Blockland et al., 2019), from personal conversations, and increasingly, SMS or social media.

As mentioned, FLEx is by far the most widely adopted tool used in LD for building and organizing lexica, grammar, and annotating (glossing) text. It has a user friendly way of data collection and glossing which can easily be entered using the interface, or it can be done by semantic concept or domain as per Moe (2003), which then automatically creates entries or the contents can be associated with existing entries as needed.

[153] An example of such a collection can be found in the DBOE (Bowers and Stöckle, 2018) which is a collection of lexical content from the Bavarian dialectal regions of Austria and the former Austro-Hungarian empire, which was converted to a TEI dictionary format from other legacy databases and whose original contents were made up of paper slips containing vocabulary and examples.

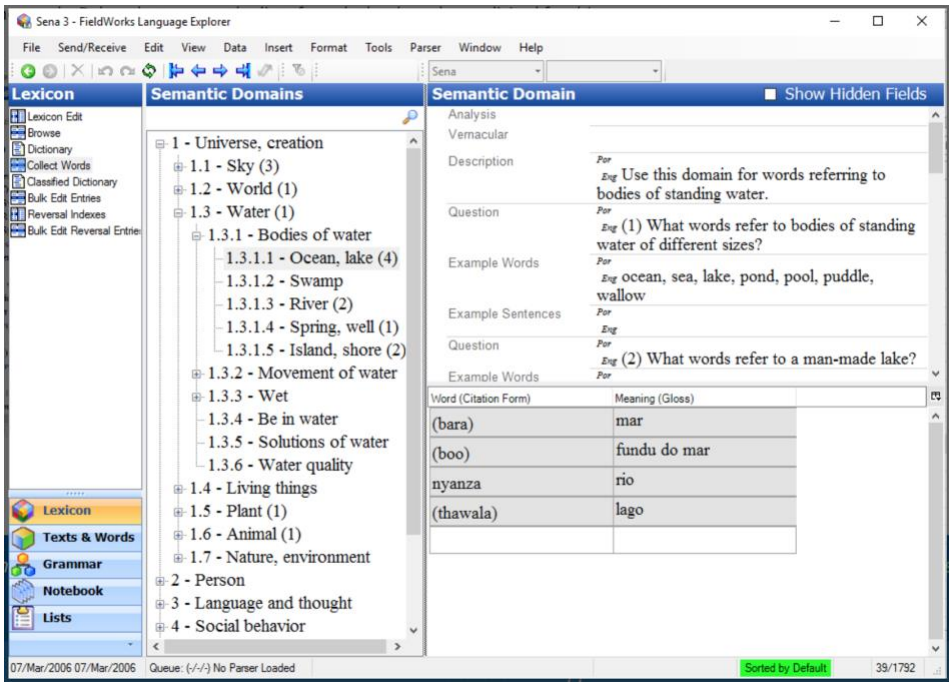


Figure 52: Semantic domain-based vocabulary collection in FLEx[154]

FLEx takes the annotated contents of glossed texts and creates lexical entries where non-existing or allows users to associate with existing entries. Grammatical information and glosses from existing entries are also used to automatically populate interlinear glossed text annotations. Grammatical rules can also be used for the programs' automatic parser in providing automatic morphological parsing of lexical forms. Additionally, FLEx allows for user friendly rendition and output of structured dictionaries from a lexicon, as well as online publishing in conjunction with SIL's Webonary[155] application (see section 4.4.4.3 for further discussion on presentation formatting).

While there are many more functions to FLEx, it is beyond the scope of this dissertation to list them all. There are however, significant downsides to FLEx, amongst the most notable, is the fact that FLEx is designed around the interface, editing and management but the data itself is only accessible as a whole project inside the application, as the contents are not visible as individual files unless exported. Thus synchronization (an important capacity in many LD

[154] Figure 52 taken from http://software.sil.org/fieldworks/resources/tutorial/lexicon/semantic-domains/
[155] https://www.webonary.org/

projects) can only be done in very limited ways, specifically: via USB, Language Depo[156], or SIL's Chorus Hub[157]. FLEx is also prone to glitches that can prevent basic functioning which require developer assistance to restore functionality[158], and it lacks the ability to import existing morpheme glosses and support for custom annotation tiers (Arkhipov and Thieberger, 2018). In fact, FLEx lacks the ability to import structured data in any formats other than SIL affiliated software services and data formats, namely" SFM (standard format) dictionaries used by Toolbox, and LinguaLinks. Another major gap in FLEx is the fact that it is only designed to allow for projects to work with a single language vernacular at a time, this means that it cannot handle projects documenting multiple language varieties beyond simple dialectal differences (e.g. while it could integrate data from both UK and US English, it wouldn't be able to handle integration of different Mixtec varieties, as they have different phonological and morphological rules and different ISO 639 language tags).

Thus, while the tool is very useful and powerful, adopting it requires users to sacrifice a lot of control as to data structure, contents as well as the freedom with which one can work with and edit contents from other sources and tools (see next section for more discussion). Finally, as is the case with most of the SIL software discussed herein, FLEx is only available for Windows and Linux operating systems.

Another SIL tool is WeSay[159] (see Perlin, 2012 for review) is a computer application designed for collaboration on dictionary creation with non-linguist users, especially those in communities with limited access to high quality computers.

Finally, as is the method used in this project, lexicon and dictionary files can be created, edited managed and linked with other project sources using XML editors such as Oxygen XML editor, which is the most common software for working with TEI data, however it can also be used for any XML data, including metadata such as OLAC, IMDI, etc. Using a tool such as Oxygen requires a more direct handling of the raw XML data and (depending on the specific

[156] https://languagedepot.org
[157] https://software.sil.org/chorushub/
[158] See FLEx user's group for recorded issues: https://groups.google.com/forum/#!forum/flex-list
[159] https://software.sil.org/wesay/

needs of the project) relies on the user having the ability to create XSLT and/or XQuery scripts for transformation, but it has many benefits, particularly that there are never any limitations to what one can include in their data based solely on the limitations of the software as can be the case with FLEx.

4.4.4.3 Presentation Formatting

For projects in which non-technical experts and community contributions are essential, usability is of course of the foremost importance, and thus the primary focus in choosing a tool can often be reduced to the ease of access, and editing, and the varieties of data types that can be used. Below is a brief list of tools that have taken the initiative to remove the burden of data modeling from projects seeking to searchable create well presented, online digital dictionaries

- Webonary, part of the SIL software ecosystem, is an online platform that allows users to publish dictionaries or grammars from Toolbox and FLEx data. The platform allows users to freely browse and search contents. Additionally, there is an accompanying mobile application Dictionary App Builder[160] which can be used to work on a dictionary which can then be published on Webonary and/or as a mobile application.

- Talking Dictionaries[161] is an online desktop and mobile application platform designed to be as user friendly as possible in order to allow collaboration between linguists and speaker communities and to have the capacity to embed all sorts of sources in entries such as posts and videos from social media. The application started as a single project on the Tuvan language (Harrison and Anderson, 2006) and has since grown to include 120 languages. It can record and playback audio, offline data access, semantic domains, privacy settings, and search the entire contents of a dictionary. Harrison et al. (2019) presents the use of the application in the context of a Zapotec Language Activism and Documentation work dealing with multiple varieties of Zapotec. Talking Dictionaries data can be imported in bulk in plain text (CSV), or JSON, although it currently doesn't have any features for exporting data.

[160] https://software.sil.org/dictionaryappbuilder/
[161] https://livingtongues.org/talking-dictionaries/

- Zahwa[162] which is an application for Android mobile devices was originally designed as an application for documenting procedures of food preparation, is a user-friendly application that allows users to create audio, image and video content to create easily usable lexical content and areas of cultural knowledge.

- SayMore[163] (for review see Moeller, 2014) is a multifaceted used for creating and organizing transcriptions and contains many of the key features of an LD project in a user-friendly, non-expert oriented way. It features progress tracking and data management components, IMDI metadata, simplified transcription and export options for ELAN[164], FLEx, Toolbox, YouTube and more.

It is of course possible to create such resources according to best practices without using such tools, particularly when working with XML data, it is fairly simple to convert a digital dictionary as well as other LD content to HTML using XSLT and/or format it in conjunction with CSS.

[162] https://zahwa.aikuma.org/
[163] https://software.sil.org/saymore/
[164] See also Pennington (2014) for a discussion on using SayMore in combination with FLEx and ELAN.

Figure 53: Example of MIX Dictionary in HTML with CSS formatting

The TAPAS project[165] (Flanders and Hamlin, 2013) was designed as a hub for both depositing, archiving and presenting TEI data and I was an early adopter having deposited the MIX TEI dictionary and a number of other resources. However, despite the main purpose being to provide a basic way to present TEI data in a user friendly way, the system was never able to properly display my data neither in the various built-in formats or in conjunction with the CSS schema made for the dictionary. This problem was never solved and in discussing the issues with the programmers involved, it was essentially communicated to me that since the MIX project represents a more niche case, they could not prioritize the types of changes to their system that would be necessary to accommodate the data. Though the work they are doing is highly useful and it is understandable that the limited resources need to be allocated to the areas and types of

[165] https://tapasproject.org/

datasets most deposited, this situation demonstrates a significant gap in the TEI community ecosystem for those of us in lexicography and linguistics. Unfortunately, this adds to the challenges of using TEI for LD tasks.

As stated by Arkhipov and Thieberger (2018), aside from archives, there is no established and easily reusable solution for publication of language documentation data in a user-friendly way. Moreover, it has been the case that those that do provide user friendly solutions, often do not provide for or accommodate the most common data types actually used or produced in LD and lexicography; or as discussed, particularly with regards to SIL, they may only accommodate data in formats produced by a certain set of software. While given the dire situation in which many languages are in, it is entirely understandable that the priority would be given to the concrete aspect of immediate output for LD work. However, as mentioned above, the reality is that the diversity of solutions, many of which do not do not address data export or formatting, significant progress remains to be made in achieving the kind of solutions for interoperability and reusability, which are of course canonical pillars of LD as per Bird and Simons (2003b). These issues are discussed in the next section.

4.4.4.4 Interoperability, Interchange and Workflows

A major complication in working with the diverse data types and sources inherent to LD is that there is often a bottleneck in processing in terms of annotating and integrating resources between tools and various sources of language data (Arkhipov and Thieberger, 2018). This problem is exacerbated by the limited varieties of workflows possible in carrying out the necessary tasks given that not all tools are equally accommodating of each other's data formats, thus the directionality of data interchange is an issue. As has been emphasized throughout this dissertation, the issue of interoperability and interchange are ubiquitously recognized as key factors in the choice of software, as well as the eventual quality of the contents produced in an LD project. First, however it is worth specifying exactly what is meant by *interoperability* and *interchange* as they are distinct yet often obfuscated.

As discussed by Unsworth (2011), *interoperable* data can be taken directly from one system and operated on directly in another, and *interchange* is a format that is an agreed upon

encoding scheme that is capable of translating between two formats. Therefore, obviously, interoperability is the ideal for LD data, whereas interchange is the fallback where the former is not possible. Furthermore, perhaps with the exception of complimentary SIL tools, there are in fact no instances of truly interoperability in the purpose-specific LD/lexicographic tools; instead there are only tools that through internal conversion processes, allow for data interchange between different data formats, most of which are specific to the given tools.

In this section, I discuss the issues of interchange between complementary tools and data in general which make up the essential components and capacities of typical LD workflows, specifically: transcription tools with different specialization (e.g. ELAN/EXMARaLDA and Praat); transcription tools and lexicon development (e.g. ELAN/EXMARaLDA and FLEx); external contents of various formats and lexicon development tools. It should of course be kept in mind that the tools in question evolve very rapidly and thus at any time there could be updates in a given tools capabilities.

A major factor enabling both sustainability and the ability to convert between the data formats of different tools as well as the ability to adhere to common data standards in LD has been the adoption of XML as either a native format or at least an option for import/export (Arkhipov and Thieberger, 2018). Thus, tools with XML as a native data format are inherently more easily capable of being able to read other data formats (both XML and plain text)[166], as well as produce data that can be transformed or input into other formats[167]. That said however, it is nonetheless the case that not all tools take full advantage of this, particularly FLEx which only reads external files native to the SIL software ecosystem.

Perhaps the most typical workflow in LD at present is to transcribe speech using ELAN and then to import the glosses into FLEx where they can be further annotated, and the lexicon is developed all within a single system. Such workflows have the advantage of making use of the

[166] While a system whose data format is XML is more easily able to read other XML as well as plain text formats (such as Praat), the inverse is not as conducive because XML data can be more complex given that it may have layered structures and attributes which cannot be converted into the simple capacity of plain text.
[167] An indication of the benefits of XML can be seen in the fact that FLEx chose to change from an SQL server to an XML model after version 7.

user-oriented output tools in the SIL software ecosystem (e.g. Dictionary App Builder and Webonary).

ELAN, as has been discussed above, is by far the most powerful tool in enabling workflows that can import and export from numerous different tools and data formats. Moreover, it should be stated that if it weren't for the developers of ELAN (The Language Archive at Max Planck Institute for Psycholinguistics, Nijmegen), the usefulness of FLEx would be much more limited, as it would only be able to import data from SIL software, whose time-aligned speech tools are much less advanced than ELAN. Thanks to ELAN, data can come both from the tools itself, as well as from the full array of speech transcription software toolkits whose data formats ELAN is able to read, which it can then export into FLEx. This of course allows users to have the ability to: integrate transcription data from external sources, including where different tools were used; export and re-import to and from other transcription programs with different specialization (e.g. Praat).

Figure 54 shows a mapping between a sample ELAN annotation and a FLEx file. Note that since ELAN annotation tiers can be freely named, they need to be mapped to the FLEx annotation tiers which are pre-determined by the tool.

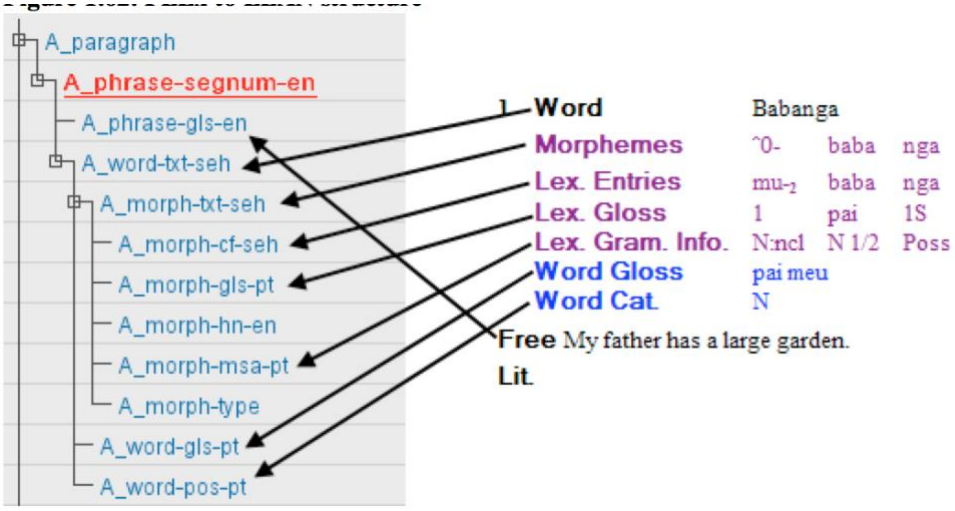


Figure 54: ELAN to FLEx mapping

The fact that ELAN can import and export to and from Praat, and that it is possible to import certain acoustic measures (e.g. ".Pitch" and ".Interval" files) from the latter into the former means that despite its limited capacity for phonetic analysis, it is still possible to use the ELAN as a project's primary method of transcription, glossing and general organization when needed and make use of the acoustic phonetic specialization in Praat. Given these issues in conjunction with the fact that both ELAN and EXMARaLDA can import or export to Praat '.TextGrid' file format, in the context of language documentation, it is entirely possible that in a project dealing with a tonal language, and/or one looking employing certain advanced functions of quantitative acoustic modeling or analysis, the transcription and acoustic data extraction could be done in Praat and then re-imported to ELAN or EXMARaLDA.

In contrast to ELAN, and to a lesser degree EXMARaLDA, as mentioned above while Praat's plain text-based ".TextGrid" files can be read into numerous other programs (including ELAN and EXMARaLDA amongst others), Praat cannot read *any* other tools' transcription files. Thus, the full burden of interchange between Praat and other software tools are entirely due the programs like ELAN and EXMARaLDA having the capacity to both read and write files to and from Praat[168].

In an ideal scenario, it may be possible to plan and execute an LD project with a strict workflow, only integrating limited types of data in specific file formats, avoiding all the issues and limitations of data interoperability and interchange for each tool discussed above. It would only be possible however were the project to only get data sources in specific formats that can be handled by these tools. In many cases, it is likely that the flow of data will not always be so linear, and it is common to come across data in a number of different formats either from other tools, or simple standalone files such as: excel spreadsheets, word files, PDF documents such as scanned legacy dictionaries, content from webpages with downloadable HTML and other contents such as social media posts and any other random instance of language use in general.

[168] Note that upon attempting to import the time-aligned TEI files from this project (originally converted from Praat), EXMARaLDA was unable to read the contents and the documentation doesn't provide any further clarity. Files with which this was attempted can be found here: https://github.com/iljackb/Mixtepec_Mixtec/tree/master/media/speech-mix/with-txtgrd

Thus, while FLEx may be convenient if all resources in a project are annotated and managed within either: SIL's software ecosystem; or ones (like ELAN) that convert data to FLEx compatible format, FLEx is highly limited in the types of data that can be integrated into a dataset and cannot import many common types of data mentioned above. Thus, in using FLEx, in order to integrate data that doesn't come from the narrow array of sources, editors either need to enter by manually copying or by converting via XSLT or some other programming means into one of the few formats FLEx can import, which is not likely considering the LIFT format is not a well-documented format made for external programmers.

As is commonly the case in DH contexts, using XML editing software such as Oxygen XML Editor[169] it is possible to create one's own conversions using XSLT programming in order to create the same kind of workflows and processes of integrating the corpus sources with one another and in extracting their contents, adding them to a digital lexicon, as well as converting them to the aforementioned output formats. Within the current project[170], Oxygen is used to edit and manage files, and of course TEI is the data format for both text and annotated speech corpus and dictionary; herein XSLT conversion are manually developed to carry out many different conversion, such as convert from: Praat TextGrid annotations to time aligned TEI; plain text and/or CSV to and from TEI; extract annotated contents from TEI corpus documents[171] to TEI dictionary, TEI to HTML, perform random transformations of data structure to multiple files. These files are all reusable and made openly available on GitHub[172], with further adoption of the time-aligned speech, general TEI annotated texts, and dictionaries for these purposes, the more likely it will be that such XSLT scripts can become the basis for a more stable means of converting various types of data between steps in the workflow.

[169] In contrast to the other LD tools described above, Oxygen is not a free software.
[170] Note that of the speech transcription and/or lexicon development tools discussed (e.g. ELAN, FLEx, Praat, EXMARaLDA, etc.), this project has only used Praat
[171] When using the term 'TEI corpus' I am not referring to the element <teiCorpus> https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-teiCorpus.html, instead I refer to any document containing annotated text or annotated speech
[172] https://github.com/iljackb/Mixtepec_Mixtec

Due of course to the fact that ELAN's EAF and FLEx LIFT data are formats made for, and used by the given specific tools, and that the workflows in LD have remained within the confines of a few tools, conversion schemas to and from TEI have not yet been established as people have just tended to keep the data in the various tool-specific formats rather than any final integrated output standard. Doing so would create a further connection between the field of LD and the data and practices common in DH. While ELAN's EAF data model is openly documented[173], which makes prospects of the development of mapping to be made to TEI much easier, this is not the case for FLEx's data models (LIFT or Flexfile). Development of such schemas between TEI, ELAN and FLEx (each for interlinear glossed text, as well as lexicon/dictionary data) will be a major effort to be undertaken in the near future.

In the absence of an all-purpose tool that would resolve the issue of interchange and interoperability, an alternative would be to settle upon a lossless interchange format and ensure the field would greatly benefit were a single data exchange format be adopted. Arkhipov and Thieberger (2018) raise the possibility of a format such as Cross-Linguistic Data Formats (CLDF)[174] (Forkel et al., 2018), which was designed for the *Cross-Linguistic Linked Data project.*

CLDF is a tabular data model that is intended to provide a simple model for exchanging datasets of certain lexical information such as parallel vocabulary (cognates, translations, dictionaries, etc.). Databases produced within the CLLD framework include: The World Atlas of Language Structure (WALS)[175]; The Comparative Siouan Dictionary[176]; Phoible[177]; Glottolog[178]; Concepticon[179].

There is however, a reason that XML has become widely adopted over tabular data/spreadsheets which is that, in resources such as dictionaries there are simply too many variables to anticipate in a table files. This fact is acknowledged by the authors, and in order to

[173] https://archive.mpi.nl/tla/elan/documentation
[174] https://cldf.clld.org/
[175] https://wals.info/
[176] https://csd.clld.org/
[177] https://phoible.org/
[178] https://glottolog.org/
[179] https://concepticon.clld.org/

express the kinds of varying, complex data types, structures, and combinations, the model recommends that in certain cases, that users turn to using multiple spreadsheets which then must be documented in metadata (Forkel et al., 2018). Inevitably, if the aim is to try to use the format to exchange data between tools like ELAN and FLEx for example, in many cases, the data will likely not necessarily be any easier to process, or convert than using a well-documented XML format. Additionally, while it is quite simple to convert XML data to a CSV tabular form, it is often not as simple to convert the other direction, which would need to be possible given the fact that such tools use XML as their working format.

Thus, rather than trying to make it an all-encompassing exchange format, CLDF is likely best suited to serve as an interchange for a limited, and focused array of linguistic datasets. Conversion between XML (including TEI) and CLDF would be quite simple, as XSLT schemas to extract and convert such datasets for output to specific databases and tools could be easily developed. Having such an established tabular data output model established would be beneficial for everyone as it would provide an extra layer in the ongoing efforts to create, documented and structured datasets in linguistics and LD.

An alternative to the adoption of a relatively novel format used purely for interchange, would of course be to make use of an established data standard such as TEI that is already adopted in countless linguistic and lexicographic projects around the world. As has been shown, many or most text corpora are based on, or compatible with TEI, the spoken language encoding is both compatible with the text corpus encoding practices, and the dictionary is widely adopted and has the capacity to express the entirety of the needs of tools like FLEx. As it is incredibly easy to convert TEI lexical data to a CSV tabular wordlist such as CLDF, and TEI is already well established and is an XML data format as are most of the LD tools, TEI, with its larger and established user community would be a sounder choice.

One criticism of TEI, particularly in the dictionary domain, has been the wide variety of options and practices has created a situation which has led to a great divergence in formatted dictionaries despite using the same standard. For this reason, the initiative of Lex-0 (Tasovac and Romary, 2018; Bański et al., 2017) has sought to create a streamlined set of recommendations

for dictionary encoding to reduce possible variation and to establish a baseline encoding that at minimum could be used as an interchange format both within TEI as well as between other systems such as the OntoLex Lemon markup vocabulary for the semantic web.

While it is clear that developers of tools such as ELAN and those in the SIL infrastructure (FLEx, Toolbox, etc.) choose to use their own unique data structures that are tailored to the needs and features of their own tools, rather than deal with the chaos of an open source standard such as TEI, there would nonetheless be benefits to the cause of data interchange and interoperability were they to adopt TEI as working format, or at very least add TEI as an import and export format. Alternatively, however, it would be possible to create external XSLT schemas to allow conversion between the given tools; the online TEI service OxGarage[180], which facilitates conversion of numerous different text documents, presentations and spreadsheets to and from TEI could potentially integrate such schemas to that web service so that users could easily carry out such conversions online. Such developments would make a large quantity of data used in DH compatible with that of LD.

4.4.4.5 On Issues Related to Choosing Data Structure and Tools for LD

The better one understands the underlying structure of the data, the easier it will be to implement a system which will be sustainable long-term (Good, 2011). However, as mentioned, some tools provide better documentation of their data structure than others. While software designed specifically for LD will have certain optimized capabilities for certain linguistic data types, it is impossible for a software toolkit to anticipate every need of a given project, and that tools designed for use by linguistic experts may not be ideal for use when the team consists of non-linguists (Ibid).

Another major point by Good (2011) is that despite the advent and widespread adoption of FLEx in LD projects, while it would be desirable to explicitly recommend a particular software for working with language data, the needs of each project are too specific and that no tool can bridge the tradeoff between: a) the needs to be able to implement any given underlying data structure; b) the kinds of formats it can work with and output; c) the ability to ensure the

[180] https://oxgarage.tei-c.org/

tool can be used by anyone, including non-experts; and d) that the data produced can be used by the target audience of the given language. Thus, Good's advice for choosing software in such an endeavor is that project leaders should clearly establish the overall goals of the LD in advance and consult with experienced individuals.

As discussed by Arkhipov and Thieberger (2018), a hypothetical all-inclusive, omni-functional software tool for LD would require a high degree of detailed insight into the diverse and complex practices of LD researcher, and would need to support a wide variety of primary data types (e.g. audio, video), content sources such as raw texts, wordlists, paradigms, questionnaires and metadata, etc. Additionally, it would need to allow for a wide array of annotations, linking between media sources and descriptions; it would also ideally feature dynamic search and analytical tools as well as visualizations and publishing. Such a tool would need to be cross-platform, which is burdensome to the developers and is likely to become heavy and slow in accommodating so many features. For these reason, the authors cast doubt on the prospects of the advent of such an all-encompassing LD tool (Ibid). For this reason, the goal of developing and promoting maximal data interoperability and interchange within the tools used at the various stages of the data collection, annotation and organization processes should remain the main priority.

4.4.5 Publishing and Obtaining of Existing Language Resources

Recording, preserving and publishing the stories and knowledge of individuals in their own language can obviously be a very important thing both on a personal level, to them their families, and communities, as well as for posterity both for linguistic and numerous other purposes. However, as especially in the context of theoretical linguistic studies, study of indigenous languages by linguists took place within what Czaykowska-Higgins (2009) refers to as a *Linguist-Focused Model* (*see below for discussion*). In such linguistic practices, the content of linguistic consultations have not always been anything more than elicitation sessions containing speech uttered for the sole benefit of the investigator, in many cases nothing more than field notes were saved as records (Thieberger, 2014, 2016). Additionally, in traditional linguistic practice, it has not been a priority to make supporting source data available or to systematically account for metadata such as speaker sources, demographics, how data was

obtained, etc. (Thieberger, 2014). The practice of not publishing source datasets was of course reinforced by a lack of scholarly/academic benefit as it was not normal practice to recognize the production of primary datasets as a valid output in and of itself (Thieberger, 2014; Thieberger et al., 2016).

It has been recognized over the last 20-25 years that in dealing with the reality that so many languages, especially indigenous languages like MIX, it is imperative that not only should these resources and records be created, they need to be done in formats that are portable (e.g. usable across software and hardware platforms, and (with appropriate informed consent and permissions) stored in archives that are stable long-term (Bird and Simons, 2003b; Himmelmann, 2006; Austin, 2006; Woodbury, 2014). Finally, in 2010 Language Documentation was recognized by the Linguistic Society of America (LSA) as a distinct field of scholarly merit[181] which was a significant legitimizing development which should hopefully contribute to a higher level of academic support for such projects (Thieberger, 2014; Austin, 2016).

In addition to preservation and reuse, concern for accountability is another major characteristic of language documentation and access to the primary (and meta-) data is imperative for others to re-use and analyze materials (Himmelmann, 2006, 2012; Austin, 2016; Gawne and Berez-Kroeker, 2018). In neglecting to produce source data from which linguistic analyses are based (or at least some kind of indirect output of it such as basic transcriptions of speech), linguists have denied others the ability to subject their interpretations to scientific scrutiny (Himmelmann, 2006; Thieberger, 2014; Gawne and Berez-Kroeker, 2018). As pointed out by Thieberger (2014), the need and utility to publish a documented collection of language materials can been illustrated by the controversy between the claims made about the Pirahã language by Everett (2005) and the chomskyan formalists who reject these claims. The lack of a published archive of source material, metadata, especially documentation of how the content was gathered makes independent verification of the claims impossible. Thus, not only is the practice of providing primary linguistic data incredibly important for issues of language heritage, reuse, etc. it has implications for scientific and theoretical empirical analysis for those in the field of Linguistics.

[181] https://www.linguisticsociety.org/resource/resolution-recognizing-scholarly-merit-language-documentation

4.5 Ethical Issues in Language Documentation and Linguistics

Because of course of the historical, social and political circumstances that indigenous peoples and communities have faced, LD and projects dealing with such subject matter need to be aware of, and adhere to different ethical principles than linguists working with other languages. Historically, the role of indigenous peoples and speakers of a documented language has been limited to being the source of speech consultation, and the roles of the linguist and speech communities were typically separated as researcher-researched, expert and non-expert. This reflects the so-called "Linguist-Focused Model" (Czaykowska-Higgins, 2009). It has only been a recent development that linguists are starting to routinely recognize that there should be an ethical responsibility to individuals whom they are working, their communities and their knowledge and that scientific study of a language should not be approached or framed in a detached way, and that such a collaborative approach can be mutually beneficial (Hale, 2001; Rice, 2006; Dwyer, 2006; Czaykowska-Higgins, 2009; Glenn, 2009). In pursuit of changing this reality, Cameron et al. (1992) defined three alternative frameworks for ethical language research which are:

- *Ethical research* in which it is the responsibility of the researcher to acknowledge the contributions of collaborators and to 'minimize damage and offset inconvenience to the researched'. This model is considered to adhere to the very minimal degree of potential engagement and advocacy of the communities and speakers as it still is based on a model of doing *research on* subjects.
- *Advocacy research* is characterized by the fact that the researcher should be committed to carrying out research *on and for* subjects, not just *on.* This can involve a wide array of different applications in practice, which could mean using their authority to defend the subjects' interest in any number of areas including health, education, political, cultural or territorial rights.
- *Empowering research* is centered on the principle of doing research '*on, for and with'* subjects. In this model, the research agenda should be in full collaboration with, or fully driven by the aims of contributors and the communities and the researcher should lend their expertise in pursuit of this. This includes the following "programmatic statements": (a) "*Persons are not objects and should not be treated as objects*," (b) "*Subjects have*

their own agendas and research should try to address them" and (c) "*If knowledge is worth having, it is worth sharing*".

Czaykowska-Higgins (2009) adds one additional framework: *Community-Based Language Research*, which goes even further than the three aforementioned from Cameron et al. (1992) in that it emphasizes the idea that a linguist should not be assumed to be the only 'expert' in the research process and that community members should also be active directors and partners in the work, as opposed to "empowered research subjects". This model is defined as: *"Research that is **on** a language**,** and that is conducted **for, with,** and **by** the language-speaking community within which the research takes place and which it affects. This kind of research involves a collaborative relationship, a partnership, between researchers and (members of) the community within which the research takes place."* This model, in its full realization could involve training members of the language-using community to carry out the research themselves, thus negating the need for linguists who are not members of the community in the research process.

Community-Based Language Research as described above are the ideal scenario, and should be considered the gold standard, there are of course many circumstances in which such a degree of collaboration may not always be possible. Nonetheless, even without full collaboration, there are other areas in which linguists working with indigenous or other minoritized languages can still collect and produce an output in an ethical way.

Specifically, the issue of what is done with the data; it should be considered an ethical priority that any linguistic knowledge about a language should be both be preserved in an archive and be accessible to the speakers and community members so that it can be repurposed for the knowledge of the community and for potential revitalization endeavors. This way, even if there is not a degree of community participation possible, at very least the language data and knowledge therefrom produced, can at least be made available for re-use by interested community members in the future.

Linguistics programs offering field methods courses can also play a significant role in promoting ethical practice for the purposes of LD and revitalization (see Campbell et al., in

press). In many graduate level field methods courses, work is carried out with a single community member in which the primary goal is linguistic analysis via translation-based elicitation in which the role of the speaker is the "subject", limited to providing linguistic information. In many cases, none of the advice or practices common to language documentation are followed (e.g. no standard data formats, no archival deposits of vocabulary or media files, etc.) and the linguistic output is hoarded in the private servers of the university and only the linguistic analyses are published. If courses were to make a policy of following deliberate practices to ensure that the work produced is not only beneficial to the students but that they also produce structured and well documented language resources that can be reused by the community this would be a significant, and overdue step in ensuring that the Linguistics field actually use their positions and resources in the best interests of the cause of the world's languages.

Such a policy would require a significant update in practice, in particular for Linguistics departments, which very rarely have any coordinated linguistic data sustainability policies and very often have no staff member who specialize in, or have significant experience with linguistic data compilation or annotation. Additionally, many Linguistics departments only have members who work with linguistic data on a single level of language (e.g. phonology, syntax, etc.), in these cases, they are often well versed in the practices of only that narrow domain of linguistics, which very rarely use any kind of data standards, archival or ethical practices. Thus, in addition to the issues pertaining to linguistic-research frameworks, the lack of integration in digital data practices across linguistic domains (which involve data standards and tools) can also have a negative effect on the cause of, and need for documentation and conservation of the world's languages.

5. Overview of Mixtecan Literature and Resources

As a major goal is to integrate all relevant Mixtecan and MIX sources into this data collecting in order to provide for the establishment of as comprehensive a basis as possible for the present and future work in MIX lexicography, and cultural documentation. In this section I introduce some key works (both historical resources as well as linguistic analyses) in Mixtecan, some of which have been integrated into this project's TEI corpus.

5.1 Codices

The earliest written Mixtec was of course the codices written in the indigenous pictographic format mostly on deerskin canvas. Unfortunately, many more were likely destroyed by Spanish missionaries, with the surviving examples having been looted and taken back to Europe, then being passed around between various nobles and monarchs, before ending up in the museums and libraries where they now are located. This has led to a gap between the Mixtec people, whose ancestors created these documents and the investigators and the institutions who possess them (Jansen and Pérez Jiménez, 2004).

Figure 55: Lady 1 Deer and Lord 1 Deer in Codex Yuta Tnoho (Vindobonensis) from Jansen and Pérez Jimenez (2018)

Mixtec codices represent the largest surviving corpus of indigenous Mexican manuscripts, they are as follows: *Codex Zouche-Nuttall, Codex Vindobonensis (aka 'Mexicanus I' or 'Codex Vienna'), Codex Bodley, Codex Selden, Codex Becker I (aka 'Codex Columbino' or 'Codex Alfonso Caso'), Codex Becker II, Codex Egerton (aka 'Sanchez Solís'), Codex Muro* and *Codex Tulane.* As can be seen in these names, intertwined with the history of colonization, the nomenclature of these (as well as other Mesoamerican) manuscripts are given to honor collectors, politicians, scholars and institutions of the western, mostly European world. Jansen and Pérez Jiménez (2004) presents a set of names that are based on the content of the codices which are aimed at removing the legacy of colonization and disappropriation from these priceless documents. The proposed revised names are as follows:

- Ñee Ñuhu: (term for codices in general*)*

  Derived from the term "sacred (deer)skin", or "book*";* this term was the original term used by Classical Mixtec speakers first documented by Francisco de Alvarado in 1593 in the first dictionary of a Mixtec variety (see 5.2 for description);
- Codex Ñuu Tnoo-Ndisi Nuu: (Codex Bodley)[182]

  The contents of this codex are a major source of history of the Mixteca Alta region, with details of dynastic records and dates, primarily about two noble houses: that of Tilatongo *Ñuu Tnoo* and that of *Ndisi Nuu;*
- Codex Iya Nacuaa I: (Columbino)

  One of two separated fragments of Codex Columbino-Becker, the contents of this codex tells the life story of the warrior king *Iya Nacuaa;*
- Codex Iya Nacuaa II (Codex Becker I)

  The other of the two separated fragments of Codex Columbino-Becker which (also) recounts the life story of the warrior king *Iya Nacuaa*;
- Codex Cochi (Codex Becker II)

  The proposed new name for this document is inspired by the ruler depicted in its contents *Iya Cochi*;
- Codex Ñuu Ñaña: (Codex Egerton/Sanchez Solís)

[182] https://www.britishmuseum.org/collection/object/EA_Am1902-Kud-Cod-8517

The contents of this document describe the dynasty of a town in the Mixteca Baja region, likely Cuyotepeji, which is represented in the codex as the Temple of the Jaguar, *Ñuu Ñaña;*

- Codex Tonindeye: (Codex Zouche-Nuttall)

  This document is two-sided, the contents of one is an (unfinished) biography of the king Lord 8 Deer Jaguar Claw with the other side used as a notebook containing notes on different dynastic histories; *Tonindeye* refers to the general theme of the contents, namely "lineage history";

- Codex Añute: (Codex Seldon)

  The revised name is based on the contents which pertain to the dynastic rulers of Añute (modern-day Magdalena Jaltepec);

- Codex Ñuu Ñaha: (Codex Muro)

  The manuscript contains genealogy of a list of ruling couples of the city state *Ñuu Ñaha* (present day San Pedro Coxcaltepec Cántaros) in the Mixteca Alta;

- Codex Yuta Tnoho: (Codex Vindobonensis)[183]

  The contents of one of the sides of this manuscript tell the legend of how the dynasties were born from the Great Mother Pochote Tree in the Sacred Valley of Yuta Tnoho (Apoala);

- Roll of Yucu Yusi: (Codex Tulane)[184]

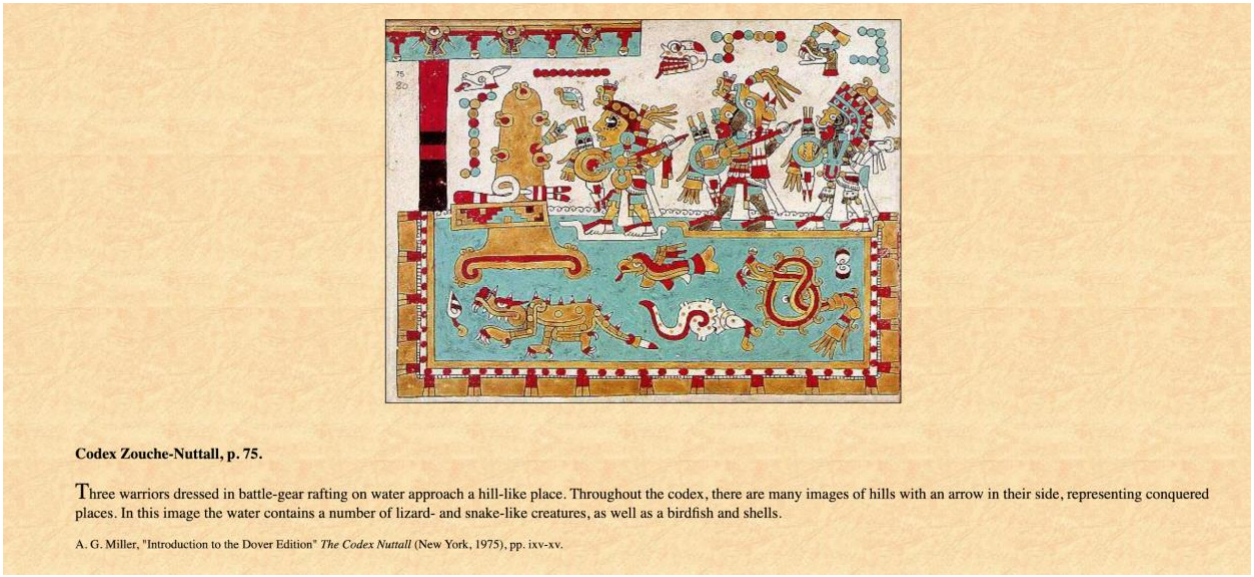
  This document is not actually a codex but a painted scroll which contains the lineages of the rulers from two of the main city-states in the Mixteca Baja: *Toavui* (Chila) and *Yucu Yusi* (Actlan) in southern Puebla;

While, as stated by and Pérez Jiménez (2004), the scholarly community is averse to changes in nomenclature, these revised indigenous based names (which are mainly derived from terms in Classical Mixtec) provide a good basis for referencing, and coining new terms for these documents in modern day Mixtec varieties as Mixtec people and scholars reclaim their heritage.

[183] https://www.britishmuseum.org/collection/object/E_Am2006-Drg-226
[184] https://digitallibrary.tulane.edu/islandora/object/tulane%3A19287

According to Jansen (1990), by the second half of the sixteenth century the use of pictographic codex style of writing had been replaced by Mixtecs both in writing Mixtec and Spanish languages though only a small number of such texts survive, example of which are the Archivo del Juzgado de Tepozcolula and the Archivo General de la Nación.

Although the in-depth study or encoding of codex material is not directly within the scope of the current state of this endeavor, I mention these documents in order to make clear that by no means does Mixtec writing begin with the arrival of the Spanish (see Jansen, 1990 for a discussion) as well as to point out resources that may be integrated into a body of digital Mixtecan resources in the future, potentially in the form of TEI digital editions annotated and described in any number of Mixtec varieties. For example, Figure 56 below shows a screenshot from a webpage of the University of Arizona Library[185] with an image from the codex Zouche-Nuttal (or Codex *Tonindeye* according to the proposed indigenous renaming as per Pérez Jiménez (2004)), there is a short description of the contents in English, simply creating Mixtec version of these descriptions would present a significant opportunity to bring the knowledge of these key items in Mixtec history and cultural heritage back into the language(s) of the people whose ancestors' deeds they describe and document.



Codex Zouche-Nuttall, p. 75.

Three warriors dressed in battle-gear rafting on water approach a hill-like place. Throughout the codex, there are many images of hills with an arrow in their side, representing conquered places. In this image the water contains a number of lizard- and snake-like creatures, as well as a birdfish and shells.

A. G. Miller, "Introduction to the Dover Edition" *The Codex Nuttall* (New York, 1975), pp. ixv-xv.

[185] http://www.library.arizona.edu/exhibits/mexcodex/nut75.htm

Figure 56: Sample image of codex with description from University of Arizona Library

5.2 Colonial Mixtec

The earliest, and most prominent use of phonetically written Mixtec unsurprisingly is in the context of religious activity. In the colonial period, the earliest sources of any Mixtec vocabulary (not including the pictographic codices) are the Doctrina en Lengua Mixteca by fray Benito Hernández, in the Ñuu Ndecu (San Miguel Achiutla) (1567) and another in the Teposcolula (1568) which were the first documents presenting Catholicism to the Mixtec people (Hollenbach, 2016). The primary Mixtec resources from this period were the *Vocabulario en lengua mixteca* from the (Alvarado, 1593) and the grammar *Arte en lengua mixteca compuesta* by Fray Antonio de los Reyes (1593) both in the Tepozcolula variety.

According to Hollenbach (2016) there are also various more manuscripts and archival documents which were almost all from the Highland Mixtec (Mixteco Alto) region. Very few materials from the Lowland Mixtec (Mixteco Bajo) regions exist and none from the Coastal Mixtec region (Mixteco de la Costa). In the later colonial period Ripalda's catechism originally published in 1719 and again in 1755 (Ripalda, 1755). By the end of the colonial period, the use of written Mixtec in the Mixteco Alto had ceased though several catechisms were published between 1834 and 1899 in Lowland Mixtec varieties. These materials represent a largely untapped historical resource for future historical linguistic and any number of other studies. A project that provides a possible roadmap on how it would be possible to make use of, and present these historical materials is Ticha[186] (Allen et al., 2016; Lillehaugen et al., 2016; Broadwell et al., in press) in which historical Zapotec texts (religious, linguistic, wills, bills of sale, etc.) from the colonial period are being digitized, transcribed, translated and presented in an Omeka-based[187] online hub which includes parallel digital editions and allows for crowdsourcing.

5.4 Brief Overview of Mixtecan Linguistics Literature

The earliest modern linguistic research in Mixtec was undertaken in the 1930's by Kenneth Pike of the Summer Institute of Linguistics (SIL) studying the San Miguel el Grande

[186] https://ticha.haverford.edu/
[187] https://omeka.org/

variety (ISO 639-3: mig). Cornelia Mak published research on MIG, and the varieties spoken in San Esteban Atatláhuca (ISO 639-3: mib), Santo Tomás Octopec (ISO 639-3: mie) as well as comparative studies of the tonal systems of MIG and MIB varieties in 1953 and the MIG, MIB, and MIE varieties in 1958 (Mak, 1953, 1958).

The PhD thesis of Robert Longacker proposed a Proto-Mixtecan system largely based on the comparative data provided by Mak (Longacre, 1957) and in 1960 Mak and Longacre co-authored a revised analysis which considered additional data that had been collected from more Mixtec varieties (Mak and Longacre, 1960). In 1961 Longacre and René Millon proposed a system of Proto-Mixtec-Amazugo bringing together comparative data linking the two closely related sub-branches of the Oto-Manguean language family. Further reconstructions of Proto-Mixtec were published based on comparative Mixtec data by Josserand (1983) which presented an in-depth description of Mixtec dialectal typology; finally, Dürr (1987) presents a reconstruction on the tonal system. These publications, especially Josserand (1983) are particularly important in the field of Mixtecan historical and comparative linguistics.

While there are too many individual publications on different varieties of Mixtec to be named herein, the studies by Brugman and Macaulay of Chalcatongo Mixtec (Brugman, 1983; Brugman and Macaulay, 1986; Macaulay, 1982, 1985, 1987a,b, 1990, 1993, 1996, 2005, 2011, 2012; see also Macaulay and Salmons, 1995) are significant both for the depth of linguistic coverage of a Mixtec variety, as well as for its origins and methodology. As pointed out by McKendry (2013), these represented a new development in the study of Mixtecan languages as the project's consultants (at least in the early stages) were residents of California and were members of an expatriate community, thus allowing them to initially conduct research outside of the home region of the speakers.

5.4.1 Other Mixtec Related Projects

There are several particularly significant initiatives that are working for the interest of the larger Mixtec community and other indigenous communities on the Central Coast of California (though their scope is well beyond language documentation). One such organization is the

Mixtec/Indígena Community Organizing Project (MICOP)[188], which is indigenous led and serves many functions in the Mixtec and other immigrant communities in Ventura county California and works to build community leadership and self-sufficiency, education, interpretation, health outreach various skills/job training programs and cultural promotion. Additionally, the MICOP organization runs a radio station *Radio Indigena*[189], which broadcasts segments in indigenous languages, including different varieties of Mixtec. MICOP coordinates with the Linguistics department of the University of California, Santa Barbara (UCSB) in creating collaborative, community-based programs aimed at fostering language maintenance, Mixtec literacy, social justice, which are collectively referred to as the Mexican Indigenous Language Promotion and Advocacy project (MILPA)[190] (Bax et al. 2019; Campbell and Bucholtz, 2017; Hernández Martínez et al., in press). Within this context, community members participate in graduate linguistics courses at UCSB and are fully involved in collaborating in the linguistic analyses, and other field-methods activities, e.g. phonological analysis, transcription of spoken language, audio and video recording, translation, grammar writing, archival, etc., (Bax et al. 2019; Campbell and Bucholtz, 2017; Hernández Martínez et al., in press). Notably, as a result of this program in 2019-2020 a grammar of Mixtepec-Mixtec is currently in progress (Salazar et al., 2020).

There are numerous web and social media based initiatives that have been increasingly active and producing new content. Conocelos (http://conocelos.mx/inicio/) is a community-led project by a number of indigenous language speakers (including several varieties of Mixtec) in Mexico which is building a tool to translate between indigenous languages and to build a collection of resources such as stories and vocabulary resources. Figure 57 shows a recent entry from Conocelos during the Covid-19 pandemic in spring 2020[191]:

[188] http://mixteco.org/about-us/
[189] http://mixteco.org/radio/
[190] The work done in MILPA is associated with the (NSF Grant #1660355) https://nsf.gov/awardsearch/showAward?AWD_ID=1660355&HistoricalAwards=false
[191] Note that resources pertaining to COVID-19 created in numerous varieties of Mixtec should be a rich source of comparative cognate data for future comparative vocabulary building.

Figure 57: Covid-19 public health advice 'stay at home' in MIX

Another initiative of note is a Facebook page "Tu'un Savi" (https://www.facebook.com/tuunsavi20/) which produces diagrams with vocabulary and often videos of different varieties of Mixtec, including Mixtepec-Mixtec. Videos produced on this page are often also shared on YouTube as well.

Another recent project in progress is Mesolex[192] (Lexicosemantic Resources for Mesoamerican Languages), which is not specific to Mixtec, but Mixtec varieties make up a significant portion of the dataset and the target languages. The primary component of Mesolex is a portal with two modules which seeks to ingest and disseminate lexical databases including dictionaries mapping the data structures of the source materials to TEI data and metadata. Also included will be the capacity to include audio and video content for the given indigenous language resources deposited therein.

[192] (DEL Grant #HAA-266482-19) https://securegrants.neh.gov/publicquery/main.aspx?f=1&gn=HAA-266482-19

5.5 Mixtepec-Mixtec Literature

The first study of any aspect of Mixtepec-Mixtec was Pike and Ibach (1978) who described the phonetic and phonological inventory. From 2004 to 2010 Mary Paster and Rosemary Beam de Azcona published a series of papers on the language's phonology, morphology and the role of lexical tone in Paster and Beam de Azcona (2004, 2005) and Paster (2005, 2010). The primary consultant for these studies was one of the two primary consultants/collaborators for this work as well, and they described the linguistic variety as 'Yucunani Mixtec' rather than Mixtepec-Mixtec. While apart from the TEI dictionary (Bowers and Romary, 2018: see section 7), there is not any other dictionary of Mixtepec-Mixtec, however Vocabulario Básico Tu'un Savi-Castellano (Galindo Sánchez, 2009) is a dictionary created for the variety of Mixtec spoken in Veracruz by descendants of a migrant community who originally came from San Juan Mixtepec in the 1940's.

Nieves (2012) discusses ceremonial speech (*El Parangón*) observable in certain civic and religious ceremonies, in which several interesting rhetorical devices are found including parallelisms, metaphor, metonymy and other which are used in ritualistic speech.

Finally, as mentioned in section 2.1, Bowers (in press) presents an in depth study of Mixtepec-Mixtec body-part terms (henceforth 'BPT') in which, in line with the theory of embodiment (Lakoff and Johnson, 1980a,b; Johnson, 1987) there is an expansive network of extended senses as the head component of a compound, in multi-word expressions and polysemous forms which have arisen in the language via metaphor and metonymy in lexical innovation and grammaticalization. These extensions pertain to part-whole terms for objects (meronymy), spatial relations, relational concepts of differing levels of abstraction, as well as grammatical functions. Bowers (in press) adds both collaborative evidence to the issues discussed for related varieties of Mixtecan (Brugman, 1983; Brugman and Macaulay, 1986; Hollenbach, 1995; Langacker, 2002), as well as bringing several previously unobserved extensions into the discussion and presenting a more granular account of the motivating cognitive and conceptual sources. Central to this work is the detailed analysis of: the schematic knowledge sources of the extended BPT; lexical and cognitive strategies responsible for certain

semantic changes, and the diachronic directionality, both on the semantic, and grammatical levels of the language.

6. On the Corpus: Encoding, Annotation, Contents

In this section I give an inventory of the major components of the corpus and a description of the tools and formatting techniques used, as well as an overview of significant document/resource typologies. The description of these resource typologies and my approach to integrating them into the corpus is particularly relevant in that they represent a wide array of lexical resources, one or more of which are likely to be found in any LD project[193]. It should be noted that the annotation process is still ongoing and thus at the time of submission, not all resources will have the annotation structures to be described in the section fully implemented.

6.1 Audio and Video Repository

The spoken language resources in this project comprise of the following:

- recordings and videos (made with or by project collaborators);
- recordings and videos found online;
- transcriptions of spoken language not recorded;

The entirely of the audio and video recordings created over the course of this work (for which written informed consent has been obtained) have been published as an archive titled: *Mixtepec Mixtec Language Resources* on Harvard's Dataverse (Bowers, Salazar, and Salazar 2019)[194].

[193] Though due to the fact that there are a practically innumerable potential number of sources that have been and continue to be acquired and integrated into the project, a definitive enumeration of the resources and the formatting practices is always subject to change.
[194] https://doi.org/10.7910/DVN/BF2VNK

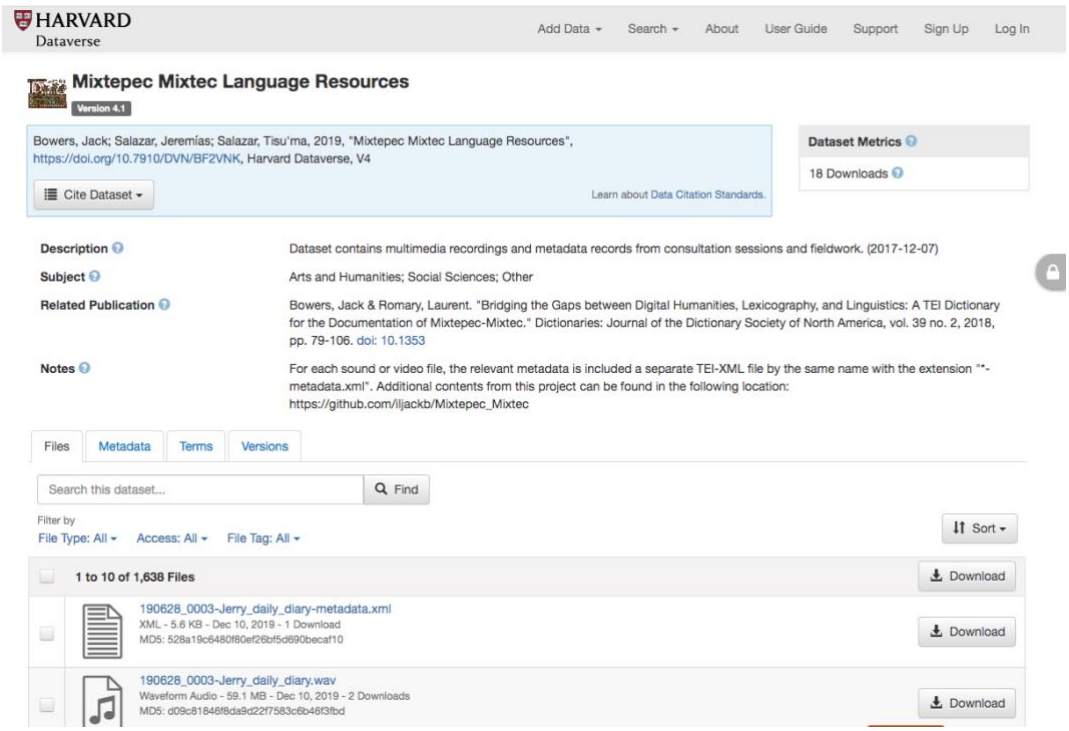


Figure 58: Screenshot of Archive or MIX media files on Dataverse

At the time of submission there are 837 audio files, 5 video and TEI metadata records for each, in which key data points are recorded. Each file (both media and metadata) can be freely downloaded and has a unique DOI, thus each can be cited individually. The need for long-term persistent identification of datasets are the underlying principles of Harvard Dataverse (King, 2007)[195]. Such resources and infrastructures as Dataverse represent a move to recognize all aspects of scientific and scholarly work, and their user friendly design reduce the barriers to making such deposits, accessing the data and with the fact that they are legally published materials with clearly stated citation information (at least in the case of Dataverse), they provide an extra professional incentive to making ones data open and accessible[196].

[195] While at present the only content archived via Dataverse is the actual recordings, videos, some fieldnotes related to consultation session and TEI files containing relevant metadata, at a later stage additional content such as transcriptions and full corpus files may be added.

[196] By design, the Dataverse interface should allow for file previewing, which would be ideal for audio and video contents (as well as for the respective corresponding metadata files) and would represent a more accessible type of repository than the major traditional archives used in LD such as AILLA, DOBES, etc. for which users must apply for access. However, at present the preview function is not working for certain types of files, including .wav thus this features is not yet available. It has been discussed with the Dataverse developers and there is hope that this can eventually be resolved.

The Harvard Dataverse repository service automatically generates metadata for the Mixtepec Mixtec Lexical Resources archive in: DCMI, OAI_ORE, Schema.org JSON, and several other formats, however it does not generate these for the actual TEI files deposited therein, which are dedicated solely to documenting the key metadata for the accompanying lexical resource files (currently mostly audio and video recordings). This latter kind of metadata and its specific instances within this, or any other project is of course the most important, and its expression is the sole purpose for the existence of the OLAC and IMDI metadata schemes. Thus, as discussed in section 4.4.1.6 defining the correspondences between these three systems is of major importance both for the field of interested communities in the present and future, as well as to the prospect of this project producing the most optimal output in terms of the best practices discussed in this section.

6.2 Text-based Resources

The sources of written materials in this project are from the following:

- booklets and papers published by the Summer Institute of Linguistics (SIL) (*roughly 27,000 tokens*);
- written material created in this project by speakers;
- documents on Mixtec containing examples from others researchers (*namely Mille Nieves*);
- a set of public safety documents published by the Mexican government[197];
- excerpts from any written communication from speakers;
- a small number of previous publications on the language[198];

[197] These have not yet been made into a corpus because of the layout, it is likely better to just study and extract the language content as needed a place in dictionary.
[198] Specifically: Pike and Ibach (1978); Paster and Beam de Azcona (2004, 2005); Paster (2005, 2010).

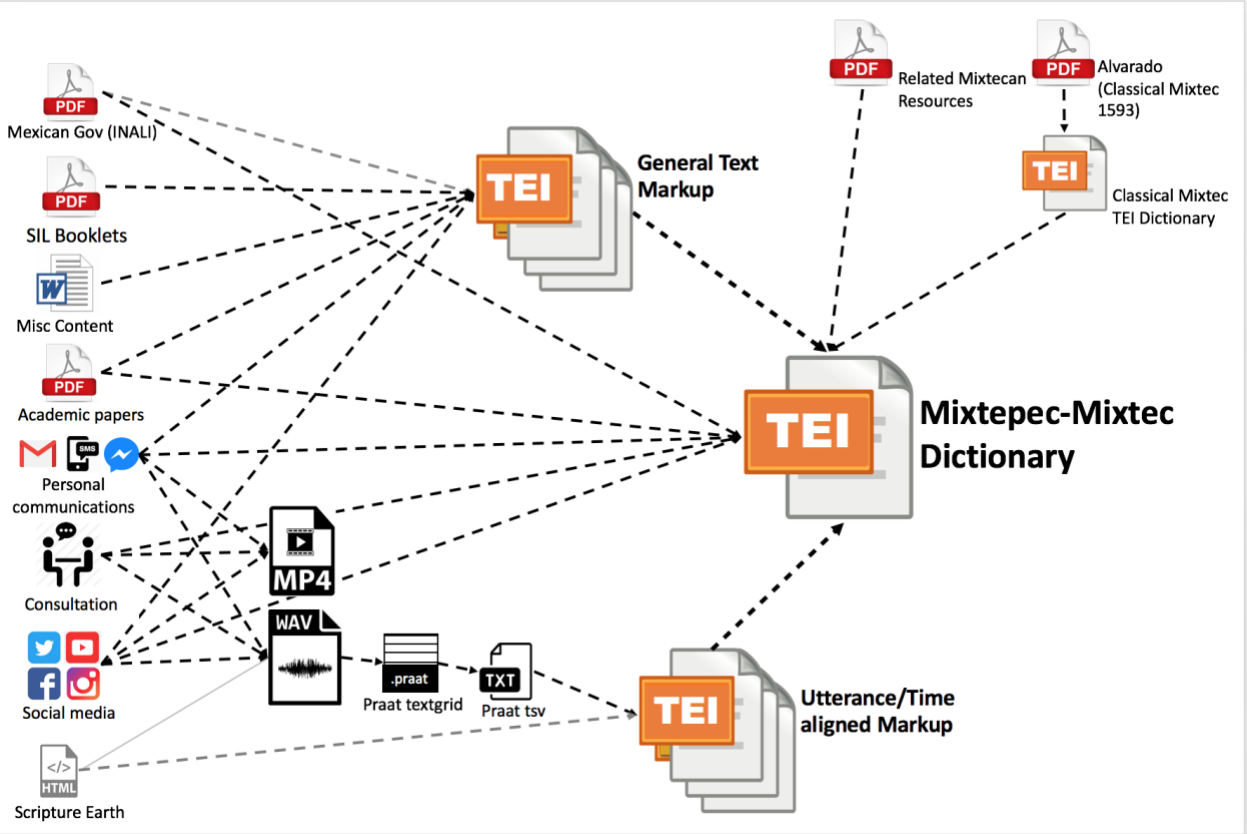


Figure 59: Workflow of sources of Mixtec-Mixtec (and related language materials) and their conversion to different TEI document types

Of these resources: only SIL booklets, the writings created by project collaborators for the purpose of this work, and a limited number of documents in which there are example sentences are encoded into TEI. Thus, with only a few exceptions[199], the content from academic papers, the pdf in the public safety documents from the Mexican government (SEGOB - Secretaría de Gobernación[200] and Sistema Nacional de Protección Civil[201]), and content obtained in personal communications are simply manually noted and recorded in the dictionary. At the time of submission, there are only two primary text-based resources[202] that are included in this project, those from the SIL booklets and publications and the diary written in Vienna by speaker/collaborator Tisu'ma Salazar in 2017.

[199] Exceptions include Nieves (2012) paper on Mixtecan ritual speech which has a significant amount of vocabulary and example sentences whose context are important and thus it was desirable to encode as part of the corpus.
[200] https://www.gob.mx/segob
[201] http://www.proteccioncivil.es/sistema-nacional
[202] Although as will be described, as part of the overall corpus there are TEI XML versions of Praat-born spoken language transcriptions, these are distinguished in this description from those that were created as text.

6.2.1 SIL Text Content and Structure

The text component of the MIX TEI corpus comprised of the SIL documents is made up of mostly publications which are booklets whose target audience is Mixtec children[203]. Structurally, these booklets are generally classifiable in the following types:

- Prose (short stories, legends, etc.);
- Activity workbooks (picture based exercises, crossword puzzles, mazes);
- Vocabulary and pedagogical reference;

It is however possible for these categories to be mixed, for example there are documents in which there is a story in prose, but then at the end contains a worksheet of some kind for readers to fill in. Additionally, there is a more recent publication[204] which is targeted to heritage speakers and learners of MIX, and it has vocabulary with accompanying audio files. Each of these types of course requires different encoding in TEI. In addition to the structural typology, there is a shallow conceptual taxonomy which has been applied, the categories are as follows:

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

In TEI, this is described in the header, within <classDecl> with the <taxonomy> element.

```
<classDecl>
  <taxonomy xml:id="tax.sil-mix">
    <category xml:id="pedagogical">
      <catDesc>PEDAGOGICAL</catDesc>
```

[203] The use and encoding of the SIL documents is done with the consent of SIL Mexico with non-exclusive re-usage permission. Note that at the time of submission, there have been several additional booklets that have been added to the SIL Mexico page but which are actually much older and the PDF's are just scans of type-written text. Given that these require additional work to integrate, I have not included these in the TEI encoded corpus (though it will be done at a later time)

[204] https://mexico.sil.org/resources/archives/82562

```
      <category xml:id="pedagogical-inter">
        <catDesc>PEDAGOGICAL:INTERACTIVE</catDesc>
      </category>
      <category xml:id="pedagogical-ref">
        <catDesc>PEDAGOGICAL:REFERENCE</catDesc>
      </category>
    </category>
    <category xml:id="fiction">
      <catDesc>FICTION</catDesc>
      <category xml:id="fiction-fantasy">
        <catDesc>FICTION:FANTASY</catDesc>
      </category>
      <category xml:id="fiction-realistic">
        <catDesc>FICTION:REALISTIC</catDesc>
      </category>
    </category>
    <category xml:id="folklore">
      <catDesc>FOLKLORE</catDesc>
    </category>
  </taxonomy>
</classDecl>
```

Figure 60: Taxonomy of SIL documents in MIX corpus as per TEI header

In the sections below, the TEI encoding of these document structures the above will be described.

6.2.2 Text Document Metadata: <teiHeader>

The foremost component of the <teiHeader> is the title statement <titleStmt>. For as many languages the given document's title is written in, it is placed in a <title> element with the given language declared in @xml:lang. The authors and editors of the original content are given the <author> or <editor> labels, in cases where the role of a participant doesn't have the appropriate built-in TEI tag, the <respStmt> and the specific role is given in the <resp> (responsibility) element. It is often the case that a person requires multiple instances of <resp> as many participants carry out multiple roles in the creation and/or annotation of any given resource. Within <respStmt>, the person's name is placed in <name>, which is given an @xml:id as it is common for a person to be tagged in various annotation functions, particularly in assigning responsibility for a given translation or interpretation. Figure 61shows a typical example of a <titleStmt>.

```
<titleStmt>
  <title xml:lang="mix">Tu'un yata tsa'a kue kaa kaxi Xnuviko</title>
  <title xml:lang="es">La leyenda de las campanas de Mixtepec</title>
```

```
<author>Francisco Mendoza Santiago</author>
<editor>Gisela Beckmann</editor>
<editor>María Gómez Hernández</editor>
<respStmt>
  <resp>TEI Encoding</resp>
  <resp>Annotation</resp>
  <resp>Glossing</resp>
  <name xml:id="JB">Jack Bowers</name>
</respStmt>
<respStmt>
  <resp>Glossing</resp>
  <name xml:id="TS">Juan "Tisu'ma" Salazar</name>
</respStmt>
</titleStmt>
```

Figure 61: TEI <titleStmt> with title, author and secondary participant information

In each document which has come from a published source (e.g. SIL documents), the necessary provenance and bibliographic details are given in <sourceDesc>, with a <bibl> elements, and a statement along with a pointer to the source of the text in the value of @target in the <ptr> element.

```
<sourceDesc>
  <bibl xml:id="bibl.156">
    <title xml:lang="mix">Ntintsitsa ntivixi</title>
    <author>Gómez Hernández, María</author>; <editor>Beckmann, Gisela</editor>;
      <editor>Nieves, María M.</editor>. <date>2008</date>. <edition>(2nd
      ed.)</edition>.<publisher>Instituto Lingüístico de Verano, A.C.</publisher>
    <pubPlace>Tlalpan, D.F., México</pubPlace> Obtained from:
  <ptr target="https://mexico.sil.org/resources/archives/55533"/>
  </bibl>
</sourceDesc>
```

Figure 62: Bibliography for SIL source document declared in the TEI header <sourcDesc>

Where the source has an abstract of the content, this is placed within the header in the <abstract> element with the language attribute @xml:lang. As (at least to date) the only instances of this is in the SIL documents and are in Spanish, thus the value of which is always "es"[205].

```
<abstract xml:lang="es" xml:id="L157-resumen">
  <p>Cuenta la leyenda que la gente de Mixtepec fue hasta Puebla a conseguir unas campanas para su iglesia.
De regreso, anunciaron su llegada desde un monte tocando las campanas. Es por eso que ese cerro se llama "Monte
```

[205] While depending on the version of the given source SIL document in which the abstract may occur in the original in the front or in the back, this content is always included in the <abstract> element which necessarily occurs in the TEI header, thus in front of the main content.

de la Campana". Además, esta leyenda explica la razón por la que las campanas de Pinotepa Nacional suenan igual que las de Mixtepec.
</abstract>

Figure 63: Example of <abstract> element from TEI encoding of SIL document

6.2.3 SIL Documents: Basic TEI Document Structure

In the corpus of text documents whose contents are prose in nature, the encoding is done in line with typical TEI practice in text segmentation. All the main content of a document is contained within the <body> element and the document type according to the taxonomy is declared on the @decls attribute on <text>. Where there are either pages and/or distinct segmentations in the original source (due to topic or other specific distinct content), these are represented by element and are given distinct @xml:id values.

Images are encoded as they appear using the <graphic> element, which is often embedded in the <head> element as (particularly in the SIL documents), the image is the head feature of the given page in the original document. The specific image is referenced using the @url attribute which points to its location in the project directory.

Where the content is organized by paragraphs, the element is used to wrap the actual MIX content which is encoded in <seg> which takes the @type attribute to distinguish between where the content is a full sentence e.g. (<seg type="S">), a phrase (<seg type="phrase">), a general lexical term in isolation (<seg type="term">), or a caption occurring in an image and not in actual text, or in interactive documents where there is a blank space (<seg type="blank">). Each <seg> is labeled with a language tag @xml:lang, and each is given a unique @xml:id. Finally, each token (except where the <seg> is a blank space) is encoded as <w>, which is also given a unique @xml:id which serves as a target for annotation. Punctuation characters are encoded as <pc> (punctuation character). Note that the contents of <w> do not necessarily represent a full lexical item as there are many compounds in MIX which are spelled with whitespace, the means with which these are joined in the annotation of the corpus is described in the following section.

Thus, a typical example from file L157-tok.xml (Mendoza Santiago, 2008) of all of the above is shown in Figure 64, in which on the left the source from the original PDF document is displayed with the given TEI encodings on the right.

Figure 64: Source content (image and text) from SIL document and TEI encoding structure

6.2.3.1 SIL Document Types: Pedagogical Reference

Documents which are pedagogical references can either be booklets with prose explanation of given themes with one or more illustrations, or they can be reference or vocabulary lists of MIX words along with an accompanying image; in some cases, there may be Mixtec-Spanish bilingual material (though I will discuss these encodings in following sections). Of the SIL resources (at the time of submission) eleven documents[206] are either full or partial

[206] The following encoded documents can be found in the projects GitHub directory under the subsequent folders by the same names minus the '.xml' extension (https://github.com/iljackb/Mixtepec_Mixtec/tree/master/SIL_docs):

pedagogical references (note documents can be both referential and interactive as many have reference in the some of the content as well as interactive contents at the end or interspersed throughout)[207]. Where the content is simply prose, the previous example (Figure 64) is typical of the TEI, thus it isn't necessary to show any further examples of the encoding (though numerous more examples will be shown in the context of explaining additional features of the encoded corpus in following sections).

Of those which are vocabulary or reference content in the source, there are generally two main TEI structural encoding approaches. The first is to use TEI <list>, which is used where vocabulary is presented in a sequential linear order. The second is where the vocabulary is presented alongside images that it corresponds to, if the layout is not linear the tag is based on <ab> (anonymous block) units[208]. In each case the vocabulary is also encoded further, as the Mixtec vocabulary is encoded as <w> elements (with unique @xml:id's) within a <seg type="term"> which is included in order to provide a wrapper in the case of compounds and multi-word expressions. In the Figures (65 and 66), I show the encodings of the two prototypical examples from document sources: L331-tok.xml (Beckmann and Nieves, 2011) and L100-tok (Beckmann and Nieves, 2012) respectively:

L093-tok.xml, L097-tok.xml, L100-tok.xml, L105-tok.xml, L144-tok.xml, L145-tok.xml, L151-tok.xml, L162-tok.xml, L331-tok.xml, Las_aves-mix.xml, Aprendamos_el_idioma_mixteco_(Mixtepec).xml.

[207] The resource titled "Aprendamos la idioma mixteco" is a unique resource in that it is a vocabuarly learning booklet but comes with audio files to accompany the text, thus the approach to encoding this resource involved spoken language annotation (Praat) as well as general text annotation.

[208] While within TEI there is a much more sophisticated system of marking up of images and text than what is done here, which is often used in annotation of manuscripts, these documents are not historical and the main purpose of this work is to make use of the language content, thus I have not chosen to use the maximum capabilities of the TEI.

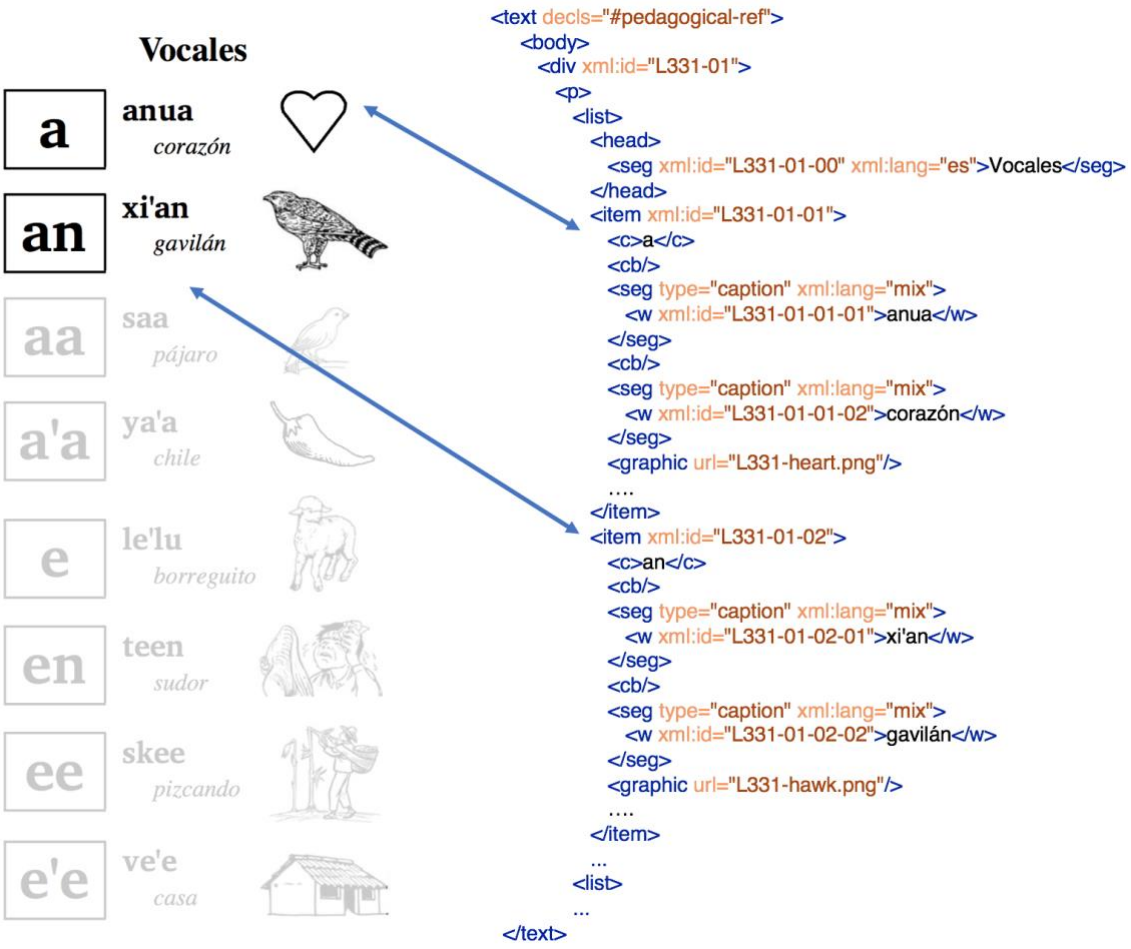
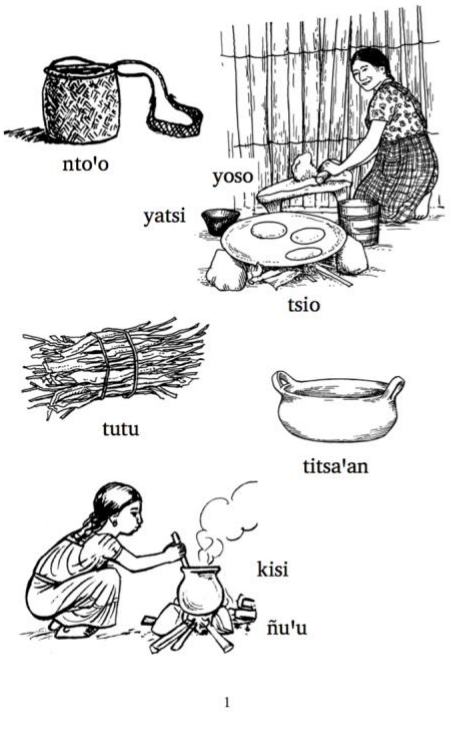


Figure 65: Side by side comparison of vocabulary from pedagogical document in TEI <list>

Figure 65 shows the encoding of the orthographic conventions used by SIL along with sample vocabulary, an image, and a Spanish translation. In the TEI, the ordering of each object is maintained in the sequence of encodings. The character element <c> is used to encode the orthographic character in question and the column break element </cb> is used along with the elements described above in order to maintain the source formatting and division of contents.




```
<text decls="#pedagogical-ref">
  <body>
    <div xml:id="L100-01">
      <ab><graphic xml:id="L100-01-01-img" url="L100-sling-basket.png"/>
        <seg type="term" xml:lang="mix">
          <w xml:id="L100-01-01" corresp="#L100-01-01-img">nto'o</w>
        </seg></ab>
      <ab><graphic xml:id="L100-01-02-img" url="L100-crock-pot.png"/>
        <seg type="term" xml:lang="mix">
          <w xml:id="L100-01-02">titsa'an</w>
        </seg></ab>
      <ab><graphic xml:id="L100-01-03-img" url="L100-bundle-sticks.png"/>
        <seg type="term" xml:lang="mix">
          <w xml:id="L100-01-03">tutu</w>
        </seg></ab>
      <ab><graphic xml:id="L100-01-04-img" url="L100-cooking_woman-yoso-yatsi-tsio.png"/>
        <seg type="term" xml:lang="mix">
          <w xml:id="L100-01-04">yoso</w>
        </seg>
        <seg type="term" xml:lang="mix">
          <w xml:id="L100-01-05">yatsi</w>
        </seg>
        <seg type="term" xml:lang="mix">
          <w xml:id="L100-01-06">tsio</w>
        </seg></ab>
      <ab><graphic xml:id="L100-01-05-img" url="L100-cooking_girl-pot-smoke.png"/>
        <seg type="term" xml:lang="mix">
          <w xml:id="L100-01-07">kisi</w>
        </seg>
        <seg type="term" xml:lang="mix">
          <w xml:id="L100-01-08">ñu'u</w>
        </seg></ab>
    </div>
    ....
</text>
```

Figure 66: Side by side image of source document with mixture of images and text with TEI encoding

In the encoding shown in Figure 66, the images are grouped in <ab> elements along with the given MIX vocabulary items that in the source is placed alongside the image. Whereas in the <list> documents, the ordering of the contents in the source is important, in these it is less so given that in the source the text is simply placed next to the item in the image[209].

6.2.3.2 SIL Document Types: Activity Books

Of the SIL resources (at the time of submission), there are eight documents which are full or partial activity booklets[210]. While these documents are not identical in their content and degree

[209] If this document were historic in nature and it was desirable to encode the relation of the non-linear text with regard to the page and image, it could be done in TEI using the elements: <surfaceGrp> (https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-surfaceGrp.html) and <surface> (https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-surface.html).

[210] The following encoded documents can be found in the projects GitHub directory under the subsequent folders by the same names minus the '.xml' extension (https://github.com/iljackb/Mixtepec_Mixtec/tree/master/SIL_docs): L094-tok.xml, L095-tok.xml, L104-tok.xml, L105-tok.xml, L144-tok.xml, L160-tok.xml, L162-tok.xml, Cruxigramas-tei.xml

of interaction, they are given the taxonomic classification “PEDAGOGICAL:INTERACTIVE” (introduced in section 6.2.1). Depending on whether the full document or just a given section is interactive or not, this feature can be tagged on the <text> or the element as: @decls="#pedagogical-inter". Several key examples of the TEI encoding applied are described below. The main feature shared in all of the interactive contents is a blank space with the purpose of the user inserting the correct vocabulary content. These are encoded within separate blocks as: //seg/span with a sequence of underscores as the value of , e.g. “_____”. The encoding of blank space, especially for this purpose is currently an unestablished area of TEI, and thus this solution may be changed in the future.

The most prototypical example of an interactive document will have an image stimulus on which the missing vocabulary should be determined. The following example from L162-tok.xml (Beckmann and Nieves, 2008a) shows a side-by-side image of the source document and its TEI encoding.

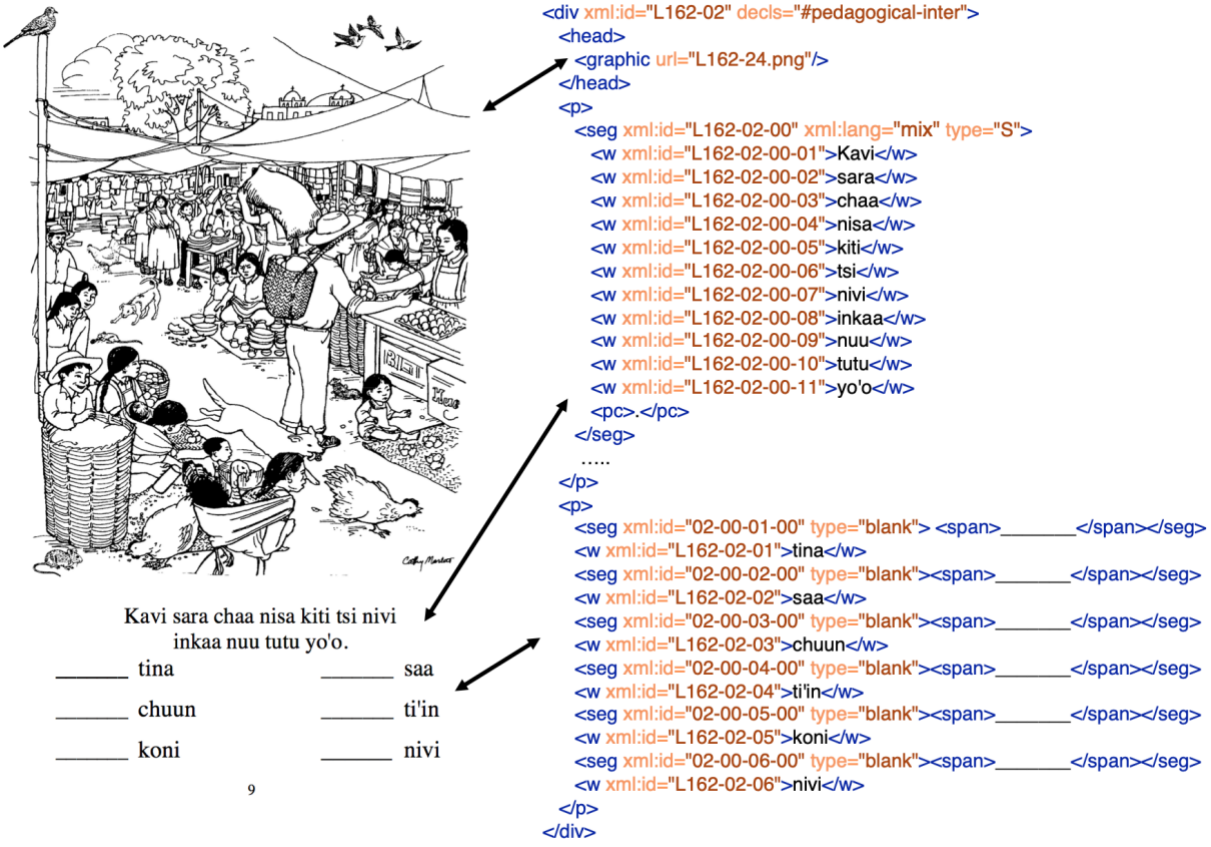
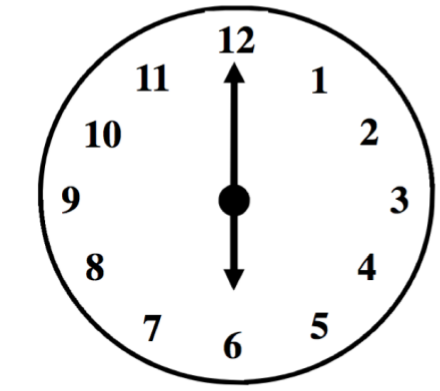


Figure 67: Left original interactive source; right TEI encoding

In the following example from L094-tok.xml (Nieves and Beckmann, 2007a), there is some additional content necessary to express the information in TEI as given that the example image is an analogue clock, there are two possible correct times being shown on the clock, and two possible thus answers as well[211]. This is encoded in TEI using the <choice> attribute, and since neither value is in fact written in the text, the values are recorded in the attribute @when in the following XML structure //seg/time[@when][212].




```
<div xml:id="L094-01">
  <head>
    <graphic url="L094-1-600.png"/>
  </head>
  <label>
    <choice>
      <seg>
        <time when="06:00:00"/>
      </seg>
      <seg>
        <time when="18:00:00"/>
      </seg>
    </choice>
  </label>
  <lb/>
  <p>
    <seg xml:id="d1e160" xml:lang="mix" type="S">
      <pc>¿</pc>
      <w xml:id="d1e163">Nchii</w>
      <w xml:id="d1e165">hora</w>
      <w xml:id="d1e167">kui</w>
      <pc>?</pc>
    </seg>

    <seg xml:id="L094-01-02" xml:lang="mix" type="blank">
      <span>_____</span>
    </seg>
    ……..
  </p>
</div>
```

Figure 68: Encoding of time choice in interactive pedagogical document L094 with ambiguous answers

While this hasn't been within the scope of this early stage of the work, and it would perhaps require more discussions with the SIL publishers, it would be possible to make interactive online versions of these workbooks as a pedagogical application in which people

[211] Though currently the answers are not being added in every document, including such information could be useful in compiling a more comprehensive dataset which could be reused for interactive pedagogical purposes.
[212] The <seg> is used due to the fact that <time> cannot occur directly in <choice>.

could insert their answers and get feedback. At very least these examples provide a format which could easily be used as a basis for the creation of interactive pedagogical content using TEI in combination with other technologies.

6.2.4 Speaker Authored Text

As part of a working research trip to Vienna, one project collaborator produced a diary of the trip to Europe from his home in California[213]. This was designed dually with the purpose of creating additional written MIX contents beyond the SIL publications, as well as for him to gain practice in writing in MIX. This document was written in a word file and converted to TEI; it contains 3,317 tokens, and roughly 1000 distinct lexical items and phrases. From the point of view of building a multilingual annotated corpus, this was strategically done with knowledge that the researcher (myself) was there with him for most of the events that are described in the text. Thus, given the combination that the vocabulary usage was mostly within my knowledge and the fact that I already knew what was being described, it provided an advantageous set of vocabulary which enabled much of the translation to be done with minimal assistance from the author or other native speakers. Though the TEI structure of the document does not differ in any significant way from the SIL prose resources, it is nonetheless a unique source of written language that is composed simply for the purpose of recording events, providing an additional type of language content to the corpus.

Though I do not claim this project to delve into the domain of language revitalization, the composition of a daily diary by a native speaker is an example of the type of expansion of the domains of language usage that would represent potential avenues of language revitalization that may be pursued by the speech community in the future. Additionally, once presented in a user-oriented output, the materials created could hopefully be used as resources in revitalization endeavors by providing a template for speakers to copy in documenting their own day to day activities. For discussions about language revitalization in general see: (Hinton and Hale, 2001; Grenoble and Whaley, 2006; Tsunoda, 2013; Galla, 2009), and for revitalization related to Mixtec specifically see (Campbell et al., in press; Hernández Martínez et al., in press; Reyes Basurto et al., in press).

[213] https://github.com/iljackb/Mixtepec_Mixtec/blob/master/misc-sources/Tisu-Vienna-Diary-201711.xml

On the linguistic, lexicographic, and perhaps even anthropological sides, this document presents a very important, though not unique issue of vocabulary. As within this document, the author is describing his trip to Vienna, throughout which he describes his trip to the airport using public transportation, his stay in the hotel and important landmarks he sees throughout his trip to Austria. Given that all of these things and places are domains which are of course non-native to the Mixtec region, they are also not native to the language and thus there is a very high quantity of Spanish loanwords.

An additional area of importance of this document has to do with important editorial decisions, specifically with regard to how to deal with spelling variations, as most Mixtec speakers do not regularly use the working orthography system, thus resulting in a significant amount of variation which need to be normalized in order for the corpus contents to be as consistent as possible.

6.2.5 Other text resources: Conocelos.MX

Another occasional source of MIX vocabulary is the Facebook page Conocelos.MX[214] which is dedicated to producing language content for Mexico's indigenous languages. This page is affiliated with an indigenous led project which has created a Google translate like tool[215]translating some basic vocabulary between Spanish and a number of different indigenous languages of Mexico, including several Mixtec varieties. There have been a few dozen entries for Mixtepec-Mixtec created as part of this work, and the content which is posted on the Facebook-based site is also available on the Traductor website. A major attribute of the materials being created as part of this project is that they generally use an image template which is used for each language and thus they have begun to compile an onomasiological dataset with multiple Mixtec varieties and other varieties of Mexico's language families.

[214] https://www.facebook.com/LenguasOriginariasDeMexico/
[215] traductor.conocelos.mx

Figure 69: Post from Conocelos.mx Facebook page with Mixtepec-Mixtec vocabulary

In the encoding of these contents, the text is structured in the same way as shown in previous sections, and the annotation which also applies to all documents will be discussed in section 6.4 From a lexicographic point of view, these posts do in some cases pose some challenges in that as is the case in most native speaker authored content, the spelling conventions used are not always consistent, nor do they follow the conventions used by SIL used in this project. Due to the issues of minimal pairs stemming from nasality, length, tone (which is mostly not included in writing), variation in the representation of any of these features creates homographs. Thus, integrating such contents introduces variation into the corpus which then needs to be normalized.

In addition to encoding the linguistic content, the metadata, namely the date and provenance of the posts as well as the location (of the speaker's residence) is given in <sourceDesc> as shown below:

```
<sourceDesc>
  <ab>Source from Conocelos.mx Facebook post <date>2018-11-16</date>
    <ref target="https://www.facebook.com/LenguasOriginariasDeMexico/posts/329300507662775"/>
    Same resource available using the Conocelos.Mx Traductor tool: <ref
target="http://conocelos.mx/traductor/index"/>
  </ab>
  <ab>
    <location>
      <placeName>Santiago Juxtlahuaca</placeName>
      <region>Oaxaca</region>
      <country>Mexico</country>
    </location>
  </ab>
</sourceDesc>
```

Figure 70: <sourceDesc> in TEI encoding of Conocelos.mx Facebook post

Additionally, the hashtags from the original post are maintained and encoded within the TEI header using the <keywords> element, with each hashtagged term represented as <term>. Also, the inclusion of <langUsg> with the value 'Mixtepec-Mixtec' specified in <language> and the ISO 639-3 code in @ident (note that the <langUsage> element is included in every MIX corpus document). See the full structure below:

```
<profileDesc>
  <langUsage>
    <language xml:lang="en" ident="mix">Mixtepec-Mixtec</language>
  </langUsage>
  <textClass>
    <keywords>
      <term>#Mixteco</term> <term>#SanJuanMixtepec</term> <term>#YoHabloMixteco</term>
    </keywords>
  </textClass>
</profileDesc>
```

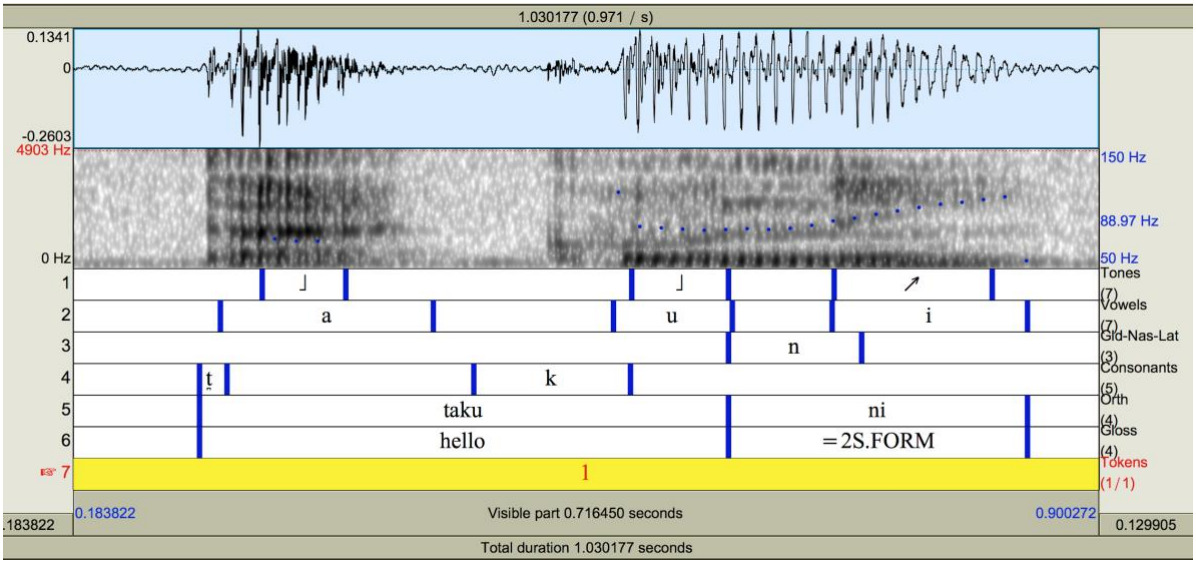
Figure 71: Representing hashtags from Facebook post containing MIX vocabulary

6.3 Spoken Language Transcriptions and TEI Encoding

Spoken language sources[216] have been annotated using Praat (Boersma and Weenik, 2020) and all MIX contents are transcribed in IPA and the working Mixtec orthography. As discussed, the reason why Praat was originally chosen, and the only major advantage of Praat over other annotation tools is that IT allows for pitch (i.e. F0) analysis, which given that MIX is a tonal language, is a necessity. Additionally, Praat has a scripting language which can greatly expedite a wide array of different functions including: annotation, file management, making modifications, qualitative and quantitative data extraction from sound files and their annotations and much more.

6.3.1 Praat Annotation Schemes

This system was designed in the initial stages of the project to be able to systematically study and extract the acoustic signal data from different categories of phonetic units as well as the tone contours in their entirety, allowing for overlap of their signals. In this schema, there were specific tiers for: tones, vowels, glides/nasals and lateral, Mixtec orthographic form, gloss (according to *Leipzig Glossing Rules*; Bickel et al., 2008), and token number (*which is necessary for the parsing of the output contents when converting the tab-separated output to TEI).*



[216] While several videos have been produced over the course of the project, none have yet been annotated, however when they are, it will be necessary to use ELAN, as Praat doesn't allow for processing of videos.

Figure 72: Example of original Praat TextGrid transcription

Due to the time needed to carry out this annotation system and the urgency to create a usable output, this methodology was changed. In fact, according to Himmelmann (2018) it can be expected that there will be a 10 to 1 temporal ratio in transcribing speech; to transcribe one minute of speech, it will take roughly 10 minutes which can be compounded in the case that more than one speaker is involved. This ratio was in fact even greater in the system implemented in the early stages of this project in which the transcription contained different tiers for each vowels, semi-vowels/glides/nasals, consonants, tones. For this reason, there is a significant need for machine learning techniques in automatic spoken language transcription. Recently there has been increasingly promising results in such methods, which if successfully applied, could greatly assist in both the rate of processing as well as enhancing the discoverability of the contents in the context of online repositories by enabling direct queries into the text contents (see: Strunk et al., 2014; Adams et al., 2017, 2018; Michaud et al., 2018, 2020; Johnson et al., 2018; Neubig et al., 2020).

While the system is for the time being not being carried out any longer, it could potentially be reused or resumed in the future. The data it produced could make use of the fairly extensive array of processes available in the Praat software toolkit including in depth quantitative study of key phonological features such as vowel quality, nasalization, voice onset time (VOT), tone contours (Figure 73), spectrograms (Figure 74, formants and much more[217]. Additionally, this detailed segmentation should prove useful as training data for automatic transcription systems in the future.

[217] See Praat guidelines: http://www.fon.hum.uva.nl/praat/

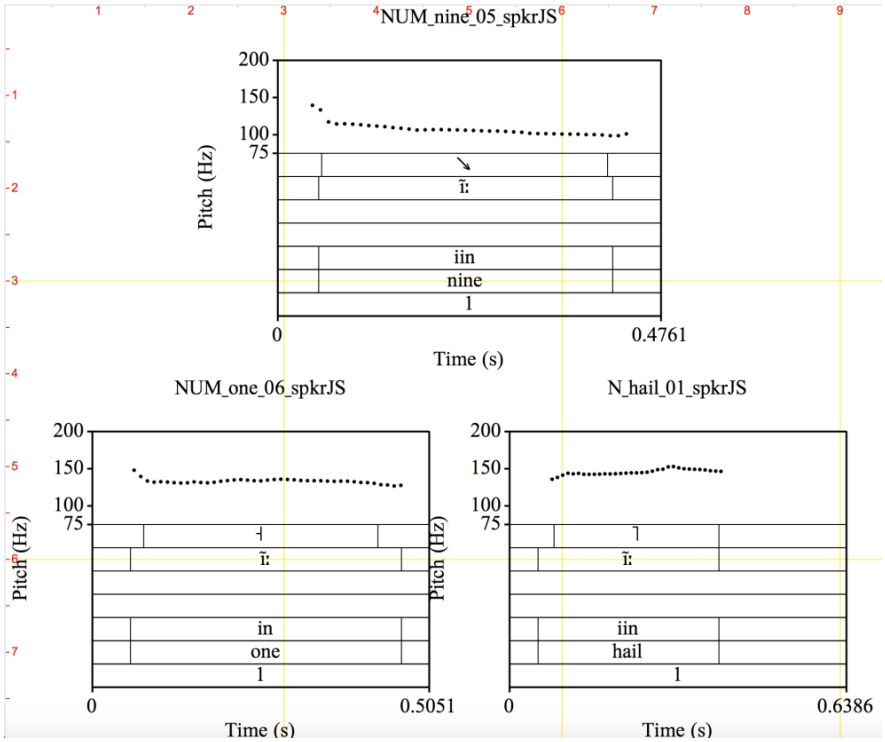


Figure 73: Plotting of F0 contour for tones of 3 transcribed MIX minimal pairs in Praat

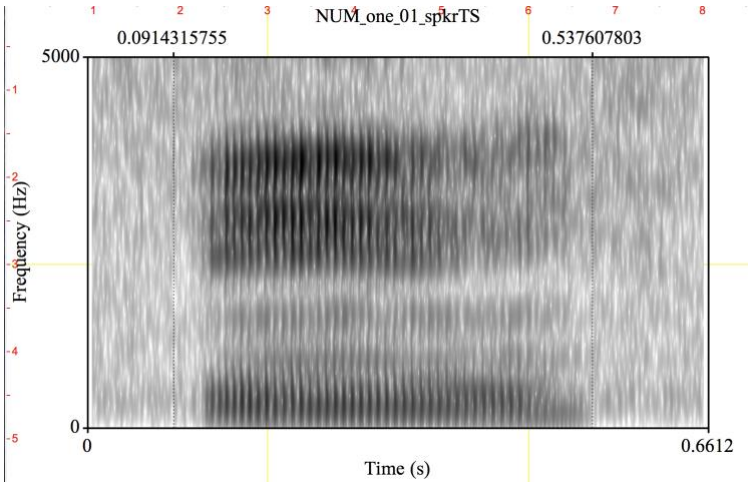


Figure 74: Plotting of spectrogram of MIX lexical item 'one' [ĩːɬ] in Praat

In the updated transcription system, the transcription schema only has tiers for the following: Mixtec orthographic form, Mixtec IPA form[218], Spanish, English and the token number representing a unique utterance in a recording (which is again, necessary for the parsing of the output contents when converting the tab-separated output to TEI). The inclusion of the separate Spanish and English tiers are for where a recording contains translations of the given Mixtec vocabulary either as glosses or potentially as elicitation prompts. For reasons of annotation speed, and the fact that not every Mixtec item or phrase can have a word for word translation or gloss, the translations are given for the full token rather than word by word which can be done at a later stage directly in TEI.

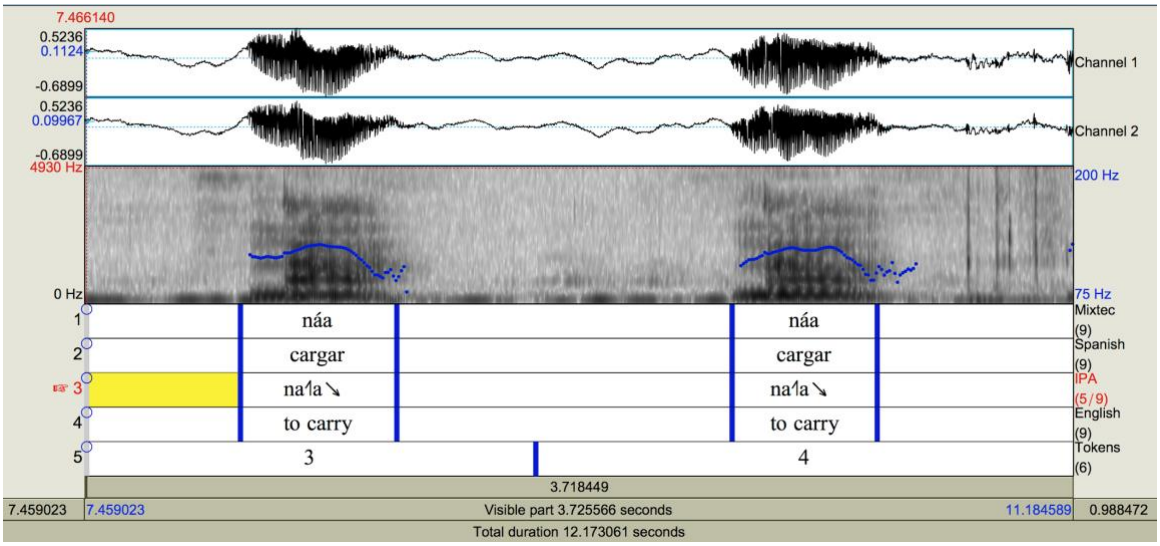


Figure 75: Example of current Praat TextGrid annotation system

Using Praat scripting, the key temporal and transcription content from the TextGrid is extracted and saved as a tab-separated text file shown in Figure 76 below:

2.04	Tokens	1	3.77
2.56	English	to end	3.18
2.56	Mixtec	naá	3.18
2.56	Spanish	terminar	3.18
2.56	IPA	naɹa↗	3.18
3.77	Tokens	2	5.08

[218] In the speech transcriptions the tones are transcribed using their Unicode stand-alone characters rather than the combining diacritics. If there is a contour and it isn't clear whether the specific onset or offset tone level is phonologically significant (as shown in Figure 75), the global-rise or fall character is used. If the tone is not known/unclear, no tone is included.

4.23	Spanish	terminar	4.91
4.23	Mixtec	naá	4.91
4.23	English	to end	4.91
4.23	IPA	naɲa↗	4.91
7.10	Tokens	3	9.16
8.05	IPA	naˀa↘	8.63
8.05	Mixtec	náa	8.63

Figure 76: Example of (partial) tab-separated file from TextGrid annotation

While it is of course not ideal to have multiple annotation schemes in the project data, the two systems contain content that is actually complimentary. In fact, with the exception that both have Mixtec orthographic forms (with the tiers named "Orth" in the earlier and "Mixtec" in the latter"), they could in fact be combined without the need to modify any of the content. It may in fact be necessary or desirable for future users to add conventions from one system to the other, particularly the individual segmentation of vowels, tones and other phone types in order to provide a full sample of these features from the entire spoken language corpus.

6.3.2 Transcribing Tones

In transcribing tones, though individual cases may vary in the specific judgment made, the policy has generally to transcribe what is heard and seen in the F0 pattern at the level it appears rather than to transcribe based on what is known (or thought to be known) about the phonological tone. Additionally, in annotating non-level tones for the most part I have chosen not to specify the start and end tone level, as I am not yet certain about the status of whether the specific levels of such tones are minimally distinctive, thus global rise or fall arrows ↗ ↘ are used. In depth study of the quantitative output from these annotations will be an area of further study moving forwards. Adams et. al 2018 discusses the use of a neural network architecture with connectionist temporal classification loss function for phonemic and tonal classification in the context of LD for the tonal languages Yongning Na and Eastern Chatino. The work described therein could provide a model that could be applied to the backlog of MIX data in the future.

6.3.3 TEI Output of Praat Transcriptions

The method for annotation all our contents adopted is based on the recommendations of ISO 24624:2016 and Schmidt (2011) (described in section 4.4.2.3). The aforementioned are used as a baseline for the standoff annotation of the spoken language transcriptions, and the method integrates the ongoing work of Bański et al. (2016) as well as the guidelines in the

Morphological Annotation Framework (MAF) (ISO/FDIS 24611:2012(E)) regarding an expansion and refinement of the TEI standoff annotation system. While specifics can vary according to the annotation scheme, the Praat output produces: a timeline, Mixtec orthographic and phonetic transcriptions, and English and/or Spanish translations; and as discussed above, the transcriptions from the earlier system produces interlinear glossed text, which is included in the grammatical annotations (see section 6.4.5). The way in which these features are represented in TEI will be described in the following sections.

6.3.3.1 Timelines and Transcriptions in TEI

The Praat TextGrid timelines for a given TextGrid and accompanying ".wav" file are represented in TEI as a <timeline> element (described also in section 4.4.2.3) which occurs as the first element within <body>. Each point throughout the timeline is where one or more of the annotation segments begins or ends. Thus, only the relevant points in the annotation timeline are represented, in TEI they are encoded as <when> elements, each with a unique @xml:id to which the transcribed language content is anchored.

```
<timeline>
  <when xml:id="T2.04" interval="2.04"/>
  <when xml:id="T3.77" interval="3.77"/>
  <when xml:id="T2.56" interval="2.56"/>
  <when xml:id="T3.18" interval="3.18"/>
  <when xml:id="T5.08" interval="5.08"/>
  <when xml:id="T4.23" interval="4.23"/>
  <when xml:id="T4.91" interval="4.91"/>
  <when xml:id="T7.10" interval="7.10"/>
  <when xml:id="T9.16" interval="9.16"/>
  <when xml:id="T8.05" interval="8.05"/>
  <when xml:id="T8.63" interval="8.63"/>
  <when xml:id="T12.17" interval="12.17"/>
  <when xml:id="T9.90" interval="9.90"/>
  <when xml:id="T10.45" interval="10.45"/>
</timeline>
```

Figure 77: Timeline for utterance annotated in Praat as represented in TEI

The points assigned to the given transcription can be used in combination with the link to the given ".wav" file by software programs to play a given utterance and display its transcription using the TEI output of the original Praat annotation.

Each separate utterance in a source recording (represented in the Praat TextGrid on the "Tokens" tier) is converted into TEI as a unique <annotationBlock> element containing an utterance <u> in which the rest of the transcription and annotations (both translations from Praat, as well as any additional annotations) are placed as well.

```
<annotationBlock>
  <u n="1" xml:id="d23e0" start="2.04" end="3.77" who="#JS">
    <seg xml:lang="mix" notation="orth" xml:id="T-seg-orth-2.04">
      <w synch="#T2.56" xml:id="T-orth2.56">naá</w>
    </seg>
    <seg xml:lang="mix" notation="ipa" xml:id="T-seg-pron-2.04" sameAs="#T-orth2.56">
      <w synch="#T2.56" xml:id="T-pron2.56" sameAs="#T-orth2.56">naɻa↗</w>
    </seg>
  </u>
  ….
</annotationBlock>
```

Figure 78: Representation of one utterance converted from Praat TextGrid in TEI

For each utterance, the full time span is explicitly stated on the @start and @end, and the initials of the speaker is labeled using @who. All contents in each the orthographic and phonetic transcriptions are encased in the <seg> element and the given transcription method is specified using the attribute @notation. Each token is represented as a <w> element which has a unique @xml:id to which annotations point, and a @synch attribute to point directly to the point (via the @xml:id value) on the timeline from which the utterance occurs. It should be noted that a <w> token (despite its definition in the TEI guidelines)[219] is not necessarily a full lexical unit or word in this project as it is simply used to wrap a string of text (see section 6.4.4 for a discussion of the specifics of how this works in the annotation scheme).

6.3.3.2 Linking and Representing Phonetic and Orthographic Forms

Another important issue in the TEI representation of the transcriptions is the need to both encode and link the phonetic (IPA) and orthographic forms. This is annotated on the <w> level as well as the <seg> level for full sentences and phrases using the @sameAs attribute which is placed on the phonetic forms which point to the @xml:id of the corresponding orthographic form. Figure 79 below shows such an example:

[219] https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-w.html

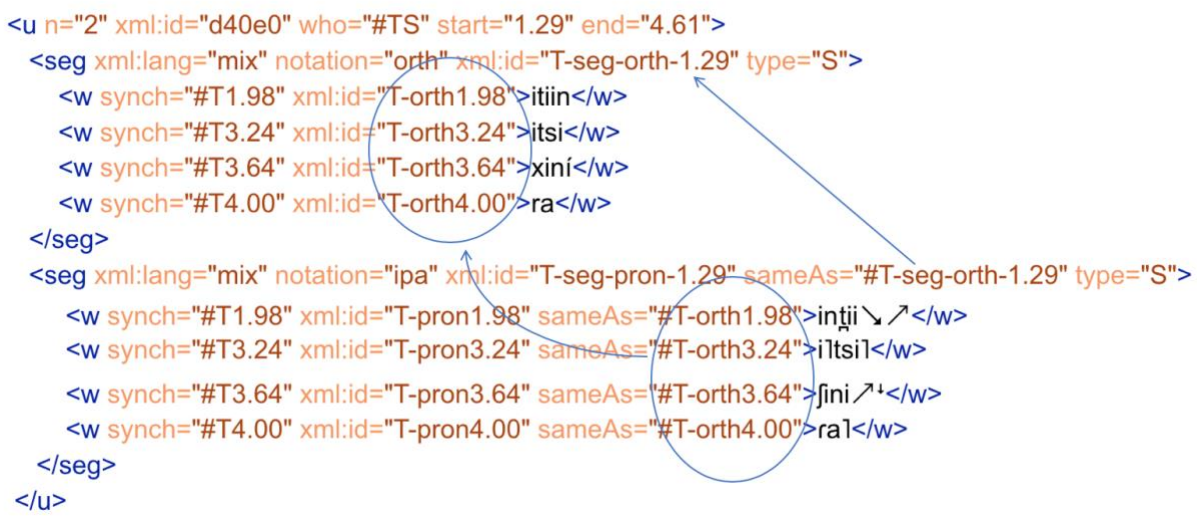


Figure 79: Shows the linking of the phonetic and orthographic transcriptions in TEI

6.3.4 Representing Spoken Resource Metadata

As in every other document, the speaker(s) who produced the language material are stated in the <titleStmt> within <respStmt>, however in cases where a speaker whose speech is in the recording also participated in the recording, interviewing process (which does indeed often occur), they be declared in both sections. So, for the recording shown in Figure 80,

```
<titleStmt>
  ….
  <respStmt>
    <resp>Transcription</resp>
    <resp>Data Modeling</resp>
    <resp>Speaker Consultation</resp>
    <name xml:id="JB">Jack Bowers</name>
  </respStmt>
  <respStmt>
    <resp>Speaker</resp>
    <name xml:id="JS">Jeremías Salazar</name>
  </respStmt>
</titleStmt>
```

Figure 80: Responsibility statement declaring speaker and primary researcher

Each of the names in the <titleStmt> are given @xml:id's which are used when it is necessary to attribute something to the given person in the document. The utterances performed by a given speaker are attributed to them explicitly in the transcription portion of the document by applying the attribute @who to the utterance <u> the value of which is the speaker's initials declared in the @xml:id. Likewise, where it is necessary to attribute some interpretive content to

myself (or another individual), this is done using the responsibility attribute @resp with the value being the @xml:id for the researcher[220].

6.3.4.1 Provenance of Corpus Files: <sourceDesc>

For each TEI record of a source originally annotated and converted from Praat (or potentially any other source), the filename of the ".wav" file and the tab-separated file exported from Praat are included in the <sourceDesc> section of the TEI header. The ".wav" files are encoded in the <media> element with the @mimeType="wav" and the file pathway declared in the @url attribute.

```
<sourceDesc>
  <p>This file was converted from the source file <ptr target="praat-
export:V_speak_01_02_03_04_05_TS.txt"/> which was extracted from the Praat TextGrid transcriptions of
the speech file <media mimeType="wav" url="soundfiles-gen:V_speak_01_02_03_04_05_TS.wav"/>
  </p>
</sourceDesc>
```

Figure 81: Example of <sourceDesc> stating the source files and their path

6.3.4.2 Pathways to Linked Files: <prefixDef>

Note that in the example above (Figure 81), both pointers use a mechanism for pointing to a file in their respective directories using a prefix ("praat-export:" and "soundfiles-gen:" respectively) defined using <prefixDef>[221]. Within the TEI data structure shown in Figure 82, a <prefixDef> is declared in the value of @ident within the header for each separate file to be referenced through a given corpus file. The prefix serves as a shortcut for a specific path within the project directory which negates the need to specify long directory locations each time a reference in placed in the dictionary. In Figure 82 the value of @matchPattern is a template for such pointers with the regular expression ([a-zA-Z0-9]+), which is replaced by the specific text of a file name. Additional uses of this mechanism will be discussed in the chapter on the Mixtepec-Mixtec TEI dictionary.

```
<listPrefixDef>
  <prefixDef ident="praat-export"
        matchPattern="([a-zA-Z0-9]+)"
        replacementPattern="../media/speech-mix/with-txtgrd/#$1"/>
```

[220] Note that at the time of submission, I am the only individual responsible for the transcription and annotation of the content, thus I do not explicitly apply @resp="#JB" to every interpretive annotation I perform, though in future stages in which other people become involved, it will likely be necessary to start such a system.
[221] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-prefixDef.html

```
    <prefixDef ident="soundfiles-gen"
            matchPattern="([a-zA-Z0-9]+)"
            replacementPattern="../media/speech-mix/with-txtgrd/#$1"/>
    <prefixDef ident="soundfiles-oax"
            matchPattern="([a-zA-Z0-9]+)"
            replacementPattern="../oaxaca/#$1"/>
    <prefixDef ident="stimuli"
            matchPattern="([a-zA-Z0-9]+)"
            replacementPattern="../media/stimuli/#$1"/>
  </listPrefixDef>
```

Figure 82: <prefixDef> list as declared in the header of corpus contents

A challenge to carrying out a proper documentation of metadata from recorded speech originally processed and annotated in Praat is that, as mentioned previously, there is no true capacity for inputting metadata in that program. Thus, in order to include the kind of key metadata required to adhere to best practices in language documentation discussed in section 4.3.2, this information must be added at the TEI stage of the data process.

6.3.4.3 Metadata for File Creation: <recordingStmt>

As discussed in section 4.4.1.3, the details of the recordings are stated in the <recordingStmt> within the header. Herein the following are stated: type of recording (audio, video) e.g. <recording type="audio">; the participants in the process (i.e. the linguistic assessors) in responsibility statement elements <respStmt>; the equipment used <equipment>; the location <location>; the date <date>. Figure 83 below shows such an example:

```
<recordingStmt>
  <recording type="audio">
    <respStmt>
      <resp>Recording</resp>
      <resp>Elicitation</resp>
      <name>Jack Bowers</name>
    </respStmt>
    <respStmt>
      <resp>Recording</resp>
      <resp>Elicitation</resp>
      <name>Andrea Guerra</name>
    </respStmt>
    <respStmt>
      <resp>Recording</resp>
      <resp>Elicitation</resp>
      <name>Larry "Kryn" Corpuz Jr.</name>
    </respStmt>
    <equipment>
      <ab>Audio recorded using a Sony PCM-D50 Linear PCM Recorder at a rate of 96kHz/24-bit.</ab>
```

```
            </equipment>
            <ab>
              <location>
                <placeName>San José State University</placeName>
                <placeName>San José</placeName>
                <region>California</region>
                <country>USA</country>
              </location>
            </ab>
            <date notBefore="2011-01" notAfter="2011-12">2011</date>
            <ab>Content was recorded using <term ana="#elicitation-translation">Translation-based
elicitation</term> using <lang>English</lang> and/or <lang>Spanish</lang>.</ab>
          </recording>
        </recordingStmt>
```

Figure 83: Example of full <recordingStmt> for recording made in 2011

6.3.4.4 Speech Event Typology: <taxonomy>

Additionally, a statement about the methodology which declares which class of speech event was captured according to the typology by Himmelmann (1998) which categorizes speech acts according to their "naturalness".

```
        <taxonomy>
          <desc>Typology of linguistic speech events captured in recordings as per: <bibl>Himmelmann
(<date>1998</date>)</bibl>. aka Typology of "naturalness".</desc>
          <category xml:id="observed">
            <catDesc>
              <term>Observed communicative event:</term> the extent of external interference is limited to the
knowledge of the speakers that the speech is being recorded or observed.</catDesc>
          </category>
          <category xml:id="staged">
            <catDesc>
              <term>Staged communicative event:</term> speech events realized for the purpose of recording (i.e.
elicited speech). Events are not really being realized for the purpose of communication but for the benefit of the
investigator.</catDesc>
            <category xml:id="staged-free-topical">
              <catDesc>
                <term>Staged-Topical</term>Prompt to speak freely about topic</catDesc>
            </category>
            <category xml:id="staged-stimuli">
              <catDesc>
                <term>Staged-Stimuli</term> events based on stimuli to be described in speakers own
words</catDesc>
            </category>
          </category>
          <category xml:id="elicitation">
            <catDesc>
              <term>Elicitation:</term> speech act for the sole purpose of linguistic investigation. (A new type of
speech event for most communities).</catDesc>
            <category xml:id="elicitation-contextualizing">
              <catDesc>
```

```
            <term>Contextualizing elicitation:</term> where native speakers are asked to provide contexts for an item or construction as prompted by the investigator.</catDesc>
          </category>
          <category xml:id="elicitation-translation">
            <catDesc>
              <term>Translation-based elicitation:</term> native speaker asked to translate item from second language</catDesc>
          </category>
          <category xml:id="elicitation-judgement">
            <catDesc>
              <term>Judgement:</term> where native speakers are asked to judge the acceptability of a given construction based on any aspect of language, e.g. grammar, etc.</catDesc>
          </category>
        </category>
      </taxonomy>
```

Figure 84: <taxonomy> in TEI header for elicitation methods used in recording

The statement specifying which category in Himmelmann's "naturalness" typology is made at the end of the <recording> element in the <term> element using both the @ana attribute and in text for human consumption. In such cases, a pointer to the stimuli is included in the statement, as shown in the following example:

```
        <recording type="audio">
          …..
          <ab>Content was recorded using <term ana="#staged-stimuli">Staged stimuli</term> using the following file: <ptr target="stimuli:frog_in_basket.jpg"/> .</ab>
        </recording>
```

Figure 85: Declaration of elicitation type with link to stimuli

In Figure 85, the @target attribute, again using the prefix "stimuli:" declared in the <prefixDef> described above, <ptr> points to the following image (Figure 86) created for the specific purpose, to prompt phrases on spatial relations with the question *¿Nchii inkaa sa'va?* 'Where is the frog?':

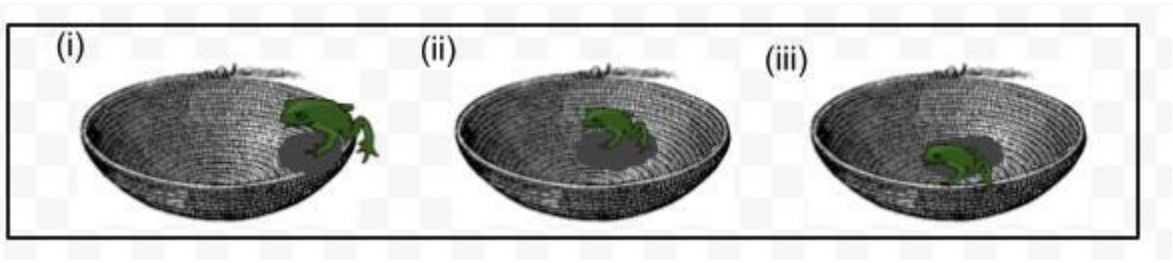


Figure 86: Stimuli used to elicit vocabulary on spatial relations

Down the line, as more diverse spoken language resources are integrated into the corpus that additional details of aspects of speech-acts recorded will need to be stated. Thus, it is likely that the array of speech acts and their sub-features will have to be expanded beyond the scope of Himmelmann's *naturalness* taxonomy as they are not entirely sufficient for all possible scenarios of data creation in LD.

6.4 Annotation Mechanisms

As a principle for best practice in language documentation is that description be kept separate from the documented resource (Himmelmann, 1998, 2006a), the content has been annotated using a multi-tiered standoff annotation. This decision ensures that the resource can be reused, reinterpreted and/or appended by people involved in this work or others without having to deconstruct major portions of the original content.

There are two methods of embedded standoff annotation used in the project: <spanGrp> is for creating new annotations, and <linkGrp> is specifically for where pre-existing translations, glosses or potentially other parallel content already exist in the source. These TEI mechanisms are used for all content (both text-based and annotated spoken sources) described below along with the specifics of application in the various sources of MIX resources.

6.4.1 Feature Structures and Annotation Inventory

The inventory for lexical features used both in the annotation of the corpus (spoken and text) as well as the dictionary are declared using TEI feature structures which are compliant with ISO 24610-1:2006 (as described in section 4.4.3.2.3) and can be used to either declare an inventory, directly annotate lexical or conceptual features for linguistic analysis, or as an abstract means of grouping and relating structured information. While feature structures can be structured using several different mechanisms including: <fs>, <fsdDecl>, <fvLib>; this project uses the former.

With the restructuring of ISOcat, these features at present are not linked to any controlled vocabulary but in the near future, this will likely change. The primary inventory of grammatical features used to annotate the corpus data is shown below:

```
<fs>
  <f name="pos">
    <vAlt>
      <symbol value="noun" xml:id="N"/>
      <symbol value="properNoun" xml:id="N-PROP"/>
      <symbol value="verb" xml:id="V"/>
      <symbol value="pronoun" xml:id="PRON"/>
      <symbol value="emphaticPronoun" xml:id="PRON-EMPH"/>
      <symbol value="demonstrative" xml:id="DEM"/>
      <symbol value="determiner" xml:id="DET"/>
      <symbol value="adposition" xml:id="ADPOS"/>
      <symbol value="interjection" xml:id="INTERJ"/>
      <symbol value="quantifier" xml:id="QNTF"/>
      <symbol value="particle" xml:id="PTCL"/>

      <symbol value="nominalizingParticle" xml:id="NMLZ"/>
      <symbol value="prefix" xml:id="PREF"/>
      <symbol value="adverb" xml:id="ADV"/>
      <symbol value="adjective" xml:id="ADJ"/>
      <symbol value="conjunction" xml:id="CONJ"/>
      <symbol value="coordinatingConjunction" xml:id="CONJ-COORD"/>
      <symbol value="subordinatingConjunction" xml:id="CONJ-SUB"/>
      <symbol value="indefiniteArticle" xml:id="ART-INDEF"/>
      <symbol value="number" xml:id="NUM"/>
    </vAlt>
  </f>
</fs>
```

Figure 87: Feature structure inventory for Mixtepec-Mixtec *part-of speech*

Currently, for each of these POS features which have sub-categories, have separate <fs> elements, e.g for the feature grammatical *number*:

```
<fs>
  <f name="number">
    <vAlt>
      <symbol value="singular" xml:id="SG"/>
      <symbol value="plural" xml:id="PL"/>
    </vAlt>
  </f>
</fs>
```

Figure 88: Feature structure inventory for Mixtepec-Mixtec *number*

The values of the features are added to the custom ODD schema for all project documents so that when annotating features in Oxygen XML Editor, the possible values from the feature structures appear as suggestions when the @ana is created or scrolled over.

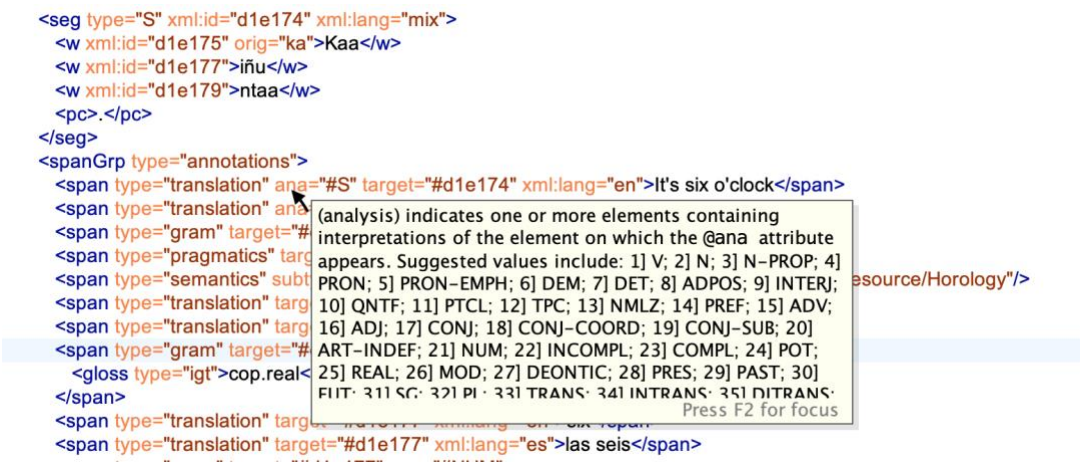


Figure 89: Pop-up value suggestions for annotation in Oxygen XML Editor from TEI ODD Schema

6.4.2 Standoff Annotation: <spanGrp>

The primary method of annotating all standoff content in TEI is the <spanGrp> element which takes any number of child elements. Annotations are placed in <spanGrp type="annotations">. In this system, one <spanGrp> is used for all levels of annotation which can occur concurrently, specifically: *translation*, *grammar* and *interlinear glossed texts* (occurring together), *semantics*[222] and where necessary, *note*[223] (used for editorial notes are usually of temporary nature and may or may not be transferred to the dictionary entry for the given content until the issue is resolved). Depending on the annotation type, the content may be specified in the text value of or via the value of @ana or @corresp.

Figure 90: Abstract model of primary features used in TEI element <spanGrp>

[222] Depending on the annotation features used in a project, *semantics* can be separated into multiple annotation categories.
[223] The categories of annotation may be extended in the future as needed, "pragmatics" in another likely candidate

Each of the major annotation levels is specified using the attribute @type (e.g.). There is a for each item (or potentially sequence of items) for which there is an annotation. The annotation is linked to the content to which it corresponds using the attribute @target which points to the item's unique identifier (i.e. the value of @xml:id). In the typology from Bański (2010) discussed in section 4.4.2.4, this would qualify as *correspondence* standoff annotation. The annotation attribute @ana contains the hashtagged referenced values of feature structures which are declared in a separate project documents which can be basic information about the language content being annotated. Where the span is for annotating translations, and the annotated item is not a simple <w> token, the @ana attribute specifies the type of content being annotated, specifically with "S" for sentence, "PHRS" for phrase, "INFL" for inflected forms, or "CMPND" compound.

The correspondence attribute @corresp can be used to link the annotation and the annotated content to some outside resources such as a url, uri, etc. In any category annotated in this system, certainty regarding some aspect of the interpretation can be expressed using the certainty attribute @cert on the given which can have the values "high", "medium", "low" or "unknown".

6.4.3 Linking Parallel Content: <linkGrp>

Where there are multilingual translations already present in a document, the TEI element <linkGrp> is used, which take any number of <link> child elements. In the scheme of Bański (2010), this use of <linkGrp> is described as *multiple-point linking.* Just as with <spanGrp>, in <linkGrp> annotations, the @type is given the value of "annotations" on the parent element and <link> is given the value of the feature (e.g. "translation"). Likewise, as with the 's in dealing with translations, the @ana is used to specify the type of lexical content annotated by the <link> with the possibilities being: "S" for sentence, "PHRS" for phrase, "INFL" for inflected forms, or "CMPND" compound.

Figure 91: Abstract model of TEI element <linkGrp>

At this point in the project the main uses of <linkGrp> based on the source content are for:

- Basic vocabulary from SIL sources with a MIX item and a Spanish (and rarely English) glosses;
- Spoken language content in which the Mixtec vocabulary is spoken along with a Spanish gloss (also from SIL) specifically the language learning resource *Na kutuꞌva ko saꞌan savi* (Bautista Martínez and Hernández Velasco., 2019)[224];

6.4.4 Translations

Translations in English and Spanish are made for both individual lexical items and whichever higher phrasal or sentence contexts they appear in which, the translations are given in the element value and linked using @target pointing to the @xml:id(s) of the Mixtec form(s) and the language of translation whose ISO 639 language tag is given in @xml:lang. Translations can be typed using the @ana attribute, though when there is a 1 to 1 relation between the and the <w>, the default is to not have an @ana, e.g. where a <w> is a full lexical item, no @ana is necessary. For full sentences, the <seg type="S"> is annotated by a single segment which is labeled as . Other translations for which the relationship between the and the annotated content <w>'s is 1 to n, can occur for: compounds, phrases[225], and inflected forms[226] which take the @ana values of "#CMPND",

[224] This file is available in the GitHub repository under the filename Aprendamos_el_idioma_mixteco_(Mixtepec).xml and at the time of submission, the TEI version of this is still in the process of encoding.
[225] While no distinction is made in the corpus annotations between phrases and multi-word expressions, they are used as a category of lexical entry in the dictionary.
[226] In classifying inflected forms which are multi-word expressions in the corpus, they are labeled as "phrase". Inflected verbs (e.g. verb phrases) are however labeled as "inflected", the purpose of which is to clearly label members of verb paradigms to which verbal multi-word expressions can also belong.

"#PHRS", "#INFL" respectively. All of these typologies are used in the automatic extraction of the contents for analysis as well as for transfer to the TEI dictionary. If there is a need to annotate a translation literally, the attribute @subtype is used with the value of "literal"[227]. The example below shows a sample sentence with several of these types annotated.

```
<seg xml:id="d1e631" n="16" xml:lang="mix" resp="#TS" type="S">
  <w xml:id="d1e632">Cha</w>
  <w xml:id="d1e634" orig="tzi'i">tsi'i</w>
  <w xml:id="d1e637">yu</w>
  <w xml:id="d1e639">soko</w>
  <w xml:id="d1e641">ra</w>
  <w xml:id="d1e643">tsa'an</w>
  <w xml:id="d1e646">yu</w>
  <w xml:id="d1e648" orig="takuan">ntakuaan</w>
  <w xml:id="d1e652">ña</w>
  <w xml:id="d1e655">katsi</w><pc>.</pc>
</seg>
<spanGrp type="annotations">
  <span type="translation" ana="#S" target="#d1e631" xml:lang="en">
    And then I was hungry so I went to buy something to eat.</span>
  <span type="translation" ana="#S" target="#d1e631" xml:lang="es">
    Y tuve hambre entonces fuí para comprar algo a comer.</span>
  <span type="translation" target="#d1e632" xml:lang="en">and then</span>
  <span type="translation" target="#d1e632" xml:lang="es">y</span>
  <span type="translation" ana="#PHRS" target="#d1e634 #d1e637 #d1e639" xml:lang="en">
   I was hungry</span>
  <span type="translation" ana="#PHRS" target="#d1e634 #d1e637 #d1e639" xml:lang="es">
   yo tenía habre</span>
  <span type="translation" ana="#INFL" target="#d1e643 #d1e646" xml:lang="en">
   I went</span>
  <span type="translation" ana="#INFL" target="#d1e643 #d1e646" xml:lang="es">fuí</span>
  <span type="translation" ana="#INFL" target="#d1e648" xml:lang="en">I bought</span>
  <span type="translation" ana="#INFL" target="#d1e648" xml:lang="es">compré</span>
  <span type="translation" ana="#PHRS" target="#d1e652 #d1e655" xml:lang="en">
    something to eat</span>
  <span type="translation" ana="#PHRS" target="#d1e652 #d1e655" xml:lang="es">
   algo a comer</span>
</spanGrp>
```

Figure 92: Sample translations of sentence from speaker authored content

The image below is a typical example of an instance necessitating the use of <linkGrp> from one of the SIL children's booklet sources (Beckmann, 2014) in which there is the MIX item corresponding to the animal pictured, along with two <link>'s, one for each unique Spanish translation equivalent. The extra tag "MEX" (the ISO 3166 country code for Mexico) is

[227] At present @subtype is not allowed in and this is done with an ODD schema customization.

appended on the language tag according to BCP 47 to distinguish where terms are specific to regional varieties.


```
<seg xml:id="d1e53" xml:lang="mix" type="term">
  <w xml:id="d1e54">chumi</w>
  <w xml:id="d1e56">xini</w>
  <w xml:id="d1e58" orig="ka'nu">ka'nu</w>
</seg>
<seg xml:id="d1e60" xml:lang="es-MEX" type="term">
  <w xml:id="d1e61">tecolote</w>
</seg>
<seg xml:id="d1e63" xml:lang="es" type="term">
  <w xml:id="d1e64">búho</w>
  <w xml:id="d1e66">cornado</w>
</seg>
<linkGrp type="annotations">
  <link type="translation" target="#d1e53 #d1e60"/>
  <link type="translation" target="#d1e53 #d1e63"/>
</linkGrp>
```

Figure 93: Example of encoding of <linkGrp> in content with existing bilingual translations

While at present, this project doesn't have a large amount of data with pre-existing parallel translations, there exist a vast amount of language corpora and historical resources that are parallel and for which annotation with <linkGrp> would be applicable for other languages. The XSLT framework developed herein could be re-used, or modified to extract translations (or other features) from such content.

6.4.5 Grammar, Information Structure and Interlinear Glossed Text

Grammar is annotated using @type="gram", and the specific feature(s) corresponding to the given @xml:id value(s) is/are placed in the value of @ana. The features

tagged in @ana are defined in the separate feature structure document (described in section 6.4.1). Additionally, the contents are annotated for interlinear glossed text (IGT), specifically according to the Leipzig Glossing Rules (Bickel et al. 2008) are combined with grammar spans as <gloss type="igt"> as a child node of . This is a convenient structure as the grammar and IGT they point to the same content, and thus it makes for a time saving and convenient system to combine the two. Where the source of the utterance is a transcribed speech file (via Praat), the value of the <gloss> can be carried over from the TextGrid annotation tier by the same name (see section 6.3). IGT is tagged separately from the other annotation as even though in some cases it may overlap with the grammatical or semantic tags in the corpus, it is not designed for consistent machine readability as in many cases, it may be desirable to display only certain information in the IGT.

The example below shows the translation, grammar and IGT annotations for sentence *kaa iñu ntaa* "it's exactly 6 o'clock" from SIL booklet L093 (Nieves and Beckmann, 2007b)[228] (discussed in example 16 in section 2.1.4). Note also that the sentence level is given the tag "DECL" (*declarative*), as well as "RESP" (*response*) as this sentence is response to a question of "what time is it".

```
<seg type="S" xml:id="d1e174" xml:lang="mix">
  <w xml:id="d1e175" orig="ka">Kaa</w>
  <w xml:id="d1e177">iñu</w>
  <w xml:id="d1e179">ntaa</w>
  <pc>.</pc>
</seg>
<spanGrp type="annotations">
  <span type="translation" ana="#S" target="#d1e174" xml:lang="en">It's six o'clock</span>
  <span type="translation" ana="#S" target="#d1e174" xml:lang="es">Son las seis</span>
  <span type="gram" target="#d1e174" ana="#DECL #RESP"/>
  <span type="translation" target="#d1e175" xml:lang="en" ana="#INFL">it is</span>
  <span type="translation" target="#d1e175" xml:lang="es" ana="#INFL">es</span>
  <span type="gram" target="#d1e175" ana="#V #INTRANS #IPFV">
        <gloss type="igt">cop.real</gloss></span>
  <span type="translation" target="#d1e177" xml:lang="en">six</span>
  <span type="translation" target="#d1e177" xml:lang="es">las seis</span>
  <span type="gram" target="#d1e177" ana="#NUM">
       <gloss type="igt">six</gloss></span>
  <span type="translation" target="#d1e179" xml:lang="en">exactly</span>
  <span type="translation" target="#d1e179" xml:lang="en">en punto</span>
```

[228] https://github.com/iljackb/Mixtepec_Mixtec/tree/master/SIL_docs/L093

```
  <span type="gram" target="#d1e179" ana="#ADV">
      <gloss type="igt">exactly</gloss></span>
</spanGrp>
```

Figure 94: Example of grammatical annotations of SIL document

Note that the annotation of IGT in the standoff mechanism is not yet in a final user-oriented format and that in order to make it presentable to users in the likeness of how it is done in FLEx or ELAN, a conversion will likely be made to extract, transform and present the IGT data in an output format. As mentioned in section 4.4.2.4 there is not yet a well-established manner of encoding IGT in TEI so in the process of developing this user-friendly output, significant attention will need to be dedicated to this issue as well as the development of conversions to and from the aforementioned dominant LD tools ELAN and FLEx.

In annotating content which already has interlinear glossed texts such as the document Bichos-SIL.xml (Beal, 2018), first the sentence level element containing the IGT is tagged is encoded as <seg type="igt"> with the appropriate @xml:lang value (which is necessary as the pre-existing sources of IGT content from SIL are in Spanish or English) and each component of the gloss is encoded as a <gloss> with @xml:id and @type="igt" containing the original glosses. Finally, the <link> element also contains @type="igt".

```
<seg xml:id="d1e1320" type="S" xml:lang="mix" n="1">
   <w xml:id="d1e1321">Yee</w>
   <w xml:id="d1e1323">in</w>
   <w xml:id="d1e1325">tintoo</w>
   <w xml:id="d1e1327">kiti</w>
   <w xml:id="d1e1329">nani</w>
   <pc>,</pc>
   <w xml:id="d1e1332">tintoo</w>
   <w xml:id="d1e1334">savi</w>
   <pc>.</pc>
 </seg>
 <seg xml:id="d1e1337" type="S" notation="igt" xml:lang="es">
   <gloss xml:id="d1e1338" type="igt">hay</gloss>
   <gloss xml:id="d1e1340" type="igt">una</gloss>
   <gloss xml:id="d1e1342" type="igt">araña</gloss>
   <gloss xml:id="d1e1344a" type="igt">animal/insecto</gloss>
   <gloss xml:id="d1e1349" type="igt">se.llama</gloss>
   <gloss xml:id="d1e1354" type="igt">araña</gloss>
   <gloss xml:id="d1e1356" type="igt">lluvia</gloss>
 </seg>
 <linkGrp type="annotations">
   <link target="#d1e1337 #d1e1320" type="igt”/> <!-- sentence -->
   <link target="#d1e1338 #d1e1321" type="igt”/> <!-- hay - yee -->
```

```
    <link target="#d1e1340 #d1e1323" type="igt”/>  <!-- una - in -->
    <link target="#d1e1342 #d1e1325" type="igt”/> <!-- araña - tintoo -->
    <link target="#d1e1344a #d1e1327" type="igt”/> <!-- insecto/animal - kiti -->
    <link target="#d1e1349 #d1e1329" type="igt”/>  <!-- se llama - nani -->
    <link target="#d1e1354 #d1e1332" type="igt”/> <!-- araña -tintoo -->
    <link target="#d1e1356 #d1e1334" type="igt”/>   <!-- lluvia - savi -->
  </linkGrp>
```

Figure 95: Example of pre-existing IGT in SIL document as marked up in TEI using <linkGrp>

6.4.6 Annotating Tone and Morphological Features

Thus far in the examples from the annotated corpus (i.e. Figures 92-95), there have been none with marked morphological information on the lexical content. Thus, in these cases the orthographic and phonetic transcriptions of the same content (as comes from the transcribed speech sources converted from Praat) contain the same number of segmentations in the XML markup and thus would be totally parallel contents. This is not always the case, as discussed in section 2, Mixtec morphological inflections which can comprise simply of tones which are often active therein, are often not expressed in the orthography. Additionally, in the encoding, the orthographic forms are by design not segmented beyond the <w> level in order to avoid complications with searching and retrieving the language content. In the annotations in which there is only orthographic content, the grammar annotates all information on the orthographic forms without further segmentation[229], e.g.

```
<seg xml:id="L147-01-01" type="S" xml:lang="mix">
  ….
  <w xml:id="d1e170">nikachi</w>
  <w xml:id="d1e172">sto'i</w>
  <pc>:</pc>
</seg>
<spanGrp type="annotations">
  ….
  <span ana="#INFL" target="#d1e170" xml:lang="en" type="translation">said</span>
  <span type="gram" target="#d1e170" ana="#V #TRANS #PFV">
    <gloss type="igt">pfv-say</gloss>
  </span>
  <span ana="#INFL" target="#d1e172" xml:lang="en" type="translation">it's owner</span>
  <span type="gram" target="#d1e172" ana="#NP #POSS #3PERS #SG">
    <gloss type="igt">owner[3sg]</gloss>
```

[229] While in the corpus, the orthographic forms are not segmented beyond the token level (i.e. where there is whitespace in the source according to the orthographic practice), the IGT does segment as normal with the expectation that in the case that the orthographic sentence is extracted for presentation/analysis, the it can easily be further segmented by the user manually as needed.

```
</span>
</spanGrp>
```

Figure 96: Example of orthographic sentence grammatically annotated without further segmentation

However, if the contents shown above were fully annotated in interlinear glossed form, they would be further segmented as follows:

ni-kachi sto'i
PFV- say owner[3SG]
'..said its owner'

Figure 97: IGT representation of previous example

Despite the reality that the linguistic content is in fact more granular than is represented in the orthographic form shown in the example above, in order to not interrupt the orthographic text, which may have negative implications for searching[230], there is no more detailed annotation applied beyond what is shown. If needed at a later point, full segmentation can be added, or possibly a duplicate of the data can be created with such segmentation.

However, where the content is transcribed from spoken language, and both phonetic and orthographic forms are present, the inflection information is further segmented in the IPA transcription with <m> (morpheme) so that these features can also be annotated grammatically. In the case of a tonal and other morphological inflections, <m> can occur on either instances of where the tone denotes some specific grammatical feature or where there are prefixes which are (mostly) not delimited in the orthography adopted in this project. The examples below show some of the particular features, in the orthographic and phonetic forms, as well as the IGT. Note the portions in grey on the IGT tier correspond to other morphological/tone features that are in fact marked with <m> themselves in the actual data, but are not in the given examples for the purpose of emphasizing another segment. To the right of each, the way that each of the

[230] While the segmentation can be searched using certain XPath expression and potentially searched if corpus is indexed in XML database, this formatting would likely be problematic for non-XML experts using the corpus in its raw current state.

orthographic and/or phonetic forms are segmented in the transcriptions in TEI is demonstrated in order to illustrate the gap in expressiveness between the former and later in the corpus[231]:

First person singular (verbal inflection or nominal possessive)[232]

with tone:

Orthography:	sketa	<w>sketa</w>
Segmented IPA:	skɛˀt̪a↘	<w>skɛˀt̪a<**m**>↘</**m**></w>
IGT:	run**1SG**	

Second person singular informal (verbal inflection or possessive)

with vowel + tone morpheme:

Orthography:	ka'**un**	<w>ka'un</w>
Segmented IPA:	kãˀʔ**ũ**↗	<w>kãˀʔ<**m**>**ũ**↗</**m**></w>
IGT:	PFV\speak**[2SG.INF]**	

Imperfective

with tone:

Orthography[233]:	sk**e**ta	<w>sketa</w>
Segmented IPA:	sk**ɛ**˦t̪a˩	<w>skɛ<**m**>˦</**m**>t̪a↘</w>
IGT:	**IPFV**\run\1SG	

Potential

with prefix:

Orthography:	**kun**kua'a	<w>kunkua'a</w>
Segmented IPA:	**kun-**kua'a	<w><**m**>**kũ**:ˀ</**m**>kwaˀa˩</w>
IGT:	**POT-**give\1SG	

[231] Note that the @xml:id's which are the targets for the standoff annotations are not shown for readability.
[232] Note that the tonal contour of first person singular is generally phonologically considered *low*, but is also realized as *falling.* In these cases, they should be considered phonetic variants or allomorphs. In the transcription of this feature in the audio files, I have generally annotated these as I observe the F0.
[233] While as displayed in the glossed examples throughout section 2, in the updated practice the marking of the onset high tone on the first vowel of imperfective verbs is included (e.g. ***ské**ta*), however this was a recent development and in the majority of the corpus this is not marked, thus in order to demonstrate this divide, it isn't included in these examples.

Perfective

with full prefix:

Orthography:	**ni**kachi	<w>nikachi</w>
Segmented IPA:	**ni**-kachi	<w><**m**>**ni**ɻ</**m**>kaɻtʃi↘</w>
IGT:	**PFV**-say\1SG	

with tone change:

Orthography:	sk**è**ta	<w>sketa</w>
Segmented IPA:	sk**ɛ**ʌ̀t̪a↘	<w>skɛ<**m**>ʌ̀</**m**>t̪a↘</w>
IGT:	**PFV**\run\1SG	

with partial (pre-nasal) prefix and tone change:

Orthography:	**n**ts**à**tsi	<w>ntsàtsi</w>
Segmented IPA:	**n**-ts**à**tsi	<w><**m**>**n**</**m**>tsa<**m**>ɻ</**m**>tsiɻ</w>
IGT:	**PFV**-eat\1SG	

Negative:

with tone change[234]

Orthography:	ku**à**’a	<w>kuà’a</w>
Segmented IPA:	kwaɻʔaˀ↘	<w>kwa<**m**>ɻ</**m**>ʔaɻ</w>
IGT:	**NEG**\give\1SG	

In the annotations the is given a @subtype, which can be "tone" in the case of the feature being realized by tone, or "morph" where morphological features such as prefixes or suffixes are present in the phonetic transcription and segmentation. The following example for the sentence *sketa ntikii* ‘I run every day’ demonstrates the way tonal features described above are annotated in an utterance transcribed from speech with both phonetic and orthographic transcriptions. Herein, there is phono-semantically relevant information on the

[234] At time of submission, this phenomena has only been observed with the verb *kua’a* ‘to give’.

tones which is not marked in the orthography[235] on the verb *skɛˈɬa*↘ 'I run' (shown above); the first (high) tone denotes *imperfective aspect* and final (global falling or low) tone denotes first person singular. Note that since the verb includes the argument as well, it is grammatically tagged as a *verb phrase VP*.

```
<u who="#TS" xml:id="d1e112" n="2" start="1.48" end="2.98" xml:lang="mix">
  <seg xml:lang="mix" xml:id="d1e113" notation="orth" type="S">
    <w xml:id="d1e114" synch="#T14">sketa</w>
    <w xml:id="d1e116" synch="#T19">ntikii</w>
  </seg>
  <seg xml:lang="mix" xml:id="d1e118" notation="ipa" type="S" sameAs="#d1e113">
    <w xml:id="d1e119" synch="#T14" sameAs="#d1e114">
      skɛ<m xml:id="d1e225">ɬ</m>ɻa<m xml:id="d1e120">↘</m>
    </w>
    <w xml:id="d1e132" synch="#T19" sameAs="#d1e116">nd̪i↘ki:↘ ↗</w>
  </seg>
</u>
<spanGrp type="annotations">
 ...
  <span type="translation" target="#d1e114" xml:lang="en" ana="#INFL">I run</span>
  <span type="gram" target="#d1e114" ana="#VP #INTRANS #IPFV #1PERS #SG">
        <gloss type="igt">ipfv\run\1s</gloss></span>
  <span type="gram" subtype="tone" target="#d1e125" ana="#IPFV"/>
  <span type="gram" subtype="tone" target="#d1e120" ana="#1PERS #SG"/>
  <span type="translation" target="#d1e116" xml:lang="en">every day</span>
  <span type="gram" target="#d1e116" ana="#ADV">
          <gloss type="igt">every.day</gloss></span>
</spanGrp>
```

Figure 98: Example of grammatical annotations of transcribed speech exported from Praat

Another, unrelated feature which deserves discussion is the fact that in some cases (most notably with the topic marker *ka*), there is grammatical content that has no translation due to the fact that they carry out purely grammatical/discourse related functions. For these the only annotation made is within the grammar annotations, e.g.

```
<seg xml:id="d1e140" n="3" xml:lang="mix" resp="#TS" type="S">
  <w xml:id="d1e141" orig="Ni kitsi">Nikitsi</w>
  <w xml:id="d1e147">Shanty</w>
  <w xml:id="d1e149">ka</w>
 ….
</seg>
<spanGrp type="annotations">
 ....
```

[235] Although in the updated orthography the high tone is in fact marked on the first vowel of imperfective verbs, it isn't marked in all cases, and in the interest of avoiding inserting elements that would interrupt the search for orthographic forms, this feature is only annotated on the phonetic form.

```
  <span target="#d1e141" xml:lang="en" type="translation">came</span>
  <span type="gram" target="#d1e141" ana="#V #INTRANS #PFV #3PERS #SG #INF">
    <gloss type="igt">pfv-arrive[3sg.inf]</gloss></span>
  <span target="#d1e147" xml:lang="en" type="translation">Shanty</span>
  <span type="gram" target="#d1e147" ana="#N-PROP">
    <gloss type="igt">Shanty</gloss></span>
  <span type="gram" target="#d1e149" ana="#PTCL">
    <gloss type="igt">=ptcl</gloss></span>
 ….
</spanGrp>
```

Figure 99: Annotation showing IGT and grammatical tagging of particle ka

It should be noted that at the time of submission the grammatical annotation system is still being implemented, and thus if searched, there are still numerous files whose annotation contents are not complete or possibly reflect earlier methods. Given that the priority is to first make basic translations of the content in order to learn and document the language, as well as to be able to further gloss and translate new content independently, the task of annotating grammar has thus far been a secondary priority. Also, of note is that, at present the decision has been to not use a pre-existing standardized tag set such as ISOcat or GOLD, as none of them, the reason being that none have all of the necessary features; thus, no matter which set is adopted, there will need to be ad-hoc additions in order to accommodate the specifics of the language and the theoretical approaches.

6.4.7 Annotating Semantics

Along the same lines as the other annotations described thus far, the basic unit of annotating semantic information is , in which the @target points to the segment(s) annotated. However, depending on the nature of the content annotated and the specific features, there are several different aspects to the system.

First, the basic annotation of sense is labeled in the value of subtype, e.g. . The annotation of domain also follows this exact pattern with the exception of the value of @subtype, e.g. . For each of these features, where available, an external uri to existing knowledge bases can be specified in the value of @corresp, these can be used to point to such resources as Wikidata[236], DBpedia (Auer et al., 2007), or geonames in the case of geographic content. The example below

[236] https://www.wikidata.org/wiki/Wikidata:Main_Page

shows the semantic annotations applied to the SIL document ‘Las Aves’ by Gisela Beckmann (2014), which contains a list of 109 bird varieties in MIX[237]. Note also that in this document, the Latin scientific species names were added by me to the translations and any inaccuracies in attribution are the fault of myself and not the original author.


```
<seg xml:id="d1e53" xml:lang="mix" type="term">
  <w xml:id="d1e54">chumi</w>
  <w xml:id="d1e56">xini</w>
  <w xml:id="d1e58" orig="ka'nu">ka'nu</w>
</seg>
<seg xml:id="d1e60" xml:lang="es" type="term">
  <w xml:id="d1e61">tecolote</w>
</seg>
<seg xml:id="d1e63" xml:lang="es" type="term">
  <w xml:id="d1e64">búho</w>
  <w xml:id="d1e66">cornado</w>
</seg>
….
<spanGrp type="annotations">
  <span xml:lang="en" target="#d1e53" type="translation">Great Horned Owl</span>
  <span xml:lang="la" target="#d1e53" type="translation">Bubo virginianus</span>
  <span type="semantics" subtype="sense" target="#d1e53"
corresp="https://www.wikidata.org/wiki/Q81515"/>
  <span type="semantics" subtype="sense" target="#d1e53"
corresp="http://dbpedia.org/resource/Great_horned_owl"/>
  <span type="semantics" subtype="domain" target="#d1e53"
corresp="http://dbpedia.org/resource/Bird"/>
  <span type="semantics" subtype="domain" target="#d1e53"
corresp="http://dbpedia.org/resource/Animal"/>
</spanGrp>
```

[237] Note that while in other corpus documents the annotation of compounds and other multi-unit terms is done by pointing to the component parts (i.e. the <w> elements) and not their <seg> wrapper, given that this document is just a list of different bird species, this document was done differently as it is a simple source of vocabulary items rather than as a corpus document with more complex uses of language. The contents of this document were just simply annotated and then extracted and entered into the TEI dictionary using XSLT.

Figure 100: Annotation of semantics of bird species from SIL document

There are several desired benefits in carrying out this additional annotation, one of which is that by utilizing the knowledge base links such as DBpedia, or Wikidata further resources for the given entry concept (in this case the specified Great Horned Owl) can be gathered and utilized in the Mixtec dictionary or other potential pedagogical resources (see Lehmann et al., (2014) for examples of knowledge extraction from DBpedia). For instance, taking the Wikidata source, the linked resource contains: numerous translation equivalents, (open-source) images of the concept as well as links to informative scientific web resources. Other potential uses is to use the definitions of the concept present in other languages as a template for additional Mixtec content either within the dictionary itself or potentially in a future resource such as a Mixtec-language encyclopedic knowledge base.

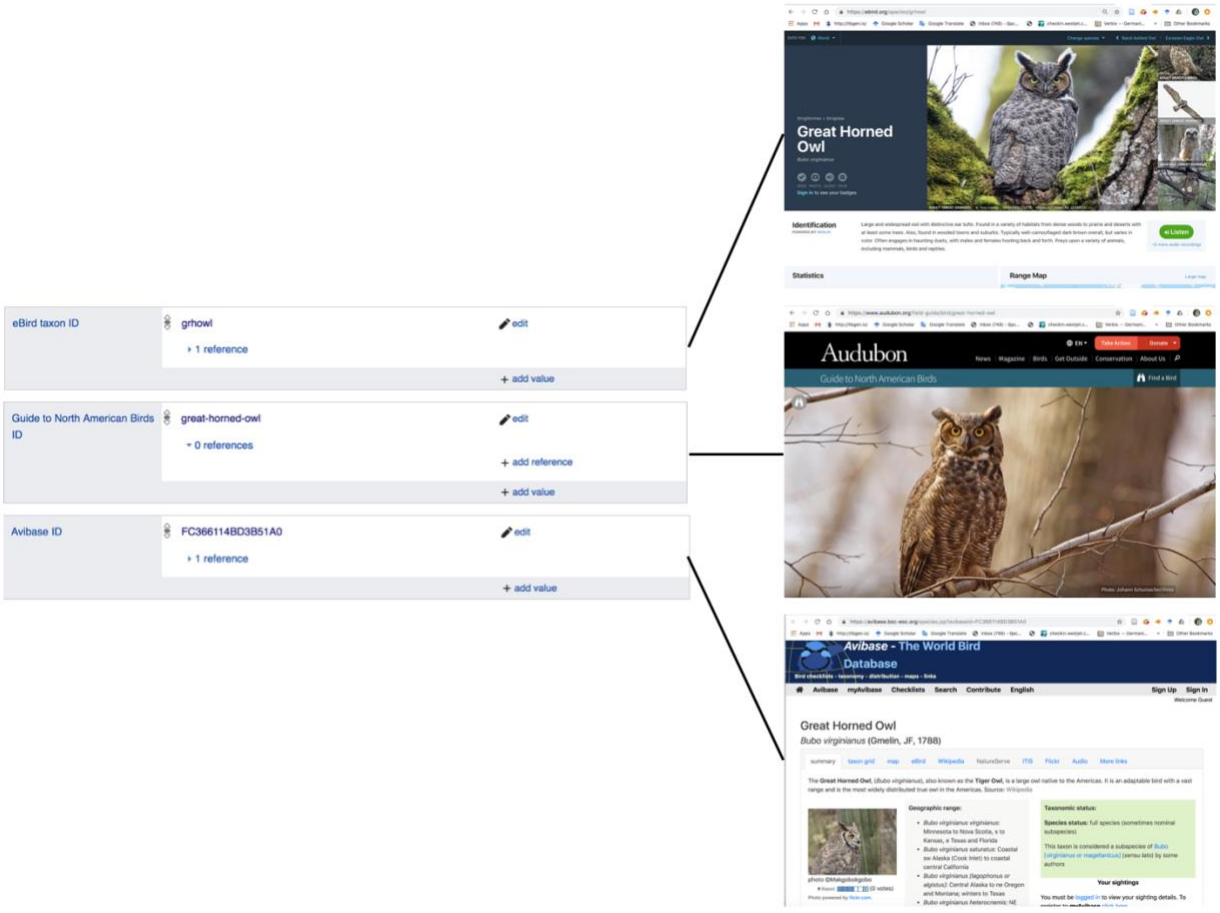


Figure 101: Links to encyclopedic sources of information via the Wikidata uri annotation

Along similar lines, semantic concept annotations may also be included for other relevant contents mentioned in the sources such as geographical locations. In the example below from a trip journal written by project collaborator Tisu'ma Salazar while in Vienna[238], the mention of the Shönbrunn palace in the Mixtec text is linked with a link to the geonames entry for the entity, also using the @corresp.

```
<seg xml:id="d1e12977" xml:lang="mix" resp="#TS" type="S">
  <w xml:id="d1e12978">Michu'ni</w><pc>,</pc>
  <w xml:id="d1e12984">mee</w><pc>,</pc>
  <w xml:id="d1e12988" orig="ku nku'un">kunku'un</w>
  <w xml:id="d1e12992">tienda</w>
  <w xml:id="d1e12994">sara</w>
  <w xml:id="d1e12996">kunku'un</w>
  <w xml:id="d1e13001" orig="yuu">yu</w>
  <w xml:id="d1e13003">kunchee</w>
  <w xml:id="d1e13005">in</w>
  <pc>“</pc> <w xml:id="d1e13008" orig="Palacio">Palasio</w><pc>”</pc>
  <w xml:id="d1e13011">ña</w>
  <w xml:id="d1e13013">nani</w>
  <w xml:id="d1e13015"" orig="Schonbrunn">Xonbrun</w><pc>.</pc></seg>
<spanGrp type="annotations">
  <span target="#d1e12977" xml:lang="en" type="translation" ana="#S">Now I'm going to go to a store
and then I'm going to go to a palace called Schönbrunn.</span>
   ….
  <span target="#d1e13015" xml:lang="en" type="translation">Schönbrunn</span>
  <span type="semantics"    corresp="https://www.geonames.org/6354998/schloss-schoenbrunn.html"
target="#d1e13015" ana="#LOC"/>
  ...
</spanGrp>
```

Figure 102: Example of semantic annotation of geographic information linking to GeoNames

4.6.7.1 Enhancing Grammatical Categories with Semantics

In providing a truly accurate description of any Mixtecan language which achieves any sort of linguistic insight, it is impossible to do so without integrating a wide array of semantic features, that are relevant both on the synchronic and diachronic levels. Thus, in designing and implementing an annotation system for this corpus, the separate spans used to annotate the grammar and semantic contents allows for the two to be included in compliment to one another without conflating the grammatical categories in order to pack in semantic features.

[238] https://github.com/iljackb/Mixtepec_Mixtec/blob/master/misc-sources/Tisu-Vienna-Diary-201711.xml

Rather than complicate the part of speech annotation inventories, by creating categories such as *verb* with possible *motion/arrival/stative* etc., even though it is quite common in linguistic literature to discuss *motion verb, stative verb* etc., which would have to occur in addition to other subtypes such as *transitive, intransitive*, etc., *or adverb* with possible subtypes of *temporal/manner/degree*, etc., this annotation method allows for tags from semantic annotations to combine with, and complement those in the grammatical annotations[239]. At this point, as discussed, there is not a fully stable ontological inventory of semantic categories and they are used in an ad-hoc manner as needed with the goal of further developing and stabilizing the categories moving forward.

While in an ideal collection of semantic tags, each feature would be part of a well-defined ontological inventory and defined in a standard vocabulary (such as ISOcat), as discussed at present, the ISOcat is in a state of flux and there is unfortunately no ontology that is sufficiently comprehensive to include all of the semantic and grammatical features needed to annotate the MIX corpus. Thus, in certain cases, some features are currently declared in the feature structure inventory as *hoc categories* for convenience until a more permanent and structured system can be implemented. The example below contains such features, many of which combine with part of speech and other tags in the annotations.

```
<fs>
  <f name="adHocCategories">
    <vAlt>
      <symbol value="temporal" xml:id="TEMP"/>
      <symbol value="manner" xml:id="MNR"/>
      <symbol value="affirmative" xml:id="AFRM"/><!-- occurs as particle' -->
      <symbol value="negative" xml:id="NEG"/><!-- occurs as: particle or tone -->
      <symbol value="degree" xml:id="DEG"/>
       <symbol value="additive" xml:id="ADD"/><!-- occurs as particle ‘ka’ -->
      <symbol value="reciprocal" xml:id="RECIP"/><!-- combines with PRON, ADV -->
      <symbol value="possessive" xml:id="POSS"/>
      <symbol value="mass" xml:id="MASS"/><!-- combines with NOUN -->
      <symbol value="count" xml:id="COUNT"/><!-- combines with NOUN -->
      <symbol value="concrete" xml:id="CONCRT"/><!-- combines with NOUN -->
      <symbol value="relative" xml:id="REL"/><!-- can combine with ADPOS or NOUN (distinction
insignificant) -->
      <symbol value="location" xml:id="LOC"/><!-- can combine with noun, can be same as "placeNoun"--
>
      <symbol value="abstract" xml:id="ABS"/><!-- combines with NOUN -->
```

[239] While in the corpus these feature are separated as described, in the dictionary, some of these combined features are annotated together such as “adv-temp” *temporal adverb.*

```
            <symbol value="collective" xml:id="COLL"/><!-- combines with NOUN -->
            <symbol value="attributive" xml:id="ATTRIB"/>
            <symbol value="predicative" xml:id="PRED"/>
            <symbol value="motion" xml:id="MTN"/>
            <symbol value="departure" xml:id="DEPT"/>
            <symbol value="arrival" xml:id="ARVL"/>
            <symbol value="source" xml:id="SRC"/>
            <symbol value="goal" xml:id="GL"/>
            <symbol value="animate" xml:id="ANIM"/>
            <symbol value="inanimate" xml:id="INANIM"/>
            <symbol value="human" xml:id="HUM"/>
            <symbol value="stative" xml:id="STAT"/>
            <symbol value="bodyPartTerm" xml:id="BPT"/>
            <symbol value="comparative" xml:id="COMPAR"/><!-- combines with any POS tag or phrase to
denote function -->
          </vAlt>
        </f>
      </fs>
```

Figure 103: Inventory of Ad-hoc features used in the corpus

4.6.7.2 Applying Semantic Theory to Corpus Annotation

While as mentioned, the number of tags in the annotation inventory is always subject to change, one major set of features from a specific linguistic theory are the two groups of semantic roles as per Role and Reference Grammar (Van Valin and Foley, 1980; Van Valin, 2005). The first group is *thematic relations* which in RRG, are defined in terms of the argument positions in the decomposed logical structure representations according to Jackendoff (1976, 1987). They are: *agent*, *patient*, *theme*, *experiencer*, *stimulus*, *cognitizer*, *perceiver*, *emoter*. The second is *semantic macroroles*, which are the two primary arguments of a transitive predication and either one of which may apply to different intransitive predictions depending on the semantics of the verb. These labels correspond to what is generally labeled grammatically as “subject” and “object”, and both in RRG and in this annotation system they are used in place of the more traditional aforementioned labels.

The example below shows a table with the given features as annotated in the corpus which shows how the features from RRG align with the grammatical features discussed in the previous section. The sentence (example 6 in section 2.1.2) is a standard transitive predication which translated to: ‘the/that priest hit me’.

	Sutu	**ka**	**ni**	**kani**	**yu**

IGT	priest	=PTCL.DEM	PFV-	hit	me
Semantics	A AGENT				U PATIENT
Gram			V TRANS COMPL		
	N	PTCL DEM	PFV		PRON 1PERS SG
Translations	*The priest hit me* *El sacerdote me golpeó*				

Table 39: Transitive sentence with RRG-based semantic annotations

This second example shows a ditransitive sentence translated as 'I will give money to Jack', which involves the standard thematic roles of an *agent, patient,* and *recipient* along with their given macroroles.

	Kunkua'a			**xu'un**	**nuu**	**Jack**
IPA	**ũ˥**	**gwã˩ʔã**	**˩**			
IGT	POT-	give	1SG	money	face	Jack
Semantics			A AGENT	U PATIENT	BPT	RECIPIENT
Gram	POT	(V)[240]	1PERS SG	N	ADPOS	N-PROP
	VP DITRANS FUT				OBLQ	

[240] The verb stem is not explicitly tagged as *verb* "V" in this case because: a) the verb stem is not marked up separately and b) it is already tagged as *verb phrase* "VP".

	1PERS SG		
Translations	*I will give money to Jack* *Voy a dar dinero a Jack*		

Table 40: Ditransitive sentence with RRG-based semantic annotations

Note that in RRG there are also verb-specific semantic roles: e.g. *runner, killer, hearer, broken,* etc. (Van Valin, 2005), however given that annotation of this information doesn't add anything to the corpus that can be of immediate usefulness to the output, these specific features are not included. It is of course entirely possible to further annotate the rest of the features should the need or desire arise in the future (along with a wide array of other features from RRG on multiple linguistic levels). It is even feasible that this last feature could be semi-automatically added using the information from the verb translations. For example, in an annotated sentence in which the *ditransitive* (DITRANS) *verb* (V) translates as "give", and in which there is an *actor* (A), who will also be the AGENT, an *undergoer,* who will also be the PATIENT, and a RECIPIENT, the verb-specific semantic roles can automatically be added to tag the AGENT as the GIVER, and the RECIPIENT as GIVEN TO.

Another major set of features to be tagged is discussed briefly in section 2.1.5 and Bowers (in press), are those having to do with relational semantics, and cognitive linguistics in general (see Grondelaers et al., (2007) for in depth discussion of the issue of 'cognitive corpora' for linguistic research) as it is relevant not only to the synchronic linguistic structure, but also is highly correlated with issues in grammaticalization, metaphor, metonymy and other types of lexical innovation which are highly relevant themes in Mixtecan languages and linguistics as well as human cognition. Of particular interest in relational semantics, is the use of extended meanings of body-part terms (BPT) in referencing spatial configurations and relations in many different human languages (Johnson, 1987; Lakoff and Johnson, 1980a; Langacker, 1986, 1987; Heine et al., 1991; Svorou, 1994). Likewise, both in MIX, and many of the world's languages BPT, spatial, functional and meronymic semantic profiles are highly productive conceptual sources motivating etymological extensions (i.e. via polysemy and compounding) notably on the level of grammar and frequently in the context of spatial and motion phrases.

A basic factor in the phenomena at hand is that in locating multiple objects with respect to one another, humans naturally exploit asymmetrical relations and extended BPT provide salient conceptual material through which these asymmetrical relationships can be communicated (Lakoff and Johnson, 1980a,b; Langacker, 1986, 1987; Talmy, 1983). The entities being designated and tagged in spatial constructions are the *trajector* (TR) which is the primary entity to be located with respect to the secondary entity, the *landmark* (LM) (Langacker, 1986, 1987, 2010). In the context of SPACE, as well as other relational constructions the relationship between the trajector and landmark is often designated by an extended BPT in MIX.

In Cognitive Grammar, in relational predicates, subject and object status can ultimately be reduced to a kind of focal prominence assigned to participants in a profiled relationship and the role of nominal subject and object specify the *trajector* and *landmark* of a profiled relationship, and while the predominant use of these concepts has been in the analysis in space, this strategy is not limited to space and they can be relevant in analyzing the semantics of non-spatial relations as well (Langacker, 1986, 1987, 2010; Svorou, 1994).

In Bowers (in press) a modified system of Universal Spatial Semantics of Zlatev (2007) and Holistic Spatial Semantics (Naidu et al., 2018) which utilizes the trajector-landmark system are combined in the analysis of various spatial and non-spatial senses of body-part terms, the categories are as follows with the tags in round brackets[241]:

Trajector (TR): static (TR-STAT) | dynamic (TR-DYN)[242]
Landmark (LM): person (PERSON) | object (OBJ)[243] | event (EVENT)[244]
Frame of Reference:

Viewpoint-centered (FOR-VC): defined through 1 or more landmarks
Geocentric (FOR-GC): involves relatively fixed, "absolute" reference points or axis

[241] While the first application of these features is in the analysis of BPT, it's application can and will be spread content that does not contain BPT.
[242] Dynamic trajector indicates *motion*, thus the latter need not be explicitly stated unless specific analysis needed.
[243] The type of landmark *object* is treated here as the default value and is not be labeled explicitly.
[244] Landmark type *event* is applicable to non-spatial applications of the trajector-landmark system in which the schema is extended.

Object-centered (FOR-OC): class of motion situations anchored at deictic center

Region: area of space usually defined in relation to LM

Path: Beginning (PATH-BEG) | Middle (PATH-MID) | End (PATH-END) | Zero (PATH-Ø)[245]

Direction (DIR): used in combination w/FoR where no LM present, multiple values possible: e.g. *Left, Backwards, Forwards*

Motion (MTN)[246]: perceivable actual motion of dynamic trajector

Manner (MNR): *multiple types possible: e.g. run, fly, jump*

In applying this annotation system, as with the semantic features described above the features are included in the @ana of which point to the respective corresponding content. A related and largely compatible approach to annotating such content was carried out for in Kordjamshidi et al. (2017), which is also implements spatial annotations of an XML corpus using a combination of holistic spatial semantics and qualitative spatial reasoning models.

In the following example, the relevant annotations of the given segments are shown in a table format for the sentient *nuu yuku inkaa yu* 'I am in the forest'[247]. This sentence is an *object-centered* (OC) construction, with a *static trajector* (TR-STAT) and the *landmark* is an *object* which is the default value (LM). The body-part term and the location are given the ad-hoc semantic tags (BPT) and (LOC) respectively and the combined span of *nuu yuku* is tagged as *region-internal* (REG-INTERN), which specifies the spatial relation of the trajector with regard to the landmark. The reason that the tag REG-INTERN is applied to the combination of the two former (*BPT and LOC*) is due to the fact that it is only in the context of the given predicate, and the particular landmark 'forest' that the sense of *nuu* denoting this particular spatial configuration is activated.

[245] The features of Path: (Begin | Middle | End) are analogous to those Source-Path-Goal image schema (Lakoff and Johnson, 1980a).
[246] The tag "MTN" for motion is equally used for the classification of the semantics of specific verbs, thus in the data this tag is only applied to the verb itself rather than the sentence level.
[247] Encoded file for the example available at the following location: https://github.com/iljackb/Mixtepec_Mixtec/blob/master/media/xml/S_LOC_I_am_in_the_wilderness_01_02_TS.xml

<table>
<tr><td></td><td>nuu</td><td>yuku</td><td>inkaa</td><td>yu</td></tr>
<tr><td>IGT</td><td>face</td><td>forest</td><td>IPFV\COP.LOC</td><td>=1SG</td></tr>
<tr><td rowspan="3">Semantics</td><td>BPT</td><td>LM
LOC</td><td></td><td>TR-STAT</td></tr>
<tr><td colspan="2">REG-INTERN</td><td colspan="2"></td></tr>
<tr><td colspan="4">FOR-OC</td></tr>
<tr><td rowspan="2">Translations</td><td colspan="4">I am in the forest
Estoy en el bosque</td></tr>
<tr><td colspan="2">in the forest
en el bosque</td><td colspan="2">I am
estoy</td></tr>
</table>

Table 41: Table showing partial annotations of a spatial phrase with extended BPT

The following example shows the annotations for a motion phrase in which the landmark entity is comprised more than one lexical item[248] where the extended body-part term *nuu* is a relative phrase meaning ‘place where’. Note that the frame of reference (FOR-OC) and the path feature of this sentence (PATH-END) are annotated on the sentence level.

<table>
<tr><td></td><td>ntsaa</td><td>kue</td><td>nuu</td><td>yee</td><td>sachu</td><td>-in</td><td>ka</td></tr>
<tr><td>IGT</td><td>PFV/arrive</td><td>=1PL.EXCL</td><td>place.where[face]</td><td>exist</td><td>work</td><td>-3SG</td><td>=PTCL</td></tr>
<tr><td rowspan="2">Semantics</td><td>MTN
ARVL</td><td>TR-DYN</td><td>BPT</td><td colspan="4">LM
LOC</td></tr>
<tr><td colspan="7">PATH-END
FOR-OC</td></tr>
<tr><td rowspan="2">Translations</td><td colspan="2">we arrived
llegamos</td><td>place where
el lugar dónde</td><td colspan="3">his work is
está su trabajo</td><td></td></tr>
<tr><td colspan="7">We arrived at the place where he works.
Llegamos al lugar dónde trabaja.</td></tr>
</table>

[248] The example can be found in the file: https://github.com/iljackb/Mixtepec_Mixtec/blob/master/misc-sources/Tisu-Vienna-Diary-201711.xml the @xml:id of the given sentence is: d1e3802

Table 42: Table showing partial annotations of a dynamic motion phrase with extended BPT

The following example (presented previously as ex. 35 in section 2.1.5) shows the combination of the spatial semantics, the macroroles from RRG and thematic relations in the annotation of a sentence in which the template for the use of the BPT *nuu* 'face' in translocative spatial constructions *transfer-of-location*, is extended into a *transfer of possession* in which the semantic RECIPIENT is analogous to the semantic goal (PATH-END).

	Kunkua'a			**xu'un**	**nuu**	**Jack**
IPA	**ũ˥**	**gwã↘ʔ ã**	**˩**			
IGT	POT-	give	\1SG	money	face	Jack
Semantics			A TR AGENT	U LM-OBJ PATIENT	BPT	RECIPIENT LM-PERS
Gram	FUT -	(V)	1PERS SG	N	ADPOS	N-PROP
	VP DITRANS FUT 1PERS SG				OBLQ	
Translations	*I will give money to Jack* *Voy a dar dinero a Jack*					

Table 43: Example showing the application of both the RRG and Cognitive Grammar-based features in annotation of transitive sentence

4.6.7.3 Final Remarks on Semantic Annotation

An important element in the inclusion of the various sets of features (RRG and the Cognitive Grammar-based features) show how the standoff annotation method allows for overlap of multiple features which can be used to apply and compare features from multiple theoretical systems. Additionally, the overlap of the annotation features clearly demonstrates the ability of

the mechanism to demonstrate and implement the overlapping of semantic and grammatical features without having to create separate corpora for each level of linguistic annotation or compromising the quality or substance of the descriptions.

While this level of detailed annotation cannot be expected to be carried out uniformly in a corpus of a size much larger than this one without a larger team of linguistically trained annotators or a purpose-specific software, implementation of such a system does allow for a systematic annotation of the linguistic data that is relevant to the study of the language and is accessible in carrying out analyses.

There are numerous other systems of semantic and pragmatic annotation that already overlap with the annotations implemented thus far that could potentially be implemented in full, or at least explored in this dataset in the future, notably the annotation of: verbal and predicate semantics, specifically as per the PropBank guidelines (Bonial et al., 2010); spatial semantics as per: Bateman and Farrar (2004a,b), Bateman et al. (2010); ISO-Space (Pustejovsky, 2017) and holistic spatial and motion semantics as per Kordjamshidi et al. (2017); motion and temporal semantics (Pustejovsky and Moszkowicz, 2008), conceptual metaphor as per Shutova (2017) among others. Romary and Salmon-Alt (2009) mapped out a model of the components and relations central to reference resolution in Cognitive Grammar (Langacker, 1986, 1987), with a small amount of additional work, these concepts could be implemented in a standoff TEI annotation system such as this. The ability to implement and integrate concepts from various theoretical systems into the existing annotation model described above without having to change it will be a test of the system's quality and durability.

It does need to be kept in mind that in working with an under-resourced language that providing concrete output usable by non-specialist community members must also factor heavily in the decisions of what major efforts are desirable to implement, as what is interesting to a theoretical linguist takes time and may not have any pragmatic usage to the people whose language I have been working with. Nonetheless, as this work is dually a language documentation of Mixtepec-Mixtec, it is also an endeavor in the use of TEI for creating a dynamic linguistic corpus which should not only meet the needs of an LD project, but should

likewise present some concrete proposals for how to implement less commonly attempted corpus linguistic endeavors it the system that should be relevant to anyone working on any language.

7. Overview of the Mixtepec-Mixtec TEI Dictionary

Aside from the annotated corpus, the archive and media files, the main output of this documentation is the trilingual TEI dictionary derived from the contents of the corpus as well as from any other manner of observation. The entries generally contain the orthographic word forms, phonetic forms (and variants), grammatical, usage, semantics/sense, etymological information and examples from the corpus. In the following sections, each of these features and their TEI encodings[249] are described in detail. At the time of submission, the dictionary has 1,139 entries. Additionally, this collection also contains the additional resource of the Classical Mixtec Dictionary (Alvarado, 1592) which was converted to TEI using GROBID Dictionaries.

The methodology and structure of the Mixtepec-Mixtec TEI dictionary has been described in Bowers and Romary (2018), which presented the in-progress resource; this section restates the content covered therein and provides numerous additional details not discussed as well as updates where necessary. The TEI dictionary of Mixtepec-Mixtec was originally compiled, and is generally edited manually in Oxygen XML Editor, though XSLT scripting methods are sometimes used as needed to both enhance the entries (i.e. with examples of an item as observed in the corpus), and to create new entries as new vocabulary is collected, annotated and identified in the data.

7.1 Metadata and Linking Resources

In addition to the prototypical lexicographic features typical in dictionaries as listed above, through links declared in the header section (TEI Guidelines, The TEI Header)[250], the TEI dictionary functions as a nexus of the linguistic (lexical feature inventories) and other referenced resources (e.g., personographic, bibliographic). The TEI guidelines allow numerous ways of linking to important information that may need to be referenced throughout a dictionary, in this

[249] Note that the dictionary is still undergoing editing and at the time of submission the formatting discussed herein is not yet universally applied and several aspects of the data collections referenced are undergoing modifications.
[250] https://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html

project several approaches based on the type of reference, and the data itself are used. In this section, several of such aspects of the dictionary are described along with discussion of how they are relevant within the context of the language documentation. Figure 104 provides an overview of the linked resources in the Mixtepec-Mixtec dictionary at the heart of this project.

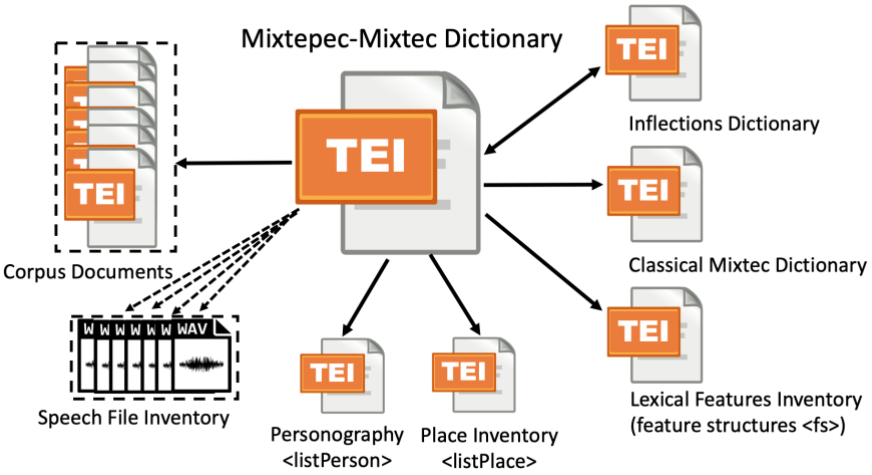


Figure 104: Diagram of dictionary and linked project resources

7.1.1 Lexical Features and Terminology Inventory

As discussed with regard to the corpus annotation features, the inventory of lexical terminology is kept in a separate document containing TEI feature structures. The feature structures document is linked to the TEI dictionary in the <teiHeader> section. Figure 105 shows the declaration of the link to the document contained in the <sourceDesc> of the header in the dictionary (left) and a sample of two particular sets of features (trajector and landmark)[251] from the document it links to.

[251] Currently there exist no registered entries for these concepts in any public terminological repository, and they are among the list of proposals to be submitted for inclusion in the future.

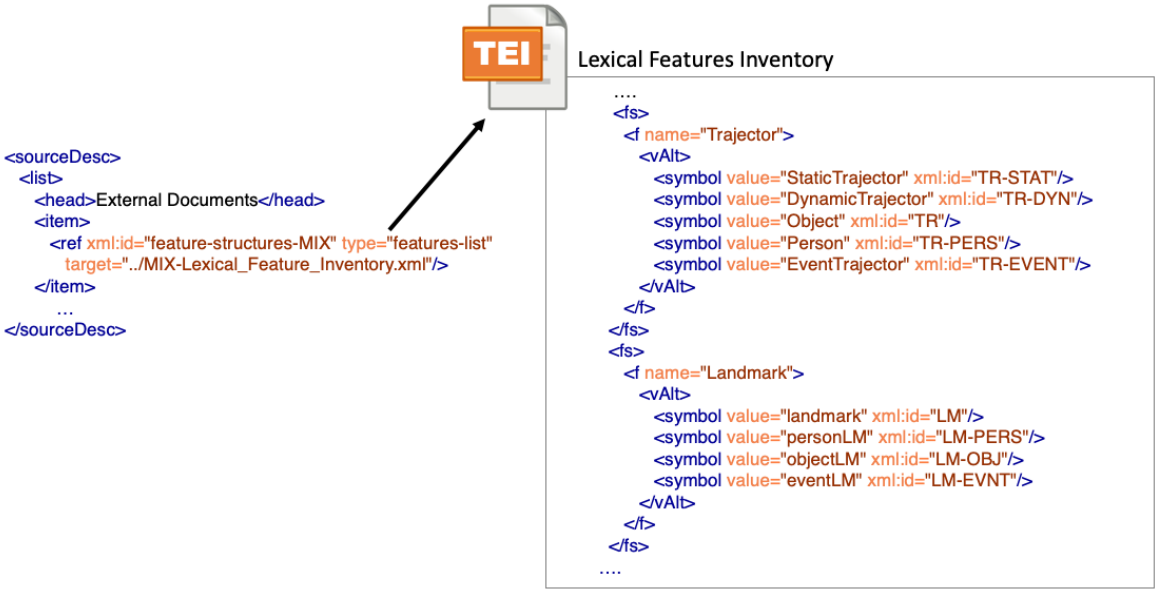


Figure 105: Feature Structures declared in TEI Header and their content in separate document bibliographic sources

7.1.2 Bibliographic Sources

As mentioned above, external data such as the documents from SIL make up a significant portion of the corpus data. Within the dictionary it is often necessary to point to these sources to attribute provenance of content such as example sentences or forms. To enable this, these sources are declared in the <sourceDesc> of the header. The pathway to the given TEI file in the corpus is declared in @corresp and a link to the external source of the file (where applicable) is placed in a pointer element <ptr/>.

```
<listBibl xml:id="SIL-MEX">
  <head>SIL Mexico Publications</head>
  <bibl xml:id="bibl.L093" corresp="SIL_docs/L093/L093-tok.xml">
    <title>Kunka'vi hora ka</title>
    <editor>Beckmann, Gisela</editor>(translator); <editor>Nieves, María M.</editor>
      (translator). <date>2007</date>. <edition>(2nd ed.)</edition>.
      <publisher>Instituto Lingüístico de Verano, A.C.</publisher>
    <pubPlace>Tlalpan, D.F., México</pubPlace> Obtained from:
     <ptr target="http://www.mexico.sil.org/resources/archives/55956"/>
  </bibl>
  <bibl xml:id="bibl.L094" corresp="SIL_docs/L094/L094-tok.xml">
    <title xml:lang="mix">Kunchau hora ka</title>
    <editor>Beckmann, Gisela</editor>; <editor>Gómez Hernández, María</editor>.
      <date>2007</date>. <edition>(2nd ed.)</edition>. <publisher>Instituto
      Lingüístico de Verano, A.C.</publisher>
    <pubPlace>Tlalpan, D.F., México</pubPlace> Obtained from:
```

```
    <ptr target="https://mexico.sil.org/resources/archives/63381"/>
  </bibl>
  …
</listBibl>
```

Figure 106: <sourceDesc> and <listBibl>

7.1.3 Personography

Additionally, as shown in Figure 107, in the header (within <particDesc> embedded in <profileDesc>), each person (speakers, editors, and researchers) who may be referred to directly in the dictionary is listed and given a unique id (with @xml:id) and their role(s) in the work are placed in the @role. This list also links to the external TEI personography document "MIX-People.xml" containing more detailed information about the participants, which in the case of speaker consultants/project collaborators is particularly relevant for the purpose of language documentation; the path to which is declared in @corresp on the element <listPerson>.

```
<particDesc>
  <listPerson corresp="MIX-People.xml">
    <person xml:id="TS" role="speaker collaborator" corresp="MIX-People.xml#TS">
      <name>Juan "Tisu'ma" Salazar</name>
    </person>
    <person xml:id="JS" role="speaker collaborator" corresp="MIX-People.xml#JS">
      <name>Jeremías Salazar</name>
    </person>
    <person xml:id="JB" role="editor researcher" corresp="MIX-People.xml#JB">
      <name>Jack Bowers</name>
    </person>
    <!-- more people here -->
  </listPerson>
</particDesc>
</profileDesc>
```

Figure 107: <particDesc> and <listPerson> for Persons declared in header

7.1.4 External Corpus and Media Files

As described in section 6.3.4.2, <prefixDef> is used to create shortcuts to linked file pathways. Within the dictionary it is uses in linking to several different external contents such as sound files and separate TEI dictionary files, namely those containing inflection paradigms as well as the TEI version of the Classical Mixtec dictionary by Fray Francisco de Alvarado (1593) (cf. Bowers and al. 2019). Within the TEI data structure shown in Figure 108, a <prefixDef> for each separate file to be referenced through the MIX dictionary is declared in the header in which a prefix is declared in the value of @ident. The value of @matchPattern is a template for such

pointers with the regular expression ([a-zA-Z0-9]+), which is replaced by the specific text of a file name (in the case of referencing whole files), or the @xml:id value of a specific entry (in the case of referencing a particular entry in a dictionary).

```
<listPrefixDef>
  <prefixDef ident="alvarado" matchPattern="([a-zA-Z0-9]+)"
    replacementPattern="../VOCESvocab-tei.xml#$1"/>

  <prefixDef ident="paradigms" matchPattern="([a-zA-Z0-9]+)"
    replacementPattern="../paradigms/#$1"/>

  <prefixDef ident="soundfiles-gen" matchPattern="([a-zA-Z0-9]+)"
    replacementPattern="../media/speech-mix/with-txtgrd/#$1"/>

  <prefixDef ident="soundfiles-oax" matchPattern="([a-zA-Z0-9]+)"
    replacementPattern="../oaxaca-oax/#$1"/>
</listPrefixDef>
```

Figure 108: Declaration of the <prefixDef> Patterns Declared in TEI Header for Linking between Documents

Thus, as indicated there are two types of paths defined here in the <prefixDef>'s: those that point to whole files (e.g. *paradigms, soundfiles-oax,* and *soundfiles-gen*) and the one that points to a specific entry in a specific file (e.g. *alvarado*). The way these paths are defined and thus referenced are different, in each case however the path is declared in @replacementPattern.

In the first case, the pathway to the directory where the various sound and other files are declared is simply the folder that the given full files are located in (e.g. "../paradigms/"). In the other case, where the desire is to be able to point to a specific entry in a given TEI file, the full directory and the file name is declared (e.g. "../VOCESvocab-tei.xml"). At the end of the value of @replacementPattern, #$1 means that any pointer with the prefix "alvarado" should point to the given path with the value of the first regular expression: ([a-zA-Z0-9]+).

To show how these mechanisms are used, the TEI version of the Classical Mixtec dictionary is referenced by prefixing "alvarado:" within the string of pointer value. The pointer in Figure 109 links to the entry for Classical Mixtec for *dzini* 'head', since it is cited as a bibliographic reference <ref type="bibl"> the pointer is placed in @source.

```
<ref type="bibl" source="alvarado:cabeza">(Alvarado)</ref>
```

Figure 109: Using Prefix Definition to Reference Entry in Colonial Mixtec Dictionary

7.2 Forms and Grammar

The lemma of a MIX form is given in the <orth> element and, if attested, in the phonetic form (IPA) as well. In MIX, the lemma is the irrealis verb stem which may be different than the realis form (see section 2.1.7). Given that in Mixtecan lexicography, it is of major importance to document patterns of the phonetic root structure (e.g. CVCV, CVV, etc.) and tone patterns (e.g. H, R, LR, etc.), these features are encoded in the dictionary using the @ana attribute on the <form> and <pron> elements respectively[252]. The contents of the @ana annotations are declared in the feature structures and referenced with the hashtag. On the lemma level, the full word structure pattern is included in a single annotation value (e.g. #CVV) with each V representing a vocalic mora, whereas on the <pron> element, each distinct tone is annotated separately Additionally, each entry minimally has the part of speech declared within <gramGrp>, and other features where applicable. The element containing the form always includes the @xml:lang attribute, the value of which is the ISO 639 language tag[253]. If an abbreviated value is used, a @norm attribute with the full form of the feature is given in order to align with terminological standards. A typical example is shown in Figure 110.

```
<form type="lemma" ana="#CVV">
  <orth xml:lang="mix">náa</orth>
  <pron xml:lang="mix" notation="ipa" ana="#H #F">náâ</pron>
</form>
<gramGrp>
  <pos>verb</pos>
  <gram type="transitivity" norm="transitive">trans</gram>
</gramGrp>
```

Figure 110: Sample of typical <form> section of entry with <gramGrp>

As discussed in section 2.17, certain MIX verbs have different stems for the realis and irrealis moods, and the inflections for: perfective, imperfective, habitual (and possibly

[252] Note that in order to save space, not every example of forms in this section will include these features annotated with @ana and this will only be shown where relevant to the content discussed.
[253] The values of English and Spanish used are ISO 639-2; that of Mixtepec-Mixtec is from ISO 639-3 as that is the only option.

progressive)[254] aspects take the *realis* stem and the inflection of potential, imperatives and modal forms. Thus, for verbs where these forms are distinct, the realis form is given following the lemma (which is the irrealis stem); this is structured in the TEI with an embedded <form> with the attribute @type="stem" and the <gramGrp> with a <gram type="aspect"> with the value of *realis*, e.g. Figure 111 shows this for the verb *kaka* 'to walk'[255]:

```
<form type="lemma">
  <orth xml:lang="mix">kaka</orth>
  <pron notation="ipa" xml:lang="mix"/>
  <form type="stem">
    <orth xml:lang="mix">tsika</orth>
    <pron notation="ipa" xml:lang="mix"/>
    <gramGrp>
      <gram type="aspect">realis</gram>
    </gramGrp>
  </form>
</form>
```

Figure 111: Realis verb stem specified in the lemma

Sound files with tokens of the given entry can be included in the entries[256], this is done by embedding a <media> element within the given form. In <media> the media type (*Multimedia Internet Mail Extension*) attribute @mimeType is used to specify the file type and @url is used to point to the file, which as described above, are located in separate directories, and whose path is defined using <prefixDef> (note, this is the identical mechanism that is used in documenting the provenance of a spoken language recording described in section 6.3.4.1, and the elements from the metadata records from the former can be semi-automatically transferred into a corresponding dictionary entry). Thus, to reference a sound file for the lexical item *inka tuku* 'again', its path is specified in the prefix "soundfiles-gen:" and the specific file is given as the other part of that string.

```
<form type="lemma">
  <orth xml:lang="mix">inka tuku</orth>
  <pron notation="ipa" xml:lang="mix">iŋk̟àà tùk̟u</pron>
```

[254] As mentioned in section 2.1.7, the issue of the distinction of *Progressive* aspect in MIX is still being investigated.
[255] Note the inclusion of the realis verb stem is a recent addition and is still being implemented. Additionally, the investigation of the tonal contours of both the realis and irrealis stems is still very much in progress, thus many lemmata in the dictionary still are without <pron> forms.
[256] Due to the need for additional editing analysis and other preliminary issues, the inclusion of sound files in entries is not yet systematically implemented throughout the dictionary. The eventual goal for this project is to produce an interactive online version of this database is to have a high quality recording for each entry.

```
    <media url="soundfiles-gen:ADV_again_01_JS.wav" mimeType="audio/wav"/>
</form>
```

Figure 112: Lemma with link to corresponding sound file

In the case that a sound file is not comprised of a single utterance as is the case of the previous example, it is possible to specify the particular span of time in the file in which the given item is uttered using the @start and @end attributes on <media>. Figure 113 shows the correspondences between the information in the TEI file containing the utterance transcription and the media file in the entry for *naá* 'to finish', in which the media file contains different vocabulary thus the start and end time are specified in the dictionary.

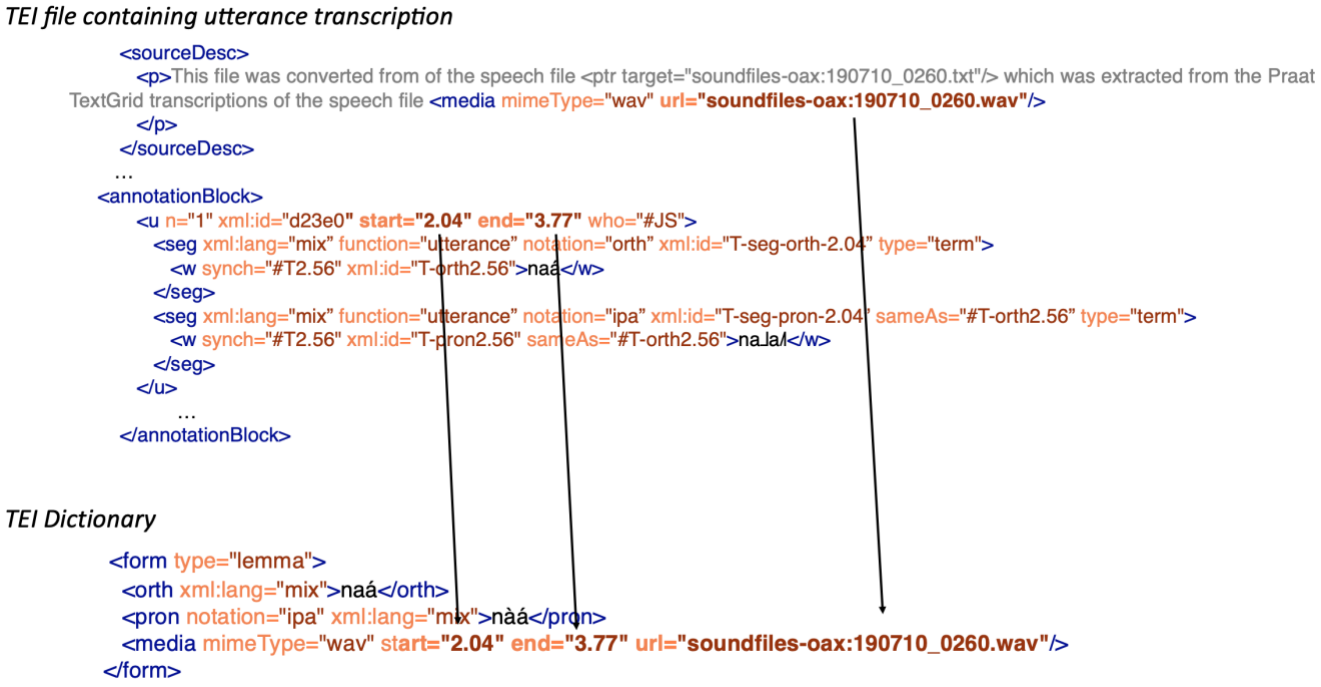


Figure 113: Correspondence between utterance transcription source file and dictionary entry for <media> file and start and end times

The grammatical categories and their values in <gramGrp> correspond to those used in the <spanGrp type="gram"> annotations (described in section 6.4.5). In the example below, the values of <span @type> correspond to the tags in <gramGrp> in the dictionary entry above and the hashtagged value of <span @ana> are the tags for their values ("V" for verb, and "TRANS" for transitive) respectively.

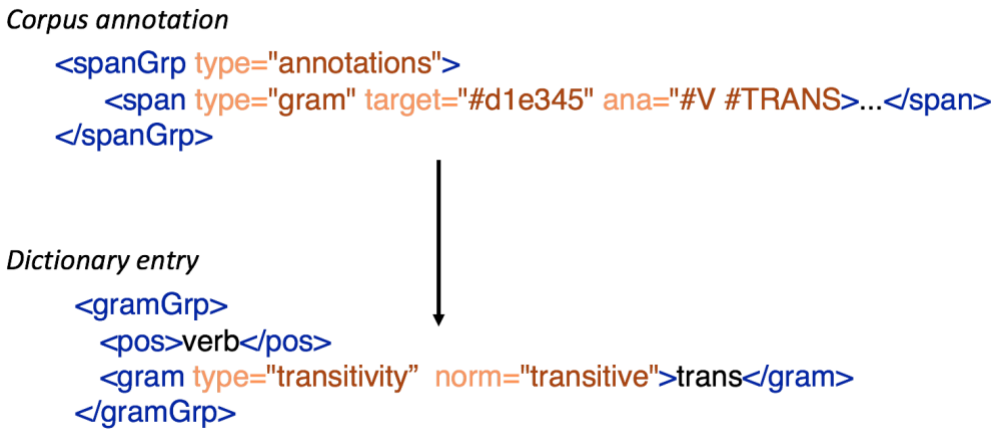


Figure 114: Correspondence between corpus annotation of grammar and representation in dictionary entry

7.2.1 Variation, Uncertain, and Conflicting Forms

As this is a language documentation project, and the language is under-resourced both in its use as a literary language and its linguistic description, it is essential that variation and areas of uncertainty of all kinds are recorded, however each type of variation is unique in the causes, possible ways of handling it conceptually as well as in the TEI modelling. These issues are described in the following sub-sections.

7.2.1.1 Orthographic Variation

Given that the MIX orthography is still under development and significant changes have been made over the last ten years in the SIL source, in addition to the fact that there are still many Mixtec people who use different spelling conventions, there are many lexical items in earlier documents with spellings that have since been changed. In these cases, both the old and up-to-date forms are included in the dictionary. In the earlier publications (encoded herein as the variant form), lexical tone was not represented in the orthography; however, this created a large number of homographs which in some cases were of the same part of speech or even within the same semantic domain. These needed to be distinguished, thus new spellings have been introduced.

The example in Figure 115 shows the updated *chuún* and antiquated *chuun* forms of the word meaning 'chicken' [ʧṵ̀ṵ̆], which is a tone-based minimal pair with the word meaning 'work' [ʧṵ̃ṵ̃], the latter retaining the original spelling while the former adds the accent above the second vowel. The old form is labeled with <form type="variant"> and the element <orth> on which the attribute @notAfter[257] denotes the point from which the new spelling was introduced and that the old spelling ceased to be used.

```
<form type="lemma">
  <orth xml:lang="mix">chuún</orth>
  <pron notation="ipa" cert="medium">ʧṵ̀ṵ̆</pron>
  <form type="variant">
    <orth xml:lang="mix" notAfter="2016">chuun</orth>
  </form>
</form>
```

Figure 115: Entry with example of spelling which has been changed in SIL sources as per a specific date

Additionally, given that the orthographic standard being developed has not been published[258], native speakers who write in the language often do not use the same spelling conventions, and thus when data from such sources are acquired, we are faced with integrating all variants into our common system. The example in Figure 116 shows the encoding of a variant spelling of the lexical item meaning 'water' which was observed in a public service publication by the Mexican government. In this orthography, the voiceless alveo-palatal affricate is represented as *ty* instead of the standard *ch*, and the long word-final vowel is represented only as a single *i*[259]. The source document of the spelling variant is provided as the value of the @source attribute, which is declared in the bibliography within <listBibl> in the header (see section 7.1 above).

[257] Note @notAfter is not currently allowed to occur in <orth> in the general TEI schema, this was done in the project via ODD schema customization by adding <orth> to att.datable.wc3.
[258] The latest known update to the orthography was obtained via personal communication with Mille Nieves of SIL Mexico in June 2017; it is upon this version that all editorial practice is based with regard to spelling normalization.
[259] Though no official documentation of the policy has been made available, according to Millie Nieves of SIL (personal communication, 2019), the variant orthography *ty* is the recommended spelling of the voiceless alveo-palatal affricate of the Mixtec Academy (Ve'e Tu'un Savi). For the Ve'e Tu'un Savi charter, see: https://goo.gl/mnLrWt)

```
<form type="lemma">
 <orth xml:lang="mix">chikuii</orth>
  <form type="variant">
    <orth xml:lang="mix" source="#infografica-308-inundaciones">tykui</orth>
  </form>
  ……
</form>
```

Figure 116: Variant Orthography from MIX Language Publication

7.2.1.2 Phonetic Variation

In our data there are certain lexical items for which pronunciation variants are observed frequently enough that alternate pronunciations are included in the dictionary entry. As shown above for orthographic variants, in pronunciation variants, the primary pronunciation[260] (<pron>) is placed as a direct child of <form>, and the variant is embedded within a separate <form>, also labeled @type= "variant".

Despite there being only a small body of linguistic literature about the language, there are cases where examples of transcribed vocabulary found in such sources are of interest and are thus integrated into the dictionary. Some instances may be the first, or only attestation of the word in the data collected, or may diverge in some way from characterizations of the item as observed from the sources in this project. Additionally, there may be divergence in the transcription conventions used to represent the content.

One such example involves the form of *iin* ‘nine’. In an earlier study, Pike and Ibach (1978) transcribe this item with an onset glottal stop, whereas all evidence from this project, as well as transcriptions from Paster and Beam de Azcona (2004) do not have an onset glottal stop. This difference is noted in the TEI dictionary as a variant form, with the source referenced in the @source as it may be evidence of an idiolect or an antiquated pronunciation. Note that these differences are captured in the different values of @ana on the respective <form type="lemma"> and <form type="variant"> elements, if there were a difference in the tone values in the variant, this would also be reflected in the variant as well.

[260] The primary pronunciation, where present, is determined by weighing the factors of observation frequency and knowledge of the language’s phonology.

```
<form type="lemma" ana= "#CVV">
  <orth xml:lang="mix">iin</orth>
  <pron notation="ipa" xml:lang="mix" ana= "#L #L">ɪ̀ɪ̀</pron>
  <form type="variant" ana= "#CVCV">
    <pron notation="ipa" xml:lang="mix" source="#bibl.pike-ibach-1978" orig="ʔi̱³i̱³">ʔɪ̀ɻɪ̃ɻ</pron>
  </form>
  </form>
</form>
```

Figure 117: Forms with varying phonetic transcriptions from different sources

Another noteworthy observation in this example is the treatment of transcription notation: unfortunately, nearly none of the past studies of MIX phonology used IPA notation in their transcriptions as is done in this project (and should indeed be done for all LD work transcribing speech as per best practices). Fortunately, TEI has the ability both to keep the original forms from the sources in @orig and to normalize the notation to IPA in the element values for compatibility.

7.2.2 Entries with Collocates

In some entries, the lemma is a phrase which to use in an utterance requires the addition of additional variable lexico-grammatical content. These are encoded in the dictionary in <colloc> which is placed directly in the <form> and is intended to be read by users in combination with the orthographic form, e.g. “in so + *(PRON)*”.

```
<form type="lemma">
  <orth xml:lang="mix">in so</orth>
  <colloc>+ (PRON)</colloc>
  …..
</form>
```

Figure 118: Entry with collocate pronoun indicated

The demonstration of the way these collocates are realized in the context of a construction is shown in the example section (see below for more discussion on usage examples).

```
<cit type="example">
  <quote xml:lang="mix">In so ko.</quote>
```

```
    <cit type="translation">
      <quote xml:lang="en">We are related</quote>
    </cit>
    <cit type="translation">
      <quote xml:lang="es">Somos parientes.</quote>
    </cit>
  </cit>
```

Figure 119: Example of entry with collocate in phrasal context

7.2.3 Inflection and Paradigms

As mentioned above, a separate inflections dictionary contains full inflectional paradigms to which entries can link using the TEI <prefixDef> strategy described earlier. This is done with the <ptr> element embedded inside the lemma as shown in Figure 120. Where the verb has a distinct realis stem, as discussed above in the main dictionary, the paradigm files also include both the lemma (irrealis stem) and the realis stem.

```
<form type="lemma" ana= "#CVCV">
  <orth xml:lang="mix">kusu</orth>
  <pron notation="ipa" xml:lang="mix" ana= "#L #L">kùsù</orth>
  <form type="stem" ana= "#CVCV">
    <orth xml:lang="mix">kixi</orth>
    <pron notation="ipa" xml:lang="mix">kìʃì</orth>
    <gramGrp>
      <gram type="aspect">realis</gram>
    </gramGrp>
  </form>
  <ptr type="inflectionParadigm" target="#sleep-V-MIX.xml"/>
</form>
```

Figure 120: Entry with pointer to external paradigm file for lemma kusu ‘sleep’

As discussed in section 2.1, in MIX, inflections can occur on verbs, nouns (for possession), the adverb *nchu’a* ‘very’ (in certain phrasal contexts), on the conjunction *tsi* ‘with’, and on certain adpositions, notably those derived from body-parts amongst others. Within the form section, full paradigms are represented as embedded blocks of inflected forms in accordance with the recommendations of TEI Lex-0 (Bański et al., 2017). Each paradigm is encoded as a sibling of the lemma in <form type="paradigm"> and the primary common feature (aspect and/or mood) is labeled as the value of @subtype, and aspect/voice/mood are encoded in <gramGrp>. In MIX, at this point, for verbs, separate paradigms are collected for: imperfective/perfective aspects and potential (note separate paradigms are not being kept for modal and habituals as their inflections are easily predictable based on the realis/irrealis forms as

well as any of the perfective, imperfective, and potential). For each inflected form, the gloss in English and Spanish are given which are aimed at making these paradigms a readily available reference resource for learners and/or researchers. In Figure 121, the first two forms of the paradigm for the imperfective forms of the verb *kusu 'sleep'* (realis stem *kixi*) are shown. The @ana values on <form> indicating root structure are inherited from the realis or irrealis forms declared on the lemma (e.g. <form type="lemma" ana= "#CVCV">) or stem (e.g. <form type="stem" ana= "#CVCV">). In the @ana on <pron>, the values for both the root and enclitic tones are included.

```
<form type="paradigm" subtype="imperfective">
 <gramGrp>
   <gram type="mood">realis</gram>
   <gram type="aspect" norm="imperfective">imperf</gram>
 </gramGrp>
 <form type="inflected">
   <orth xml:lang="mix">kíxi yu</orth>
   <pron xml:lang="mix" notation="ipa" ana= "#H #L #L">kíʃì jù</pron>
   <gramGrp>
     <per>1</per>
     <number norm="singular">sg</number>
   </gramGrp>
   <gloss xml:lang="en">I'm sleeping</gloss>
   <gloss xml:lang="es">estoy durmiendo</gloss>
 </form>
 <form type="inflected">
   <orth xml:lang="mix">kíxu</orth>
   <pron xml:lang="mix" notation="ipa" ana= "#H #H">kíʃú</pron>
   <gramGrp>
     <per>2</per>
     <number norm="singular">sg</number>
     <gram type="register" norm="informal">inf</gram>
   </gramGrp>
   <gloss xml:lang="en">you are sleeping</gloss>
   <gloss xml:lang="es">estas durmiendo</gloss>
 </form>
  ...
<form>
```

Figure 121: Partial paradigms for kusu 'sleep' in imperfective aspect

Note that there is a <gramGrp> as a direct child of <form type="paradigm">, and this contains the grammatical information common to all the inflected forms in the paradigm and inherited via the inheritance principle (Ide et al., 2000). Variants can be included in the paradigms, where included they are formatted according to the same principles described above.

With regard to the workflow and methodology, the inflection paradigms themselves are created as separate files from customized document templates in Oxygen XML Editor for each new paradigm. For each verb, noun or predicating adjective (and in some cases inflecting adverbs, conjunctions, adpositions, etc.), a separate TEI document is created. There are two templates, one for verbs which includes empty paradigms for *imperfective, potential,* and *perfective inflections*[261]. The file is appropriately named and placed in a folder with other inflection documents.

7.3 Related Entries

Where an entry has given rise to derivatives, compounds or other lexical items, these are represented as related entry <re> elements and embedded within the main entry. Related entries can contain anything the main entry can contain.

```
<entry xml:id="money-MIX">
  <form type="lemma">
    <orth xml:lang="mix">xu'u</orth>
     ...
  </form>
  ….
  <re xml:id="paper-money-MIX">
    <form type="lemma">
      <orth xml:lang="mix">xu'un tutu</orth>
        …..
    </form>
        ….
    <sense corresp="http://dbpedia.org/resource/Cash">
     ….
      <cit type="translation">
        <form><orth xml:lang="en">paper money</orth></form>
      </cit>
      <cit type="translation">
        <form><orth xml:lang="en">bill</orth></form>
      </cit>
      <cit type="translation">
        <form><orth xml:lang="es">billete</orth></form>
      </cit>
   </sense>
  ….
  </re>
  <re xml:id="coins-MIX">
```

[261] While there are other inflection paradigms that can be made, particularly the *modal* and *habitual* which can also be used for expressing future and conditional, this isn't included at present because the difference between that and other inflections is simply a prefix. These can and will be automatically generated and added at a later point when the collection of paradigms is further developed.

```
…
</re>
</entry>
```

Figure 122: Related entry in MIX dictionary

7.4 Sense

The <sense> section of course contains information pertaining to meaning, including definitions, translations, examples of usage in context, domain classification, and a number of other data fields pertaining to semantic relations. An entry may have any number of senses.

7.4.1 Links to External Knowledge Sources

As discussed in section 6.4.7 in the context of the corpus, semantics can be tagged with uri's of open source knowledge resources, these annotations can be transferred to the dictionary using the @corresp attribute within the <sense> element, as shown in Figure 123

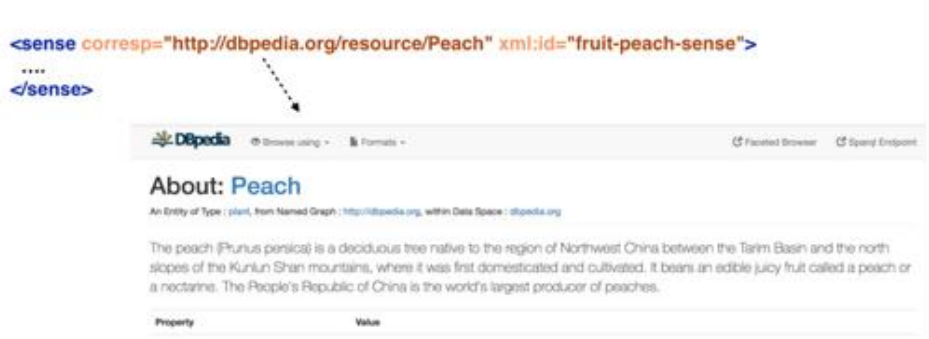


Figure 123: Visualization of use of uri link to DBpedia in sense

This is done with several benefits in mind: one is that they provide a link between a structured body of human knowledge and the Mixtepec-Mixtec language. Currently there are no Mixtec language wiki resources, and these links to DBpedia could provide a template upon which a MIX version of wiki-type entries could be based. Additionally, the multilingual definitions of the concepts found in the entries could serve as a systematic reference point upon which to base MIX definitions of the senses, which are only currently available for a small number of entries. Finally (with the inclusion of @xml:id) they enable (at least partial)

compatibility of the data to semantic web-based linked data formats such as OntoLex-Lemon (McCrae et al., 2017).

7.4.2 Translations

The most basic facet of the sense section is the multilingual translations into English and Spanish. Translations of lemmas are placed <cit type="translation"> within the <form><orth> element block. If the Mixtec item has more than one specific translation in the translation language, the others are listed in separate <cit> elements. The following example shows the English and Spanish translations for the MIX entry *ne'e* 'to scrape'.

```
<cit type="translation">
  <form>
    <orth xml:lang="en">to scrape</orth>
  </form>
</cit>
<cit type="translation">
  <form>
    <orth xml:lang="es">raspar</orth>
  </form>
</cit>
```

Figure 124: Sample of English and Spanish translations

In certain cases, such as for lexical entries for animals and some plant species, scientific names may also be included, which are given the ISO 639-2 language tag for Latin "la". Additionally, literal translations may be included by using the @subtype="literal" in <cit type="translation">. Also, as per TEI Lex0, in order to clearly distinguish the term as being scientific (rather than just a Latin translation) <usg type="domain">scientific</usg> is included in the <cit>. The following example shows both scientific and literal translations for the lexical item *chumi xini ka'nu* 'great horned owl' which is literally translatable as: 'big headed owl'.

```
<cit type="translation">
  <form>
    <orth xml:lang="en">Great Horned Owl</orth>
  </form>
</cit>
<cit type="translation" subtype="literal">
  <form>
    <orth xml:lang="en">big head owl</orth>
  </form>
</cit>
...
```

```
<cit type="translation">
  <form>
    <orth xml:lang="es">búho cornado</orth>
  </form>
</cit>
<cit type="translation">
  <usg type="domain">scientific</usg>
  <form>
    <orth xml:lang="la">Bubo virginianus</orth>
  </form>
</cit>
```

Figure 125: Sample of English and Spanish and Latin (scientific name) translations

7.4.3 Definitions

Entries can include definitions (<def>) in Mixtec, Spanish, and English. A major goal is to have definitions in Mixtec, however at present, most entries do not. English and Spanish definitions may be included where a Mixtec entry doesn't have an exact translation and/or where supplemental information about the translation is needed. In the case of the following example for the MIX word *tise'e*, for which no Spanish or English translation equivalent is known, <def> is used in the latter two languages until the species can be positively identified.

```
<entry xml:id="mosquito-small-MIX">
  <form type="lemma">
    <orth xml:lang="mix">tise'e</orth>
    <pron notation="ipa" xml:lang="mix">tìsēʔé</pron>
  </form>
  ...
  <sense n="1">
    ...
    <def xml:lang="en">Small mosquito that doesn't make noise.</def>
    ...
    <def xml:lang="es">Zancudo chico que no hace ruído.</def>
    ...
  </sense>
 ...
</entry>
```

Figure 126: Example of use of English and Spanish definitions for Mixtec lexical item tise'e for which exact translations don't exist or aren't known

7.4.4 Examples

Any number of examples of the usage of an item in the context of the source data can be included within sense; as per canonical TEI practice, these are also encoded as <cit> with the @type="example" and the wrapper <quote> which has the language tagged in @xml:lang.

English and Spanish translations of the example sentences are included and placed within the <cit type="example">. Examples can also include sound files in which the given entry occurs, as described with its use in <form>, this is done by embedding <media> inside of <cit type="example">[262]. As discussed above, the file directory is abbreviated by the use of the prefix "soundfiles-gen:" which precedes the file name and whose full path is declared in the header using <prefixDef>.

```
<cit type="example">
  <quote xml:lang="mix">¿Nchii nikuu?</quote>
  <media url="soundfiles-gen:S_Q_what_happened_02_spkrTS.wav" mimeType="audio/wav"/>
  <cit type="translation">
    <quote xml:lang="en">What happened?</quote>
  </cit>
  <cit type="translation">
    <quote xml:lang="es">¿Qué pasó?</quote>
  </cit>
</cit>
```

Figure 127: Usage example illustrating instance of item in corpus with linked <media> file

7.4.5 Images

In certain entries (often ones that correspond with certain theoretical interests pertaining to metaphor- and metonymy-driven sense change), images showing the concept denoted in the sense may be included. In TEI this is done with <graphic @url>, within which the <desc> element describes the content of the image. As in <def>, English and Spanish (not shown here) are included along with an empty tag for a future Mixtec description to be added. These images could be used for a pictographic or multimedia learning resource (e.g., a children's dictionary), and in future stages of the dictionary, images may be more systematically added for certain concepts for purposes of pedagogy and/or for use by children. Figure 128 shows a visualization of the given sense of the word for 'face', which in this sense means 'front of' something.

[262] At the time of writing, <media> is not allowed within <cit>, thus this is done by altering the schema via ODD, adding <cit> to model.graphicLike. A proposal to adapt the general TEI schema to allow this has been submitted via the TEI GitHub system (https://github.com/TEIC/TEI/issues/1914).

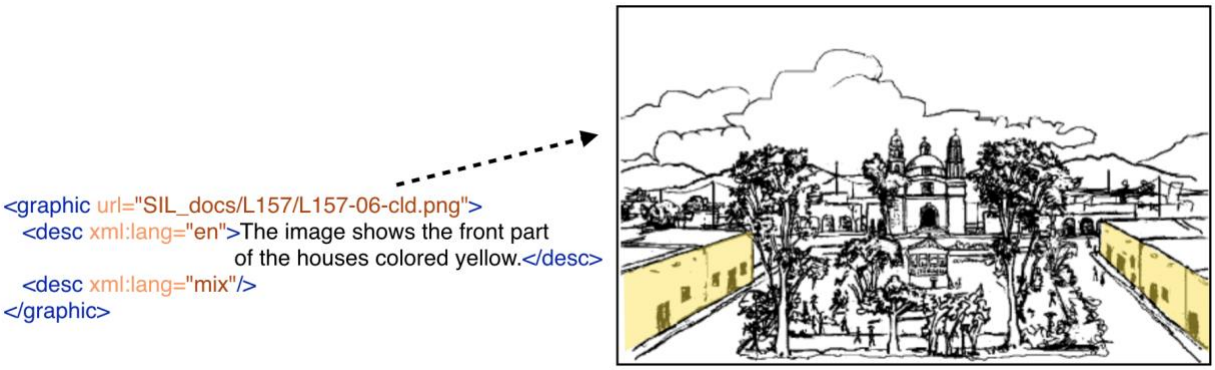


Figure 128: Use of image in sense using <graphic>

7.4.6 Semantics and Cultural Issues in Language Documentation

Especially in a language documentation project, it is important and necessary to include other notes on various specifics of an entry. An example is the lexical item *sa'an ntavi*, one of two terms referring to the Mixtepec-Mixtec language itself and whose components translate literally as 'poor language'[263]. Two of the native speaker collaborators (understandably) find this term offensive and derogatory, and although it must be included in the dictionary because it is still in frequent use in the language, they wanted it marked as dispreferred in the dictionary and the issue to be recorded in prose. This is encoded with a combination of the TEI <note> and <usg> elements, with the @type value of "attitude" and the @resp specifying the initials of those responsible for recording this information, as shown in Figure 129. The initials are the @xml:id value of the individuals and are declared in the header (as discussed in section 7.1.3 under above).

```
<usg type="attitude" resp="#TS #JS">dispreferred</usg>
<xr type="synonymOf">
  <ref xml:lang="mix" target="#MIX-language-rain">sa'an savi</ref>
</xr>
<note resp="#TS #JB">This term which translates as "poor language" is dispreferred by
  speakers consulted as it is derogatory. This is so particularly in contrast to
  the term for the Spanish language <xr type="crossReference">
  <ref corresp="#language-spanish">sa'an xchila</ref></xr> which translates as
  "fancy language".</note>
```

[263] Despite the derogatory or offensive etymology, this is actually the term used most often by native speakers.

Figure 129: Specifying information pertaining to speaker's attitude towards dispreferred lexical item

7.4.7 Semantic Relations and Domain

In addition to sense, translations, and definitions, the dictionary includes information on semantic relations and domain. While the former is commonly utilized in structuralist linguistic approaches and computational linguistics such as WordNet (Miller, 1995; Fellbaum, 1998), the latter is typical of theoretical approaches based in cognitive linguistics (Langacker, 1987; Clausner and Croft, 1999).

While theoretically, these features are a mixture of structuralist and encyclopedic models of semantics (Geeraerts, 2010), for the purposes of the project, including these features in the annotation brings significant benefits both functionally for potential users as well as for linguistic analysis. From the point of view of potential Mixtec users of this resource, these features can be harnessed to facilitate collection and generation of focused sets of vocabulary to be used for the creation of further, more focused resources such as children's books and thesauri. Below the content and implementation of these features in our dataset are described.

Semantic relations in the dictionary are encoded within specific senses of an entry within the external relation element <xr> in accordance with the recommendations of TEI Lex0 (Tasovac and Romary, 2018). The given typology is encoded in @type[264] with an embedded <ref> that takes the @xml:lang as Mixtec and English versions are provided, with English being the metalanguage for computational purposes. Where cross-references point to other entries within the dictionary, the @target attribute is used on <ref>. In the dictionary only, the members/subclasses are tagged, not the top nodes; thus, in the entry for the lexical item *kui'i* 'fruit', the semantic relation 'hypernym' for every specific fruit species is not included as this would be inefficient and burdensome. Instead, this collection can be inferred and built up from the body of items tagged "hyponym" of 'fruit'.

[264] Originally, the attributes @type and @subtype were not available on <xr> within the TEI schema, to rectify this an ODD customization was made and a proposal to add it to the general schema was submitted to the TEI via GitHub (https://github.com/TEIC/TEI/issues/1810). This proposal was accepted 2019-08-19.

Hyponymy is realized as <xr type="hyponymOf">. This category is extremely useful for generating taxonomical vocabulary lists. For the semantic relations hyponymy and meronymy, an additional <ref type="sense"> is included with the @corresp, the value of which is the same as occurs on that item's sense element. Thus, for the entry for 'peach' or other type of fruit, the <ref type="sense">[265] contains the same DBpedia URL as does the <sense @corresp> entry for *kui'i* 'fruit' itself, as shown in Figure 130.

```
<xr type="hyponymOf">
  <ref target="#fruit-MIX" xml:lang="en">fruit</ref>
  <ref target="#fruit-MIX" xml:lang="mix">kui'i</ref>
  <ref type="sense" corresp="http://dbpedia.org/resource/Fruit"/>
</xr>
```

Figure 130: Cross-reference to hyponym

Meronymy is realized as <xr type="meronymOf">[266]. As discussed by Geeraerts (2010), meronymy and hypernymy are central to the realization and analysis of metonymy. Synonymy and antonymy are encoded as <xr type="synonymOf "> and <xr type="antonymOf">. There are limits, however, to semantic relations both functionally and theoretically, as not all relevant semantic correlations in vocabulary or (more importantly) human knowledge can be defined or linked together in terms of hierarchical or pure opposition or identical senses. In order to fill some of that gap, the semantic domain is used.

In addition to semantic relations, which in lexicography are more immediately useful in computational applications, where applicable, semantic domain (Langacker, 1987; Clausner and Croft, 1999) is assigned to the sense of certain entries, a fairly common practice in compiling dictionaries. In lexicographic practice, however, the use of domains in a dictionary is often limited to technical subject classes (e.g., medicine, zoology, literature) though the FLEx software does offer a more expansive, yet nonetheless incomplete system of semantic domains as per Moe (2003) (see section 4.4.3.2.3.4). Domains are fundamental cognitive concepts according to which

[265] The inclusion of both the English translations of the related entries (<ref xml:lang="en">) and the referenced senses (<ref type="sense">) are part of the data design but the systematic implementation of which is not prioritized at this point. Later (<ref xml:lang="es">) will also have to be added as it is likely community members will prefer to have the Spanish as well.

[266] While meronymy can be and in a number of theoretical sources is subtyped according to different conceptual paradigms, there are theoretical conflicts (Geeraerts, 2010) as to the soundness of these distinctions; until further research and evaluation of this question can be carried out, then, such sub-typologies will not be assigned.

humans organize, understand, and represent experience and knowledge of the world (Langacker, 1987; Clausner and Croft, 1999), and this is a particularly enriching perspective in approaching language documentation.

In cases of polysemy, semantic domain is often a key distinction between the various senses. In Figure 131 we show the senses in the entry for *kani* 'long' (domain of SPACE), which can also be used in the sense of the domain TIME. In TEI, domain is encoded as <usg type="domain">[267]. Note this example will be further discussed in the next section 7.5 in the context of etymology.

```
<sense n="1" xml:id="long-space">
  <usg type="domain">Space</usg>
  <cit type="translation">
    <form><orth xml:lang="en">long</orth></form>
  </cit>
  <cit type="translation">
    <form> <orth xml:lang="es">lungo</orth></form>
  </cit>
 <sense n="2" xml:id="long-time">
    <gramGrp>
      <pos>adv</pos>
    </gramGrp>
    <usg type="domain" corresp="http://dbpedia.org/resource/Time">Time</usg>
    <cit type="example">
      <quote xml:lang="mix" resp="#TS">
       <xr type="crossReference"><ref>Kani</ref></xr> nchu'a ntsi ra.</quote>
      <cit type="translation">
        <form>
          <orth xml:lang="en">He lived a long time.</orth>
        </form>
      </cit>
      <cit type="translation">
        <form>
          <orth xml:lang="es">Vivió mucho tiempo.</orth>
        </form>
      </cit>
    </cit>
   <!-- <etym> here -->
  </sense>
</sense>
```

Figure 131: Embedded senses for kani meaning 'long' (SPACE) or (TIME)

[267] Where available, like <sense> and cross-references (<xr>), domain (<usg type="domain">) may also include URLs from external ontologies or sources such as dbpedia.

The inclusion of semantic domain potentially enables an alternate system of organization of a dictionary from the typical alphabetical ordering, or a derived domain-specific dictionary, and it can help with both manual and automatic word sense disambiguation (WSD)[268]. Finally, domains enable us to encode and provide more dynamic analyses of sense-based etymological processes in keeping with cognitive linguistic theory. This latter is particularly important to the description of Mixtecan languages, as discussed in the following section.

7.5 Etymology

In addition to the general documentation of the language, the dictionary is being created as a structured database of etymological information. In the data the full array of etymological processes has been observed, including: borrowing (mostly from Spanish, some from Nahuatl); inheritance (from a posited Proto-Mixtecan language inferred by comparing cognates); form changes such as compounding, derivation, onomatopoeia, phonological change; various types of sense change such as metaphor, metonymy, and grammaticalization, as well as numerous instances of combinations of these processes. The topic of encoding etymological information in TEI as applied to this project has been discussed by Bowers and Romary (2018b), and the conventions used are in line with the recommendations of TEI Lex-0 Etym (Bowers et al., 2018) and Bowers and Romary (2016). Additionally, the entire vocabulary from the 1593 Classical Mixtecan dictionary by the Dominican fray Francisco de Alvarado (1593) has been converted into TEI (Bowers, Khemakhem, and Romary 2019) using GROBID Dictionaries (Khemakhem et al., 2017) and the contents from are to be integrated into the dictionary as an important historical reference[269].

7.5.1 Inheritance, Cognates and Cross-references

As is common in the practice of philology and historical linguistics, by comparing and cataloguing other varieties of Mixtec it is possible to make conclusions about where lexical items share an etymological source. In such cases, the etymological process “inheritance” is assigned to the given entry: <etym type="inheritance"> (as per Bowers and Romary, 2016).

[268] Word sense disambiguation is particularly important given that the MIX orthography represents tone only on a small percentage of words.
[269] The vocabulary from Alvarado (1593) was converted in GROBID Dictionaries from an edited PDF version of the contents by Jansen and Perez Jiménez (2009).

7.5.1.1 Reconstructed Forms

As mentioned in section 5.4, there are several literary sources in which Proto-Mixtecan forms were presented through a comparative study of multiple Mixtec varieties: these are Longacre (1957), Mak and Longacre (1960), Longacre and Millon (1961), Josserand (1983) and Dürr (1990). These reconstructed forms are being integrated into the dictionary as etymons as shown in the following three examples in which the source of the form is duly cited in the <ref> element with the attribute value pairs of type="bibl" and the attribute @target points to the @xml:id value of the source as declared in the header.

```
<cit type="etymon">
  <form>
    <pron notation="ipa" orig="*ⁿdu³ⁿdi⁴" xml:lang="und-PMx">*ⁿdu˧˩ⁿdi˥</pron>
  </form>
  <ref type="bibl" target="#LongacreMillon1961">(Longacre and Millon, 1961)</ref>
</cit>
```

Figure 132: Encoding of reconstructed Proto-Mixtecan etymon from Longacre and Millon (1961)

```
<cit type="etymon">
  <form>
    <pron xml:lang="und-x-PMx">*sawiʔ</pron>
  </form>
  <ref type="bibl" target="#Josserand1983">(Josserand, 1983)</ref>
</cit>
```

Figure 133: Encoding of reconstructed Proto-Mixtecan etymon from Josserand (1983)

```
<cit type="etymon">
  <lang>Proto-Mixtec</lang>
  <form>
    <pron xml:lang="und-x-PMx">*túùtù</pron>
  </form>
  <ref type="bibl" target="#Dürr1987">(Dürr, 1987)</ref>
</cit>
```

Figure 134: Encoding of reconstructed Proto-Mixtecan etymon from Dürr (1987)[270]

[270] Note, this entry shows tone represented as diacritics whereas the example above from Longacre and Millon uses the tone characters (˧˩), this is because the language variety depicted by Longacre and Millon (1961)has a tone level 4 which can't be represented by IPA combining diacritics. Eventually these may all be normalized into a single system using the separate tone characters instead of the combining ones.

Due to the age of these resources, the extraction of the Proto-Mixtecan vocabulary from these original publications which, in the best of cases are available in PDF with text recognition and in the worst cases, simply scanned documents from pages written with a typewriter which is not even searchable. This makes the integration of the language content from these sources an incredibly burdensome process which must be manually done. The expansion of an OCR technology such as GROBID Dictionaries (Khemakhem et al. 2017) in order to be able to process data found in linguistic papers would fill an important gap in data digitization and preservation and would save researchers time and enable the easy access of linguistic data (see Maxwell and Bills 2017 for a discussion on another approach to retro-digitization of a legacy dictionary for an endangered language and Blockland et al. 2019 for a discussion on retrodigitization of non-dictionary text collections).

7.5.1.2 Historically Attested Forms from Alvarado Yucu Ndaa Vocabulary (1593)

Content from the Classical Mixtec dictionary is represented as a cross-reference <xr> with a date and a <lang> tag. Even though it is clearly a historically, and etymologically closely related form and language variety, the relation is not one of direct inheritance from Yucu Ndaa to Mixtepec-Mixtec, thus it is just represented as a referenced form. In referencing the Colonial Mixtec era dictionary, <prefixDef> (described above in sections 6.3.4.2 and 7.1.4) is used in the @source attribute of <ref type="bibl">. In the value of @source[271], the prefix “alvarado:” is placed before the value of the @xml:id for the entry that is being pointed to, e.g. “cabeza”.

```
<etym type="inheritance">
  ….
  <xr type="crossReference">
    <lang>Yucu Ndaa</lang>
    <date>1593</date>
    <ref type="entry" xml:lang="und-x-cmx">dzini</ref>
    <ref type="source" source="alvarado:cabeza">(Alvarado, 1593)</ref>
  </xr>
  ….
</etym>
```

Figure 135: Example of cross-referenced form from classical Mixtec dictionary

[271] Note that while the @source is used differently in each of the previous example, with the former citing Longacker and Millon (1961) (e.g. @source="#longacker1961”) and the latter using the prefixDef prefix (e.g. @source="alvarado:cabeza”), the reason is that the first is pointing only to the ID of the bibliographic source declared in the header, whereas the second is pointing directly to an entry in a separate TEI document.

7.5.1.3 Cognates

As cognates from related Mixtec varieties are observed and collected they can be integrated into the dictionary. The structure of cognates in TEI is represented in the same way as etymons, e.g. <cit type="cognate">. The language variety is specified in <lang> as well as in the @xml:lang value on the forms, possibly the location, and the bibliographic source is encoded as <ref type="bibl"> with the text of the source cited in the element value and the pointer to the resource in the attribute value @source. Given that it has been common in the previous Mixtec literature for authors to not use any standardized transcription system, it is necessary to record which system is represented in the text form. This is done with the @notation attribute on the <pron> element, if the source doesn't use IPA then a distinct string containing the author's name and the language variety is used. The following examples show two such cognates, one in which the source literature transcribed the form in IPA and the other which did not.

```
<cit type="cognate">
  <lang>Coatzospan Mixtec</lang>
  <form>
    <pron notation="trans-smll-miz" xml:lang="miz">rkɨ</pron>
  </form>
  <ref type="source" target="#Small-CoatzospanMix-1990">(Small, 1990)</ref>
</cit>
<cit type="cognate">
  <lang>San Martín Duraznos</lang>
  <form>
    <pron notation="ipa" xml:lang="smd">ʃĩɲī</pron>
  </form>
  <ref type="source" target="#Padgett-2017">(Padget, 2017)</ref>
</cit>
```

Figure 136: Two cognates from related Mixtec varieties

7.5.2 Borrowing

As mentioned, the vast majority of loanwords are from Spanish, however there are also some from Nahuatl as well (shown in the following example). Where present, the process of *borrowing* is labeled as <etym type="borrowing">, within which the etymon is given the ISO 639 language tag on the forms and the <lang> tag for human consumers. In this example the use of <seg type="desc"> is shown which is also used for the benefit of humans reading the material; as the working language is currently English, and the desire is to also provide such descriptive and narrative prose content in Spanish and eventually in Mixtec, the ISO 639-2 tag for English is

included on the element as well. The following example shows the etymology section for the MIX entry *tekiu*, which is a social custom of community labor common to Mixtec and many other indigenous people of Mexico.

```
<etym type="borrowing">
  <seg type="desc" xml:lang="en">Loanword from:</seg>
  <cit type="etym">
    <lang>Nahuatl</lang>
    <form>
      <orth xml:lang="nah">tequitl</orth>
    </form>
  </cit>
</etym>
```

Figure 137: Etymology section from entry for loanword *tekiu* from Nahuatl *tequitl*

7.5.3 Onomatopoeia

In the vocabulary for birds and insects there are several identifiable instances of onomatopoeia, these are encoded quite simply with the <seg type="desc"> as follows:

```
<entry xml:id="bumblebee-MIX">
  <form type="lemma">
    <orth xml:lang="mix">tirri</orth>
    <pron notation="ipa" xml:lang="mix" cert="medium">tiríí</pron>
  </form>
  ...
  <etym type="onomatopoeia">
    <seg type="desc" xml:lang="en">Onomatopoeia based on buzzing sound made by
bumble bee.</seg>
  </etym>
</entry>
```

Figure 138: Form and etymology sections for MIX entry *tirri* 'bumble bee'

7.5.4 Phonological Changes and Multiple Etymological Processes

As mentioned, it is quite common for multiple etymological processes to be evident in a given entry, often this involves compounding and sense related changes. However, in other entries such as in the following example for the entry for the bird *lachacha* 'chacalaca', the item which was clearly borrowed from Spanish underwent a phonological change via metathesis in which the phonological components of the word are scrambled.

```
<entry xml:id="Chachalaca">
  <form type="lemma">
    <orth xml:lang="mix">lachacha</orth>
```

```
    ..
  </form>
  …
  <etym type="borrowing">
    <seg type="desc" xml:lang="en">Altered pronunciation of loanword from:</seg>
    <etym type="metathesis">
      <lang>Spanish</lang>
      <cit type="etymon">
        <form>
          <orth xml:lang="es">chachalaca</orth>
        </form>
      </cit>
    </etym>
  </etym>
</entry>
```

Figure 139: Complex etymology containing a phonological altering (via metathesis) of loanword from Spanish for the bird *lachacha* 'chacalaca'

7.5.5 Sense-related Etymologies

As a major point of emphasis in this project is the semantics of MIX, specifically the strategies of lexical innovation, particularly from the perspective of cognitive linguistics. As mentioned throughout, there exists a significant body of literature discussing the evidence of metaphor and metonymy in lexical innovation in related varieties of Mixtecan (Hollenbach, 1995a; Brugman and Macaulay, 1986; Langacker, 2002); the dataset for MIX provides ample content that enriches such linguistic discussions (Bowers, in press). Although there are limitations to the degree of cognitive nuance and granularity of the synchronic and diachronic semantics of the language in a semasiological dictionary structure, it is possible to represent a significant enough portion of such information to be useful both in terms of producing: a dictionary that is etymologically informative for the community about their language, and a well-structured machine readable resource that systematically keeps track of key linguistic information relevant to theoretical research.

7.5.5.1 Metaphor

Figure 140 shows the etymology for MIX *kani* 'long' in the sense of the domain of TIME (discussed in the previous section).

```
<sense n="1" xml:id="long-space">
  <usg type="domain">Space</usg>
  <cit type="translation">
    <form><orth xml:lang="en">long</orth></form>
```

```
</cit>
<cit type="translation">
  <form><orth xml:lang="es">lungo</orth></form>
</cit>
<sense n="2" xml:id="long-time">
  <gramGrp>
    <pos>adv</pos>
  </gramGrp>
  <usg type="domain" corresp="http://dbpedia.org/resource/Time">Time</usg>
  <cit type="example">
    <quote xml:lang="mix" resp="#TS">
          <xr type="crossReference"><ref>Kani</ref></xr> nchu'a ntsi ra.</quote>
    <cit type="translation">
      <form><orth xml:lang="en">He lived a long time.</orth></form>
    </cit>
    <cit type="translation">
      <form><orth xml:lang="es">Vivió mucho tiempo.</orth></form>
    </cit>
  </cit>
  <etym type="metaphor" cert="high">
    <seg type="desc">Active zone of source profile (aka ontological
      knowledge/impetus) motivating the metaphor is QUANTITY. The domain mapping
      directionality of the sense change is: QUANTITY of SPACE (SIZE or DISTANCE)
      → QUANTITY of TIME. The domain shift is thus: SPACE → TIME. This
      directionality is predictable as it follows the pattern of: CONCRETE →
      ABSTRACT; and of which, the foremost is SPACE → TIME.</seg>
    <cit type="etymon" corresp="#long-space">
      <usg type="domain">Space</usg>
    </cit>
  </etym>
</sense>
</sense>
```

Figure 140: Example of metaphor in embedded sense

Despite having no written evidence of this lexical item in earlier stages of the language in the Alvarado dictionary (1593) or any other source, it's nonetheless possible to assert the directionality of this relationship between these senses, as the metaphorical process of SPACE > TIME is a predictable mapping that follows the general pattern of utilizing concrete conceptual structures to describe and understand abstract concepts (Kövecses, 2010; Gentner et al., 2002; Boroditsky, 2000). Herein the sense of 'long' (TIME) is embedded within the first spatial sense, which in this dictionary is done where one sense is clearly derived from another. When there is one or more embedded <sense> elements, the respective etymologies within should be considered sequential, stemming from the highest sense. In the example, they are also numbered using @n.

As these semantic topics are of major linguistic interest in the MIX language (and for other Mixtecan languages), a prose linguistic description of the analysis of the given process is given in the <seg type="desc"> element. Given that it is a polysemy, and is the same form as the source sense, the etymon <cit type="etymon"> does not have a form in this case as it does in other types of etymological processes. The @corresp attribute points to the source of the sense change that is the first sense. In addition to the @ type="metaphor", the data structure contains the key information for that process in the <usg type="domain">, which are in both senses, and copied within the <cit type="etymon">. Together with the embedding of senses and etymology, the contrast in the domain values from the first sense to the second provides a set of structured data that can be computationally searched and summarized when analyzing such phenomena as metaphor and domain directionality.

7.5.5.2 Metonymy

As metonymy provides mental access to a target entity in a single domain via the highlighting of various aspects of part-whole (meronymy) or class-member (hyponymy)[272], the specifics of an instance of metonymy are specified in the attribute @subtype. In the following example which shows the encoding of the entry *kiti* 'animal' or 'horse', the etymological details of the latter sense are given in that portion of the entry.

Most notably, the specifics of the etymological process (metonymy) and its subprocess (category for member) are labeled as attributes of <etym>, as @type="metonymy" and @subtype="categoryForMember" respectively. As in the previous example with metaphor, a prose description containing the rationale for the analysis in which the directionality that the term *kiti* originated as the categorical term for 'animal' and then, upon introduction via the Spanish, was extended to also denote 'horse' due to a natural process in which the new animal was simply referred to as 'animal'. The date is given as per historical sources (Spores and Balkansky, 2013) which put the first Spanish incursions into La Mixteca as occurring sometime in 1520; the imprecision of the coining of the term is denoted by the @notBefore attribute on the <date> element.

[272] In a metonymy where the process is *category for member* (such as in Figure 141) the domain is actually changed as the source sense becomes the domain.

Once again, as in the previous example, the form of the extended sense and that of the entry are the same so there is no need to include a copy of it in the <cit type="etymon">. Instead a pointer @corresp points to the source sense “#animal” (the first sense of the entry meaning ‘animal’), and within the etymon (<cit type="etymon">) a copy of the source lexical semantic profile is included in <xr> (hyponymOf/*is a: chaku* ‘living being’) in order to provide a contrast with that of the second sense (hyponymOf/*is a: kiti* ‘animal’) which is the higher level semantic change.

```
<entry xml:id="animal-horse">
   <form type="lemma">
     <orth xml:lang="mix">kiti</orth>
     <pron xml:lang="mix" notation="ipa">kìt̪ĭ</pron>
   </form>
   ...
   <sense xml:id="animal" corresp="http://dbpedia.org/resource/Animal" n="1">
     <usg type="domain">Living Being</usg>
     <xr type="hyponymOf">
       <ref xml:lang="mix" target="#living-being-MIX">chaku</ref>
     </xr>
     <cit type="translation">
       <form><orth xml:lang="en">animal</orth></form>
     </cit>
     ...
     <sense xml:id="horse" corresp="http://dbpedia.org/resource/Horse" n="2">
       <usg type="domain" corresp="http://dbpedia.org/resource/Animal">Animal</usg>
       <xr type="hyponymOf">
         <ref xml:lang="mix" target="#animal">kiti</ref>
       </xr>
       <cit type="translation">
         <form> <orth xml:lang="en">horse</orth> </form>
       </cit>
       <cit type="translation">
         <form> <orth xml:lang="es">caballo</orth> </form>
       </cit>
       <etym type="metonymy" subtype="categoryForMember">
         <seg type="desc" xml:lang="en">In this lexical item, the language reflects
           the history, since there were no horses in Mexico until the arrival of the Spanish
           in the Mixteca (sometime in <date notBefore="1520">1520</date>), there was
           naturally no Mixtecan word for 'horse'. Thus, it is clear that the categorical noun meaning
           'animal' was used to describe the unnamed animal and this term lexicalized into
           the language.</seg>
         <cit type="etymon" corresp="#animal">
           <xr type="hyponymOf">
             <ref xml:lang="mix" target="#living-being-MIX">chaku</ref>
           </xr>
         </cit>
       </etym>
     </sense>
```

```
  </sense>
</entry>
```

Figure 141: Example of *category for member* metonymy in etymological entry

7.5.6 Complex Etymologies: Derivation and Metonymy

As discussed in section 2.1.8, MIX contains several productive derivational prefixes that are used to create new lexical items (often verbs). These can be combined with other prefixes as well as other forms with complex etymologies. The following example shows such an instance in the entry for *ntasaxeen* 'to sharpen' (section 2.1.8.4), in which two derivational prefixes; the iterative *nta-*, and the causative *sa-*, are attached to the base *xeen* 'dangerous'. This example shows the use of the function word 'From' in <seg type="desc">[273], as well as plus characters "+" to indicate for the human viewer the combination of the given etymons. Note in this example the etymons with the derivational prefixes are placed as direct children to the first level <etym type="derivation">, and the portion containing *xeen* 'dangerous' is embedded in another <etym> with @type="metonymy and subtype="partForWhole" (as it is metonymy in that the single aspect of a sharp object; that it is *dangerous*, is used to represent the whole concept of sharp).

```
<entry xml:id="sharpen">
  <form type="lemma">
    <orth xml:lang="mix">ntasaxeen</orth>
  </form>
  ...
  <etym type="derivation">
    <seg type="desc" xml:lang="es">De:</seg>
    ...
    <cit type="etymon">
      <form><orth xml:lang="mix">nta-</orth></form>
      <gramGrp>
        <pos>prefix</pos>
        <gram>iterative</gram>
      </gramGrp>
    </cit>
    <pc>+</pc>
    <cit type="etymon">
      <form><orth xml:lang="mix">sa-</orth></form>
      <gramGrp>
        <pos>prefix</pos>
        <gram>causative</gram>
      </gramGrp>
    </cit>
    <pc>+</pc>
```

[273] In the <seg type="desc"> Spanish is shown in the example but English and Mixtec also included in the actual file but are not shown here to save space.

```
      <etym type="metonymy" subtype="partForWhole">
        <cit type="etymon">
          <form><orth xml:lang="mix">xeen</orth></form>
          <gramGrp>
            <pos>adj</pos>
          </gramGrp>
          <def xml:lang="en">dangerous</def>
          <def xml:lang="es">peligroso</def>
        </cit>
      </etym>
    </etym>
  </entry>
```

Figure 142: Example of complex etymology combining multiple derivational prefixes and metonymy

7.6 Human Oriented Output

The dictionary is converted to HTML (done using XSLT) which is formatted with CSS, from the HTML a PDF can also be derived. These versions of the output have the capacity to contain images and play media files as well. At present these files are only available on the GitHub repository until a more long-term online location can be established the allows users to access both the dictionary and the corpus contents. Note however that the formatting is not finalized, see a current sample in Figure 143.

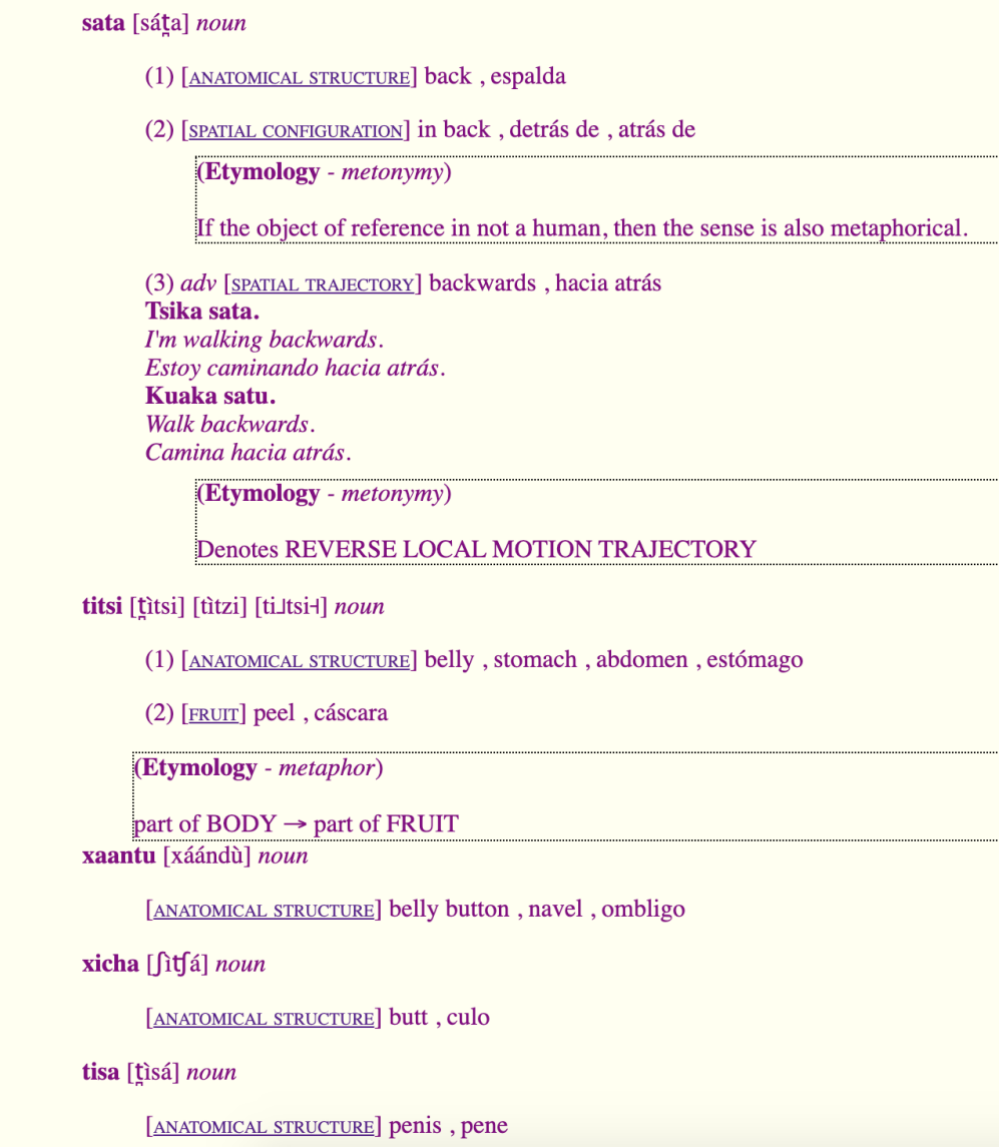


Figure 143: Screenshot of HTML version of MIX TEI Dictionary

Additionally, at present using an XSLT script, the contents are regularly exported to CSV and Excel[274] to make the data available to those who do not work with XML. Further work will need to be done in order to develop a conversion between FLEx's LIFT format and TEI as well as a script producing CLDF from TEI dictionary contents as well.

[274] See contents of datasets converted to tsv and HTML from XML in the following directory: https://github.com/iljackb/Mixtepec_Mixtec/tree/master/exports

8. Conclusion

In this dissertation I have described the work done over the course of the last ten years documenting the Mixtepec-Mixtec variety, in which the primary output has been an open source body of reusable and extensible multimedia language resources including: a multilingual TEI Dictionary, a collection of audio recordings published and archived on Harvard Dataverse; a corpus of texts derived from transcriptions of the spoken language and written language encoded and annotated in TEI; a preliminary grammatical description of basic aspects of inflection, morphology, and derivation; as well as a publication of linguistic analysis of the semantics of body part terms in the language (Bowers, in press).

Aside from the creation of the LR and the study of the language itself, a major focus and achievement of this work has been in the articulation of the many fundamental ways in which the pursuit of an LD project spans across an array of linguistic and other academic fields including: digital humanities, descriptive linguistics, digital lexicography, computational linguistics, as well as most every other subfield of linguistics. A primary thread that is relevant to both this work and the aforementioned disciplines centers around data, including: metadata, all various types of primary linguistic data, markup standards, annotation, analysis and archival as well as the tools that are used to create and manage data. Over the course of this project and dissertation, it has been a priority to identify the current limitations to the necessary workflows in the creation and management of the aforementioned due to a lack of sufficient capacity for data interchange and interoperability between the tools (with the exception of ELAN), as well as a lack of established mappings and conversion schemas between the key data formats created and standards used at different stages of the LD process.

To the best of my knowledge this project marks the first instance that the TEI guidelines have been used in carrying out the full array of central components of LD and thus, represents a step towards both ensuring that the standard is sufficiently capable of encoding all the necessary contents and establishing a precedent for the given practices for potential future adopters for similar projects. While TEI is indeed mostly capable of handling most of the many facets inherent to LD, namely the representation of spoken language, linking of media, dictionaries/lexicon development, annotated text corpora and various types of metadata, there are

a few minor areas that have been identified in need of improvement in the system (many of which I have already taken steps to change)[275]. One major one in particular is that the TEI severely needs an established practice is in interlinear glossed text, which is the primary method of text annotation in LD but has very little precedent in TEI; while in this project I do apply and present a method of IGT in the corpus annotation, the method I use is in combination with my standoff annotations which is not displayable in a user oriented way without further transformation. In Figure 28 (section 4.4.2.4) I present a possible TEI version of IGT that would be a likely candidate to become canonical practice as it is structurally highly compatible with the EAF and FLEx equivalents.

Another issue central to this project that has up to the present not been entirely settled in TEI is standoff annotation. Herein I chose to apply a multi-layer standoff annotation method to this corpus due to the combination of facts that it is both considered best practice in language documentation to keep the analysis separate from the source content and because it offers the best means of expressing overlapping features and a potentially infinite number of separate features that do not have to be applies evenly throughout the entire corpus. Despite these benefits, there remains the fact that there is little precedent or support for either searching and retrieving data in this particular format (which has been done herein with custom XSLT and XQuery scripts) or for display of data in this format, which requires further transformations and custom programming. At the moment, I have not fully achieved the level of retrievability desired for this dataset partially because the annotation and in some cases the encoding is still in progress. Finally, the time necessary to carry out the manual standoff annotations is significant and while Oxygen XML Editor does offer some assistance with the burden, it is still quite slow. While carrying out such an annotation process and creating the necessary custom scripts to search, retrieve and/or transform the annotated data into a user friendly presentation format is possible for myself, it would not be possible to do so for someone who does not have any expertise in programming.

[275] https://github.com/TEIC/TEI/issues?q=is%3Aissue+author%3Ailjackb

Thus while this work makes significant progress in furthering the capacity of TEI to be used in an LD context and furthers the precedent of using standoff annotation in a TEI corpus, other than providing detailed examples of the encoding of every aspect of LD data, significant work needs to be done by the TEI community, particularly in the domain of the development of annotation and management tools as well interchange schemas that convert between the formats used commonly in LD tools such as EAF (ELAN) as well as LIFT and Flexfiles (FLEx).

As discussed in various points of this dissertation, there remains significant work to be done in order to carry out this project to a degree that it truly meets its potential in terms of producing a reusable, extensible and openly available user friendly output for the Mixtepec-Mixtec community, learners and non-technologists, namely:

- Transcription of several dozen remaining hours of recordings;
- Continue corpus annotation: apply all fundamental features described herein to all files;
- Creation of stable website with search interface for dictionary and corpus contents with multi-media capacity;
- Establish infrastructure for parallel display of digital editions of encoded historical Mixtec resources;
- Creation of additional schemas for re-formatting the annotated corpus documents into more user friendly documents, ideally moving the translation content into an annex which can be used as learner's reference;
- Obtain funding to engage/employ native speaker(s) as co-editor(s) of the dictionary and to assist with the transcription;
- Collaborate with computational phonologist to test and apply machine learning methods in transcription backlog and tone classification;
- Deposit TEI dictionary with Mesolex
- Build relationships, including data and analysis sharing with community organizations working to support Mixtec community and language in Mexico and the various diaspora communities;
- Produce more linguistic analyses and basic language descriptions based on data;

As part of this project, in order to convert the spoken language transcriptions from Praat to the common corpus structure in TEI XSLT conversions were created, which likely represent the first schema between Praat and TEI. This is just one of numerous steps needed in order to ensure the level of data interchange that is truly needed in both the fields of DH and LD. Though not specific to the MIX project, moving forwards, a more interoperable data ecosystem is needed in LD and DH is to ensure the compatibility of TEI with the most commonly used standards and data formats in the various levels of LD, namely:

- *Metadata*: IMDI, CMDI, OLAC
- *Spoken language transcription (including IGT)*: EAF, EXMARaLDA
- *Corpora (and other IGT)*: FLEx, Toolbox
- *Dictionaries and Lexica*: FLEx

The pre-existing online conversion tool OxGarage[276] which converts between TEI to and a number of different data formats would be an obvious potential candidate into which to add the additional conversion schemas.

Moving forwards, there is significant work to be done in creating a body of openly available, accessible and interoperable Mixtec resources both for the use in the context of work being done in Mixtepec Mixtec as well as for related varieties, including:

- Integrate all vocabulary from Vera Cruz Mixtec dictionary (Galindo Sánchez, 2009) into the MIX dictionary;
- Create digital editions of the Mixtec codices: ideally with descriptions in one or more varieties of Mixtec;
- Create TEI encoded documents of the data from seminal works in Proto-Mixtecan, and Proto-Mixtec-Amazugo specifically from:
    - Longacker (1957)
    - Josserand (1983)

[276] https://oxgarage.tei-c.org/#

- Dürr (1987)
- Mak and Longacre (1960)
- Longacre and Millon (1961)

While as discussed throughout this dissertation more work is needed in terms making the editing and searching of certain aspects (particularly standoff annotations in the corpus) of TEI data more accessible for non-experts (ideally community project members), the model used in the TEI digital dictionary used for MIX could easily be expanded in creating a pan-Mixtec digital corpus which would have immediate use in academic, government and community context.

9. Bibliography

Abiteboul, S., Buneman, P., & Suciu, D. (2014). *Data on the Web: From relations to semistructured data and XML*. Morgan Kaufmann.

Adams, O., Makarucha, A., Neubig, G., Bird, S., & Cohn, T. (2017). Cross- lingual word embeddings for low-resource language modeling. *15th Conference of the European Chapter of the Association for Computational Linguistics*, *1*, 937–947.

Adams, O., Cohn, T., Neubig, G., Cruz, H., Bird, S., & Michaud, A. (2018). *Evaluating phonemic transcription of low-resource tonal languages for language documentation*.

Allen, L., Lillehaugen, B. D., Broadwell, G. A., Oudijk, M. R., & Zarafonetis, M. (2016). *Ticha: The Story of an International, Community Engaged, Digital Humanities Project*. Keystone DH, Pittsburg. https://ticha.haverford.edu/static/zapotexts/ticha-slides.pdf

Alvarado, F. de. (1593). *Vocabulario en Lengua Mixteca*. Casa de Pedro Bailli.

Aristar-Dry, H., & Simons, G. (2006). Good, Better, and Best Practice. *Proceedings of the Deutsche Gesellschaft Fur Sprachwissenschaft*. Bielefeld. http://emeld.org/documents/Bielefeld-Dry-Simons.pdf

Arkhipov, A., & Thieberger, N. (2018). *Reflections on software and technology for language documentation* (pp. 76–85). University of Hawai'i Press. http://scholarspace.manoa.hawaii.edu/handle/10125/24821

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). *Dbpedia: A nucleus for a web of open data*. Springer. http://link.springer.com/chapter/10.1007/978-3-540-76298-0_52

Austin, P. K. (2006). Data and language documentation. *Essentials of Language Documentation*, *178*, 87–112.

Austin, P. K., & Grenoble, L. A. (2007). Current trends in language documentation. *Language Documentation and Description*, *4*, 12–25. London:SOAS.

Austin, P. K. (2013). Language documentation and meta-documentation. *Keeping Languages Alive: Documentation, Pedagogy and Revitalization*, 3–15.

Austin, P. K. (2016). Language documentation 20 years on. In L. Filipović & M. Pütz (Eds.), *Endangerment of languages across the planet* (pp. 147–170). John Benjamins.

Bański, P. (2010). Why TEI stand-off annotation doesn't quite work. *Balisage: The Markup Conference*. http://www.balisage.net/Proceedings/vol5/print/Banski01/BalisageVol5-Banski01.html

Bański, P., & Przepiórkowski, A. (2009). Stand-off TEI annotation: The case of the National Corpus of Polish. *Proceedings of the Third Linguistic Annotation Workshop*, 64–67. http://dl.acm.org/citation.cfm?id=1698392

Bański, P., Gaiffe, B., Lopez, P., Meoni, S., Romary, L., Schmidt, T., Stadler, P., & Witt, A. (2016). *Wake up, standOff!* TEI Conference and Members' Meeting, Vienna.

Bański, P., Bowers, J., & Erjavec, T. (2017). *TEI-Lex0 guidelines for the encoding of dictionary information on written and spoken forms*. 485 – 494. http://nbn-resolving.de/urn:nbn:de:bsz:mh39-64273

Barras, C., Geoffrois, E., Wu, Z., & Liberman, M. (2001). Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication*, *33*(1–2), 5–22.

Bateman, J., & Farrar, S. (2004a). Modelling models of robot navigation using formal spatial ontology. In *Spatial Cognition IV. Reasoning, Action, Interaction* (pp. 366–389). Springer. http://link.springer.com/chapter/10.1007/978-3-540-32255-9_21

Bateman, J., & Farrar, S. (2004b). Towards a generic foundation for spatial ontology. *Formal Ontology in Information Systems: Proceedings of the Third Conference (FOIS-2004)*, 237.

Bateman, J. A., Hois, J., Ross, R., & Tenbrink, T. (2010). A linguistic ontology of space for natural language processing. *Artificial Intelligence*, *174*(14), 1027–1071.

Bautista Martínez, J., & Hernández Velasco, O. (Eds.). (2019). *Na kutuˈva ko saˈan savi* (2nd ed.). Instituto Lingüístico de Verano, A. C. https://mexico.sil.org/resources/archives/82562

Bax, A., Bucholtz, M., Campbell, E. W., Fawcett, A., Mendoza, G., Peters, S., & Basurto, G. R. (2019). *MILPA: A Community-Centered Linguistic Collaboration Supporting Indigenous Oaxacan Languages in California*. 6th International Conference on Language Documentation and Conservation (ICLDC), Honolulu, HI. http://scholarspace.manoa.hawaii.edu/handle/10125/44768

Beal, H. (Ed.). (2018). *Bichos*. (ILV-México Ensayos Preliminares Electrónicos No. 24) Instituto Lingüístico de Verano, A.C. https://mexico.sil.org/es/resources/archives/72837

Beckmann, G., & Nieves, M. M. (Eds.). (2007). *Ña xiko xikui* (2nd ed.). Instituto Lingüístico de Verano, A.C. https://mexico.sil.org/resources/archives/13389

Beckmann, G., & Nieves, M. M. (Eds.). (2008a). *Kue Numeru mancha Iin Sientu* (2nd ed.). Instituto Lingüístico de Verano, A.C. https://mexico.sil.org/resources/archives/51077

Beckmann, G., & Nieves, M. M. (Eds.). (2008b). *Nakutuˈa ko kaˈhcha sava kue ñaˈa* (2nd ed.). Instituto Lingüístico de Verano, A.C. https://mexico.sil.org/resources/archives/55298

Beckmann, G., & Nieves, M. M. (Eds.). (2008c). *Ma sana ino sa'an ko* (2nd ed.). Instituto Lingüístico de Verano, A.C. https://mexico.sil.org/resources/archives/55535

Beckmann, G., & Nieves, M. M. (Eds.). (2011). *Tu'un Sa'an Savi* (1st ed.). Instituto Lingüístico de Verano, A.C. https://www.sil.org/system/files/reapdata/55/20/45/55204593282311818306317216520811313523/Alfabeto_del_mixteco_de_Mixtepec_mix.pdf

Beckmann, G., & Nieves, M. M. (Eds.). (2012). *Ntusu Sa'an Savi* (2nd ed.). Instituto Lingüístico de Verano, A.C. http://www.mexico.sil.org/resources/archives/51083

Beckmann, G. (2014). *Kue Kiti Ntava* (1st ed.). Instituto Lingüístico de Verano, A.C. http://www.mexico.sil.org/resources/archives/59860

Bickel, B., Comrie, B., & Haspelmath, M. (2008). *The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses*. Max Planck Institute for Evolutionary

Anthropology & Department of Linguistics of the University of Leipzig. https://www.eva.mpg.de/lingua/resources/glossing-rules.php

Bickford, A. J., & Marlett, S. A. (1988). The semantics and morphology of Mixtec mood and aspect. 32. https://doi.org/10.31356/silwp.vol32.01

Bird, S., & Liberman, M. (2001). A Formal Framework for Linguistic Annotation. *Speech Communication*, *33*(1-2), 23–60. https://doi.org/doi: 10.1016/S0167-6393(00)00068-6

Bird, S., & Simons, G. (2001). The OLAC metadata set and controlled vocabularies. *Proceedings of the ACL 2001 Workshop on Sharing Tools and Resources-Volume 15*, 7–18. http://dl.acm.org/citation.cfm?id=1118065

Bird, S., & Simons, G. (2003a). Extending Dublin Core metadata to support the description and discovery of language resources. *Computers and the Humanities*, *37*(4), 375–388.

Bird, S., & Simons, G. (2003b). Seven dimensions of portability for language documentation and description. *Language*, 557–582.

Black, A. H. (2009). Writing Linguistic Papers in the Third Wave. *SIL Forum for Language Fieldwork 2009-004*. SIL International.

Black, C. A., & Black, H. A. (2012). *Grammars for the people, by the people, made easier using PAWS and XlingPaper* (pp. 103–128). University of Hawai'i Press. http://scholarspace.manoa.hawaii.edu/handle/10125/4532

Blokland, R., Partanen, N., Rießler, M., & Wilbur, J. (2019). Using computational approaches to integrate endangered language legacy data into documentation corpora: Past experiences and challenges ahead. *Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 24.

Boersma, P. (2014). Acoustic analysis. In R. Podesva & D. Sharma (Eds.), *Research methods in linguistics* (p. 375). Cambridge University Press. doi:10.1017/CBO9781139013734

Boersma, P., & Weenik, D. (2020). *Praat: Doing phonetics by computer* (Version 6.1.14) [Computer software]. http://www.praat.org/

Bonial, C., Babko-Malaya, O., Choi, J. D., Hwang, J., & Palmer, M. (2010). Propbank annotation guidelines. *Center for Computational Language and Education Research Institute of Cognitive Science University of Colorado at Boulder*.

Boroditsky, L. (2000). Metaphoric structuring: Understanding time through spatial metaphors. *Cognition*, *75*(1), 1–28.

Bowers, J. (2015). Using the TEI P5 for Creating Mixtepec-Mixtec Lexical Resources. *2nd International Conference on Mesoamerican Linguistics*.

Bowers, J. (in press). *Pathways and Patterns of metaphor and metonymy in Mixtepec-Mixtec body-part terms: Vol. The Grammar of Body-Part Expressions in Amerindian Languages* (R. Zariquiey & P. Valenzuela, Eds.). Oxford University Press. https://hal.inria.fr/hal-02075731

Bowers, J., & Romary, L. (2016). Deep Encoding of Etymological Information in TEI. *Journal of the Text Encoding Initiative*, *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* (pp. 21–23). Doshisha University. https://hal.inria.fr/hal-01744813

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). https://hal.inria.fr/hal-01968871/document

Bowers, J., & Romary, L. (2018b). *Encoding Mixtepec-Mixtec Etymology in TEI*. TEI Conference and Members' Meeting, Tokyo, Japan. https://hal.inria.fr/hal-02003975

Bowers, J., & Romary, L. (2019). *TEI and the Mixtepec-Mixtec corpus: Data integration, annotation and normalization of heterogeneous data for an under- resourced language*. 6th International Conference on Language Documentation and Conservation (ICLDC), University of Hawai'i at Mānoa. https://hal.inria.fr/hal-02075475

Bowers, J., & Stöckle, P. (2018). TEI and Bavarian dialect resources in Austria: Updates from the DBÖ and WBÖ. In A. U. Frank, C. Ivanovic, F. Mambrini, M. Passarotti, & C. Sporleder (Eds.), *Second workshop on Corpus-Based Research in the Humanities (CRH-2)* (Vol. 1). Gerastree Proceedings, GTP 1.

Bowers, J., Khemakhem, M., & Romary, L. (2019). TEI Encoding of a Classical Mixtec Dictionary Using GROBID-Dictionaries. In I. Kosem, C. Tiberius, M. Jakubíček, J.

Kallas, S. Krek, & V. Baisa (Eds.), *ELex 2019 conference*. Lexical Computing CZ, s.r.o. https://hal.inria.fr/hal-02264033

Bowers, J., Stöckle, P., Breuer, L. M., & Breuer, H. C. (2019). *Native-TEI dialectal dictionary for Bavarian dialects in Austria: Data structure, software and workflow*. TEI Conference and Members' Meeting, Graz, Austria. https://doi.org/10.5281/zenodo.3452702

Bowers, J., Salazar, J., & Salazar, T. (2019). *Mixtepec Mixtec Language Resources* [Data set].https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/BF2VNK

Boynton, J., Moran, S., Aristar, A., & Aristar-Dry, H. (2006). E-MELD and the School of Best Practices: An ongoing community effort. *Sustainable Data from Digital Fieldwork. Proceedings of the Conference Held at the University of Sydney, 4-6 December 2006*.

Broeder, D., Schuurman, I., & Windhouwer, M. (2014). Experiences with the isocat data category registry. *LREC 2014: 9th International Conference on Language Resources and Evaluation*, 4565–4568.

Broadwell, G. A., García Guzmán, M., Lillehaugen, B. D., & Lopez, F. H. (in press). Ticha: Collaboration with indigenous communities to build digital resources on Zapotec language and history. *Digital Humanities Quarterly*.

Brugman, C. (1983). The use of body-part terms as locatives in Chalcatongo Mixtec. *Survey of California and Other Indian Languages*, *4*, 239–290.

Brugman, C., & Macaulay, M. (1986). Interacting semantic systems: Mixtec expressions of location. *Annual Meeting of the Berkeley Linguistics Society*, *12*, 315–327.

Brugman, H., & Russel, A. (2004). Annotating Multimedia/ Multi-modal resources with ELAN. *Proceedings of LREC 2004*.

Brugman, Hennie, & Wittenburg, P. (2001). The application of annotation models for the construction of databases and tools: Overview and analysis of MPI work since 1994. *IRCS Workshop on Linguistic Databases*.

Cameron, D., Frazer, E., Harvey, P., Rampton, M. B. H., & Richardson, K. (1992). *Researching language: Issues of power and method*. Routledge.

Campbell, E. W., Basurto, G. R., & Hernandéz Martinez, C. (In press). Language revitalization and academic institutions: Refocusing linguistic field methods courses. In J. Olko & J. Sallabank (Eds.), *Revitalizing endangered languages: A practical guide*. Cambridge University Press.

Cassidy, S., & Schmidt, T. (2017). Tools for multimodal annotation. In N. Ide & J. Pustejovsky (Eds.), *Handbook of linguistic annotation* (pp. 209–227). Springer.

Chiarcos, C., Dipper, S., Götze, M., Leser, U., Lüdeling, A., Ritz, J., & Stede, M. (2008). A flexible framework for integrating annotations from different tools and tagsets. *Traitement Automatique Des Langues*, *49*(2), 271–293.

Chiarcos, C. (2012). Interoperability of corpora and annotations. In *Linked Data in Linguistics* (pp. 161–179). Springer. http://link.springer.com/chapter/10.1007/978-3-642-28249-2_16

Chiarcos, C., & Sukhareva, M. (2015). Olia–ontologies of linguistic annotation. *Semantic Web*, *6*(4), 379–386.

Cimiano, P., Buitelaar, P., McCrae, J., & Sintek, M. (2011). LexInfo: A declarative model for the lexicon-ontology interface. *Journal of Web Semantics*, *9*(1), 29–51.

Clausner, T. C., & Croft, W. (1999). Domains and image schemas. *Cognitive Linguistics*, *10*, 1–32.

Corpuz Jr., L. (2012). *The RRG Linking Algorithm and Machine Translation As applied to the Yucannany dialect of Mixtec to the Californian dialect of English* [M.A.]. San José State.

Cox, C. (2011). Corpus linguistics and language documentation: Challenges for collaboration. In *Corpus-based studies in language use, language learning, and language documentation* (pp. 239–264). Brill Rodopi.

Czaykowska-Higgins, E. (2009). Research models, community engagement, and linguistic fieldwork: Reflections on working within Canadian indigenous communities. *Language Documentation & Conservation*, *3*(1), 182–215.

Czaykowska-Higgins, E., & Holmes, M. (2013). Technology in documentation: TEI and the Nxa'amxcín Dictionary. *3rd International Conference on Language Documentation and Conservation (ICLDC)*. http://scholarspace.manoa.hawaii.edu/handle/10125/26113

Czaykowska-Higgins, E., Holmes, M. D., & Kell, S. M. (2014). Using TEI for an Endangered Language Lexical Resource: The Nxaʔamxcín Database-Dictionary Project. *Language Documentation and Conservation*, *8*, 1–37. http://hdl.handle.net/10125/4604

Dipper, S. (2005). XML-based stand-off representation and exploitation of multi-level linguistic annotation. *Berliner XML Tage 2005 (BXML 2005)*, 39–50.

DuBois, J., Schuetze-Coburn, S., Cumming, S., & Paolino, D. (2003). Outline of Discourse Transcription. In J. Edwards & M. Lampert (Eds.), *Talking Data: Transcription and Coding in Discourse Research* (pp. 45–89). Erlbaum.

Dürr, M. (1987). A preliminary reconstruction of the Proto-Mixtec tonal system. *Indiana 11*, 19–61.

Dwyer, A. M. (2006). Ethics and practicalities of cooperative fieldwork and analysis. In J. Gippert, N. P. Himmelmann, & U. Mosel (Eds.), *Fundamentals of Language Documentation: A Handbook* (pp. 31–66). Berlin & New York: Mouton de Gruyter. http://hdl.handle.net/1808/7058

Ellioitt, T., Bodard, G., & Cayless, H. *et al.* (2006-2017). *EpiDoc: Epigraphic Documents in TEI XML*. http://epidoc.sf.net

Ehlich, K. (2003). HIAT: A Transcription System for Discourse Data. In J. Edwards & M. Lampert (Eds.), *In Talking Data: Transcription and Coding in Discourse Research* (pp. 123–148). Erlbaum.

Ehlich, K. and J. Rehbein. (1976). "Halbinterpretative Arbeitstranskriptionen (HIAT)." Linguistische Berichte 45: 21–41.

Evans, V., & Green, M. (2006). *Cognitive linguistics*. Edinburgh University Press.

Everett, D. (2005). Cultural constraints on grammar and cognition in Pirahã: Another look at the design features of human language. *Current Anthropology*, *46*(4), 621–646.

Farrar, S., & Langendoen, D. T. (2003). Markup and the GOLD ontology. *Proceedings of Workshop on Digitizing and Annotating Text and Field Recordings*.

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. MIT press.

Flanders, J., & Hamlin, S. (2013). TAPAS: Building a TEI publishing and repository service. *Journal of the Text Encoding Initiative*, *5*.

Forkel, R., List, J.-M., Greenhill, S. J., Rzymski, C., Bank, S., Cysouw, M., Hammarström, H., Haspelmath, M., Kaiping, G. A., & Gray, R. D. (2018). Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, *5*(1), 1–10. https://doi.org/10.1038/sdata.2018.205

Galindo Sánchez, B. (2009). *Vocabulario Básico Tu'un Savi-Castellano* (Primera edicíon). Academia Veracruzana de las Lenguas Indígenas.

Galla, C. (2009). Indigenous language revitalization and technology from traditional to contemporary domains. *Indigenous Language Revitalization: Encouragement, Guidance & Lessons Learned*, 167–182.

Gawne, L., & Berez-Kroeker, A. L. (2018). Reflections on reproducible research. In B. McDonnell, A. L. Berez-Kroeker, & G. Holton (Eds.), *Reflections on language documentation on the 20 year anniversary of Himmelmann 1998* (pp. 22–32). University of Hawai'i Press. http: //hdl.handle.net/10125/24805

Geeraerts, D. (2010). *Theories of lexical semantics*. Oxford University Press.

Gentner, D., Imai, M., & Boroditsky, L. (2002). As time goes by: Evidence for two systems in processing space→ time metaphors. *Language and Cognitive Processes*, *17*(5), 537–565.

George, M. (2013). LMF in U.S. Government Language Resource Management. In G. Francopoulo & P. Paroubek (Eds.), *LMF Lexical Markup Framework*. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118712696.ch17

Glenn, A. (2009). Five dimensions of collaboration: Toward a critical theory of coordination and interoperability in language documentation. *Language Documentation & Conservation*, *3*(2), 149–160.

Gómez Hernández, M. (2007a). *Chuun luu ka ña sachuun xeen ti* (G. Beckmann & M. M. Nieves, Eds.; 2nd ed.). Instituto Lingüístico de Verano, A.C https://mexico.sil.org/resources/archives/55474

Gómez Hernández, M. (2007b). *Nixi ntsikoo sa'ma yakui ka* (G. Beckmann & M. M. Nieves, Eds.; 2nd ed.). Instituto Lingüístico de Verano, A.C https://mexico.sil.org/resources/archives/13229

Gómez Hernández, M. (2008a). *Ntintsitsia ntivixi* (G. Beckmann & M. M. Nieves, Eds.; 2nd ed.). Instituto Lingüístico de Verano, A.C. https://mexico.sil.org/resources/archives/55533

Gómez Hernández, M. (2008b). *Ña niyu ʉ nchu ʼa* (G. Beckmann & M. M. Nieves, Eds.; 2nd ed.). Instituto Lingüístico de Verano, A.C https://mexico.sil.org/resources/archives/55300

Gómez Hernández, M. (2008c). *Sa'va tsi Lochi* (M. M. Nieves & G. Beckmann, Eds.; 2nd ed.). Instituto Lingüístico de Verano, A.C. https://mexico.sil.org/resources/archives/55437

Gómez Hernández, M. (2008d). *Staa Nti'i* (M. M. Nieves & G. Beckmann, Eds.; 2nd ed.). Instituto Lingüístico de Verano, A.C. http://www.sil.org/mexico/mixteca/mixtepec/L099-TortillaGorda-mix.htm

Good, J. (2011). *Data and language documentation*. Cambridge University Press.

Gracia, J., Vila-Suero, D., McCrae, J. P., Flati, T., Baron, C., & Dojchinovski, M. (2014). Language Resources and Linked Data: A Practical Perspective. In *Knowledge Engineering and Knowledge Management* (pp. 3–17). Springer. http://link.springer.com/chapter/10.1007/978-3-319-17966-7_1

Grenoble, L. A., & Whaley, L. J. (2006). *Saving languages: An introduction to language revitalization*. Cambridge University Press.

Gries, S. T., & Berez, A. L. (2017). Linguistic annotation in/for corpus linguistics. In *Handbook of linguistic annotation* (pp. 379–409). Springer.

Grondelaers, S., Geeraerts, D., & Speelman, D. (2007). A case for a cognitive corpus linguistics. *Methods in Cognitive Linguistics*, *18*, 149.

Hale, K. (2001). Ulwa (Southern Sumu): The beginnings of a language research project. In *Linguistics Fieldwork* (pp. 76–101). Cambridge University Press.

Harrison, K. D., & Anderson, G. D. S. (2006). *Tuvan Talking Dictionary*. http://tuvan.swarthmore.edu/

Harrison, K. D., Lillehaugen, B. D., Fahringer, J., & Lopez, F. H. (2019). Zapotec Language Activism and Talking Dictionaries. *Electronic Lexicography in the 21st Century (ELex 2019): Smart Lexicography*, 96.

Heine, B., Claudi, U., & Hünnemeyer, F. (1991). *Grammaticalization: A Conceptual Framework*. Chicago University Press.

Hernández, F. B. (1567). *Doctrina Christiana en Lengua Mixteca* (in the Tlaxiaco-Achiutla variant). Casa de Pedro Ocharte.

Hernández Martínez, C., Campbell, E. W., & Reyes Basurto, G. (In press). MILPA (Mexican Indigenous Language Promotion and Advocacy): A community-centered linguistic collaboration supporting Indigenous Mexican languages in California. In J. Olko & J. Sallabank (Eds.), *Revitalizing endangered languages: A practical guide*. Cambridge University Press.

Hills, R. A. (1990). A Syntactic Sketch of Ayutla Mixtec (C. H. Bradley & B. E. Hollenbach, Eds.; Vol. 2, pp. 1–206). Summer Institute of Linguistics and the University of Texas Arlington.

Himmelmann, N. P. (1998). *Documentary and descriptive linguistics*. Walter de Gruyter, Berlin/New York.

Himmelmann, N. P. (2006). Language documentation: What is it and what is it good for. In G. Jost, N. P. Himmelmann, & U. Mosel (Eds.), *Essentials of language documentation* (pp. 1–30). Mouton de Gruyter.

Himmelmann, N. P. (2012). Linguistic data types and the interface between language documentation and description. *Language Documentation & Conservation*, *6*, 187–207.

Himmelmann, N. P. (2018). Meeting the transcription challenge. In B. McDonnell, A. L. Berez-Kroeker, & G. Holton (Eds.), *Reflections on Language Documentation 20 Years after Himmelmann 1998. Language Documentation & Conservation* (pp. 33–40). University of Hawai'i Press.

Hinton, L., & Hale, K. (2001). *The green book of language revitalization in practice*. Brill.

Hollenbach, B. E. (1995a). Semantic and Syntactic Extensions of Body-Part Terms in Mixtecan: The Case of "Face" and "Foot." *International Journal of American Linguistics*, 168–190.

Hollenbach, B. E. (1995b). Cuatro morfemas funcionales en las lenguas mixtecanas. *Vitalidad e Influencia de Las Lenguas Indígenas En Latinoamérica: II Coloquio Mauricio Swadesh*, 284–293.

Hollenbach, B. E. (2016). *Notes on Mixtec Terms for Supernatural Beings*. http://barbaraelenahollenbach.com/PDFs/MxTrmSup.pdf

Ide, N., Kilgarriff, A., & Romary, L. (2000). A formal model of dictionary structure and content. *Euralex*, 113–126. http://arxiv.org/abs/0707.3270

Ide, N., & Romary, L. (2004a). International standard for a linguistic annotation framework. *Natural Language Engineering*, *10*(3–4), 211–225.

Ide, N., & Romary, L. (2001). Standards for language resources. In S. Bird, P. Buneman, & M. Liberman (Eds.), *IRCS Workshop on Linguistic Databases*. https://hal.inria.fr/inria-00100589

Ide, N., & Romary, L. (2004). A registry of standard data categories for linguistic annotation. *4th International Conference on Language Resources and Evaluation-LREC'04*, 135–138. http://hal.archives-ouvertes.fr/inria-00099858/

IMDI. (2001). *Mapping IMDI Session Descriptions with OLAC.* Version 1.04. Max-Planck-Institute for Psycholinguistics Nijmegen. https://tla.mpi.nl/wp-content/uploads/2012/06/IMDI-to-OLAC-Mapping-1.04.pdf

IMDI. (2003). *Metadata Elements for Session Descriptions.* Version 3.0.4. Max-Planck-Institute for Psycholinguistics Nijmegen. https://tla.mpi.nl/wp-content/uploads/2012/06/IMDI_MetaData_3.0.4.pdf

IMDI. (2009). *Metadata Elements for Catalogue Descriptions.* Version 3.0.13. Max-Planck-Institute for Psycholinguistics Nijmegen. https://archive.mpi.nl/forums/uploads/short-url/jQsltlORAACT3EOvepd1AoQ5AGu.pdf

Instituto Nacional de Lenguas Indígenas. (2008). *Catálogo de las Lenguas Indígenas Nacionales: Variantes Lingüísticas de México con sus autodenominaciones y referencias geoestadísticas.* (R.-261246). https://www.inali.gob.mx/pdf/CLIN_completo.pdf

ISO 24611:2012 *Language resource management — Morpho-syntactic annotation framework (MAF)*. International Organization for Standardization. https://www.iso.org/standard/51934.html

ISO 24612:2012 *Language resource management – Linguistic Annotation Framework (LAF)*. International Organization for Standardization. https://www.iso.org/standard/37326.html

ISO 24624:2016 *Language resource management —Transcription of spoken language*. https://www.iso.org/fr/standard/37338.html

Jackendoff, R. (1976). Toward an explanatory semantic representation. *Linguistic Inquiry*, *7*(1), 89–150.

Jackendoff, R. (1987). The Status of Thematic Relations in Linguistic Theory. *Linguistic Inquiry*, *18*(3), 369–411.

Jansen, M. (1990). The search for history in Mixtec codices. *Ancient Mesoamerica*, *1*(1), 99–112.

Jansen, M., & Pérez Jiménez, G. A. (2004). Renaming the Mexican Codices. *Ancient Mesoamerica*, *15*(2), 267–271.

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*.

Jansen, M. E., & Pérez Jiménez, G. A. (2018). Reading Mixtec Manuscripts as Ceremonial Discourse: Historical and Ideological Background of Codex Añute (Selden). In *Mesoamerican Manuscripts* (pp. 416–459). Brill.

Johnson, L. M., Di Paolo, M., & Bell, A. (2018). *Forced Alignment for Understudied Language Varieties: Testing Prosodylab-Aligner with Tongan Data*. 194–203.

Johnson, M. (1987). *The body in the mind: The bodily basis of meaning, imagination, and reason*. University of Chicago Press.

Johnson, A. F. (1988). A syntactic sketch of Jamiltepec Mixtec. In C. H. Bradley & B. E. Hollenbach (Eds.), Studies in the syntax of Mixtecan languages (Vol. 1, pp. 11–150). Summer Institute of Linguistics and the University of Texas Arlington.

Josserand, J. K. (1983). *Mixtec Dialect History*. (Doctoral Dissertation, Tulane University).

Kemps-Snijders, M., Windhouwer, M. E., Wittenburg, P., & Wright, S. E. (2008). ISOcat: A revised ISO TC 37 data category registry. *Presentation at the Conference on Terminology and Information Interoperability–Management of Knowledge and Content (TII 2008), Moscow*.

Kemps-Snijders, Marc, Windhouwer, M., Wittenburg, P., & Wright, S. E. (2009). ISOcat: Remodeling metadata for language resources. *International Journal of Metadata, Semantics and Ontologies (IJMSO)*, *4*(4), 261–276.

Kemps-Snijders, M., Windhouwer, M., Wittenburg, P., & Wright, S. E. (2008). ISOcat: Corralling data categories in the wild. *6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Khemakhem, M., Foppiano, L., & Romary, L. (2017). Automatic extraction of TEI structures in digitized lexical resources using conditional random fields. *Electronic Lexicography, ELex 2017*.

King, G. (2007). An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. *Sociological Methods & Research*, *36*(2), 173–199. https://doi.org/10.1177/0049124107306660

Kipp, M. (2001). Anvil-a generic annotation tool for multimodal dialogue. *Seventh European Conference on Speech Communication and Technology*.

Kordjamshidi, P., van Otterlo, M., & Moens, M.-F. (2017). Spatial role labeling annotation scheme. In *Handbook of Linguistic Annotation* (pp. 1025–1052). Springer.

Kovecses, Z. (2010). *Metaphor: A practical introduction*. Oxford University Press.

Kuiper, Albertha., & Merrifield, William. R. (1975). Diuxi Mixtec Verbs of Motion and Arrival. International Journal of American Linguistics, 41(1), 32–45.

Kuiper, A., & Oram, J. (1991). A syntactic sketch of Diuxi-Tilantongo Mixtec. In C. H. Bradley & B. E. Hollenbach (Eds.), *Studies in the syntax of Mixtecan languages* (Vol. 3, pp. 179–408). Summer Institute of Linguistics and the University of Texas Arlington.

Ladefoged, P. (1996). *Elements of acoustic phonetics*. University of Chicago Press.

Ladefoged, P., & Maddieson, I. (1996). *The Sounds of the World's Languages*. Wiley.

Lakoff, G., & Johnson, M. (1980a). *Metaphors we live by*. University of Chicago press.

Lakoff, G., & Johnson, M. (1980b). The metaphorical structure of the human conceptual system. *Cognitive Science*, *4*(2), 195–208.

Langacker, R. W. (1986). An introduction to cognitive grammar. *Cognitive Science*, *10*(1), 1–40.

Langacker, R. W. (1987). *Cognitive grammar: A basic introduction*. Oxford University Press.

Langacker, R. W. (2002). A study in unified diversity: English and Mixtec locatives. *Ethnosyntax: Explorations in Grammar and Culture*, 138–161.

Langacker, R. W. (2010). Conceptualization, symbolization, and grammar. *International Journal of Cognitive Linguistics*, *1*(1), 31–63.

Langendoen, D. T., & Simons, G. F. (1995). A rationale for the TEI recommendations for feature-structure markup. *Computers and the Humanities*, *29*(3), 191–209.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., & others. (2014). DBpedia–A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*.

Lehmberg, T., & Kai, W. (2008). Annotation Standards. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics* (Vol. 1). Walter de Gruyter.

Lillehaugen, B. D., Broadwell, G. A., Oudijk, M. R., Allen, L., Zarafonetis, M., & Helen Plumb, M. (2016). *Humanidades digitales multilingual y multicultural: El caso de Ticha, un explorador digital de texto para el zapoteco colonial*. 3er Encuentro de Humanistas Digitales, Mexico City. http://bit.ly/3EHDTicha

Longacre, R. E. (1957). Proto-Mixtecan. *International Journal of American Linguistics*, *23*(4).

Longacre, R. E., & Millon, R. (1961). Proto-Mixtecan and Proto-Amuzgo-Mixtecan vocabularies: A preliminary cultural analysis. *Anthropological Linguistics*, 1–44.

Loper, E., & Bird, S. (2002). NLTK: The natural language toolkit. *ArXiv Preprint Cs/0205028*.

López Santiago, S. (2008). *Tuykuku Ka'nu* (M. M. Nieves & G. Beckmann, Eds.; 2nd ed.). Instituto Lingüístico de Verano, A.C. https://mexico.sil.org/resources/archives/55906

Macaulay, M. (1982). Verbs of motion and arrival in Mixtec. *Annual Meeting of the Berkeley Linguistics Society*, *8*, 414–426.

Macaulay, M. (1985). The semantics of "come", "go", and "arrive" in Otomanguean languages. *Kansas Working Papers in Linguistics*, *10*. https://doi.org/10.17161/KWPL.1808.503

Macaulay, M. (1987a). Cliticization and the morphosyntax of Mixtec. *International Journal of American Linguistics*, *53*(2), 119–135.

Macaulay, M. (1987b). *Morphology and cliticization in Chalcatongo Mixtec* [Doctoral Disseration, University of California, Berkeley]. http://escholarship.org/uc/item/5qb9x714.pdf

Macaulay, M. (1990). Negation and mood in Mixtec: Evidence from Chalcatongo. *Anthropological Linguistics*, 211–227.

Macaulay, M. A. (1996). *A grammar of Chalcatongo Mixtec.* University of California Publications in Linguistics (Vol. 127). University of California Press.

Macaulay, M. (2005). The Syntax of Chalcatongo Mixtec. *Verb First: On the Syntax of Verb-Initial Languages*, *73*, 341.

Macaulay, M. (1993). Argument status and constituent structure in Chalcatongo Mixtec. *Annual Meeting of the Berkeley Linguistics Society*, *19*, 73–85.

Macaulay, M. (2011). Verbs of motion and arrival in Mixtec. *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, *8*. http://elanguage.net/journals/bls/article/download/2284/2246

Macaulay, M. (2012). Argument status and constituent structure in Chalcatongo Mixtec. *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, *19*. http://elanguage.net/journals/bls/article/download/3011/2946

Macaulay, M., & Salmons, J. C. (1995). The phonology of glottalization in Mixtec. *International Journal of American Linguistics*, *61*(1), 38–61.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. transcription format and programs* (Vol. 1). Psychology Press.

Mak, C. (1953). A comparison of two Mixtec tonemic systems. *International Journal of American Linguistics*, *19*(2), 85–100.

Mak, C. (1958). The tonal system of a third Mixtec dialect. *International Journal of American Linguistics*, *24*(1), 61–70.

Mak, C., & Longacre, R. (1960). Proto-Mixtec phonology. *International Journal of American Linguistics*, *26*(1), 23–40.

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*, 85–91.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., & Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. *Proceedings of ELex 2017 Conference, September*, 19–21.

McCrae, J. P., Tiberius, C., Khan, A. F., Kernerman, I., Declerck, T., Krek, S., Monachini, M., & Ahmadi, S. (2019). The ELEXIS interface for interoperable lexical resources. *Proceedings of the Sixth Biennial Conference on Electronic Lexicography (ELex)*.

McKelvie, D., Brew, C., & Thompson, H. S. (1997). Using SGML as a basis for data-intensive natural language processing. *Computers and the Humanities*, *31*(5), 367–388.

McKendry, I. (2013). *Tonal association, prominence and prosodic structure in South-Eastern Nochixtlán Mixtec* [Doctoral Dissertation, University of Edinburgh]. https://era.ed.ac.uk/bitstream/handle/1842/33079/McKendry_dissertation_2013_UnivEdinburgh_for_web.pdf?sequence=1

Meakins, F. (2007). Review of Computerized Language Analysis (CLAN). *Language Documentation & Conservation*, *1*(1), 107–111.

Mendoza Santiago, F. (2008). *Tu'un yata tsa'a kue kaa kaxi Xnuviko* (G. Beckmann & M. M. Nieves, Eds.; 2nd ed.). Instituto Lingüístico de Verano, A.C. http://www.mexico.sil.org/resources/archives/51095

Mendoza Santiago, F. (2009). *Tuyuku Xnuviko* (M. M. Nieves & G. Beckmann, Eds.; 2nd ed.). Instituto Lingüístico de Verano, A.C. http://mexico.sil.org/resources/archives/55901

Michaud, A., Adams, O., Cohn, T. A., Neubig, G., & Guillaume, S. (2018). Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone toolkit. *Language Documentation & Conservation*, *12*, 393–429.

Michaud, A., Adams, O., Cox, C., Guillaume, S., Wisniewski, G., & Gaillot. (2020). *La transcription du linguiste au miroir de l'intelligence artificielle: Réflexions à partir de la transcription phonémique automatique*. https://halshs.archives-ouvertes.fr/LACITO/halshs-02881731

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Moe, R. (2003). Compiling dictionaries using semantic domains. *Lexikos*, *13*(1), 215–223.

Moeller, S., Kazeminejad, G., Cowell, A., & Hulden, M. (2019). Improving low-resource morphological learning with intermediate forms from finite state transducers.

Proceedings of the Workshop on Computational Methods for Endangered Languages, *1*(1), 11.

Moeller, S. R. (2014). Review of SayMore, a tool for Language Documentation Productivity. *Language Documentation & Conservation*, *8*, 66–74.

Mosel, U. (in press). Corpus building for under-researched language- a practical guide. In F. Ahoua, D. Gibbon, & S. Skopeteas (Eds.), *Linguistic Fieldwork and Language Documentation A Course Book on Foundational Skills*.

Mous, M. (2007). Language documentation as a challenge to description. *Talk Presented at Annual Conference on African Linguistics*, *38*.

Naidu, V., Zlatev, J., Duggirala, V., Van De Weijer, J., Devylder, S., & Blomberg, J. (2018). Holistic spatial semantics and post-Talmian motion event typology: A case study of Thai and Telugu. *Cognitive Semiotics*, *11*(2).

Nakhimovsky, A., Good, J., & Myers, T. (2012). Interoperability of Language Documentation Tools and Materials for Local Communities. *DH*, 280–282.

Nathan, D., & Austin, P. K. (2004). Reconceiving metadata: Language documentation through thick and thin. *Language Documentation and Description*, *2*, 179–187.

Neubig, G., Rijhwani, S., Palmer, A., MacKenzie, J., Cruz, H., Li, X., Lee, M., Chaudhary, A., Gessler, L., & Abney, S. (2020). A summary of the first Workshop on Language Technology for Language Documentation and Revitalization. *1st Joint SLTU (Spoken Language Technologies for Under-Resourced Languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) Workshop*.

Nieves, M. M., & Beckmann, G. (Eds.). (2007a). *Kunchau hora ka* (2nd ed.). Instituto Lingüístico de Verano, A.C. https://mexico.sil.org/resources/archives/63381

Nieves, M. M., & Beckmann, G. (Eds.). (2007b). *Kunka'vi hora ka* (2nd ed.). Instituto Lingüístico de Verano, A.C. https://mexico.sil.org/resources/archives/55956

Nathan, D. (2010). Sound and unsound practices in documentary linguistics: Towards an epistemology for audio. *Language Documentation and Description*, *7*(1), 1–17.

Nieves, M. M. (2012). El parangón en San Juan Mixtepec, Juxtlahuaca: El día de la boda y el día del cambio de autoridades. *Ponencia presentada en el 5to Coloquio de Lenguas Otomangues y Vecinas, Antonio de los Reyes el 22 de abril de 2012. Centro Cultural San Pablo, Oaxaca*, *13*, 16. https://mexico.sil.org/resources/archives/52395

Ogrodniczuk, M. (2011). *The Packaged TEI P5-based Stand-off Annotation Format⋆.* http://nlp.ipipan.waw.pl/Bib/ogr:11:format.pdf

Ohala, J. J., & Ewan, W. G. (1973). Speed of pitch change. *The Journal of the Acoustical Society of America*, *53*(1), 345–345.

Ohala, J. J. (1978). Production of tone. In V. A. Fromkin (Ed.), *Tone: A linguistic survey* (pp. 5–39). Elsevier.

Ostler, N. (2008). Corpora of less studied languages. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: Vol. 29.1* (pp. 457–483). Walter de Gruyter.

Paster, M. (2005). Tone Rules in Yucanani Mixtepec Mixtec. *SSILA Meeting*, 13.

Paster, M. (2010). The role of homophony avoidance in morphology: A case study from Mixtec. In D. Rosenlum & S. Gamble Morse (Eds.), *13th Annual Workshop on American Indian Languages* (Vol. 21, pp. 29–39).

Paster, M., & Beam de Azcona, R. (2004). *Aspects of tone in Yucunany Mixtepec Mixtec*. Conference on Otomanguean and Oaxacan Languages.

Paster, M., & Beam de Azcona, R. (2005). A Phonological Sketch of the Yucunany Dialect of Mixtepec Mixtec. In L. Harper & C. Jany (Eds.), *The 7the Annual Workshop on American Indian Languages* (pp. 61–76). UC Santa Barbara.

Penfield, S. D. (2018). Interdisciplinary research in language documentation. In B. McDonnell, A. L. Berez, & G. Holton (Eds.), *Reflections on Language Documentation 20 Years after Himmelmann 1998* (p. 76). http://hdl.handle.net/10125/24810

Pennington, R. (2014). *Producing time-aligned interlinear texts: Towards a SayMore–FLEx–ELAN workflow*. SIL International.

Perlin, R. (2012). Review of WeSay, A Tool for Collaborating on Dictionaries with Non-Linguists. *Language Documentation & Conservation*, *6*, 181–186.

Pike, E. V., & Ibach, T. (1978). The phonology of the Mixtepec dialect of Mixtec. In M. Jazayery, E. C. Polomé, & W. Winter (Eds.), *Linguistic and Literary Studies in Honor of Archibald A. Hill* (Descriptive Linguistics, Vol. 2, pp. 271–285). Mouton.

Przepiórkowski, A., & Bański, P. (2009). XML text interchange format in the National Corpus of Polish. *The Proceedings of Practical Applications in Language and Computers PALC 2009, Frankfurt Am Main: Peter Lang*.

Pustejovsky, J. (2017). ISO-Space: Annotating static and dynamic spatial information. In *Handbook of Linguistic Annotation* (pp. 989–1024). Springer.

Pustejovsky, J., & Moszkowicz, J. (2008). Integrating motion predicate classes with spatial and temporal annotations. *Coling 2008: Companion Volume: Posters*, 95–98.

Ramos Hernández, H. (2007). *To'lo ña ma nikuii ka kana* (M. M. Nieves & G. Beckmann, Eds.; 2nd ed.). Instituto Lingüístico de Verano, A.C. https://mexico.sil.org/resources/archives/55469

Ramos López, M., López Hernández, W., & Mendoza Santiago, F. (2008). *Tutu ña sna'a nixi skua'a ñayivi* (M. M. Nieves & G. Beckmann, Eds.; 2nd ed.). Instituto Lingüístico de Verano, A.C. https://mexico.sil.org/resources/archives/55299

Reyes Basurto, G., Hernández Martínez, C., & Campbell, E. W. (In press). What is community? Perspectives from the Mixtec diaspora in California. In J. Olko & J. Sallabank (Eds.), *Revitalizing endangered languages: A practical guide*. Cambridge University Press.

Reyes, A. de los. (1593). *Arte en lengua mixteca conpuesta*. Casa de Pedro Balli.

Rice, K. (2006). Ethical issues in linguistic fieldwork: An overview. *Journal of Academic Ethics*, *4*(1–4), 123–155.

Ringersma, J., Drude, S., & Kemps-Snijders, M. (2010). Lexicon standards: From de facto standard Toolbox MDF to ISO standard LMF. *LRT Standards Workshop, Seventh Conference on International Language Resources and Evaluation [LREC 2010]*.

Ringersma, J., & Kemps-Snijders, M. (2007). Creating multimedia dictionaries of endangered languages using LEXUS. *Interspeech 2007: 8th Annual Conference on the International Speech Communication Association*, 65–68.

Ripalda, P. G. de. (1755). *De la sagrada compañia de Jesús: Vol. Cathecismo y explicación de la doctrina christiana Traducido en lengua mixteca por el M.R.D.* (M. R. D. Fr. A. Gonzales, Trans.; Reprint of a 1719 publication).

Rojas Santiago, M., Rojas Santiago, R., Bautista Marroquín, J., Ramos Hernández, D., Santiago Velasco, G., Santiago Gómez, J., & Hernandéz López, G. (2014). *Nuu ntakitaꞌan kue tuꞌun* (Primera edición). Instituto Lingüístico de Verano, A.C. http://www.mexico.sil.org/resources/archives/59743

Romary, L. (2011). Stabilizing knowledge through standards-A perspective for the humanities. In K. Grandin (Ed.), *Going Digital: Evolutionary and Revolutionary Aspects of Digitization*. Science History Publications. https://hal.inria.fr/inria-00531019

Romary, L. (2015a). TEI and LMF crosswalks. *Journal for Language Technology and Computational Linguistics*, GSCL (Gesellschaft für Sprachtechnologie und Computerlinguistik) *30*(1). https://hal.inria.fr/hal-00762664v4

Romary, L. (2015b). Standards for language resources in ISO–Looking back at 13 fruitful years. *Edition - Die Fachzeitschrift Für Terminologie, Deutscher Terminologie-Tag e.V. (DTT)*. https://hal.inria.fr/hal-01220925

Romary, L., Khemakhem, M., Khan, F., Bowers, J., Calzolari, N., George, M., Pet, M., & Bański, P. (2019). LMF Reloaded. In M. Gürlek, A. Naim Çiçekler, & Y. Taşdemir (Eds.), *Proceedings of the 13th International Conference of the Asian Association for Lexicography* (pp. 533–539). ASOS. https://www.asialex.org/pdf/Asialex-Proceedings-2019.pdf

Romary, L., & Tasovac, T. (2018). TEI Lex-0: A Target Format for TEI-Encoded Dictionaries and Lexical Resources. *TEI Conference and Members' Meeting*.

Romary, L., & Wegstein, W. (2012). Consistent modelling of heterogeneous lexical structures. *Journal of the Text Encoding Initiative*, *3*. https://doi.org/10.4000/jtei.540

Salazar, J., Alfaife, S., Belmar Viernes, G., Campbell, E. W., Mendoza, G., Olguín Martínez, J., Reyes Basurto, G., Scanlon, C., Troiani, G., & Vásquez Aguilar, A. (2020). *Yucunani Mixtepec Mixtec: Collaborative Grammar Sketch*. [manuscript]. University of California, Santa Barbara.

Salmon-Alt, S., & Romary, L. (2009). Reference resolution within the framework of cognitive grammar. *International Colloqium on Cognitive Science*, 16. https://hal.inria.fr/inria-00100430

Schmidt, T. (2011). A TEI-based Approach to Standardising Spoken Language Transcription. *Journal of the Text Encoding Initiative*, *Issue 1*. https://doi.org/10.4000/jtei.142

Schmidt, T., Duncan, S., Ehmer, O., Hoyt, J., Kipp, M., Loehr, D., Magnusson, M., Rose, T., & Sloetjes, H. (2008). An exchange format for multimodal annotations. *International LREC Workshop on Multimodal Corpora*, 207–221.

Schmidt, T., & Wörner, K. (2009). EXMARaLDA–Creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics*, *19*(4), 565–582.

Schwartz, L., Chen, E., Schreiner, S., & Hunt, B. (2019). *Bootstrapping a Neural Morphological Analyzer for St. Lawrence Island Yupik Nouns from a Finite-State Transducer*.

Seifart, F., Evans, N., Hammarström, H., & Levinson, S. C. (2018). Language documentation twenty-five years on. *Language*, *94*(4), e324–e345.

Selting, M., Auer, P., Barth-Weingarten, D., Bergmann, J. R., Bergmann, P., Birkner, K., Couper-Kuhlen, E., Deppermann, A., Gilles, P., & Günthner, S. (2009). Gesprächsanalytisches transkriptionssystem 2 (GAT 2). *Gesprächsforschung: Online-Zeitschrift Zur Verbalen Interaktion*. http:// www.gespraechsforschung-ozs.de/heft2009/px-gat2.pdf

Shutova, E. (2017). Annotation of linguistic and conceptual metaphor. In *Handbook of linguistic annotation* (pp. 1073–1100). Springer.

Silverman, D. (2003). Phonetics and function in diachronic conflict: The case of rising tones. *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, *39*(1), 690–701.

Simons, G., & Bird, S. (2003a). Building an open language archives community on the OAI foundation. *Library Hi Tech*, *21*(2), 210–218.

Simons, G., & Bird, S. (2003b). The Open Language Archives Community: An infrastructure for distributed archiving of language resources. *Literary and Linguistic Computing*, *18*(2), 117–128.

Simons, G. F., & Bird, S. (2008). *Toward a global infrastructure for the sustainability of language resources*. https://minerva-access.unimelb.edu.au/handle/11343/25076

Simons, G. F., & Black, H. A. (2009). Third wave writing and publishing. *SIL Forum for Language Fieldwork*, 2009-005. SIL International.

Simons, G. F., & Fenning, C. D. (Eds.). (2018). *Ethnologue: Languages of the World* (21st ed.). SIL International. https://www.ethnologue.com/subgroups/mixtec

Sloetjes, H., Somasundaram, A., & Wittenburg, P. (2011). ELAN—Aspects of Interoperability and Functionality. *Conference of the International Speech Communication Association (Interspeech 2011)*, 3249–3252.

Spores, R., & Balkansky, A. K. (2013). *The Mixtecs of Oaxaca: Ancient Times to the Present*. University of Oklahoma Press.

Strunk, J., Schiel, F., & Seifart, F. (2014). Untrained Forced Alignment of Transcriptions and Audio for Language Documentation Corpora using WebMAUS. *LREC*, 3940–3947.

Stührenberg, M. (2012). The TEI and Current Standards for Structuring Linguistic Data. *Journal of the Text Encoding Initiative, Issue 3*. https://doi.org/10.4000/jtei.523

Svorou, S. (1994). *The Grammar of Space*. John Benjamins Publishing.

Talmy, L. (1983). *How language structures space*. Springer. http://link.springer.com/chapter/10.1007/978-1-4615-9325-6_11

Thieberger, N. (2010). *Anxious Respect for Linguistic Data: The Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) and the Resource Network for Linguistic Diversity (RNLD)* (M. Florey, Ed.; pp. 151–158). Oxford University Press. http://hdl.handle.net/11343/31158

Thieberger, N. (2012). Using language documentation data in a broader context (F. Seifart, G. Haig, N. P. Himmelmann, D. Jung, A. Margetts, & P. Trilsbeek, Eds.; pp. 129–134). University of Hawai'i Press. https://scholarspace.manoa.hawaii.edu/handle/10125/4527

Thieberger, N. (2014). Digital humanities and language documentation. In L. Gawne & J. Vaughan (Eds.), *Selected papers from the 44th conference of the Australian Linguistic Society, 2013* (pp. 144–159). University of Melbourne. http://hdl.handle.net/11343/40961

Thieberger, Nick. (2016). Documentary Linguistics: Methodological Challenges and Innovatory Responses. *Applied Linguistics*, *37*(1), 88–99.

Thieberger, Nick, Margetts, A., Morey, S., & Musgrave, S. (2016). Assessing annotated corpora as research output. *Australian Journal of Linguistics*, *36*(1), 1–21.

Tsunoda, T. (2013). *Language endangerment and language revitalization: An introduction* (Vol. 148). Walter de Gruyter.

Tyers, F. M., Washington, J. N., Kavitskaya, D., Gökırmak, M., Howell, N., & Berberova, R. (2019). A biscriptual morphological transducer for Crimean Tatar. *Proceedings of the Workshop on Computational Methods for Endangered Languages*, *1*(1), 10.

Unsworth, J. (2011). Computational Work with Very Large Text Collections. Interoperability, Sustainability, and the TEI. *Journal of the Text Encoding Initiative, Issue 1*, Article Issue 1. https://doi.org/10.4000/jtei.215

Van Uytvanck, D., Stehouwer, H., & Lampen, L. (2012). Semantic metadata mapping in practice: The Virtual Language Observatory. *LREC 2012: 8th International Conference on Language Resources and Evaluation*, 1029–1034. http://pubman.mpdl.mpg.de/pubman/faces/viewItemOverviewPage.jsp?itemId=escidoc:1454694

Van Valin, R. D. (2005). *Exploring the syntax-semantics interface*. Cambridge University Press.

Van Valin, R. D., & Foley, W. A. (1980). Role and reference grammar. In *Current approaches to syntax* (pp. 329–352). Brill.

Windhouwer, M., Kemps-Snijders, M., & Wright, S. E. (2010). Referencing ISOcat data categories. *LREC10-W4*, 27.

Windhouwer, M., & Wright, S. E. (2012). Linking to linguistic data categories in ISOcat. In *Linked Data in Linguistics* (pp. 99–107). Springer.

Withers, P. (2012). Metadata management with Arbil. *LREC 2012: 8th International Conference on Language Resources and Evaluation*, 72–75.

Woodbury, A. (2003). *Defining Language documentation* (P. K. Austin, Ed.; Vol. 1). SOAS.

Woodbury, A. C. (2011). Language documentation. In P. K. Austin & J. Sallabank (Eds.), *The Cambridge Handbook of Endangered Languages* (pp. 159–186). Cambridge University Press.

Woodbury, A. C. (2014). Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire. *Language Documentation and Description*, *12*, 19–36.

Wörner, K. (2009). *Werkzeuge zur flachen Annotation von Transkriptionen gesprochener Sprache* [Universität Bielefeld]. http://nbn-resolving.de/urn:nbn:de:hbz:361-16696

Wright, S. E., Windhouwer, M., Schuurman, I., & Kemps-Snijders, M. (2013). Community efforts around the ISOcat Data Category Registry. In *The People's Web Meets NLP* (pp. 349–373). Springer.

Zlatev, J. (2007). Spatial semantics. In D. Geeraerts & H. Cuyckens (Eds.), *The Oxford handbook of cognitive linguistics* (pp. 318–350). Oxford University Press.

Documentation linguistique et standards dans le champ des humanités numériques : la TEI et la documentation du mixtèque de Mixtepec

1. Introduction au projet

Cette thèse décrit le projet de documentation concernant la langue mixtèque de Mixtepec[277] (MIX) *sa'an savi* « rain language (langue de la pluie) » utilisant la TEI (Text Encoding Initiative, en français « Initiative pour l'encodage du texte » www.tei-c.org) comme format d'encodage. Ces travaux ont pour objectifs de rendre compte de la façon dont la TEI et les technologies XML associées peuvent être utilisées comme format principal pour l'encodage, les métadonnées et l'annotation dans le cadre de projets linguistiques pluridimensionnels incluant des langues avec peu de sources primaires ; d'évaluer les outils, normes et standards, et pratiques actuellement utilisés en documentation linguistique ; et de créer un ensemble de ressources linguistiques pour la langue et la communauté mixtèques. En raison de l'étendue des données et ressources diverses produites, ce projet est composé d'éléments qui entrent aussi bien dans le champ des humanités numériques, que dans ceux de la documentation linguistique, de la linguistique descriptive et de la linguistique de corpus. Du fait de la pertinence de ces chevauchements disciplinaires, et dans le but de respecter les meilleures pratiques en vigueur dans chacune des disciplines, ces travaux ont mis en évidence la possibilité et la nécessité d'identifier plus concrètement, de discuter et de faire converger davantage encore les intérêts, technologies, pratiques et standards liés à chacune d'elles.

Le résultat principal du projet est la création d'un ensemble de ressources linguistiques multimédias réutilisables et évolutives en source ouverte (open source) incluant un dictionnaire TEI multilingue, une collection d'enregistrements audio publiés et archivés sur Harvard Dataverse (Bowers, Salazar et Salazar, 2019)[278], et un corpus de textes dérivés d'un ensemble composé de transcriptions du langage parlé et de textes écrits encodés et annotés en format TEI, et de descriptions et d'analyses linguistiques et lexicographiques de la langue mixtèque de Mixtepec[279]. La langue MIX étant dotée de peu de sources primaires, l'objectif était d'intégrer autant de

[277] Mixtepec-Mixtec Iso 639-3 [mix]; Glottolog [mixt1425]
[278] https://doi.org/10.7910/DVN/BF2VNK
[279] Le répertoire GitHub (https://github.com/iljackb/Mixtepec_Mixtec) contient les fichiers annotés qui composent le corpus et le dictionnaire TEI.

ressources disponibles dans cette langue que possible dans le corpus TEI avec un schéma d'encodage et d'annotation commun, dont la réalisation nécessite, en fonction de la source, des degrés divers d'opérations manuelles, de scriptage et d'utilisation d'outils digitaux. Les ressources linguistiques créées sont à leur tour utilisées pour faire progresser la connaissance de tous les aspects de la langue elle-même sur les plans linguistique et lexicographique, permettant de réaliser empiriquement des descriptions grammaticales basées sur le corpus et d'analyser les caractéristiques de la langue. Toutefois, comme nous le verrons, si ces travaux ont permis de produire des analyses et descriptions linguistiques (partie 2), en particulier sous la forme d'une analyse sémantique de termes de parties du corps (Bowers, sous presse), le résultat principal, et point central de la présente thèse, est la description de la structure, des sources et du contenu du corpus, des archives et du dictionnaire.

Dans la phase de collecte des données, d'annotation et d'encodage, j'ai cherché à recueillir le contenu relatif à chaque niveau linguistique, à la fois phonétique, sémantique et étymologique, ainsi que les potentielles variantes sous-dialectales et même idiolectales. Du fait de la complexité des données et du très vaste champ d'application des recherches linguistiques et lexicographiques entreprises, il est fondamental, à la fois pour mes propres travaux actuels et pour anticiper une réutilisation future, de disposer d'un moyen permettant d'organiser l'ensemble des divers composants des ressources langagières au sein d'un système dynamique, flexible et non tributaire d'un logiciel. En outre, étant donné le manque de dictionnaires consacrés à la langue[280], il est particulièrement important que les ressources créées soient réutilisables et évolutives, et puissent continuer à être développées, avec la possibilité de pouvoir être facilement exportées ou converties dans d'autres formats et rendues accessibles sous une forme conviviale pour les utilisateurs, notamment les membres de la communauté mixtèque.

La portée de ces travaux étant multiforme et recouvrant divers domaines d'études, je me suis heurté, au cours de leur réalisation, à des questionnements importants concernant un certain

[280] Alors qu'au moment de la présentation de ces travaux il n'y avait pas d'autre dictionnaire relatif à la langue mixtèque de Mixtepec à proprement parler, il existe un petit dictionnaire (Galindo Sánchez, 2009) pour la variété de mixtèque Abasolo del Valle parlée dans la municipalité de Playa Vicente (État de Veracruz) par une communauté qui a migré de la région de San Juan Mixtepec en plusieurs fois entre les années 1930 et les années 1950. Cette variété est généralement acceptée comme étant globalement identique au mixtèque de Mixtepec.

nombre de disciplines différentes, et j'ai dû en permanence trouver des moyens de les traiter de façon à servir la langue et à fournir un résultat de qualité pour la communauté mixtèque, en respectant les meilleures pratiques éthiques et en produisant au final un résultat conforme aux règles de l'art sur le plan des humanités numériques, de la TEI et de la documentation linguistique.

Pour atteindre ces objectifs, le modèle TEI a été choisi comme format pour l'encodage et l'annotation du corpus, pour le dictionnaire d'origine numérique et pour les métadonnées, ce format étant susceptible de satisfaire au mieux l'ensemble des objectifs de recherche précédemment mentionnés et d'obtenir le résultat souhaité. Comme nous le verrons plus tard, contrairement à tout l'éventail d'outils et, dans certains cas, de formats de données tributaires d'un outil existants pour chacun des principaux composants utilisés en documentation linguistique et en linguistique computationnelle, l'utilisation du modèle TEI permet d'encoder et d'annoter l'intégralité des données dans le même format. La TEI est très largement acceptée dans la communauté lexicographique numérique comme le standard de facto pour l'encodage des dictionnaires rétronumérisés et des dictionnaires d'origine numérique, et est de plus en plus utilisée pour les corpus de textes lexicaux annotés. En outre, chaque fichier intègre de très nombreuses caractéristiques liées aux métadonnées, ce qui permet de créer des structures d'éléments pour les champs linguistiques, les personnes et les lieux, et de faire des liens entre le contenu linguistique et les médias associés sans avoir à produire et à éditer séparément les métadonnées et le contenu.

Alors que la TEI est bien établie et de plus en plus largement adoptée pour les projets et ressources concernant les principales langues mondiales, en particulier les langues européennes et nord-américaines, elle l'est beaucoup moins pour ceux en lien avec des langues indigènes. En dehors des publications relatives au projet actuel (Bowers, 2015 ; Bowers et Romary, 2017 ; 2018 a, b ; 2019), Czaykowska-Higgins et Holmes (2013) et Czaykowska-Higgins et al. (2014) décrivent la création d'un dictionnaire TEI et d'une interface applicative à partir de sources héritées pour la langue indigène moses-columbia salish (« Nxaʔamxcín »). On citera également le récent projet Mesolex (DEL Grant #HAA-266482-19)[281] dont l'objectif principal est de collecter des ressources lexicales pour certaines langues indigènes mésoaméricaines (incluant des variétés de

[281] https://securegrants.neh.gov/publicquery/main.aspx?f=1&gn=HAA-266482-19

mixtèque), et de les convertir en un format TEI couramment accessible. Un avantage majeur du recours au modèle TEI pour traiter une langue dotée de peu de sources primaires est qu'il permet d'encoder des documents susceptibles d'être utilisés aussi bien comme un corpus linguistique annoté, qui (avec des schémas simples) est rendu accessible à un usage humain, que par des chercheurs dans d'autres domaines. Alors que la création de telles ressources polyvalentes et flexibles est, comme nous le verrons, au cœur de la mission des humanités numériques, elle n'a pas toujours été une priorité essentielle pour la plupart des cercles de linguistes.

Dans certains cas, l'utilisation de la TEI pour les travaux de documentation a nécessité le recours au vocabulaire de balisage pour des applications nouvelles, ou moins courantes, afin de prendre en compte les subtilités particulières des données. Ils requièrent en outre l'utilisation de diverses combinaisons de composants et de caractéristiques TEI qui sont moins souvent associés et pour lesquels il n'existe ainsi que peu (voire pas) d'exemples dans les directives, aucun cas d'usage ne figurant par ailleurs dans les publications antérieures (les gloses interlinéaires sont un cas particulièrement flagrant de telles omissions). On ne peut pas nier que l'adoption de cette approche a, contrairement à d'autres outils logiciels majeurs comme FLEx de SIL[282], ELAN[283], Toolbox[284], etc.[285]., parfois été lourde, tant du fait du temps requis pour annoter manuellement, organiser le contenu, écrire les scripts de conversion, que parce que je ne suis pas capable de tirer profit des fonctionnalités de productivité orientées utilisateur des outils précédemment décrits. Néanmoins, comme j'ai pris le temps de travailler sur les différents points, ma démarche n'a pas seulement été utile à ce projet en définissant la façon d'intégrer de nouvelles combinaisons uniques de caractéristiques pour une langue indigène non indo-européenne, mais a également servi à étudier de manière exhaustive les lacunes en matière de TEI, mais aussi de normalisation, d'interopérabilité et d'échange des données.

[282] https://software.sil.org/fieldworks/
[283] https://tla.mpi.nl/tools/tla-tools/elan/
[284] https://software.sil.org/toolbox/
[285] Bien que certains éléments des outils plus communément utilisés puissent sembler plus conviviaux pour l'utilisateur, ces logiciels n'étaient pas adaptés à ce projet pour un certain nombre de raisons. Ces points seront traités dans la présente thèse.

De plus, on peut espérer que l'adoption de la TEI pour ces travaux, associée à l'étude des outils et formats de données les plus couramment utilisés en documentation linguistique contribuera à la mise en œuvre de nouvelles mesures afin : d'augmenter la facilité d'utilisation de la TEI pour de futurs utilisateurs potentiels cherchant à mener des projets similaires, à la fois pour développer de nouveaux outils pour des non-experts et pour créer un précédent susceptible d'être imité ; d'élaborer un ensemble de scripts et de feuilles de style pour des conversions entre des formats de données différents ; et enfin de faire progresser les standards et échanges de données.

Lorsque l'on travaille sur des langues dotées de peu de sources primaires, il est impératif de pouvoir intégrer *n'importe quelle* source de données potentielle contemporaine ou historique susceptible de provenir d'un grand nombre de formats numériques ou analogiques différents. Et pour créer la capacité nécessaire pour intégrer et traiter de telles données, il est essentiel de développer des outils logiciels comme GROBID-Dictionaries (Khemakhem et al., 2017). GROBID-Dictionaries scanne et traite des ressources lexicales au format PDF pour générer un dictionnaire TEI. C'est un élément majeur dans le développement d'outils permettant aux chercheurs de numériser et de créer des corpus structurés à partir de ressources existantes (quand elles existent) (Khemakhem et al., 2017). En outre, comme ce type de tâches et de démarches tend à se développer, on peut espérer que cela générera une demande pour favoriser le développement de fonctionnalités logicielles toujours plus faciles à utiliser pour réaliser ces tâches, et/ou l'adaptation des outils logiciels existants pour rendre ces tâches possibles.

Si je présente les éléments, résultats et perspectives positifs de ces travaux, j'expose également les aspects qui méritent d'être améliorés, notamment mon approche méthodologique ou technique, ou le résultat en lui-même, et pour lesquels certains points doivent encore être traités pour aller de l'avant. Pour finir, cette thèse présente uniquement la base des questions méthodologiques et, bien entendu, du résultat linguistique. Mon objectif étant que toutes les dimensions de ces travaux continuent à progresser, je présente ici les résultats préliminaires des éléments techniques et de quelques éléments linguistiques de ce projet.

2. Introduction à la langue

La langue mixtèque de Mixtepec est parlée dans les 72 communautés, quartiers et colonies (« colonias ») de la municipalité de San Juan Mixtepec[286]. Dans les données du gouvernement mexicain[287], cette langue est désignée en tant que mixtèque du centre-ouest (« *mixteco de oeste central* »). Josserand (1983) classe cette variété dans la région de dialecte « Southern Mixteca Baja »[288], limitrophe de la région « Mixteca Alta »[289], comme une branche de dialecte distincte[290], bien qu'il soit probable que cette classification nécessite d'être révisée étant donné que davantage de variétés (notamment celles dans la région Juxtlahuaca) sont documentées. Au Mexique, la langue MIX est également parlée par plusieurs milliers de locuteurs vivant dans l'État de Basse-Californie (« Baja California »), dans la ville de Tlaxiaco, dans la municipalité de Santiago Juxtlahuaca, et, aux États-Unis, par des populations significatives habitant en particulier autour de Santa Maria (où les deux autres collaborateurs de ce projet ont grandi et résident encore) et d'Oxnard en Californie, dans l'Oregon, en Floride, et dans l'Arkansas.

Le nombre de variétés de mixtèque varie de 52 selon le site Ethnologue[291] (Simons et Fennig, 2018) à 81 selon l'INALI (Institut national mexicain des langues indigènes) (2008). Les sources de site Ethnologue étant toujours recensées auprès du gouvernement mexicain, l'INALI est probablement la source la plus fiable[292]. Aucune statistique concernant les données démographiques des locuteurs et le statut de la langue mixtèque de Mixtepec n'a été collectée depuis 2000 (un recensement réalisé en 2010 a recueilli des informations par familles de langues

[286] Bien que cette information ne figure pas dans les sources gouvernementales publiques, un document non officiel contenant une liste des lieux et de leurs habitants connus dans la municipalité de San Juan Mixtepec, établie par Gisela Beckmann, chercheuse à l'organisation SIL, est disponible à l'adresse suivante : https://github.com/iljackb/Mixtepec_Mixtec/blob/master/misc-sources/Pueblosy%20su%20estatus%20alfabetico.doc (source : Gisela Beckmann, communication personnelle, juillet 2020)

[287] https://www.inali.gob.mx/clin-inali/html/v_mixteco.html#47

[288] L'expression « région de dialecte » est utilisée conformément aux classifications référencées selon Josserand (1983). Il est à noter que le terme « dialecte » a traditionnellement été utilisé pour parler de manière méprisante des langues indigènes au Mexique, et est considéré comme péjoratif. Le terme « variété » est ainsi généralement employé pour parler de différentes langues mixtèques (ou autres langues indigènes).

[289] Malgré les classifications, j'ai entendu des locuteurs natifs de la langue MIX décrire cette variété comme appartenant au groupe « Mixteco Alto ».

[290] https://glottolog.org/resource/languoid/id/mixt1425 (consulté le 29/12/ 2019)

[291] https://www.ethnologue.com/subgroups/mixtec (consulté le 20/08/2019)

[292] Il convient de remarquer que depuis octobre 2019, le site Ethnologue est désormais un service payant pour les « pays à revenu élevé », et son accès est ainsi limité en l'absence d'abonnement. Les sources sur lesquelles sont basés les chiffres ne peuvent plus être vérifiées.

uniquement), date à laquelle le nombre de locuteurs était de 9 166[293]. Une évaluation actualisée de ses locuteurs est nécessaire du fait des informations contradictoires concernant le statut de mise en péril. Selon l'ELDP (Programme de documentation sur les langues en péril)[294], le statut est « Langue menacée », alors que le site Ethnologue[295] donne le statut de « Langue stable »[296].

Sur la base des observations directes et de discussions sur le sujet avec des locuteurs MIX, le statut de « Langue menacée » est certainement le plus exact. La combinaison de facteurs comme l'usage de l'espagnol beaucoup plus répandu dans les loisirs, sur internet, à l'école, et le nombre important de locuteurs MIX vivant hors de la zone où la langue est parlée et dont les enfants ne sont pas en contact avec la langue en dehors de la maison, notamment les enfants dont les parents parlent l'espagnol ou l'anglais, réduit en effet visiblement le nombre de nouveaux locuteurs. Outre ces questions d'ordre pragmatique/démographique, et comme c'est le cas dans de nombreuses sociétés indigènes post-coloniales, que ce soit dans l'histoire ou de nos jours, les locuteurs de langues indigènes sont victimes de racisme et de discrimination au Mexique comme ailleurs dans le monde où il existe des communautés issues de la diaspora. Cet état de fait, associé à la croyance que parler des langues indigènes ne présente aucun avantage, a sans aucun doute influencé l'attitude de certains parents qui ne transmettent pas la langue à leurs enfants (Basurto, Hernández Martínez, et Campbell, sous presse).

De plus, les enfants des locuteurs MIX qui vivent en zones urbaines ont de plus en plus tendance à avoir seulement une connaissance réceptive du mixtèque, étant donné que, dans leur vie quotidienne, ils interagissent avec des personnes qui ne sont pas susceptibles de parler le mixtèque (y compris d'autres personnes indigènes), l'espagnol devenant ainsi la seule langue de communication pratiquée. À cela s'ajoute le fait que, même parmi ceux qui parle le mixtèque, il existe un phénomène de diglossie qui fait que leur usage du mixtèque est limité à certains contextes, et surtout à certains sujets de discussion. Cette situation a pour effet de limiter les domaines de la vie quotidienne pour lesquels le mixtèque dispose d'un vocabulaire. Pour les

[293] https://www.ethnologue.com/subgroups/mixtec (consulté le 20/08/2019)
[294] http://www.endangeredlanguages.com/lang/10531 (consulté le 20/08/2019)
[295] https://www.ethnologue.com/language/mix (consulté le 20/08/2019)
[296] Cette divergence est particulièrement curieuse étant donné que la page du projet ELP (Projet sur les langues en péril) (qui indique le statut « Langue menacée ») cite le site Ethnologue comme source donnant le statut « Langue vigoureuse ».

domaines dans lesquels le mixtèque n'est pas employé, des mots empruntés à l'espagnol seront utilisés, ou la discussion se tiendra en espagnol (au moins dans le cas des locuteurs bilingues).

2.1 Bref aperçu de la typologie et des caractéristiques de la langue mixtèque de Mixtepec

Le sujet principal de cette thèse portant sur la documentation linguistique et sur l'approche spécifique des moyens technologiques, il n'est pas essentiel de donner ici une description exhaustive de la langue mixtèque de Mixtepec sur le plan linguistique. La priorité a été donnée à la collecte et à l'annotation des ressources pour s'assurer que celles-ci soient préservées et correctement documentées. Toutefois, cette Partie I fournit une description élémentaire de certaines caractéristiques majeures de la langue MIX, qui serviront de référence pour certains des exemples linguistiques présentés dans ce document, à la fois dans le corpus et dans le dictionnaire, et qui constitueront la base d'une grammaire plus complète qui sera élaborée dans un avenir proche.

Il convient également de noter qu'une grammaire de la langue est en cours d'écriture par Salazar et al. (2020), avec Jeremías Salazar, dans le cadre d'un cours de méthodologie de terrain dispensé par Eric Campbell à l'Université de Californie à Santa Barbara[297]. Après ces indications, et afin de fournir un contexte linguistique à de nombreuses caractéristiques linguistiques qui sont discutées dans les exemples développés dans cette thèse, je donne ici un bref aperçu de la structure de la langue MIX et de ses particularités les plus notables.

2.1.1 Tonalités lexicales

La langue MIX est une langue tonale avec cinque tons lexicales: *haut, moyen, bas, montant* et *descendant.*

[297] Le titre de la grammaire de Salazar et al. (2020) fait référence au « Yucunani Mixtepec Mixtec ». En effet, l'un de mes collègues dans le cadre à ce projet, Jeremías Salazar, a également été le consultant et collaborateur du cours de l'UCSB, et est l'auteur principal de cette grammaire en cours d'élaboration.

Tones	examples
Bas	[sùt̪ù] sutu *'priest'* [òkò] oko *'vignt'*
Moyen	[vēʔē] ve'e *'maison'* [jāʧī] yachi *'près'*
Haut	[kóní] koni *'dinde* [lóʧǐ] lochi *'vautour'*
Montant	[jǒsō] yóso *'metate'* [jōsǒ] yosó *'une plaine (herbeuse)'* [t̪ǐnà] tina *'chien'* [jǔt̪ī] yuti *'sable'* [t̪ínānă] tinana *'tomate'*
Descendant	[súkû] suku *'haut'* [kōt̪ô] koto *'sarape'* [āʔâ] a'an *'non'* [sâʔvà] sa'va *'frog'* [sâʔmă] sa'ma *'vêtements'*

Table 44: Les cinques tons lexicaux avec examples

Sur les syllabes bimoriques, il y a aussi un certain nombre de combinaisons de tons niveaux (Table 45), ainsi que plusieurs combinaisons de tons niveaux avec contours montants et descendants (Table 46) qui peuvent s'appliquer[298]:

Tones	examples
Bas Bas	[ʧu̍ů̍] chuun *'étoile'* [nt̪a̰à] ntaa *'plat', 'la verité'* [i̍i̍] iin *'neuf'*

[298] Un inventaire complet des combinaisons de niveaux de tons possibles étant encore à l'étude actuellement, il se peut que des exemples d'autres combinaisons de contours soient découverts, ou que certaines combinaisons décrites ici s'avèrent incorrectes.

Bas Moyen	[vèē] vee *'lourd'* [ʧàā] chaa *'homme'* [kàā] kaa *'métal'*
Moyen Bas	[ɲn̄ù̄] ñuu *'ville', 'village'* [yōò] yoo *'verre, tasse* [sāà] saa *'oiseau'*
Moyen Moyen	[ĩ̄ĩ̄] in *'un'* [lūū] luu *'petit'*
Moyen Haut	[mēé] très [kʷēé] kuee *'non'*
Haut Bas	[ʧáì] chai 'chaise' [mpáà] mpaa *'parrain (de fils)', 'compadre'*
Haut moyen	[ĩ́ĩ̄] iin *'exister', 'il y a'* [kʷíī] kuii *claire'*
Haut Haut	[ĩ́ĩ́] íin *'grêle'* [ndʒáá] nchaa *'bleu'*

Table 45: Combinaisons de tons niveaux sur les syllabes bimoraiques avec examples

Tones	**examples**
Bas Montant	[xe̋ĕ̋] xeen *'tranchant', 'dangereux'* [ĩ̋ĭ̋] iin *'sel'* [ìĭ] ii *'sacré'* [nù̋ŭ̋] nuu '*visage'* [nàă] naá *'terminer*
Moyen Montant	[vīĭ] vii *'beau', 'aspect sain'* [nāă] naa *'foncé'*
Haut Montant	[kʷíĭ] kuii '*vert'*

	[kʷi̋ĩ̋] kuiin *'étroit'* [íĩ] ii *'mari'* [i̋ĩ̋] iin '*peau'*
Bas Descendant	[ʃìô] xio *'robe', 'jupe'* [kʷàâ] kua *'environ'*
Haut Descendant	[páâ] paa *'père'* (emprunt d'espagnole *padre* [ˈpa.dɾe]) [hwa̋â̋] 'Juan' (emprunt d'espagnole *Juan* [ˈhwan]) [kwáâ] kuaa *'blind'* [kʷa̋â̋] kuaan *'jaune'* [náâ] náa *'porter*
Montant Moyen	[t͡să̄ā] tsaa *'nouveau'* [ŋk̰ʷĩ̄ī] nkuii *'renard*
Descendant Moyen	[tẫā̃] taan *'tremblement de terre'*

Table 46: Combinaisons de tons niveaux et tons contours sur les syllabes bimoraiques avec examples

Pour plus de détails sur la phonologie, voir Paster (2005, 2010) ; Paster et Beam de Azcona (2004, 2005) et Pike et Ibach (1978).

2.1.2 Principes fondamentaux de la structure de l'information

Sur le plan syntaxique, à l'instar d'autres langues mixtèques, MIX est une langue VSO (exemples (1)-(3)), même si cela peut varier dans le cas de changements de focus pragmatiques comme les formes interrogatives (exemple 4), les réponses à des questions ouvertes (« WH questions ») (exemple 5), les phrases emphatiques (exemple 6). De plus, comme dans d'autres variétés de mixtèque, il n'y a pas de règle et l'ordre des mots joue un rôle majeur dans les fonctions syntaxique et pragmatique.

(1) TOURNURE INTRANSITIVE

tsátsi chaa

IPFV\eat (manger) man (homme)

« the man is eating » (l'homme mange)

(2) TOURNURE TRANSITIVE

tsátsi chaa kuñu

IPFV\eat (manger) man (homme) meat (viande)

« the man is eating meat » (l'homme mange de la viande)

(3) TOURNURE DITRANSITIVE

kun-kua'a xu'un nuu Jack

POT-give (donner)\1SG money (argent) face (face/visage) Jack (Jack)

« I will give money to Jack » (je vais donner de l'argent à Jack)

(4) QUESTION OUVERTE (WH)-CHANGEMENT DE FOCUS ÉTROIT

nchíí yee =ni

Where (où) live (vivre) =2SG.FORM

« Where do you live? » (Où vis-tu ?)

(5) RÉPONSE À UNE QUESTION OUVERTE (WH)-FOCUS ÉTROIT

nuu chuun inkaa =yu

Face (face/visage) work (travailler) COP.LOC =1SG

« I'm at work » (je suis au travail)

(6) EMPHASE PAR LE DÉMONSTRATIF

sutu =ka ni-kani =yu

priest (prêtre) =PTCL.DEM PFV-hit (frapper) =1SG

« that priest hit me » (ce prêtre m'a frappé)

2.1.3 Marque de la personne et pronoms

Pour les verbes, adjectifs prédicatifs (attributs), noms, adverbes, appositions et, dans certains cas, pour les conjonctions (pour les fonctions comitatives), la marque de la personne est

donnée par une déclinaison morphologique (qui peut être une voyelle et/ou un changement de ton), un enclitique ou un pronom. On peut toutefois noter que les verbes ne prennent la marque de la personne que lorsque le sujet nominal n'est pas explicitement mentionné. Lorsqu'il existe deux verbes consécutifs, comme dans les modes volitifs (exemple 7), le premier et le second verbes prennent tous les deux la marque de la personne, mais le second utilise la racine irréelle alors que le premier utilise la racine réelle (voir le paragraphe 2.1.7 pour la description des racines et modes des verbes dans la langue MIX) :

(7) tsátsi chaa
IPFV\eat (manger) man (homme)
« the man is eating » (l'homme mange)

(8) kúni =yu katsi
IPFV\want (vouloir) =1SG eat[IRREAL] (MANGER[IRRÉEL]) \1SG
« I want to eat » (je veux manger)
(littéralement) « I want I eat » (je veux je mange)

L'usage de morphèmes par opposition aux enclitiques comme présenté ci-dessus pour marquer l'argument primaire d'un verbe est conditionné par les propriétés phonologiques de la racine, en particulier les environnements de tons et de voyelles (pour plus de détails sur ces facteurs phonologiques, voir : Paster et Beam de Azcona (2004, 2005) ; Paster (2005)). En outre, dans certains cas, la pragmatique peut également jouer un rôle. La langue MIX comprend au moins trois groupes de pronoms : les pronoms enclitiques dépendants, les pronoms emphatiques indépendants, et les pronoms démonstratifs. Le Tableau 1 fait l'inventaire des clitiques/pronoms, morphèmes et pronoms emphatiques.

Personne	Genre/Entité	Clitique/Pronom	Morphèmes	Emphatique
1.	(sg)	yu	(ton bas)	mee
	Exclusif (pl)	kue		meekue
	Inclusif (pl)	ko, yóó	-o	meekueko

2.	Familier (sg)	ku	-u ~ -un	meu
	Familier (pl)	kueyu, koyu		meekueyu
	Formel (sg)	ni		meeni
	Formel (pl)	kueni		meekueni
3.	Général (sg, pl)	ña, kui ~ vi	-i, -a	meeña
	Informel (pl)	kueyi, koyi		meekueyi
	Formel : masculin (sg)	ra		meera
	Formel : masculin (pl)	kuera		meekuera
	Formel : féminin (sg)	ñá, ná	-í, -á	meeñá, meená
	Formel : féminin (pl)	kueñá, kuená		meekueñá, meekuená
	Formel : humain (sg)	na		meena
	Formel : humain (pl)	na		meekuena
	Animal	ti		meeti
	Divinité/Saint	ya		meeya
	Bois	tu		meera
	Sphérique	ti		meeti
	Enfant	tsi		meetsi
	Liquide	ra		meera

Tableau 47: Pronoms enclitiques et emphatiques MIX[299]

Les pronoms emphatiques sont employés dans les formes réfléchies pour insister, apporter un contraste et dans les changements de thèmes. Ils associent la forme emphatique de base *mee* à un pronom enclitique ou au morphème correspondant. Les deux premiers groupes de pronoms présentés dans le Tableau 1 peuvent être utilisés comme sujets (exemples (4), (5), (8) ci-dessus), ou objets (exemple (6) ci-dessus) dans des phrases transitives et intransitives, et peuvent également servir à marquer la possession (voir le paragraphe 2.1.5). Certains des pronoms figurant dans le Tableau 1 sont dérivés des noms qu'ils remplacent, comme indiqué dans le Tableau 2 :

[299] Il est à noter que pour les formes animal, bois, sphérique, enfant et liquide, il existe également des versions plurielles des pronoms enclitiques et emphatiques qui suivent les mêmes modèles (par exemple pour les pronoms enclitiques : kue+PRON et pour les pronoms emphatiques : meekue+PRON), mais elles n'ont pas été intégrées ici pour des raisons de place.

Forme complète du nom	Signification	Enclitique/Pronom
ña'á	« Woman » (femme)	ñá
kiti	« Animal » (animal)	ti
tutú	« Wood » (bois)	tu

Tableau 48: Forme complète des noms sources et pronoms enclitiques correspondants

2.1.3.1 Pronoms démonstratifs et composants

Les pronoms démonstratifs sont composés de certains pronoms enclitiques et de la particule démonstrative *-ka* (exemple : *ñaká*) qui peut signifier « that » (ce), « there » (sujet impersonnel comme dans « il y a »), « these » (ces, ceux-ci), « those » (ces, ceux-là). *Ñáká* veut ainsi dire « that woman » (cette femme) (venant du pronom féminin formel ñá)[300]et *naka* « those people » (ces gens) (le même *na* que pour le pronom enclitique général formel à la 3e personne). Il existe aussi le pronom distal *ika* signifiant « there » (sujet impersonnel comme dans « il y a »). Ces pronoms ont également une fonction emphatique et peuvent être employés pour distinguer des personnes auxquelles il est fait référence conjointement dans un discours.

(8) **ñaká** n-tsatsi cha n-tsi'i chikuii
those PFV-eat (manger)\1SG et PFV-drink (boire)\1SG water (eau)
« I ate those and drank water' (j'ai mangé ceux-là et bu de l'eau)

La particule *ka* que l'on retrouve dans ces formes est utilisée principalement pour produire l'effet emphatique démonstratif, et suit la plupart du temps les sujets et objets nominaux, et même

[300] D'autres variétés de mixtèque, comme le mixtèque Chalcatongo (Macaulay, 1996), le Diuxi-Tilantongo (Kuiper et Oram, 1991), le mixtèque Jamiltepec (Johnson, 1988), le mixtèque Ayutla (Hills, 1990), parmi bien d'autres, attestent de l'existence de pronoms indépendants de « forme libre » incluant la 1re, la 2e ainsi que d'autres personnes. Il se pourrait que les pronoms MIX *yo* (2e personne du singulier, informel) et *yóó* (1ère personne du pluriel, inclusif) figurant dans le Tableau 2 en soient en fait des exemples, car il existe manifestement des termes apparentés dans de nombreuses autres variétés, comme *yòò'* (inclusif) dans Ayutla (Hills, 1990), *yò'ó* (inclusif) dans Jamiltepec (Johnson, 1988) et *yo̲'ó/yò* (2e personne du singulier, informel) dans Diuxi-Tilatongo (Kuiper et Oram, 1991). Dans toutes les données MIX observées, ceux-ci apparaissent uniquement en tant qu'objets d'un verbe transitif. Il pourrait ainsi exister un autre groupe de pronoms indépendants des 1re et 2e personnes qui seraient le pendant des noms complets des formes de la 3e personne, dont des pronoms enclitiques comme *ñá, tu, ti,* (exemples : *ña'á* « woman » (femme), *tutú* « wood » (bois), *kiti* « animal » (animal) respectivement). Des recherches plus approfondies sont néanmoins nécessaires.

les obliques. C'est également un composant actif dans les changements pragmatiques et de la structure de l'information qui permettent certaines extensions grammaticalisées de termes de parties du corps (BPT/body-part terms) (voir Bowers (sous presse) pour la discussion). Il est à noter qu'il existe une autre particule *ka* rencontrée dans d'autres variétés dont le mixtèque Chalcatongo (Macaulay, 1996), qui la décrit comme la particule additive[301] (voir les exemples (10), (42)).

(9) PARTICULE DÉMONSTRATIVE

chaa **=ka**

man (homme) **=PTCL.DEM**

« that man' (cet homme)

(10) PARTICULE ADDITIVE

ma= kua'a **=ka** staa katsi-a

NEG=GIVE (donner)\1SG **=PTCL.ADD** tortilla (tortilla) eat (manger)-3SG.INF

« I will not give him anything more to eat » (je ne lui donnerai pas quelque chose de plus à manger)

Il existe en outre un autre pronom démonstratif proximal *ño'o,* « this » (ce/cet/cette) (exemple 11), qui semble être le pendant pronominal de *yo'o* (voir l'exemple (12), ainsi que les exemples (19) et (24)), et peut avoir la fonction de déterminant démonstratif proximal, par exemple « this (X) » (ce/cet (X)), ou de pronom locatif proximal signifiant « here » (ici).

(11) nchii kuu **ño'o**

what COP **PRON.DEM.PROX**

« what is this? » (qu'est-ce que c'est ?)

[301] De plus amples recherches sont nécessaires à ce sujet, mais il est vraisemblable que les tons soient différents entre les deux. Si, dans le premier cas de la particule *ka* démonstrative, le ton est haut [ḵá], je ne suis pas sûr de celui de la particule additive, étant donné que toutes les occurrences de celle-ci apparaissant actuellement dans le corpus sont issues de sources écrites.

(12) staa **yo'o**

tortilla **DET.DEM.PROX**

« this tortilla » (cette tortilla)

2.1.4 Copules et mots apparentés (cognats)

La langue MIX possède plusieurs copules verbales qui suivent les mêmes modèles de déclinaison que les verbes réguliers, et certains adjectifs peuvent être employés comme prédicats (attributs)[302]. Les deux copules principales en langue MIX sont *kaa* et *kuu,* et dans nombre d'autres variétés de mixtèque, notamment Chalcatongo (Macaulay, 1996), Diuxi-Tilatongo (Kuiper et Oram, 1991) et Ayutla (Hills, 1990), les cognats de ces formes sont classifiés comme *réels* et *potentiel*s. Bien qu'il existe, comme le montrent les exemples (16) et (17), de bons usages des deux copules, leur distribution n'est pas conforme à une telle classification distincte en *réelle* et *potentielle*[303].

(12) ka'nu ta **ku-**i

big (grand) very (très) **COP-3**

« it is very big » (c'est très grand)

(13) nchii **kuu** ño'o

what (que/quoi) **COP** PRON.DEM.PROX

« what is this? » (qu'est-ce que c'est ?)

(14) che'e **kaa** xini patsa'nu

beautiful (magnifique) **COP** hat (chapeau) grandfather (grand-père)

« Grampa's hat is nice » (le chapeau de grand-père est très beau)

(15) nixi **ka-**u

[302] Il est à noter que les facteurs précis selon lesquels des adjectifs peuvent prendre la fonction de prédicats (attributs) n'ont pas encore été déterminés.

[303] Le fait que la copule *kuu* puisse se décliner en une forme potentielle *kun-kuu* et en une forme perfective *ni-kuu* constitue un élément de preuve complémentaire que la copule *kuu* n'est pas en elle-même « potentielle ». En outre, la copule *kaa* peut également se décliner en une forme potentielle *kun-kaa.*

how (comment) **COP**-2SG.INF

« How are you?' (comment vas-tu ?)

On note une dichotomie intéressante entre les deux en comparant la paire question-réponse suivante (exemple 16 et exemple 17) où *kuu* est employé dans la question et *kaa* dans la réponse :

(16) nchii hora **ku-i**

what (quel/quelle) time (heure) **COP-3S**

« what time is it? » (quelle heure est-il ?) (Nieves et Beckmann, 2007b)

(17) **kaa** iñu ntaa

COP six (six) o'clock (heures)

« It's six o'clock » (il est six heures) (Nieves et Beckmann, 2007b)

Dans le corpus, on rencontre également souvent la copule *« kaa'* dans des phrases signifiant « look like' (ressembler) :

(18) tono **kaa** ti'in+ita

Look.like (ressembler) **COP** skunk[rat+flower] (mouffette/bête puante + shunk/cannabis)

« It looks like a skunk » (cela ressemble à une bête puante/du shunk) (Rojas Santiago et al., 2014)

Toutefois, dans une phrase signifiant « to be similar to » (être semblable à), l'ordre est inversé :

(19) yutu yo'o tsá'-i kui'i ña **kaa** tono limu

Tree (arbre) this (cet) IPFV/give (donner)-3 fruit (fruit) that (que/qui) **COP** like (être comme) lime (citron vert)

« This tree produces fruit that is similar to limes » (cet arbre donne un fruit semblable aux citrons verts) (Rojas Santiago et al., 2014)

Il existe aussi un autre verbe analogue à une copule *iin* qui peut être employé dans des sens différents, notamment comme une copule existentielle « there is » (il y a), « to be » (être).

(20) COPULE EXISTENTIELLE : *iin*

iin ve'e na'nu

exist (exister) building (bâtiment) very.big (très grand)

« there is a very big building » (il existe/y a un très grand bâtiment)

Même si ce n'est pas encore clair, si des critères sémantiques ou lexicaux déterminent si un adjectif peut être attribut, lorsque c'est possible, ils se déclinent de manière identique aux verbes avec les mêmes pronoms/enclitiques, ou morphèmes :

(21) NOM-ADJECTIF

yutu suku

tree (arbre) tall (grand)

« tall tree » (grand arbre)

(22) ADJECTIVE ATTRIBUT

suku =yu

tall (grand) =1SG

« I am tall » (je suis grand)

2.1.5 Syntagmes nominaux, expression de la possession et notions apparentées

Dans la langue MIX, comme dans d'autres variétés de mixtèque, les syntagmes nominaux précèdent les adjectifs qualificatifs (exemple 23) et les déterminants démonstratifs (exemple 24). Dans les constructions possessives (exemple 26 et exemple 27) et partitives (partie-tout) (exemple 25), les noms sont exprimés dans le même ordre syntaxique que les compléments de nom, le premier nom (la partie) précédant le terme principal (le tout), par exemple sous la forme N

(partie/possédée)-N (tout/possesseur). L'article indéfini *in* (et les nombres en général)[304], ainsi que la marque du pluriel *kue*, précèdent tous les deux le nom qu'ils qualifient.

(23) NOM-ADJECTIF

yutu **suku**

Tree (arbre) **tall (grand)**

« tall tree » (grand arbre)

(24) NOM-DÉTERMINANT DÉMONSTRATIF

yutu **yo'o**

tree (arbre) **DET.DEM.PROX**

« this tree » (cet arbre)

(25) NOM-GÉNITIF (COMPLÉMENT DE NOM)/PARTIE-TOUT

xiní **chaa**

hat (chapeau) **man** (homme)

« the man's hat » (le chapeau de l'homme)

(26) POSSESSIF

maa **=yu**

mother (mère) **=1SG**

« my mother » (ma mère)

(27) POSSESSIF ET TERMES DE PARTIES DU CORPS

nu**u̱**

Face (visage)**1SG**

« my face » (mon visage)

(28) ARTICLE INDÉFINI

[304] L'article indéfini *in* est le nombre « one » (un). L'orthographe le représente clairement car le nombre neuf *iin* est également une voyelle nasale antérieure longue avec un ton bas ou descendant [î].

in chaa

ART.INDEF.SG man (homme)

« a man » (un homme)

(29) MARQUE DU PLURIEL

kue= chaa

PL= man (homme)

« the men » (les hommes)

En outre, les phrases obliques avec appositions reflètent aussi la même structure, ce qui, comme montré par Brugman (1983), Brugman et Macaulay (1986) et Bowers (sous presse), n'est pas fortuit étant donné que de nombreuses prépositions sont des extensions métaphoriques de noms relationnels, plus particulièrement de termes de parties du corps qui sont, dans leur sens le plus primitif, des syntagmes nominaux partitifs (de type partie-tout). Exemple :

(30) **nuu** + ve'e

face (face/visage) + house (maison)

« front of the house » (devant de la maison)

(31) TERMES DE PARTIES DU CORPS UTILISÉS DANS DES APPOSITIONS STATIQUES

ntú'u saa =ka **nuu** ve'e

IPFV\sit (être assis) bird (oiseau) =PTCL.DEM **face** (face/visage) house (maison)

« that bird is sitting in front of the house » (cet oiseau est assis devant la maison)

(32) inká-i **tsa'a** yutu

IPFV\COP.LOC-3 **foot** (pied) tree (arbre)

« It is under the tree » (il est sous l'arbre)

Mais la sémantique de la partie du corps particulière apparaît de manière évidente dans l'usage d'un sens appositionnel élargi donné qui dépend du terme associé, comme on le voit dans

l'exemple (33). Lorsqu'il s'agit d'objets qui sont semblables, sur le plan physique, à un animal à quatre pattes, le terme de partie du corps *titsi* (« stomach » – ventre) est utilisé à la place de « foot » (pied). Dans la traduction « under the table » (sous la table), la configuration de l'objet situé sous la table ressemble plus à celle d'un objet situé sous un animal à 4 pattes, alors que lorsqu'il s'agit d'un objet assis au pied d'un arbre, cela ressemble plus à quelque chose situé au pied d'un humain :

(33) ntú'-i **titsi** mesa
IPFV\SIT-3 **stomach** (ventre) table (table)
« It is sitting under the table » (il est assis sous la table)

(34) TERMES DE PARTIES DU CORPS UTILISÉS DANS DES APPOSITIONS DYNAMIQUES
ntsaa =kue **nuu** chuun
PFV\arrive (arriver) =1PL.EXCL **face** (face/visage) work (travailler)
« We arrived at work » (nous arrivions au travail)

Ces termes de parties du corps au sens élargi s'étendent dans des appositions au-delà de la notion d'espace et de mouvement, comme on le voit dans les exemples (35) et (36) avec *nuu* « face » (face/visage), et dans l'exemple (37) avec *tsa'a* « foot » (pied) dans des phrases ditransitives obliques avec objets indirects :

(35) FACE/VISAGE DANS LE TRANSFERT DE POSSESSION
kun-kua'a xu'un **nuu** Jack
POT-give (donner)\1SG money (argent) **face** (face/visage) Jack
« I will give money to Jack" (je vais donner de l'argent à Jack)

(36) FACE/VISAGE DANS LE TRANSFERT D'INFORMATION
ntakani =na **nu̱** ña ntivi karru =ku
PFV\tell (dire) =3PL.FORM.GEN **face** (face/visage)\1SG REL PFV\break (casser) car (voiture) =2SG.INF

« Someone told me your car broke down » (quelqu'un m'a dit que ta voiture était tombée en panne)

(37) FOOT UTILISÉ POUR « EN ÉCHANGE DE/CONTRE »

kun-cha'vi =yu **tsa'**-i

POT-pay (payer) =1SG **foot** (pied) -3

« I'm going to pay for it" (je vais payer pour ça)

À partir des exemples ci-dessus, on peut noter que même dans le sens élargi (exemples 35-37) dans lequel la signification va, par grammaticalisation, bien au-delà du sens nominal original, la structure informative BPT-N est conservée. Les extensions des termes de parties du corps (BPT), en particulier dans le cas des phrases avec des notions d'espace et de mouvement, peuvent être mieux analysées à l'aide des concepts de trajecteur et de repère issus de la Cognitive Grammar (Grammaire Cognitive) (Langacker, 1986, 1987), voir Bowers (sous presse) pour cette analyse.

2.1.6 Conjonctions et adverbes

Lorsqu'elle porte la marque de la personne, la structure des adjectifs, adverbes et conjonctions prédicatifs (attributs) reflète également celle de V-PERS$_{(SUBJ)}$, prenant par exemple la forme ADJ-PERS, ADV-PERS, CONJ-PERS. La conjonction *tsi* « with » (avec), « and » (et) (que l'on peut parfois rencontrer sous la forme d'une apposition « to » (à/de)), se décline en *tsi-an* « with him/her/it »(avec lui/elle) (informel) :

(38) ntuu **tsi** tsikuaa

day (jour) **and** (et) night (nuit)

« day and night » (jour et nuit)

(39) ni-kitsaa =kuera **tsi-an** ñuu yo'o

PFV-arrive (arriver) =3PL.M.FORM **with** (avec) **-3SG.INF** town (ville) this (ce/cette/ça)

« they arrived in this town with it » (ils sont arrivés dans cette ville avec lui/elle/ça) (Mendoza Santiago, 2008)

Lorsqu'ils sont déclinés, certains adverbes se placent entre la base et la déclinaison ou clitique. On notera, dans l'exemple (40), que le terme de partie du corps *sata* est employé dans un sens adverbial élargi signifiant « backwards » (à reculons/en arrière) (voir Bowers (sous presse) pour une analyse et une discussion approfondies). En outre, l'exemple (41) montre à la fois une conjonction déclinée et la présence de l'adverbe *ta* « very » (très) situé entre le verbe et l'enclitique *yu* (1e personne du singulier).

(40) tsíka **sat̲a**
IPFV\walk (marcher) **back** (dos)**\1SG**
« I'm walking backwards » (je marche en arrière/à reculons)

(41) kúni **=ta =yu** káka+nuu tsi-an
IPFV\want (vouloir) **=very** (très) **=1SG** stroll [walk+face] (se promener [marcher+face/visage]) with (avec)-3SG.INF
« I really want to take a stroll with him" (je veux vraiment aller me promener avec lui) (Gómez Hernández, 2008a)

Dans l'exemple qui suit, la particule additive *ka* suit l'adverbe *so* et précède le pronom enclitique du sujet *ti,* ce qui représente également un exemple de comparaison :

(42) luu **so =ka =ti**
small (petit) **very** (très) **=PTCL.ADD =3SG.ANML**
« It is so much smaller » (il est tellement plus petit – animal) (Rojas Santiago et al., 2014)

Il convient toutefois de noter que dans la structure standard de l'information VSO, la plupart des adverbes ne sont pas déclinés et sont situés en position finale dans la phrase :

(43) ni-kuun savi **takuni**
PFV-fall (tomber) rain (pluie) **yesterday** (hier)
« it rained yesterday » (il a plu hier)

2.1.7 Déclinaisons verbales : aspect et mode

Selon Bickford et Marlett (1988), dans les langues mixtèques, la déclinaison des verbes porte sur l'aspect et le mode, et non purement sur le temps, et bien que les différents aspects puissent faire référence à des événements ayant lieu dans le présent, le passé et le futur, ils concernent la structure temporelle interne d'une situation par opposition à un positionnement spécifique dans le temps. Bickford et Marlett (1988), Macaulay (1996) et bien d'autres travaux ont montré qu'il existe une distinction majeure entre le mode de la réalité (ou réel/indicatif) et le mode de l'irréel, qui se reflète dans une dichotomie entre les racines verbales dans les langues mixtèques. En conséquence, beaucoup de verbes MIX (mais pas tous) possèdent une forme réelle (indicative) et une forme irréelle[305], par exemple :

Verbe	Forme réelle	Forme irréelle
« walk » (marcher)	tsika	kaka
« sing » (chanter)	tsita	kata
« cry » (pleurer)	tsaku	kuaku
« give » (donner)	tsa'a	kua'a
« sleep » (dormir)	kixi	kusu

Tableau 49: Formes verbales réelles et irréelles dans la langue MIX

Comme décrit par Macaulay pour le mixtèque Chalcatongo, certains verbes dont les racines réelle et irréelle sont différentes présentent divers types d'alternances entre les formes données, la plus courante étant l'alternance entre la forme réelle *ts* et la forme irréelle *k*. Mais il en existe d'autres, notamment l'alternance x- et k- (*ts* et *k* en langue MIX), l'alternance x- et k- plus

[305] On notera qu'en lexicographie mixtèque, la forme des verbes utilisée dans les gloses est la forme irréelle.

alternance des tons, l'alternance x- et k- plus alternance des voyelles, l'alternance x- et k^w-, l'alternance des tons (uniquement), et plusieurs autres[306].

Les formes réelles sont employées avec les aspects Perfectif (également appelé *Complétif* [307]), Imperfectif (également appelé *Incomplétif,* or *Continuati*f), Habituel et Progressif[308]. Les formes irréelles sont utilisées pour l'aspect Potentiel, les impératifs et la tournure Modale[309]. La déclinaison des verbes MIX porte donc sur l'aspect et le mode et est marquée par une combinaison de racines verbales (le cas échéant) en complément de préfixes, et/ou du ton.

2.1.7.1 Imperfectif

L'aspect imperfectif est employé pour parler de situations présentes, et n'est pas décliné à l'aide d'un préfixe, mais par l'intermédiaire d'un ton haut appliqué sur la voyelle initiale[310] de la forme réelle du verbe[311].

Verbe (forme irréelle)	Imperfectif
katsi *« eat » (manger)*	**tsátsi [tzátsí]** *« I am eating » (je mange)*
ko'o *« drink » (boire)*	**tsí'i [tsíʔî]** *« I am drinking » (je bois)*
ka'an *« speak » (parler)*	**ká'an yu** *« I am speaking » (je*

[306] Dans la langue MIX, en raison du manque de données, en particulier pour les formes irréelles, et plus spécifiquement les tons, les détails et l'étendue des alternances est toujours à l'étude.
[307] Parmi les autres études des variétés de mixtèque qui emploient les termes *Complétif* et *Incomplétif,* on peut citer : Paster et Beam de Azcona (2005) pour le mixtèque de Mixtepec Yucunani, Macaulay (1996) pour le mixtèque Chalcatongo, Kuiper et Oram (1991) pour le mixtèque Diuxi-Tilatongo, Hills (1991) pour le mixtèque Ayutla (bien que les deux derniers emploient le *Continuatif* plutôt que l'*Incomplétif*).
[308] Kuiper et Merrifield (1975), Macaulay (1996), Bickford et Marlett (1988), entre autres, ont parlé de l'aspect Progressif dans d'autres variétés de mixtèque, qui sont caractérisées par des racines verbales additionnelles, en complément de la différence standard entre Réel et Irréel, bien que cela apparaisse uniquement dans les phrases verbales exprimant un mouvement. Ce point est lié à la sémantique du mouvement et de l'arrivée. Le comportement spécifique des racines verbales pour l'aspect progressif dans la langue MIX comparée à des variétés apparentées nécessite toutefois une analyse plus approfondie qui fera l'objet de travaux ultérieurs.
[309] Le terme *Modal* est employé conformément à Macaulay (1996) pour décrire la fonction apparentée pour le mixtèque Chalcatongo.
[310] Alors que dans l'orthographe de travail le ton bas marquant l'aspect perfectif n'est pas représenté, le ton haut marquant l'imperfectif est repéré par un signe diacritique de ton haut au-dessus de la voyelle.
[311] Il convient de remarquer que l'étude des modèles de tons des lemmes verbaux n'avance pas beaucoup car, dans de nombreux cas, certains verbes ne sont apparus que dans des sources écrites, dans lesquelles le ton n'est représenté qu'à l'imperfectif, et dans quelques paires minimales. Ainsi, lorsque je montre ces formes, j'utilise l'orthographe de travail dans laquelle le ton n'est marqué qu'à l'aspect imperfectif et dans certains éléments lexicaux peu distinctifs.

	parle)
kuaku *« cry » (pleurer)*	**tsakui** *« he/she is crying » (il/elle pleure)*
kusu *« sleep » (dormir)*	**kíxi yu** *« I am sleeping » (je dors)*

Tableau 50: Verbes dans leurs formes irréelle (glose) et imperfective

(44) tsí'i ntixi michuni

IPFV\drink (boire)\1SG pulque (pulque) right.now (en ce moment)

« I'm drinking pulque right now » (je bois du pulque en ce moment/je suis en train de boire du pulque)

(45) ká'an =kuená sa'an savi

IPFV\speak (parler) =3PL.FEM.FORM Mixtepec-Mixtec (mixtèque de Mixtepec)

« They (elder women) are speaking Mixtepec-Mixtec » (elles (les vieilles femmes) parlent le mixtèque de Mixtepec)

(46) tsáku vari kúni =ta =yu tanta'a cha koo xu'un

IPFV\cry (pleurer)\1SG because (parce que) IPFV\want (vouloir) =very (très) =1SG get.married (se marier)\1SG and NEG.exist (exister) money (argent)

« I'm crying because I really want to get married but there's no money » (je pleure parce que je veux vraiment me marier, mais il n'y a pas d'argent)

(47) tsátsi =na tikoo tsi ntuchi

IPFV\eat (manger) =3PL.FORM tamale (tamal) and (et) bean (haricot)

« they're eating tamales and beans » (ils mangent des tamales et des haricots)

2.1.7.2 Perfectif

L'aspect perfectif est employé normalement pour des événements isolés du passé. Comme décrit par Paster et Beam de Azcona (2004) et par Paster (2005), il est habituellement marqué par

le préfixe verbal ni- [nì] (48), et, sur les verbes avec des arrêts pré-nasalisés et consonnes affriquées (*nt-, nts-)* sur le début, par un ton bas sur la première voyelle du radical (50). En outre, dans certaines conditions tonales et phonologiques, il peut être marqué soit par la combinaison d'un n- pré-nasal et d'un changement de ton (ton bas) sur la première voyelle (49), soit simplement par un changement de ton (ton bas-montant) sur la première voyelle (51).

Verbe (forme irréelle)	**Imperfectif**	**Perfectif**
ya'a *« approach, cross, pass » (approcher, traverser, passer)*	**yá'a** *« I'm approaching » (j'approche)*	**ni-ya'a** *« I approached » (j'ai approché)*
ko'o *« drink » (boire)*	**tsí'i** *« I'm drinking » (je bois)*	**ntsii'i [ntzìʔì]** **«** *I drank » (j'ai bu)*
ntava *« fly » (voler)*	**ntava** *« it (animal) is flying, it flies » (il (animal) vole)*	**nt̲ava ti [ndàva]** *« it (animal) flew » (il(animal) a volé)*
sketa *« run » (courir)*	**skéta** *« I am running » (je cours)*	**sk̲eta [skɛ̆tâ]** *« I ran » (j'ai couru)*

Tableau 51: Différence entre les formes irréelle, imperfective et perfective des verbes

(48) ni-ya'a uvi hora
PFV-pass (passer) two (deux) hour (heure)
« two hours passed » (deux heures sont passées)

(49) n-tsi'i chikuii tsi luluu kafé
PFV-drink (boire)\1SG water (eau) and (et) little.little (un peu/petit) coffee (café)
« I drank water and a very small coffee » (j'ai bu de l'eau et un tout petit peu de café)

(50) nta̲va taka =ka xini =yu

PFV\fly (voler) woodpecker (pic) =PTCL.DEM head (tête) =1SG

« the woodpecker flew over my head » (le pic a volé au-dessus de ma tête)

(51) ske̲ta nuu chuun takuni

PFV\run (courir)\1SG face (face/visage) work (travail) yesterday (hier)

« I ran to work yesterday » (j'ai couru jusqu'au travail hier)

2.1.7.3 Potentiel

Le potentiel est généralement utilisé pour les situations non réelles et futures relatives, et est marqué par le préfixe (ku- ~ kun-[312]). Dans tous les exemples rencontrés dans les données observées, le préfixe est employé avec une consonance nasale, et sa variante avec voyelle nasalisée apparaît lorsque les racines verbales commencent par *k.*

Imperfectif	**Potentiel**
skéta *« I am running » (je cours)*	**ku-sketa** *« I will run » (je courrai)*
tsí'i na *« they are drinking » (ils boivent)*	**kun-ko'o na** *« they will drink » (ils boiront)*
kí'vi na *« they are entering » (ils entrent)*	**kun-ki'vi na** *« they will enter » (ils entreront)*
tsá'i *« he/she is giving » (il/elle donne)*	**kun-kua'i** *« he/she will give » (il/elle donneront)*

Tableau 52 : Différences entre les formes verbales Imperfectif et Perfectif

[312] Il existe deux variantes de forme pour le préfixe du futur : *un-* [ṹ], et [ŋ́], les deux étant généralement représentées par l'orthographe *kun-*. Il est à noter que le préfixe du potentiel est probablement dérivé de ce que d'autres variétés de mixtèque appellent la « copule du potentiel » *kúu.* Macaulay (1996) note que dans le mixtèque Chalcatongo, le terme apparenté (cognat) de la copule du potentiel précitée (également *kúu*) comporte aussi une variante courante composée uniquement de la voyelle ú.

(52) ku-sketa xchaan

POT-run (courir)\1SG tomorrow (demain)

“I will run tomorrow » (je courrai demain)

(53) i’iin ñachaa ku-ntuta’an =ra kumi chika

each the.men (chaque/chacun hommes) POT-receive (recevoir) =3SG.MASC four (quatre) plantain (plantain)

« the men will each receive four plantains » (les hommes recevront chacun quatre plantains) (Beckman et Nieves, 2008b)

(54) kun-ku’u =yu ntuku iki katsi

POT-go (aller) =1SG look.for (chercher)\1SG calabaza (citrouille) eat (manger)\1SG

« I will go look for calabaza to eat » (j’irai chercher une citrouille pour manger) (Gómez Hernández, 2007a)

(55) kun-ko’o =kuera ntixi tsini vichi

POT-drink (boire) =3PL.MASC.FORM pulque (pulque) tonight (ce soir)

« they (elder men) will drink pulque tonight » (ils (les hommes âgés) boiront du pulque ce soir)

2.1.7.4 Impératifs

Les tournures impératives emploient la forme irréelle du verbe, et sont souvent réalisées avec un modèle de ton moyen-moyen. Alors que l’on utilise uniquement la racine irréelle pour les ordres informels, on emploie aussi la déclinaison formelle *=ni* pour donner un ordre/conseil à une personne âgée ou respectée.

Imperfectif	**Impératif**

tsika *« walk » (marcher) (informel)*	**kaka [kākā]** *« walk! » (marche!)*
tsátsi ni *« you are eating » (vous mangez (formel)*	**katsi ni** *« eat! » (mangez/veuillez manger!) (formel)*
tsá'a ni *« you are giving » (vous donnez (poli))*	**kua'a ni** *« give » (veuillez donner!) (formel))*

Tableau 53: Comparaison des formes verbales Imperfectif et Impératif

(56) kaka chinu inkaa =yu
walk[IMP] (marcher/venir [IMP]) over.to (vers/jusqu'à) COP.LOC =1SG
« walk over to me » (marche(z)/viens(venez) vers moi)

(57) Kuntu'u nuu̱
sit[IMP] (s'asseoir [IMP]) face (face/visage)\1SG
« sit down in front of me » (assieds-toi/asseyez-vous en face de moi)

(58) katsi =ni
eat[IMP] (manger [IMP]) =2SG.FORM
« eat! » (mange, s'il te plaît!) (poli)

(59) kua'a =ni ntaku
give[IMP] (donner [IMP]) =2SG.FORM broom (balai)
« give me the broom » (donne-moi le balai, s'il te plaît!) (poli)

2.1.7.5 Habituel

L'aspect habituel est marqué par le préfixe (ntsi-) appliqué à la racine réelle. Il peut exprimer une habitude passée ou des actions en cours dans le passé :

(60) che'e ta ntsi-kana =ti

Beautiful (beau) so (si/tellement) HAB-sing (chanter) =3SG.ANML

« it was so beautiful when it sang » (il était si beau quand il chantait (animal)) (Ramos Hernández, 2007)

(61) ntsi-kuntu'un =ti nta'a in yutu

HAB-sit (s'asseoir) =3SG.ANML hand (main) ART.DEF.SG tree (arbre)

« it was sitting on the branch of a tree » (il était assis sur la branche d'un arbre (animal)) (Gómez Hernández, 2008b)

(62) tsini =na tu'un yutu ña ntsi-kaa ñuu yo'o

know (connaître) =3PL.FORM story (histoire) tree (arbre) REL HAB-stand (se trouver) town (ville) this (ce/cette)

« they know the story of the tree that used to stand in this town » (ils connaissent l'histoire de l'arbre qui se trouvait dans cette ville) (Mendoza Santiago, 2009)

(63) ntsi-tsatsi staa

HAB-eat (manger)\1SG tortilla (tortilla)

« I was eating tortillas » (je mangeais des tortillas)

2.1.7.6 Modaux

L'aspect modal est marqué par le préfixe (na-) appliqué à la racine réelle (quand elle est distincte), et peut avoir de nombreuses fonctions. Il peut notamment exprimer des incitations, des intentions, la nécessité, des hypothèses, des possibilités, et des modes de type subjonctif.

(64) na-ko'on

MOD-go (aller)[1PL.INCL]

« let's go! » (allons-y!)

(65) kua'a sa'mu na-kiku na-chinchee yo

give[IMP] (donner [IMP]) clothes (vêtements) MOD-sew (coudre)\1SG MOD-help (aider)\1SG 2SG.INF

« Give (me) the clothes, I can help you sew » (donne (moi) les vêtements, je peux t'aider à coudre) (Gómez Hernández, 2007b)

(66) na-tsinu sa'mu ra na-ko'on viko

MOD-be.finished (être fini/terminé) clothes (vêtements) CONJ MOD-go (aller)[1PL.INCL] party (fête/soirée)

« when the clothes are done, let's go to the party » (dès que les vêtements sont terminés, allons à la fête) (Gómez Hernández, 2007b)

(67) ta ni-ne'e xu'un na-ntakuaan ntivi

when (quand) PFV-get (avoir/recevoir)\1SG money (argent) MOD-buy (acheter)\1SG egg (œuf)

« when I get money, I'll buy eggs » (quand j'aurai de l'argent, j'achèterai des œufs) (Beckmann et Nieves, 2007)

(68) ntsi-ntu'un nchatu nuu avión = a-kitsa-i

HAB-sit (s'asseoir)\1SG wait (attendre)\1SG face (face/visage) airplane (avion) =PTCL.DEM MOD-arrive (arriver)-3SG

« I was sitting down, waiting for the airplane to arrive » (j'étais assis, attendant que l'avion arrive)

(69) takua na-kuu ki'in avión

so.that (de façon à) MOD-be.able (pouvoir/être capable de) catch (prendre)\1SG plane (avion)

« so that I could catch the plane » (de façon à ce que je puisse prendre l'avion)

(70) ku-yakua nta'a tatu na- ke'e nuu sta-u

POT-get.dirty hand (avoir les mains sales)\1SG if (si) MOD-touch (toucher)\1SG face (face/visage) tortilla (tortilla)-2SG.INF

« I'll get my hands dirty if I touch your tortilla » (j'aurai les mains sales si je touche ta tortilla) (Gómex Hernández, 2007a)

2.1.7.7 Négation

Dans la langue MIX, la négation est exprimée essentiellement par le préfixe verbal *ma-*, ou le préfixe adverbial *kue,* qui permettent de modifier les adjectifs et les verbes. Dans le mixtèque Chalcatongo, Macaulay décrit le cognat *ma-* (qui prend la même forme) comme un *marqueur de mode négatif*, dont le sens est l'opposé de *na-* (également cognat de même forme).

(71) ma- sana + in-o sa'an =ko

NEG- forget (oublier) -1PL.INCL language (langue) =1PL.INCL

« we must not forget our language » (nous ne devons pas oublier notre langue) (Beckmann et Nieves, 2008c)

(72) ma- tsíni =na tu'un + yata ñ-oo

NEG- IPFV\know (connaître) =3PL.GEN legend (légende) town (ville)-1PL.INCL

« they don't know the legend of our town » (ils ne connaissent pas la légende de notre ville) (López Santiago, 2008)

(73) A ma- kuu chinche-u yu

QNEG- be.able (pouvoir/être capable de) help (aider) -2SG.INF PRON.1SG

« Can you not help me? » (Ne peux-tu pas m'aider ?) (Gómez Hernández, 2007a)

(74) Kue va'a kíku =ku

NEG well (bien) IPFV\sew (coudre) =2SG.INF

« You're not sewing well » (tu ne couds pas bien) (Gómez Hernández, 2007b)

(75) Kue kúni =yu sachuun

NEG IPFV\want (vouloir) =1SG IPFV\work (travailler)\1SG

« I don't want to work » (je ne veux pas travailler) (Gómez Hernández, 2007a)

(76) Kue tsitsini =yu michu'ni in libru ka'vi =yu
NEG eat.breakfast (prendre le petit-déjeuner) =1SG right.now (en ce moment) ART.INDEF.SG book (livre) read (lire) =1SG
« Right now, I'm not eating breakfast, I'm reading a book » (en ce moment, je ne suis pas en train de prendre le petit-déjeuner, je lis un livre)

(77) kue nchichi
NEG difficult (difficile)
« easy » (facile)

Dans la langue mixtèque Chalcatongo (Macaulay, 1996), on ne rencontre le préfixe *ma-* qu'avec des verbes au mode Potentiel, et dans les quelques exemples observés dans le corpus, il apparaît qu'il s'agit dans la plupart de cas avec des racines verbales *irréelles*[313]. Toutefois, dans la langue MIX, on peut également le rencontrer avec des verbes au perfectif, qui, pour rappel, utilise la racine réelle (pour les verbes pour lesquels les deux racines sont distinctes) :

(77) ma- ni- kuu sketa =ti
NEG- PFV-be.able (pouvoir/être capable de) run (courir) =3SG.ANML
« it could not run » (il n'a pas pu courir (animal))

(78) ma- ni-ntakuaan =kue nchii nchai
NEG- PFV-buy (acheter) =1PL.EXCL any (du, de la) food (nourriture)
« We did not buy any food » (nous n'avons pas acheté de nourriture)

(79) ma- n-tsini lochi =ka
NEG- PFV-know (savoir) vulcher (vautour) =PTCL.DEM
« the vulture didn't know » (le vautour ne savait pas) (Gómez Hernández, 2008c)

[313] Il convient de noter que certains verbes sont intrinsèquement « potentiels », et ne comportent que des formes irréelles, comme *kuu* 'to be able to' (pouvoir/être capable de) et *kuni* 'to want' (vouloir).

(80) ma- n-tsa’ -i mii katsi ña’a =ka

NEG- PFV-allow (permettre) -3SG PRON.EMPH.3SG [IRREALIS]eat (manger) woman (femme) =PTCL.DEM

« He didn’t allow himself to be eaten by that woman » (il ne se permettait pas d’être mangé/de se faire manger par cette femme) (Gómez Hernández, 2008d)

En outre, il n’a été observé qu’un seul exemple de négation marquée uniquement par un changement de ton. Il s’agit de la forme potentielle du verbe « give » (donner), dans laquelle la première voyelle de la racine change pour prendre un contour de ton bas-montant. La négation standard *ma-* peut toutefois aussi être utilisée sans changement de ton. L’utilisation du changement de ton comme marque de la négation est documentée dans le mixtèque Ayutla (Hills, 1990). Dans cette variété, c’est le principal moyen de marquer la négation :

Forme potentielle (affirmative)	Forme potentielle (négative)
kun-kua’a [ǘɻ̨wàʔà] *« I will give » (je donnerai)*	**kua’a [ɻ̨wă̋ʔà]** (ou) **ma-kun-kua’a** *« I will not give » (je ne donnerai pas)*

Tableau 54 : Négation du verbe kua’a « to give » (donner)

2.1.8 Dérivation

La langue MIX, comme beaucoup de variétés de mixtèque, possède une série de préfixes dérivationnels qui peuvent être combinés à des verbes ou à des noms pour créer de nouveaux éléments lexicaux. Ils sont décrits ci-après :

2.1.8.1 Causalité

Le préfixe causal *sa-* est de toute évidence dérivé de *sa’a* « to do, make » (faire), et s’utilise pour exprimer des notions de causalité ou certains types d’activités. Il existe également des variantes comportant simplement *s-* ou *x-* [ʃ] :

Source	Verbe causal
va'a *« good » (bon/beau)*	**sava'a** *« to construct, « build » (construire)*
chuun *« work » (travail)*	**sachuun** *« to work » (travailler)*
na'a *« appear » (apparaître)*	**sna'a** *« to show, teach » (montrer, enseigner)*
núu *« come down » (descendre/baisser/tomber)*	**xnuu** *« to bring down » (faire baisser/abattre/renverser)*
tutsi *« hurt » (douleur/souffrance)*	**stutsi** *« to hurt (someone) » (blesser/faire du mal à (quelqu'un))*
tsio *« side » (côté/partie)*	**satsio** *« to separate » (séparer)*

Tableau 55: Verbes causaux et leurs sources lexicales

On remarquera que la forme causale peut être observée dans le nom de la principale ville où l'on parle le mixtèque de Mixtepec (San Juan Mixtepec) *Xnubiko,* ou *Snubiko* que l'on peut analyser comme suit : *xnuu* « bring down from » (faire descendre de) + *biko* « clouds » (nuages).

2.1.8.2 Itération

Le préfixe itératif *nta-* s'utilise pour exprimer la répétition ou le recommencement. Dans d'autres variétés de mixtèque, l'itération est aussi désignée par la notion de *répétitiv*ité (voir : Macaulay, 1996 : Chalcatongo Mixtec), et prend la forme *na-* :

Source	Verbe itératif
kaka *« walk » (marcher)*	**ntakaka** *« to walk again » (remarcher/marcher à nouveau)*

kana *« to yell, call » (hurler/appeler)*	**ntakana** *« to tell » (dire/raconter)*
tu'u *« word » (mot)*	**ntatu'u** *« to discuss, talk over » (discuter)*
kuni *« know » (savoir/connaître)*	**ntakuni** *« to recognize » (reconnaître)*

Tableau 56: Verbes itératifs et leurs sources lexicales

2.1.8.3 Inchoation

Il existe deux formes de préfixes inchoatifs : *ntu-* (dérivé de *ntu'u* « to become » (devenir)) et *ku-* (dérivé de la copule du potentiel *kuu*). Ils expriment la notion de transition (entrée dans un état)[314]:

Source	Verbe inchoatif
tsaa *« new » (nouveau)*	**ntutsaa** *« to renew » (renouveler/reprendre)*
va'a *« good » (bon/bien)*	**ntuva'a** *« feel better » (se sentir mieux)*
vii *« clean, beautiful » (propre, beau)*	**ntuvii** *« to become clean » (redevenir propre)*
yachi *« close » (près/proche)*	**kuyachi** *« to approach » ((s')approcher)*
kuaa *« blind » (aveugle/store)*	**kukuaa** *« to go blind » (devenir aveugle/perdre la vue)*

Tableau 57 : Verbes inchoatifs et leurs sources lexicales

2.1.8.4 Combinaisons de formes dérivatives

On remarquera qu'il existe au moins un exemple observé d'élément lexical combinant les préfixes de causalité et d'itération, et aussi que l'ordre dans lequel ils sont associés est le suivant :

[314] La publication de Mille Nieves (communication personnelle, 26 juillet 2017) est une source d'information sur les formes inchoatives.

le préfixe causal *sa-* est attaché directement à la base lexicale, et le préfixe itératif *nt-* au préfixe causal. Cette combinaison résulte probablement du fait que l'acte d'affûtage entraîne un mouvement répété, et que le résultat final est que l'objet affûté est devenu dangereux.

Source	Itération + Causalité
xeen *« dangerous »* *(dangereux)*	**ntasaxeen** *« to sharpen »* *(affûter)*

Tableau 58: Dérivation combinant causalité et itération

2.2 Remarques finales sur la description linguistique

Je le répète, la description linguistique très limitée présentée ici est très loin d'être exhaustive, et elle ne constitue pas l'objectif principal de la présente thèse (qui est de présenter les ressources de la langue MIX, le corpus, le dictionnaire et les méthodes d'annotation dans un contexte d'interconnexion entre les domaines de la documentation linguistique et des humanités numériques). Les points et caractéristiques linguistiques présentés ci-dessus, ainsi que de nombreux autres non inclus dans ce document, seront discutés en détail dans des publications ultérieures avec des analyses comparatives de phénomènes connexes présentés dans la littérature mixtèque. En outre, étant donné que l'encodage du corpus et des sources audio non annotées collectés jusqu'ici est traité, des analyses quantitatives de corpus pourront être réalisées. Voir également Bowers (sous presse) pour une discussion approfondie de la sémantique des termes de parties du corps dans la langue MIX, pour la présentation d'ensemble des principes de base concernant le relatif et nominalisateur *ña* (voir Hollenbach, 1995b, pour une discussion des fonctions parallèles dans plusieurs langues mixtèque connexes), et une introduction à la sémantique du langage spatial.

3. Origines du projet de documentation de la langue mixtèque de Mixtepec et méthodes appliquées

Comme déjà indiqué, cette thèse présente un projet qui a apporté une contribution importante à la fois en matière de documentation linguistique de la langue MIX, et pour les humanités numériques/la lexicographie numérique dans la mesure où il a permis de dépasser le

cadre d'utilisation traditionnel de la TEI. Toutefois, en raison de la façon dont ces travaux ont débuté (comme une coopération informelle dans la poursuite d'objectifs communs), des problématiques liées à la disponibilité des données sur la langue, et des aspects logistiques du travail avec des collaborateurs, jusqu'à ces dernières années, ils n'ont pas forcément été menés comme le serait un projet de documentation linguistique prototypique, car ils n'avaient pas été initialement pensés comme un projet de documentation linguistique. En outre, l'aspect technologique s'est développé à la fois sur la base des besoins analytiques (linguistiques) et sur celle des nécessités pratiques (méthode d'annotation du corpus, gestion des métadonnées, etc.), et il a donc été traité de manière ad hoc, particulièrement au début. Dans ce chapitre, je présente un bref aperçu des origines du projet et de son développement. J'analyserai ensuite, dans les parties suivantes, des points issus de publications antérieures sur les sujets qui nous intéressent, particulièrement ceux qui ont trait à la documentation linguistique et aux humanités numériques, et la façon dont ces travaux abordent des questions majeures.

Le projet de documentation de la langue MIX est né dans le cadre d'un cours sur les méthodes de de terrain suivi lorsque je préparais mon Master en linguistique à l'Université d'État de San José (Californie) en 2010, et s'est poursuivi progressivement. Jeremías Salazar, le consultant qui intervenait pendant ce semestre d'études, est originaire de la ville de Yucunani[315] dans le district de San Juan Mixtepec[316]. Il s'est ensuite installé avec sa famille à Santa Maria en Californie, qui est aujourd'hui un foyer démographique important pour les mixtèques de Mixtepec et beaucoup d'autres peuples mixtèques (voir Reyes Basurto et al., sous presse). Une grande partie de cet enseignement était axé sur des sujets tels que la phonétique, la phonologie et les principes de base de structure de l'information. Dans ce contexte, nous avons décidé avec quelques collègues, d'organiser et de réunir des enregistrements réalisés pendant de séances de consultation. La plupart d'entre eux ont été réalisés au moyen d'un enregistreur PCM linéaire Sony PCM-D50 avec une fréquence de 96 kHz/24-bit. Pour l'annotation, nous avons utilisé le logiciel Praat (Boersma et Weenik, 2020). De notre propre initiative, Jeremías, deux collègues et moi-même

[315] https://www.geonames.org/8880392/yucunani.html
[316] http://www.geonames. org/3518634/san-juan-mixtepec.html

avons continué la démarche de consultation une fois la séquence de cours achevée[317]. Dans l'année qui a suivi, Jeremías a quitté l'État, mais nous[318] avons poursuivi le travail avec son frère, Tisu'ma Salazar, qui habitait également dans la région de la baie de San Francisco. À partir de là, il est devenu mon principal consultant et collaborateur dans ce projet. Tisu'ma avait précédemment travaillé comme linguiste-consultant lorsqu'il étudiait à l'Université de Californie à Berkeley, où ont été produites plusieurs descriptions des aspects phonologiques et morphologiques de la langue (Paster, 2005, 2010; Paster et Beam de Azcona, 2004, 2005). Après l'obtention de mon diplôme en 2012, j'ai continué à travailler avec Tisu'ma.

Pendant environ trois ans de travaux (poursuivis à mi-temps à titre officieux), l'objectif principal et la portée de nos recherches ont consisté à apprendre les particularités linguistiques de la langue, en particulier la phonétique, la phonologie, la structure de l'information, ainsi que des questions liées à la sémantique, principalement la métaphore, la métonymie et la grammaticalisation. Quand j'ai commencé à m'intéresser plus en détails à ces questions, la nécessité de tenter de mettre en œuvre un système pour pouvoir stocker, annoter et retrouver l'ensemble des niveaux d'information linguistique avec leurs interfaces s'est imposée. Au même moment, après avoir discuté avec mes collègues mixtèques de notre collaboration, il est apparu clairement que le but de leur implication dans nos travaux communs était que ceux-ci devaient conduire à un résultat qui soit également susceptible d'être utile à la communauté. C'est à partir de là que nos travaux se sont orientés sciemment vers un projet de création de corpus et de documentation linguistique, et cela constituait un défi sur un certain nombre de plans.

À l'époque, je n'avais aucune formation réelle en documentation linguistique, et mon approche antérieure avait consisté à trouver des méthodes de linguistique computationnelle et de corpus pour gérer, stocker et traiter les données. Toutefois, étant donné que pratiquement chaque sous-domaine linguistique a ses propres pratiques distinctes pour stocker, annoter et rechercher

[317] Les locuteurs collaborateurs n'ont pas été rémunérés et ont participé bénévolement à ce projet. Les seules conditions « formelles » de cette participation concernaient les déplacements, dans le but précis de travailler ensemble, comme décrit plus bas.

[318] Mes deux collègues du Master de linguistique de l'Université d'État de San Jose et moi-même avons assisté aux séances de consultation volontaires jusqu'à l'obtention de notre diplôme en 2012. Après cette date, j'ai continué seul les travaux avec un locuteur collaborateur. Voir (Corpuz, 2012) pour un résultat des travaux de collaboration présenté par mon collègue Larry Corpuz Jr.

des données (bien que manifestement aucune n'ait été uniformément adoptée, ni particulièrement conviviale pour les utilisateurs), il n'existait pas de pratiques établies pour représenter la structure des données d'interfaces linguistiques ou les ambiguïtés, ni de représentation suffisante des métadonnées importantes. En outre, la plupart des approches existantes basées sur Python, comme NLTK (Natural Language Toolkit/logiciel de langage naturel) (Loper et Bird, 2002) n'étaient pas axées sur la production de données faciles à utiliser nécessaire dans le cadre d'un projet de documentation linguistique.

De plus, comme c'est souvent le cas avec les langues indigènes dotées de peu de sources primaires, les variations (phonétiques, orthographiques ou autres) étaient omniprésentes dans le jeu de données et il était important pour moi de les conserver, alors que la plupart des outils et pratiques de linguistique computationnelle ont été développés à partir de grandes langues internationales (occidentales) (notamment l'anglais, l'allemand, le français et l'espagnol). Il n'existait pas non plus de support Unicode approprié pour prendre en compte les caractères avec des diacritiques (ce qui est nécessaire avec la langue mixtèque). Il existait ainsi une lacune fondamentale dans la capacité des systèmes existants à gérer et à utiliser les données.

Dans le même temps, il devenait de plus en plus nécessaire d'aller au-delà du corpus texte intégral/séparation par tabulation que j'utilisais pour stocker le vocabulaire, et de mettre en place une structure de données plus dynamique. C'est ce qui m'a conduit à la TEI, qui a établi des modules et des directives pour l'encodage structuré à la fois de corpus de textes et de dictionnaires. En 2013, j'ai commencé à compiler un dictionnaire TEI pour stocker le vocabulaire et les informations étymologiques[319]. Alors qu'il était évident que la technologie TEI et XML constituait le meilleur choix pour mes besoins spécifiques, lorsque j'ai approfondi mon travail de création d'un dictionnaire, il est apparu que, dans de nombreux domaines, elle n'était pas suffisamment développée pour prendre en compte les types de détails et de caractéristiques que je souhaitais inclure, en particulier pour effectuer une vraie analyse étymologique[320], et pour d'autres

[319] https://github.com/iljackb/Mixtepec_Mixtec/blob/master/MIX-Lexicon-TEI-Dict.xml

[320] Comme l'un des axes majeurs de l'étude linguistique de la langue MIX était centré sur les facteurs cognitifs impliqués dans l'étymologie des termes de parties du corps, comme la métaphore et la métonymie, entre autres processus majeurs, la nécessité de mettre en place des moyens plus stables et expressifs pour encoder cette information dans la TEI a motivé les travaux décrits dans Bowers et Romary, 2016.

spécificités particulièrement pertinentes pour traiter une langue indigène dotée de peu de sources primaires (voir les détails au Chapitre 7). Ces lacunes viennent du fait que la TEI, et en particulier le module Dictionnaire, ont essentiellement été conçus pour et par des lexicographes, et non des linguistes, et qu'elle est adoptée en grande majorité pour des projets concernent des langues européenne (Bowers et Romary, 2018a).

En outre, comme je souhaitais à la fois créer une collection de ressources aussi vaste que possible, et que j'avais besoin d'accroître ma propre connaissance de la langue afin de pouvoir effectuer sans supervision la traduction, l'annotation et l'élaboration des gloses, je devais rassembler davantage de données linguistiques. Ainsi, avec l'accord de l'éditeur, des versions de livrets SIL (existants à l'origine sous le forme de fichiers PDF) encodés sur la base de la TEI ont été créés et ajoutés au corpus annoté[321]. Avec les transcriptions des enregistrements originaux, ces documents émanant du SIL représentent l'essentiel des sources textuelles dans le corpus de ce projet, et constituent aujourd'hui la plus grande partie du contenu écrit de la langue qui a été publié.

Le fait que la langue MIX soit dotée de peu de sources primaires, n'ait pas fait l'objet d'analyses linguistiques antérieures au-delà du système phonologique (voir Pike and Ibach, 1978 ; Paster, 2005, 2010 ; Paster et Beam de Azcona, 2004, 2005), et ne dispose pas de corpus, ou même d'un système orthographique bien établi, signifiait qu'il n'y avait pas d'autre moyen de traduire ou d'annoter le corpus autre qu'un traitement manuel. Comme c'est souvent le cas avec des telles langues dans lesquelles le nombre de participants potentiels est extrêmement limité (essentiellement parce que ces travaux ne bénéficiaient d'aucun financement), il y avait très peu d'approches possibles pour annoter le corpus (voir Thieberger et al., 2016). Ainsi, la démarche retenue pour le corpus de textes a été dans un premier temps de générer les traductions, puis, en attendant la disponibilité d'un ou deux collaborateurs, de les parcourir, de les corriger et de les compléter pour chaque document, selon les besoins. Des annotations plus approfondies seront ajoutées plus tard.

[321] Les sources originales sont tirées de : http://mexico.sil.org/resources/search/code/mix?sort_order=DESC&sort_by=field_reap_sortdate et les contenus encodés et annotés via la TEI sont disponibles sur : https://github.com/iljackb/Mixtepec_Mixtec/tree/master/SIL_docs

Étant donné que j'ai surtout travaillé avec un seul locuteur à une certaine période, en dehors de la communauté parlant la langue, j'avais peu d'occasions de collecter dans une mesure importante le langage parlé dans des contextes naturels. Au cours des premières années de ce projet, je me suis ainsi surtout attaché à recueillir du vocabulaire essentiellement par élicitation de traductions[322]. Même si ce n'était bien entendu pas la meilleure pratique pour la documentation linguistique (voir Himmelmann, 1998 ; Woodbury, 2003), cela a permis de collecter la plupart du vocabulaire le plus important, et pour moi plus particulièrement, d'étudier les phénomènes spécifiques qui m'intéressaient. Il y a eu quelques exceptions à cette pratique lorsque des locuteurs collaborateurs ont à certaines occasions enregistré des conversations ayant lieu dans leur vie quotidienne, ou sont allés en voyage dans la région[323].

J'ai continué à travailler sur ce projet lorsque je me suis installé à Paris (2014-2015), puis à Vienne (2015-aujourd'hui) pour des motifs professionnels. Pendant cette période, les problèmes rencontrés pour poursuivre mes travaux avec mes collègues mixtèques résidant aux États-Unis ont généré un ensemble de facteurs et de contraintes spécifiques qui ont impacté la façon dont ceux-ci ont été menés jusqu'à présent, bien qu'une communication assez régulière ait été rendue possible grâce à la messagerie mobile, aux réseaux sociaux et aux outils de visioconférence comme Skype, Google Hangouts, etc.

En 2017, grâce à des fonds octroyés par DARIAH ((Digital Research Infrastructure for the Arts and Humanities - Organisation européenne pour les sciences humaines et sociales), Tisu'ma a pu venir passer deux semaines à Vienne pour m'assister sur certains aspects du projet. En outre, durant l'été 2019, grâce à des fonds obtenus de EPHE (École Pratique des Hautes Études) et de l'INRIA (Institut National de Recherche en Informatique et en Automatique), j'ai finalement pu passer trois semaines dans la région[324], avec mes deux collaborateurs de longue date Jeremías et Tisu'ma Salazar. Nous avons séjourné chez leurs parents dans la ville de Santiago Juxtlahuaca. Tous les contenus audio obtenus pendant ce dernier voyage ont été réalisés avec un enregistreur

[322] Bien que la plupart du vocabulaire ait été obtenu par élicitation, dans l'étude des configurations spatiales, plusieurs séries d'images ont été créées à cette fin.
[323] Concernant le contenu recueilli par le biais d'enregistrements réalisés par des locuteurs, leur consentement éclairé à enregistrer diverses conversations a été obtenu pour la plupart des enregistrements (malheureusement pas pour tous), et, en raison de la mauvaise qualité du matériel utilisé, une grande partie de ces enregistrements n'était pas utilisable.
[324] D'abord dans la région de la baie de San Francisco en Californie (USA), et à Vienne (Autriche) depuis 2015.

PCM linéaire Tascam DR-05X à une fréquence de 96 kHz/24-bit [325]. L'ensemble des enregistrements réalisés et toutes les métadonnées (TEI) des contenus créés pendant ces voyage et le reste du projet, sont disponibles sur notre répertoire Dataverse sous le nom « Mixtepec Mixtec Lexical Resources » (Ressources lexicales du mixtèque de Mixtepec [326] (Bowers, Salazar, et Salazar, 2019).

Afin de fonder une base permettant de constituer un jeu de données lexicographiques le plus exhaustif possible, les travaux réalisés ne se limitent pas simplement à la documentation et au traitement de la langue MIX, et des ressources issues de variétés de mixtèque historiques associées sont intégrées au projet, notamment dans le dictionnaire. En outre, comme décrit dans Bowers, Khemakhem, et Romary (2019), en utilisant le logiciel d'OCR (reconnaissance optique de caractères) de GROBID-Dictionaries (Khemakhem et al., 2017), un dictionnaire TEI issu d'un jeu de données de mixtèque classique (mixtèque historique) [327] initialement publié par le frère dominicain Francisco Alvarado en 1593 a été créé et ajouté aux résultats du projet. L'intégration de telles ressources fournit une source riche de données historiques comparatives, qui non seulement améliore la qualité du dictionnaire mixtèque de Mixtepec, mais peut être réutilisée par les personnes qui travaillent sur d'autres variétés de mixtèque.

4. Interactions et divergences de la documentation linguistique, de la description linguistique, des humanités numériques et de la linguistique de corpus

Étant donné que ces travaux sont à l'interface entre de nombreux sous-domaines comme les humanités numériques/la lexicographie digitale, la documentation linguistique, la linguistique de corpus, notamment, il existe un large éventail de publications issues de ces divers domaines qui concernent différents aspects de ces travaux, mais il y en a très peu couvrant chacun des aspects essentiels. L'une des nécessités fondamentales de tout projet de documentation linguistique est de fournir un ensemble documenté de données langagières primaires, accompagnées des informations lexicales relatives, potentiellement, à tous les niveaux du langage (par exemple

[325] Comme nous le verrons dans les parties qui suivent, les métadonnées de tous les fichiers multimédias créés indiquent l'équipement spécifique utilisé pour les enregistrements, la méthode d'élicitation et plusieurs autres facteurs importants.
[326] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/BF2VNK
[327] Le mixtèque classique est également appelé « mixtèque colonial ».

phonétique/phonologie, morphosyntaxe, sémantique, informations de lexique ou dictionnaire, etc.), souvent avec des transcriptions et annotations (par exemples des textes avec gloses interlinéaires). En outre, il est impératif d'organiser les données, d'en permettre l'accès, de les publier, et d'analyser l'information, c'est-à-dire de garantir leur réutilisation optimale, et si possible, une vérification empirique dans les règles de l'art, en utilisant dans l'idéal les normes sur les données (Bird et Simons, 2003b ; Thieberger, 2010, 2012, 2014 ; Gawne et Berez-Kroeker, 2018). Ces points revêtent bien entendu la même pertinence pour tout projet multiforme concernant la linguistique, la lexicographie et/ou la linguistique de corpus (la distinction entre ces disciplines pouvant dans certains cas être assez arbitraire) (voir Cox (2011) pour une discussion approfondie du chevauchement et des divergences entre linguistique de corpus et documentation linguistique). Ce vaste champ d'application présente des défis technologiques et logistiques particulièrement complexes en termes de logiciels, formats de données, balisage et flux de travail.

Dans cette partie, je traite des points clés, principes et fondements théoriques issus des publications essentielles relatives aux différents domaines qui sont au cœur de ces travaux, notamment : les humanités numériques, la documentation linguistique, la description linguistique, l'interaction entre humanités numériques et documentation linguistique, la conception et la gestion des données, les meilleures pratiques et questions éthiques en documentation linguistique, ainsi que des problématique liées au traitement des langues dotées de peu de sources primaires.

4.1 Documentation linguistique et humanités numériques

Les humanités numériques présentent une singularité dans la mesure où elles ne recouvrent pas réellement un seul domaine, mais sont plutôt un moyen de traiter, d'encoder, d'annoter et de présenter des travaux portant sur différents sujets des sciences humaines (comme l'histoire, la littérature, la linguistique, la lexicographie, etc.), ces travaux étant généralement réalisés dans, lors de ou par des départements universitaires, bâtiments, conférences et revues distincts. Les humanités numériques ont évolué pour former un domaine séparé, mais multidisciplinaire, principalement parce qu'au sein des limites traditionnelles des cultures et pratiques académiques de chacun des domaines des sciences humaines, l'utilisation d'outils technologiques n'était pas privilégiée que ce soit dans l'enseignement ou dans leurs programmes et départements institutionnels respectifs.

La numérisation des données héritées et la création de nouvelles données *d'origine numérique* sont cruciales pour leur préservation et leur réutilisation, et augmentent de manière exponentielle les capacités de recherche, d'extraction et d'analyse des matériaux sources, aussi bien au profit des chercheurs que des divers publics potentiels. Dans les sciences humaines, il est fréquent que le contenu d'une source soit pertinent pour plusieurs domaines (par exemple en littérature historique, épigraphie et numismatique), alors que toutes les études spécialisées sont, en tant que telles, également des sources primaires majeures de données linguistiques historiques. Ainsi, leurs contenus et analyses, de même que leur provenance, etc., sont tous potentiellement pertinents pour les linguistes historiques, voire les historiens et les anthropologues, entre autres. Sur ces bases, les personnes travaillant dans ces domaines ont nécessairement dû chercher à développer et à échanger des méthodes et des connaissances à partir des diverses technologies existant en dehors de leurs propres domaines d'intervention, pour permettre également la création de normes et standards permettant l'échange et l'analyse de jeux de données numériques.

De la même manière, la documentation linguistique est fondamentalement pluridisciplinaire. En effet, selon Himmelmann (1998), les orientations de la documentation linguistique vont nécessairement bien au-delà de celles des sous-disciplines de la description/de l'analyse linguistique, étant donné qu'elles peuvent concerner aussi bien :

- les approches sociologiques et anthropologiques du langage (sociolinguistique variationniste, analyse conversationnelle, anthropologie linguistique et cognitive, contact linguistique, etc.) ;
- la linguistique « pure » (théorique, comparative, descriptive) ;
- l'analyse du discours, la recherche sur le langage parlé, la rhétorique ;
- l'acquisition du langage ;
- la phonétique ;
- l'éthique, les droits linguistiques et la planification linguistique ;
- la méthodologie de terrain ;
- la littérature et l'histoire orales ;
- la linguistique de corpus ;

- la pédagogie des langues.

Austin (2013) ajoute à cela :

- l'ethnographie
- la psychologie
- la bibliothéconomie
- l'archivage
- les arts médiatiques, les arts et sciences de l'enregistrement
- la pédagogie.

4.1 Documentation linguistique et humanités digitales

En outre, Himmelmann (1998) indique que le défi théorique majeur rencontré par les linguistes en documentation linguistique est d'arriver à synthétiser un cadre cohérent à partir de l'ensemble des disciplines listées précédemment, ce qui est également un point central des humanités digitales (voir Penfield (2014) pour une vue approfondie des problématiques clés rencontrées dans les études interdisciplinaires universitaires). Même si les projets portant sur la lexicographie et la linguistique[328] ne sont pas rares, ce qui l'est davantage est de parler de documentation linguistique[329] dans le contexte des humanités digitales. De plus, ceux qui travaillent dans le domaine de la documentation linguistique considèrent rarement que cela entre dans le champ des humanités digitales. Mais cet état de fait est en train d'évoluer avec la tendance actuelle qui voit les méthodes mises en œuvre en documentation linguistique s'aligner sur les objectifs et les approches qui sont au cœur des humanités digitales, en mettant notamment l'accent sur les possibilités de réutilisation, la compatibilité et l'évolutivité, et sur la capacité à répliquer la recherche et les données de recherche (Bird et Simons, 2003b; Thieberger, 2010,

[328] Il est en fait plus rare qu'un projet relevant des humanités digitales soit qualifié de « linguistique », étant donné que lorsqu'il fait appel à des méthodes digitales, ce domaine se décrit généralement lui-même comme étant de la « linguistique computationnelle » ou de la « linguistique de corpus ».

[329] Pourtant, les travaux réalisés dans nombre de projets de dialectologie européens sont très similaires, sur de nombreux plans, à la documentation linguistique (voir Bowers et Stöckle (2018) qui donne un exemple de travaux effectués dans le domaine des humanités digitales sur des dialectes bavarois en Autriche dans le cadre du projet d'héritage culturel à long terme « Datenbank der bairischen Mundarten in Österreich » (Banque de données des dialectes bavarois en Autriche)).

2012, 2014 ; Gawne et Berez-Kroeker, 2018). Bird et Simons (2003b) constitue une publication phare citée aussi bien dans le contexte des humanités digitales que dans celui de la documentation linguistique. Elle traite des questions majeures de la documentation et de la description linguistiques en matière de contenu, de format, de découverte, d'accès, de citation, de préservation et de droits. Alors que le public cible concerné par ces travaux et ce sujet était composé des chercheurs en documentation et description linguistiques, de nombreux principes et problématiques exposés dans cette publication s'appliquent également dans la pratique des humanités digitales en général.

En matière de documentation linguistique assistée par la technologie, deux projets déterminants reflètent clairement le lien intrinsèque entre les humanités digitales et la documentation linguistique : le projet DOBES (« Dokumentation bedrohter Sprachen ») (Documentation des langues menacées) (2000-2011)[330], qui a abouti à la création d'une archive sur des langues en péril, et le projet E-MELD (« Electronic Metastructure for Endangered Languages Documentation » (Métastructure électronique pour la documentation des langues en péril) (Boynton et al., 2006)[331]. Leur objectif était d'identifier les problématiques principales et de faire des recommandations visant à mettre en œuvre les meilleures pratiques pour traiter les points clés relevant à la fois de la documentation linguistique, des humanités digitales et aussi de la linguistique de corpus, afin de faciliter les processus et d'accroître la durabilité et l'interopérabilité du résultat produit. Les sujets traités couvrent : les formats de données et d'archivage, les métadonnées, l'annotation, l'analyse, les standards, les outils, les flux de travail et la gestion[332].

[330] http://dobes.mpi.nl/ (consulté le 31/12/2019)
[331] http://emeld.org/ (consulté le 31/12/2019)
[332] Parmi d'autres projets importants dans le cadre du développement des meilleures pratiques et de la documentation linguistique comme un champ d'intervention distinct, on peut citer le programme de documentation des langues en péril « Endangered Languages Documentation Programme (ELDP) » (2002-aujourd'hui) (https://www.eldp.net/), et l'initiative interorganisations des États-Unis « Documenting Endangered Languages (DEL) » (Documentation des langues en péril) de la National Science Foundation (Fondation nationale américaine pour les sciences) et du National Endowment of the Humanities (Fonds de dotation national américain pour les sciences humaines) https://www.nsf.gov/pubs/2005/nsf05590/nsf05590.htm) (2005-2020).

4.2 Description linguistique versus documentation linguistique

Un point majeur à clarifier est la distinction entre la collecte, la description et l'analyse de *données primaires*, l'objectif de la documentation étant l'enregistrement et la production d'enregistrements de la langue naturelle parlée, et la description linguistique un simple sous-produit (Himmelmann, 1998, 2006 ; Austin, 2006 ; Woodbury, 2003 ; Mous, 2007 ; Good, 2011). Plus fondamentalement, le but principal de la documentation linguistique est la collecte de données, leur représentation et leur diffusion via la production de grammaires et de dictionnaires, la création de nouveaux matériels linguistiques, ainsi que l'annotation et les analyses étant secondaires.

Étant donné que le public cible d'un projet de documentation linguistique est potentiellement beaucoup plus large et inclut (en particulier) des membres de la communauté, des chercheurs dans d'autres domaines, des anthropologues, des ethnologues, etc., les spécialistes en documentation linguistique sont confrontés au défi majeur consistant à développer un cadre cohérent ou un ensemble de principes pour recueillir et représenter le contenu relevant de toutes ces disciplines différentes d'une façon qui ne soit pas susceptible d'exclure un domaine ou un objectif au profit des autres.

> *« Une séparation claire de la documentation et de la description permet de garantir que la collecte et la présentation des données primaires reçoivent l'attention théorique et pratique qu'elles méritent. » (Himmelmann, 1998, p.164)*

La distinction la plus fondamentale entre documentation et description linguistiques est peut-être le rôle des données en lien avec les objectifs et motivations des travaux : alors que, comme décrit plus haut, l'objectif de la première est la création de supports bien documentés et d'autres ressources linguistiques primaires en vue de la préservation et de la réutilisation, la seconde vise quant à elle essentiellement à produire des analyses grammaticales et (dans certains cas) des dictionnaires, son public cible privilégié étant les linguistes, qui l'utilisent comme aide pour mener certaines analyses linguistiques (Himmelmann, 1998, 2006 ; Woodbury, 2003 ; Austin, 2006 ; Austin et Grenoble, 2007).

Cette distinction marquée entre les deux domaines a été contestée dans Nathan et Austin (2004), et Austin et Grenoble (2007), qui soutiennent que la création d'une documentation qui soit la plus exploitable, qualitative et complète possible (sous la forme de « points d'entrée » multiples comme des transcriptions, traductions et annotations) dépend nécessairement de l'analyse linguistique, et que celle-ci est absolument nécessaire pour découvrir et évaluer les contenus lexicaux de l'ensemble de ressources issues de la documentation. Himmelmann (2006, 2012) indique lui-même que, même si la documentation et la description linguistiques peuvent être séparées assez clairement sur les plans méthodologique et épistémologique, cela n'implique pas forcément qu'elles puissent, ou doivent, effectivement être séparées dans la pratique. L'analyse linguistique est par exemple nécessaire pour identifier et déterminer la présence, ou l'absence, dans les sources de styles de discours, formes lexicales, paradigmes, constructions de phrase, etc. fondamentaux. Lorsque l'analyse est nécessaire pour des tâches de cette nature, il est indispensable de documenter les caractéristiques et les bases de leur identification et de leur traitement, par exemple en segmentant les périodes et particularités linguistiques qui sont susceptibles d'affecter la signification de base, etc. La documentation de ces questions revient à faire de la description linguistique, et cela a des incidences à la fois sur le plan de la découverte et sur celui de la réutilisation potentielle.

Si la distinction entre documentation et description linguistiques relève d'une différence dans l'angle d'approche entre les données primaires (par exemple les enregistrements audio/vidéo, les transcriptions, etc.) pour l'une, et les résultats et ressources analytiques (par exemple dictionnaires, grammaires et analyses) pour l'autre, dans la plupart des cas, il est probable qu'un projet portant sur la documentation comprendra aussi une partie description sous une forme ou une autre (Good, 2011). Comme indiqué au point 3, cela se vérifie dans les présents travaux. En effet, cette étude linguistique était motivée au début par l'envie de connaître la langue, et la production d'un dictionnaire et d'un corpus digital était initialement considérée comme un moyen d'arriver à cette fin. Dans un second temps seulement, les travaux ont été sciemment poursuivis sous la forme d'un projet de documentation linguistique, mais avec pour objectif de produire des ressources susceptibles d'être utilisées par la communauté des locuteurs.

4.3 Ressources linguistiques et données

Les ressources utilisées en documentation linguistique sont de facto des corpus de langues avec peu de sources primaires et/ou de langues à l'étude, ce qui implique forcément qu'elles diffèrent des corpus de langues majeures en termes d'objectifs, de production, de contenu, de sources et d'ampleur (Mosel, sous presse). Les aspects spécifiques de ces différences varient bien entendu en fonction de la situation et de l'histoire de la langue considérée, mais, alors qu'une langue majeure va de toute évidence disposer d'un éventail complet de ressources écrites et parlées préexistantes qui serviront dans toutes sortes de domaines et de registres, d'un groupe beaucoup plus vaste de locuteurs et d'un ensemble de sources qui augmentent de manière naturelle, certains projets de documentation linguistique peuvent n'avoir absolument aucune ressource préexistante de quelque mode ou genre que ce soit. Les sources de données peuvent ainsi être sporadiques, et un groupe de sources diverses non uniformes peut même potentiellement comprendre l'intégralité des ressources linguistiques existantes pour une langue donnée.

Dans un projet de documentation linguistique type, la principale source de contenu sera vraisemblablement composée de fichiers d'enregistrements audio ou vidéo réalisés auprès de locuteurs natifs. Ces fichiers sont ensuite transcrits dans un format aligné dans le temps à l'aide d'un logiciel de type Praat, ELAN (Brugman et Russel, 2004), or EXMARALDA (Schmidt et Wörner, 2009). Outre l'audio ou la vidéo, on peut disposer de textes intégrés dans un corpus, sous la forme de documents originaux écrits par des locuteurs, ou de sources préexistantes de toute nature.

Il existe une autre différence majeure dans la finalité : alors que les corpus des langues majeures sont universellement utilisés en linguistique, et pour d'autres niveaux de recherche éventuels (voire l'apprentissage d'outils technologiques), les corpus de documentation linguistique peuvent servir à des fins très diverses, notamment comme héritage culturel et linguistique, matériel pédagogique, ainsi que pour la recherche. En outre, dans les cas où le projet repose sur la création de collections de ressources linguistiques (création de corpus), en particulier quand il s'agit de langues indigènes ou menacées, on se heurte à des défis majeurs qui concernent : a) la création du contenu original (séances de consultation, etc.) ; b) l'accumulation

de ressources provenant de sources externes pour la création du corpus ; c) l'intégration de ces ressources dans des formats de données courants de façon à ce qu'elles puissent être consultées à partir d'une interface de recherche commune, et leur production éventuelle dans un format de présentation destiné à être utilisé par la communauté. Sur le plan des données, les facteurs clés pour relever ces défis sont l'interopérabilité, l'échange, les standards, ainsi que les outils.

Pour finir, la création et la gestion des métadonnées sont d'une importance cruciale, à la fois à court et à long terme à des fins d'archivage et de préservation, et sur le plan de la recherche, de la réutilisation, de l'analyse, etc.

4.4 Standards et outils

Outre le choix du bon formatage des fichiers, les standards de données constituent un élément de plus en plus essentiel en matière de métadonnées, de balisage des corpus, de ressources descriptives comme les dictionnaires, d'annotation des corpus, et de descriptions et inventaires grammaticaux. Selon Romary (2011), la normalisation des données devrait permettre de stabiliser les connaissances contenues dans les données, et de les structurer de façon à prévenir d'éventuels blocages en travaillant à l'avenir avec des données normalisées. De plus, l'utilisation de standards de données facilite leur échange entre les utilisateurs et les outils, et permet aux utilisateurs de profiter du fait qu'elles sont déjà documentées, ce qui leur évite de perdre du temps pour concevoir et décrire leur propre système de balisage (Romary, 2011). Dans la version intégrale de cette thèse, je parle des standards, des formats spécifiques aux outils et des questions liées à la compatibilité et à l'échange de données pour tous les aspects principaux de la documentation linguistique et de la lexicographie, notamment :

- des métadonnées (OLAC, IMDI, TEI, AILLA, CMDI) ;
- de la transcription du langage parlé (Praat, ELAN, EXMARaLDA, ISO 24624:2016) ;
- du balisage des corpus (XML, TEI) ;
- de l'annotation des corpus (annotation à distance, annotation en ligne, bases de données relationnelles) ;
- des ressources descriptives (FLEx; ELAN, TEI) ;
- des dictionnaires (FLEx, TEI, LMF, Toolbox) .

- des descriptions grammaticales (FLEx, TEI, ELAN, ISO).

5. Aperçu général des publications et ressources mixtèques

L'objectif majeur est d'intégrer toutes les sources mixtèques et MIX pertinentes dans la collecte des données afin d'élaborer une base aussi exhaustive que possible pour les travaux actuels et futurs en lexicographie et documentation culturelle MIX.

5.1 Manuscrits

La source la plus ancienne de mixtèque écrit est bien entendu constituée des manuscrits écrits sous la forme de pictogrammes indigènes représentés principalement sur des toiles en peau de daim. Malheureusement, beaucoup d'autres ont probablement été détruits par les missionnaires espagnols. Les exemplaires subsistants ont été pillés et ramenés en Europe, puis sont passés entre les mains de différents nobles et monarques avant de finir dans les musées et bibliothèques dans lesquels ils se trouvent aujourd'hui. Ce phénomène, auquel s'ajoute le fait que ces trésors culturels portent quasiment tous le nom des Européens qui les ont achetés ou qui les possèdent actuellement, a créé un fossé entre le peuple mixtèque, dont les ancêtres sont à l'origine de ces documents, et les experts et institutions qui les détiennent (Jansen et Pérez Jiménez, 2004). Dans la version intégrale de cette thèse, j'évoque les propositions faites par Jansen et Pérez Jiménez (2004) pour changer les noms de ces documents par des termes mixtèques (essentiellement de la variété classique) dérivés de leur contenu, afin que le peuple mixtèque puisse récupérer le patrimoine culturel qui lui revient de droit.

5.2 Mixtèque colonial

L'utilisation la plus ancienne et la plus marquante de mixtèque écrit phonétiquement se retrouve sans surprise dans le contexte religieux. Pendant la période coloniale, les premières sources de vocabulaire mixtèque en général (sans parler des manuscrits pictographiques) sont la « Doctrina en Lengua Mixteca » (Doctrine en langue mixtèque) du frère Benito Hernández de la ville de San Miguel Achiutla (Ñuu Ndecu) (1567), et une autre provenant de Teposcolula (1568), qui sont les premiers documents présentant le catholicisme au peuple mixtèque (Hollenbach,

2016). Les ressources mixtèques primaires remontant à cette période sont le « *Vocabulario en lengua mixteca* » (Vocabulaire de langue mixtèque) issu de Alvarado, 1593, et la grammaire « *Arte en lengua mixteca compuesta* » (Art de langue mixtèque composée) élaborée par le frère Antonio de los Reyes (1593), qui concernent toutes les deux la variété Tepozcolula.

Selon Hollenbach (2016), il existe également divers autres manuscrits et documents d'archives qui proviennent quasiment tous de la région Mixteco Alto (« mixtèque des montagnes »). Il y a très peu de sources issues des régions du Mixteco Bajo (« mixtèque des plaines »), et aucune de la région du Mixteco de la Costa (« mixtèque de la côte »). Plus tard dans la période coloniale, le catéchisme de Ripalda a été publié une première fois en 1719, puis de nouveau en 1755 (Ripalda, 1755). À la fin de la période coloniale, l'utilisation du mixtèque écrit a cessé dans la région du Mixteco Alto, même si plusieurs catéchismes ont été publiés entre 1834 et 1899 dans des variétés de Mixteco Bajo (« mixtèque des plaines »). Ces documents constituent une ressource historique en grande partie inexploitée pour des travaux ultérieurs de linguistique historique et autres études. Le projet Ticha[333] (Allen et al., 2016 ; Lillehaugen et al., 2016 ; Broadwell et al., sous presse), dans lequel des textes zapotèques historiques (religieux, linguistique, testaments, actes de vente, etc.) datant de la période coloniale ont été numérisés, transcrits, traduits et présentés sur une plateforme en ligne basée sur Omeka[334], qui inclut des éditions digitales parallèles et permet le crowdsourcing (externalisation ouverte ou production participative), constitue une feuille de route éventuelle sur la façon dont il serait possible d'utiliser et de présenter ces ressources historiques.

5.4 Bref aperçu des publications linguistiques mixtèques

La plus ancienne étude linguistique moderne du mixtèque a été entreprise sur la variété San Miguel el Grande (ISO 639-3: mig) par Kenneth Pike du Summer Institute of Linguistics ou SIL (Institut d'été de linguistique) dans les années 1930. Cornelia Mak a publié des recherches sur la langue MIG et les variétés parlées à San Esteban Atatláhuca (ISO 639-3: mib) et Santo Tomás Octopec (ISO 639-3: mie), ainsi que des études comparatives des systèmes de tons des variétés MIG et MIB en 1953, et des variétés MIG, MIB et MIE en 1958 (Mak, 1953, 1958).

[333] https://ticha.haverford.edu/
[334] https://omeka.org/

Dans sa thèse de doctorat, Robert Longacker a proposé un système de proto-mixtèque fondé en grande partie sur les données comparatives fournies par Mak (Longacre, 1957), et, en 1960, Mak et Longacre ont co-écrit une analyse révisée prenant en compte des données complémentaires collectées dans d'autres variétés de mixtèque (Mak et Longacre, 1960). En 1961, Longacre et René Millon ont proposé un système de proto-mixtèque-amazugo qui rassemble des données comparatives reliant les deux sous-branches étroitement apparentées de la famille des langues otomangues. D'autres reconstitutions du proto-mixtèque ont été publiées sur la base de données mixtèques comparatives par Josserand (1983), qui a présenté une description approfondie de la typologie dialectale du mixtèque. Pour finir, Dürr (1987) présente une reconstitution du système tonal. Ces publications, notamment celle de Josserand (1983), sont particulièrement importantes pour le domaine de la linguistique historique et comparative du mixtèque.

Alors que cette thèse mentionne beaucoup de publications individuelles sur différentes variétés de mixtèque, les études de Brugman et Macaulay sur le mixtèque Chalcatongo (Brugman, 1983 ; Brugman et Macaulay, 1986 ; Macaulay, 1982, 1985, 1987a,b, 1990, 1993, 1996, 2005, 2011, 2012 ; voir aussi Macaulay et Salmons, 1995) sont importantes d'une part pour la profondeur de la couverture linguistique d'une variété mixtèque, d'autre part sur le plan des origines et de la méthodologie. Comme l'a souligné McKendry (2013), elles sont représentatives d'un nouveau développement dans l'étude des langues mixtèques car les consultants sur le projet (tout au moins au début) étaient des membres d'une communauté expatriée résidant en Californie, ce qui leur a permis de mener initialement les recherches hors de la région d'origine des locuteurs.

5.4.1 Autres projets relatifs au mixtèque

Il existe plusieurs initiatives particulièrement importantes qui œuvrent dans l'intérêt de la communauté mixtèque au sens plus large et d'autres communautés indigènes sur la côte centrale de la Californie (même si leur champ d'application va bien au-delà de la documentation linguistique). L'une de ces organisations est le *Mixtec/Indígena Community Organizing Project*

(MICOP) (Projet d'organisation de la communauté mixtèque/indigène)[335], qui est dirigé par des indigènes et assure divers fonctions dans la communauté mixtèque et dans d'autres communautés immigrées dans le comté de Ventura en Californie. Son action vise à développer le leadership et l'indépendance communautaire, l'enseignement, la lecture, les services de santé, ainsi que différents programmes de formation professionnelle et de promotion de la culture. L'organisation MICOP exploite en outre la station de radio *Radio Indigena*[336], qui diffuse des émissions en langues indigènes, dont différentes variétés de mixtèque. MICOP travaille de concert avec le département de linguistique de l'Université de Californie à Santa Barbara (UCSB) pour créer des programmes collaboratifs communautaires dont l'objectif est de favoriser le maintien des langues, l'alphabétisation mixtèque et la justice sociale, qui sont désignés collectivement sous l'appellation de « Mexican Indigenous Language Promotion and Advocacy project (MILPA) » (Projet de promotion et de défense des langues indigènes mexicaines)[337] (Bax et al. 2019 ; Campbell et Bucholtz, 2017 ; Hernández Martínez et al., sous presse). Dans ce contexte, des membres de la communauté participent à des cours universitaires de linguistique à la UCSB, et collaborent pleinement aux analyses linguistiques et à d'autres activités liées à la méthodologie de terrain, comme des analyses phonologiques, la transcription du langage parlé, la réalisation d'enregistrements audio et vidéo, la traduction, l'écriture d'une grammaire, l'archivage, etc. (Bax et al. 2019 ; Campbell et Bucholtz, 2017 ; Hernández Martínez et al., sous presse). Ce programme va en particulier aboutir en 2019-2020 à une grammaire du mixtèque de Mixtepec, qui est actuellement en cours d'élaboration (Salazar et al., 2020).

De nombreuses initiatives ont vu le jour sur internet et les réseaux sociaux, et ont été de plus en plus actives dans la production de nouveaux contenus. Conocelos (http://conocelos.mx/inicio/) est un projet communautaire mené par un groupe de locuteurs de langues indigènes (incluant plusieurs variétés de mixtèque) au Mexique, qui travaille à la création d'un outil permettant des traductions entre langues indigènes, et à la constitution d'un ensemble de ressources comme des contes et récits, ainsi que des recueils de vocabulaire. Le

[335] http://mixteco.org/about-us/
[336] http://mixteco.org/radio/
[337] Les travaux réalisés dans le projet MILPA sont associés à la subvention 1660355 de la NSF (Fondation nationale américaine pour la science). https://nsf.gov/awardsearch/showAward?AWD_ID=1660355&HistoricalAwards=false

figure 57 représente une publication récente de Conocelos pendant l'épidémie de Covid-19 au printemps 2020[338]:

Figure 144: Conseil de santé publique « Stay at home » (Restez chez vous) en langue MIX dans le cadre de l'épidémie de Covid-19

Une autre initiative intéressante est la création de la page Facebook « Tu'un Savi » (https://www.facebook.com/tuunsavi20/), qui publie des schémas accompagnés de vocabulaire, et souvent aussi des vidéos de différentes variétés de mixtèque, dont le mixtèque de Mixtepec. Ces vidéos sont, dans bien des cas, également partagées sur YouTube.

Un autre projet récent en cours est le projet Mesolex[339] (Lexicosemantic Resources for Mesoamerican Languages - Ressources lexico-sémantiques pour les langues mésoaméricaines), qui n'est pas spécifique au mixtèque, mais dans lequel les variétés mixtèques forment une partie importante de l'ensemble de données et des langues cibles. L'élément principal de Mesolex est un portail composé de deux modules dont le but est d'incorporer et de diffuser des bases de données lexicales, dont des dictionnaires, pour cartographier les structures de données des matériaux sources en données et métadonnées TEI. Ce projet permettra également d'intégrer du contenu audio et vidéo pour les ressources linguistiques indigènes déposées sur le portail.

[338] Il convient de noter que les ressources relatives au Covid-19, qui ont été créées dans diverses variétés de mixtèque, pourraient être une source de termes apparentés très riche pour la constitution future d'un vocabulaire comparé.
[339] Subvention DEL « Documenting Endangered Languages » (Documentation des langues en péril) n° HAA-266482-19) https://securegrants.neh.gov/publicquery/main.aspx?f=1&gn=HAA-266482-19

5.5 Publications sur le mixtèque de Mixtepec

La première étude d'un aspect du mixtèque de Mixtepec a été réalisée par Pike et Ibach (1978), qui ont décrit l'inventaire phonétique et phonologique. Entre 2004 et 2010, Mary Paster et Rosemary Beam de Azcona ont publié une série d'articles sur la phonologie et la morphologie de la langue, et sur le rôle du ton lexical (dans Paster et Beam de Azcona (2004, 2005) et Paster (2005, 2010)). Le consultant principal dans ces recherches était également l'un des deux consultants/collaborateurs majeurs dans les travaux décrits ici. Paster et Beam de Azcona ont décrit la variété linguistique comme « Yucunani Mixtec » (mixtèque Yucunani), plutôt que mixtèque de Mixtepec. Alors qu'il n'existe, en dehors du dictionnaire TEI (Bowers et Romary, 2018), pas d'autre dictionnaire de mixtèque de Mixtepec, le dictionnaire « Vocabulario Básico Tu'un Savi-Castellano » (Vocabulaire de base Tu'un Savi-Castellano) (Galindo Sánchez, 2009) a été créé pour la variété de mixtèque parlé à Veracruz par des descendants d'une communauté de migrants venus de San Juan Mixtepec dans les années 1940.

Nieves (2012) parle du discours solennel (« *El Parangón* » - Le parangon) que l'on retrouve dans certaines cérémonies civiles et religieuses, et qui contient plusieurs procédés rhétoriques intéressants, comme des parallélismes, métaphores, métonymies, ainsi que d'autres utilisés dans le langage rituel.

Pour finir, comme déjà mentionné, Bowers (sous presse) présente une étude approfondie des termes de parties du corps dans la langue mixtèque de Mixtepec (désignés par l'abréviation « BPT » pour « body-part terms »), selon laquelle, en accord avec la théorie de personnification (Lakoff et Johnson, 1980a,b ; Johnson, 1987) il existe un vaste réseau de sens élargis utilisés pour le terme principal d'un mot composé, dans des expressions comprenant plusieurs mots et dans des formes polysémiques, qui sont apparus dans la langue via des métaphores et des métonymies selon des processus d'innovation lexicale et de grammaticalisation. Ces extensions de sens s'appliquent à des termes partitifs (de type partie-tout) pour des objets (méronymie), des relations spatiales, des concepts relationnels avec des niveaux d'abstraction différents, et à des fonctions grammaticales. Bowers (sous presse) apporte des exemples probants aux questions discutées pour des variétés apparentées de mixtèque (Brugman, 1983 ; Brugman et Macaulay, 1986 ; Hollenbach, 1995 ; Langacker, 2002), et propose en outre plusieurs extensions qui

n'avaient pas été observées précédemment, ainsi qu'une description plus fine des sources cognitives et conceptuelles expliquant ces phénomènes. La partie centrale de ces travaux est l'étude détaillée des sources de connaissance schématiques pour les termes de parties du corps élargis, des stratégies lexicales et cognitives à l'origine de certaines évolutions sémantiques, et de la directionnalité diachronique, au niveau tant sémantique que grammatical de la langue.

6. Corpus : encodage, annotation et contenus

Dans cette partie, je présente un inventaire des composants principaux du corpus, et une description des outils et techniques de formatage utilisés, et je donne un aperçu des typologies de documents/ressources importantes. La description de ces typologies de ressources et ma démarche pour les intégrer dans le corpus sont particulièrement pertinentes dans la mesure où elles représentent un vaste éventail de ressources lexicales, que l'on peut, pour l'une ou plusieurs d'entre elles, retrouver dans tout projet de documentation linguistique[340]. Il convient de remarquer que le processus d'annotation est toujours en cours, et qu'au moment de la présentation de ces travaux, les structures d'annotation qui vont être décrites ici ne seront ainsi pas complètement appliquées à toutes les ressources.

6.1 Répertoire audio et vidéo

Les ressources de langage parlé intégrées dans ce projet comprennent :

- des enregistrements et vidéos (réalisées avec ou par des collaborateurs du projet) ;
- des enregistrements et vidéos trouvés sur internet ;
- des transcriptions de langage parlé n'ayant pas fait l'objet d'enregistrements.

L'intégralité des enregistrements audio et vidéo créés au cours de ces travaux (pour lesquels un consentement éclairé écrit a été obtenu) ont été publiés dans une archive intitulée : *Mixtepec Mixtec Language Resources* (« Ressources sur la langue mixtèque de Mixtepec » sur Harvard Dataverse (Bowers, Salazar, et Salazar 2019)[341].

[340] Mais, étant donné qu'un nombre potentiel quasiment incalculable de sources ont été, et continuent à être acquises et intégrées dans le projet, le recensement définitif des ressources et les pratiques de formatage restent sujets à des modifications.
[341] https://doi.org/10.7910/DVN/BF2VNK

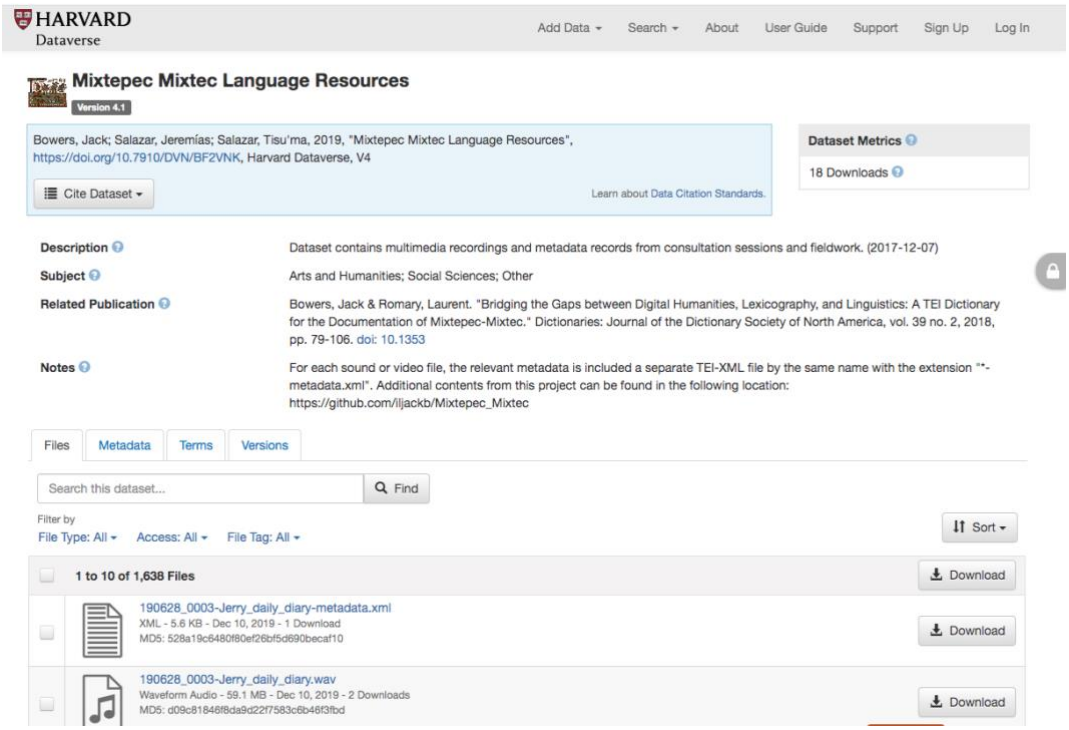


Figure 145: Capture d'écran de l'archive ou des fichiers multimédias MIX sur Dataverse

Au moment de la présentation des travaux, il existe 837 fichiers audio, 5 vidéos et, pour chacune de ces sources, des dossiers de métadonnées TEI dans lesquels sont enregistrées des données clés. Chaque fichier (à la fois multimédias et de métadonnées) peut être téléchargé librement, et possède un DOI (Digital Object Identifier, « identifiant d'objet numérique ») unique, chacun pouvant donc être cité individuellement. La nécessité de mettre en place une identification des jeux de données capable de persister sur le long terme est l'un des principes fondamentaux de Harvard Dataverse (King, 2007)[342]. Les ressources et infrastructures telles que Dataverse sont un moyen de reconnaître tous les aspects des travaux scientifiques et académiques, et leur conception conviviale pour l'utilisateur réduit les obstacles susceptibles d'entraver leur dépôt et à l'accès aux données. En outre, le fait qu'il s'agisse de ressources publiées légalement et accompagnées d'informations de référence clairement indiquées (du

[342] Alors que pour l'instant, les enregistrements réels, les vidéos, quelques notes de terrain concernant les séances de consultation et des fichiers TEI contenant les métadonnées pertinentes constituent le seul contenu archivé par l'intermédiaire de Dataverse, des contenus additionnels comme des transcriptions et tous les fichiers du corpus pourront être ajoutés ultérieurement.

moins dans le cas de Dataverse) constitue, pour les chercheurs, une motivation professionnelle supplémentaire pour rendre leurs travaux publics et accessibles[343].

Le service de répertoire Harvard Dataverse génère automatiquement des métadonnées pour l'archive des ressources lexicales du mixtèque de Mixtepec en formats DCMI, OAI_ORE, Schema.org JSON, et dans plusieurs autres formats. Toutefois, il n'en génère pas pour les fichiers TEI proprement dits qui sont déposés dans ce système, qui sont destinés seulement à documenter les métadonnées clés pour les fichiers de ressources lexicales les accompagnant (actuellement principalement les enregistrements audio et vidéo). Ce dernier type de métadonnées, avec ses exemples spécifiques contenus dans le présent projet (ou dans tout autre), est bien entendu le plus important, et l'existence des schémas de métadonnées OLAC et IMDI a pour unique objectif de permettre leur expression. Ainsi, comme discuté au paragraphe 4.4.1.6, il est d'une importante fondamentale de définir les correspondances entre ces trois systèmes, à la fois pour les communautés intéressées dans le présent et le futur, et, dans la perspective de ce projet, pour produire le résultat les plus optimal possible sur le plan des bonnes pratiques évoquées dans cette partie.

6.2 Ressources linguistiques dans le corpus

Les sources textuelles intégrées dans le corpus TEI sont issues :

- des livrets et articles publiés par le Summer Institute of Linguistics (SIL - Institut d'été de linguistique) (*à peu près 27 000 tokens*) ;
- de documents écrits créés par des locuteurs dans le cadre de ce projet ;
- de transcriptions de langage parlé (conversion à partir du logiciel Praat)
- de documents sur le mixtèque comportant des exemples donnés par d'autres chercheurs (*notamment Mille Nieves*) ;

[343] L'interface Dataverse devrait par sa conception permettre la prévisualisation des fichiers, ce qui serait idéal pour les contenus audio et vidéo (ainsi que pour les fichiers de métadonnées correspondants), et constituerait un type de répertoire plus accessible que les principales archives traditionnelles utilisées en documentation linguistique comme AILLA, DOBES, etc., auxquelles les utilisateurs doivent demander l'accès. Pour l'instant, la fonction de prévisualisation ne marche toutefois pas pour certains types de fichiers, dont les ceux en format .wav, et elle n'est donc pas encore disponible. Ce problème a été discuté avec les développeurs de Dataverse, et il pourrait peut-être être résolu.

- d’une série de documents de sécurité publique publiés par le gouvernement mexicain[344];
- d’extraits de communications écrites produites par des locuteurs ;
- d’un petit nombre de publications antérieures sur la langue[345].

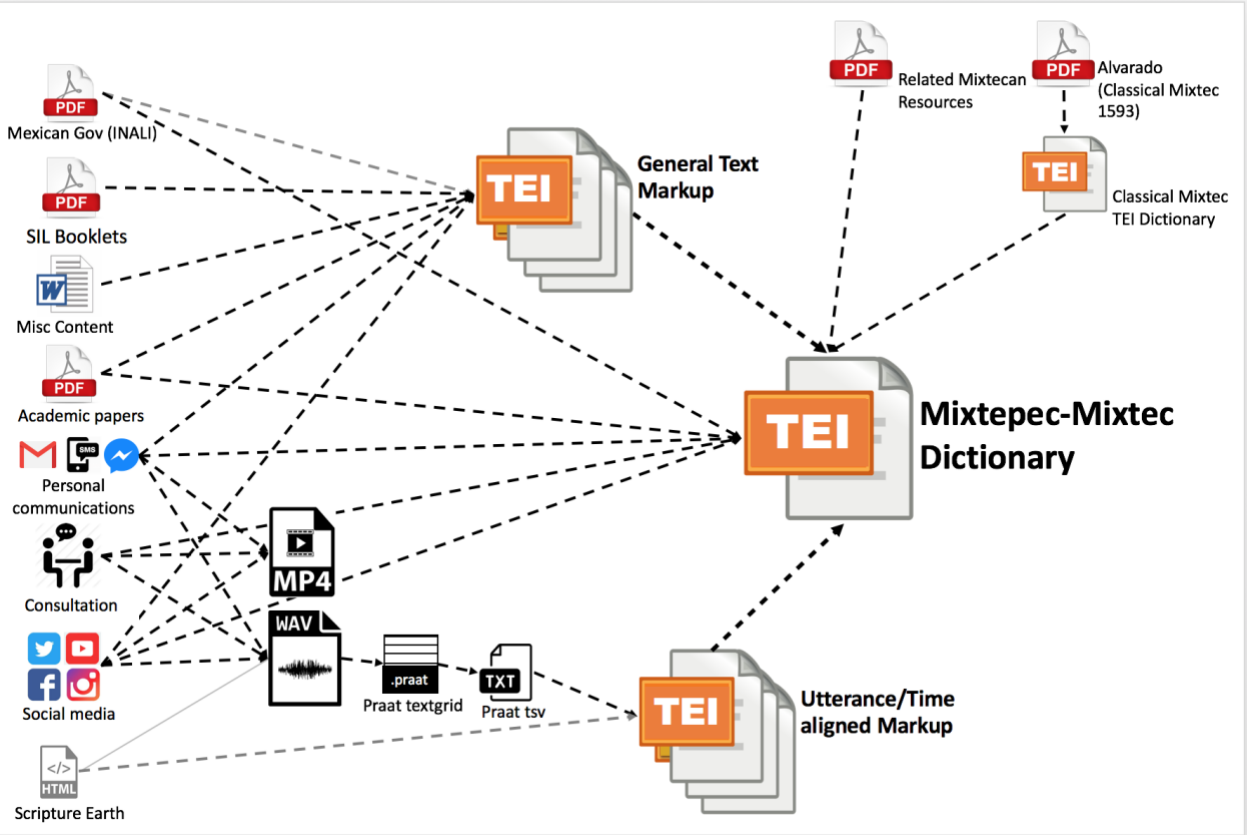


Figure 146: Flux des sources mixtèque-mixtèque (et documents linguistiques associés) et leur conversion dans différents types de documents TEI

Parmi ces ressources, seuls les livrets du SIL, les écrits produits par les collaborateurs du projet pour les besoins de ces travaux et un nombre limité de documents contenant des exemples de phrases sont encodés en format TEI. À quelques exceptions près[346], le contenu des articles universitaires, du fichier pdf des documents de sécurité publique émanant du gouvernement

[344] Ces documents n'ont pas encore été mis en corpus en raison de leur mise en page. Il vaut sûrement mieux se contenter de les étudier et d’en extraire le contenu linguistique qui mérite une place dans le dictionnaire.
[345] Plus spécifiquement : Pike et Ibach (1978) ; Paster et Beam de Azcona (2004, 2005) ; Paster (2005, 2010).
[346] Parmi ces exceptions, on peut citer la publication de Nieves (2012) sur le discours rituel mixtèque qui contient une quantité significative de vocabulaire et d’exemples de phrases dont le contexte est important, et qu’il était souhaitable d’encoder pour l’intégrer dans le corpus.

mexicain (SEGOB - Secretaría de Gobernación[347] and Sistema Nacional de Protección Civil[348](Secrétariat du gouvernement et du système national de protection civile), et des communications personnelles a donc seulement été noté manuellement et intégré dans le dictionnaire. Au moment de la présentation de ces travaux, seules deux ressources textuelles primaires[349] sont incluses dans ce projet, celles provenant des livrets et des publications du SIL, et le journal écrit à Vienne par mon locuteur collaborateur Tisu'ma Salazar en 2017.

6.2.1 Sources textuelles

À part quelques exceptions, l'encodage de la plupart des sources textuelles utilise les mêmes structures TEI de base. À titre d'exemple, lorsque le contenu est organisé en paragraphes, l'élément <p> est utilisé pour entourer le contenu MIX réel qui est encodé en tant que <seg> et prend l'attribut @type pour faire la distinction avec un contenu se présentant sous la forme d'une phrase complète (<type seg="S">), d'un groupe de mots (<type seg=" groupe de mots">), d'un terme lexical général isolé (<type seg="terme">), ou d'une légende se trouvant dans une image, et non dans le texte lui-même, ou dans des documents interactifs comportant un espace blanc (<type seg="blanc">). Chaque <seg> est marqué avec une balise linguistique @xml:lang, et reçoit un identifiant unique @xml:id. Enfin, chaque token (excepté lorsque le <seg> est un espace blanc) est encodé sous la forme <w>, qui reçoit également un identifiant unique @xml:id servant de cible pour l'annotation. Les signes de ponctuation sont encodés sous la forme <pc> (signes de ponctuation). Il convient de noter que les contenus de <w> ne représentent pas nécessairement un élément lexical complet, car il existe beaucoup de mots composés dans la langue MIX qui sont orthographiés avec un espace blanc. La façon dont ils sont reliés dans l'annotation du corpus est décrite dans la partie qui suit.

La figure 4 montre ainsi un exemple type des explications qui précèdent extrait du fichier L157-tok.xml (Mendoza Santiago, 2008). Sur cette figure, on trouve à gauche la source extraite du document PDF original, et sur la droite les encodages TEI correspondants.

[347] https://www.gob.mx/segob
[348] http://www.proteccioncivil.es/sistema-nacional
[349] Bien que, comme je vais le décrire, le corpus général contienne des versions TEI XML de transcriptions de langage parlé issues de Praat, celles-ci sont, dans cette description, distinguées de celles créées sous forme de texte.

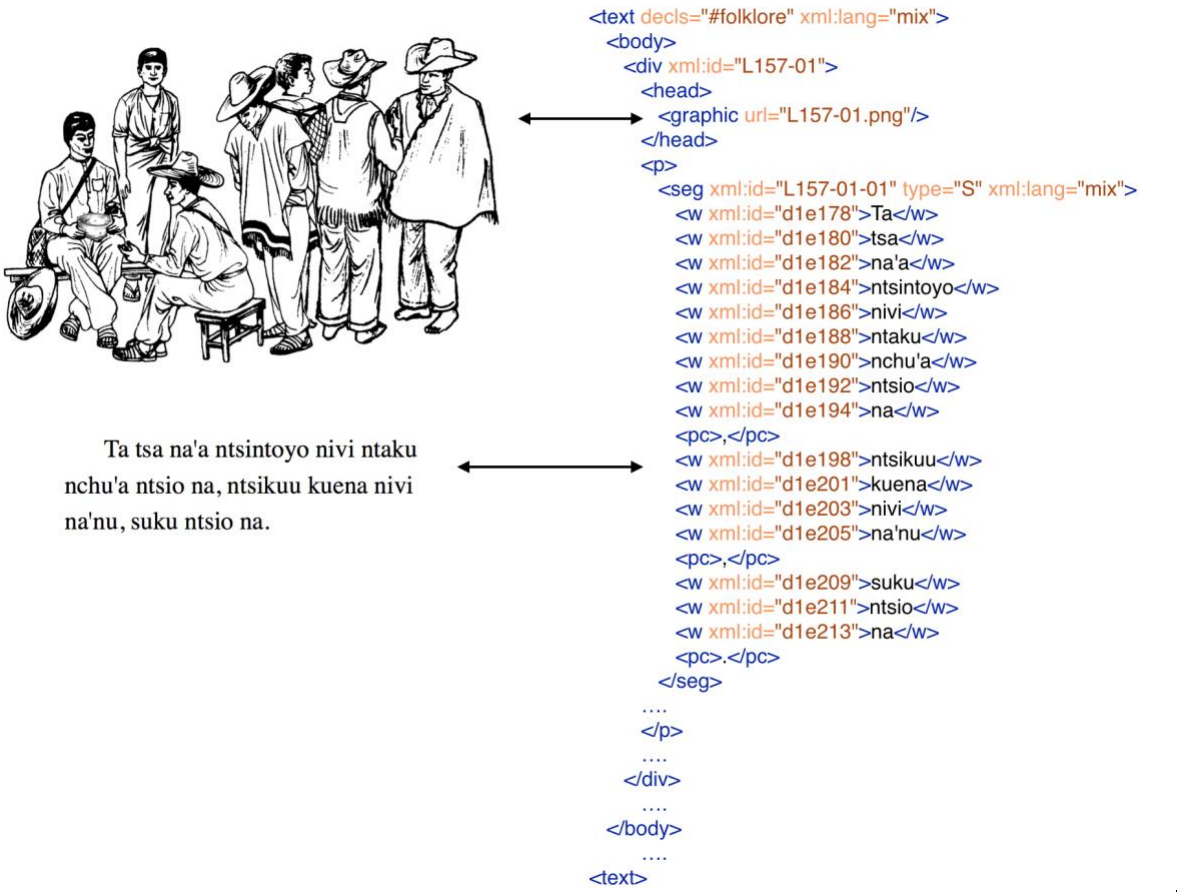


Figure 147: Contenu source (image et texte) issu d'un document SIL et structure d'encodage TEI

6.2.2 Transcriptions du langage parlé

Les sources de langage parlé[350] ont été annotées à l'aide du logiciel Praat (Boersma et Weenik, 2020), et tous les contenus MIX sont transcrits en IPA (International Phonetic Alphabet - Alphabet phonétique international) et en orthographe de travail mixtèque.

[350] Même si plusieurs vidéos ont été réalisées au cours du projet, aucune n'a encore été annotée. Toutefois, lorsqu'elles le seront, il sera nécessaire d'utiliser le système ELAN, étant donné que le logiciel Praat ne permet pas le traitement de vidéos.

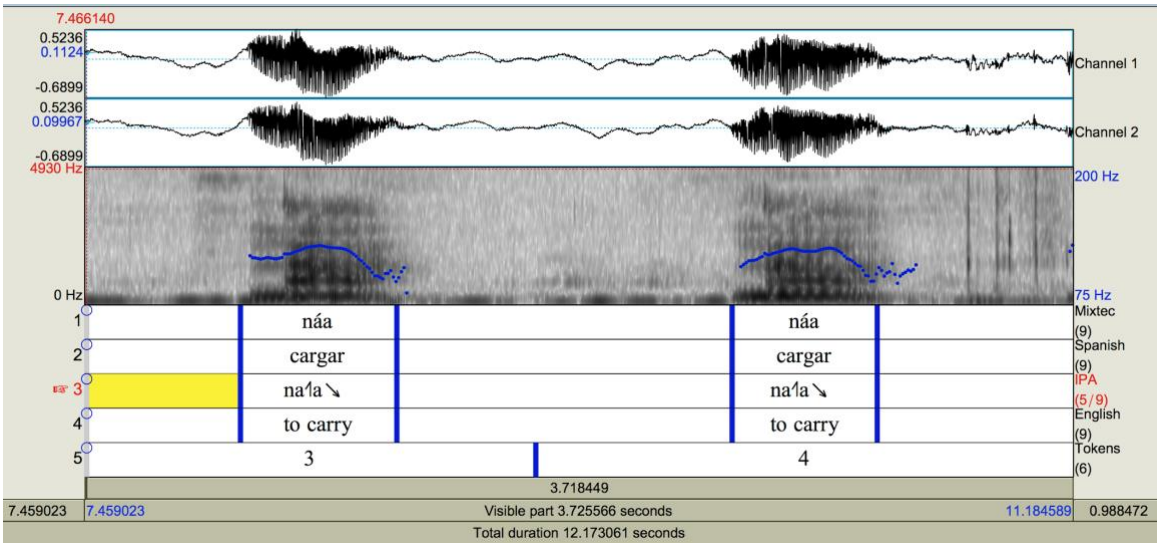


Figure 148: Exemple du système d'annotation Praat TextGrid actuel

La raison pour laquelle Praat a initialement été choisi, et son seul avantage majeur sur d'autres outils d'annotation, est qu'il permet l'analyse des hauteurs de voix (c'est-à-dire F0), ce qui est indispensable étant donné que la langue MIX est une langue tonale. En outre, Praat dispose d'un langage de script qui facilite grandement de multiples fonctions comme l'annotation, la gestion des fichiers, les modifications, l'extraction de données qualitatives et quantitatives à partir de fichiers audio, ainsi que leurs annotations, et bien d'autres encore. En utilisant le script XSLT, le résultat des transcriptions Praat est converti du format séparé par tabulation en une structure TEI qui est compatible avec les données textuelles décrites précédemment, et suit les recommandations de la norme ISO 24624:2016 et de Schmidt (2011).

Les chronologies Praat TextGrid pour une grille de texte donnée et le fichier « .wav » qui l'accompagne sont représentées dans le système TEI sous forme d'élément <timeline> (« chronologie ») qui apparaît comme élément premier au sein d'un <body> (« groupe »). Chaque point présent dans la chronologie correspond à l'endroit où un ou plusieurs segments d'annotation commence ou se finit. Ainsi, seuls les points importants de la chronologie d'annotation sont représentés. Ils sont encodés dans le système TEI sous forme d'éléments <when> (« quand »), chacun possédant un identifiant unique @xml:id auquel est attaché le contenu linguistique transcrit.

<chronologie>

```
<quand xml:id="T2.04" intervalle="2.04"/>
<quand xml:id="T3.77" intervalle="3.77"/>
<quand xml:id="T2.56" intervalle="2.56"/>
<quand xml:id="T3.18" intervalle="3.18"/>
<quand xml:id="T5.08" intervalle="5.08"/>
<quand xml:id="T4.23" intervalle="4.23"/>
<quand xml:id="T4.91" intervalle="4.91"/>
<quand xml:id="T7.10" intervalle="7.10"/>
<quand xml:id="T9.16" intervalle="9.16"/>
<quand xml:id="T8.05" intervalle="8.05"/>
<quand xml:id="T8.63" intervalle="8.63"/>
<quand xml:id="T12.17" intervalle="12.17"/>
<quand xml:id="T9.90" intervalle="9.90"/>
<quand xml:id="T10.45" intervalle="10.45"/>
</chronologie>
```

Figure 149: Chronologie de l'énoncé annoté dans Praat tel que représenté dans TEI

Les points assignés à une transcription donnée peuvent être utilisés en association avec le lien vers le fichier « .wav » correspondant par les logiciels pour lire l'énoncé et afficher sa transcription en utilisant le résultat TEI de l'annotation Praat originale.

Chaque énoncé séparé présent dans un enregistrement source (représenté dans Praat TextGrid sur le niveau « Tokens ») est converti dans le système TEI comme un élément unique <annotationBlock> (« bloc annotation ») contenant un énoncé <u>, dans lequel le reste de la transcription et des annotations (à la fois les traductions issues de Praat, et toute annotation complémentaires) est également placé.

```
<bloc annotation>
  <u n="1" xml:id="d23e0" début="2.04" fin="3.77" qui="#JS">
    <seg xml:lang="mix" notation="orth" xml:id="T-seg-orth-2.04">
      <w synch="#T2.56" xml:id="T-orth2.56">naá</w>
    </seg>
    <seg xml:lang="mix" notation="ipa" xml:id="T-seg-pron-2.04" identique="#T-orth2.56">
      <w synch="#T2.56" xml:id="T-pron2.56" identique="#T-orth2.56">naˀa↘</w>
    </seg>
  </u>
  ….
</Bloc annotation>
```

Figure 150: Représentation d'un énoncé converti de Praat TextGrid en TEI

Pour chaque énoncé, l'intervalle de temps complet est indiqué explicitement sur @start (« début ») et @end (« fin »), et les initiales du locuteur sont marquées à l'aide de @who (« qui »). Tous les contenus de chacune des transcriptions orthographiques et phonétiques sont intégrés dans l'élément <seg> et la méthode de transcription utilisée est indiquée à l'aide de

l'attribut @notation (« notation »). Chaque token est représenté sous forme d'élément <w> possédant un identifiant unique @xml:id auquel renvoient des annotations, et un attribut @synch qui indique directement le point (via la valeur @xml:id) de la chronologie (<timeline>) à partir duquel l'énoncé se produit. On remarquera qu'un token <w> (malgré la définition donnée dans les directives TEI)[351] n'est pas nécessairement un élément lexical complet ou un mot dans ce projet, étant donné qu'il est utilisé simplement pour entourer une chaîne de caractères.

6.2.3 Annotation du corpus

Dans le corpus, les contenus sont extraits et entrés dans le dictionnaire digital au moyen du script XSLT, et les annotations portent sur les caractéristiques suivantes :

- Traductions anglaises et espagnoles
- Caractéristiques grammaticales et autres caractéristiques lexicales
- Textes avec gloses interlinéaires.

Selon les bonnes pratiques appliquées en documentation linguistique, la description doit être séparée de la ressource documentée (Himmelmann, 1998, 2006a), et, conformément à ce principe, le corpus a été annoté par l'intermédiaire d'un procédé d'annotation multi-niveaux à distance, cette approche spécifique étant désignée par Bański (2010) sous l'appellation *annotation de correspondance à distance*. Elle garantit que la ressource peut être réutilisée, réinterprétée et/ou ajoutée par les personnes intervenant dans ces travaux, ou par d'autres, sans qu'il soit nécessaire de déstructurer des parties importantes du contenu original. Deux méthodes sont utilisées dans le projet pour intégrer des annotations à distance : <spanGrp> (figure 8) permet de créer de nouvelles annotations, et <linkGrp> (figure 9) est utilisé spécifiquement lorsque la source comprend déjà des traductions, gloses ou autres contenus parallèles préexistants. Ces deux procédés emploient des attributs XML pour relier les valeurs cibles issues du corpus encodé à leurs annotations en indiquant la(les) valeur(s) @xml:id correspondante(s).

La principale méthode pour annoter tous les contenus distants dans le système TEI utilise l'élément <spanGrp> qui accepte n'importe quel nombre d'éléments enfants . Les annotations sont placées dans <spanGrp type="annotations">. Dans ce système, un élément

[351] https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-w.html

<spanGrp> est utilisé pour tous les niveaux d'annotation que l'on peut rencontrer simultanément, en particulier *la traduction, la grammaire* et *les textes avec gloses interlinéaires* (qui apparaissent ensemble), *la sémantique*,[352] et, si nécessaire, *les notes*[353] (utilisées pour les commentaires éditoriaux, elles sont généralement temporaires, et peuvent, ou non, être transférées dans l'entrée de dictionnaire pour le contenu concerné jusqu'à ce que la question soit résolue). En fonction du type d'annotation, le contenu peut être spécifié dans la valeur textuelle de ou par l'intermédiaire de la valeur @ana ou @corresp.

```
<seg xml:id="L147-01-01" type="S" xml:lang="mix">
 ….
  <w xml:id="d1e170">nikachi</w>
  <w xml:id="d1e172">sto'i</w>
  <pc>:</pc>
</seg>
<spanGrp type="annotations">
  ….
  <span ana="#INFL" cible="#d1e170" xml:lang="en" type="traduction">a dit</span>
  <span type="gram" cible="#d1e170" ana="#V #TRANS #PFV">
    <glose type="igt">pfv-dire</glose>
  </span>
  <span ana="#INFL" cible="#d1e172" xml:lang="en" type="traduction">son propriétaire</span>
  <span type="gram" cible="#d1e172" ana="#NP #POSS #3PERS #SG">
    <glose type="igt">propriétaire[3sg]</gloss>
  </span>
</spanGrp>
```

Figure 151: Exemple de phrase orthographique annotée grammaticalement sans autre segmentation

Lorsqu'il existe déjà des contenus bilingues ou multilingues parallèles dans les documents sources, le procédé d'annotation <linkGrp> sert à définir les relations entre ces caractéristiques :

[352] Selon les caractéristiques d'annotation utilisées dans un projet, la *sémantique* peut être séparée en plusieurs catégories d'annotation.
[353] Les catégories d'annotation peuvent être étendues ultérieurement en cas de besoin, la « pragmatique » étant un autre candidat probable.


```
<seg xml:id="d1e53" xml:lang="mix" type="terme">
  <w xml:id="d1e54">chumi</w>
  <w xml:id="d1e56">xini</w>
  <w xml:id="d1e58" orig="ka'nu">ka'nu</w>
</seg>
<seg xml:id="d1e60" xml:lang="es-MEX" type="terme">
  <w xml:id="d1e61">hibou</w>
</seg>
<seg xml:id="d1e63" xml:lang="es" type="terme">
  <w xml:id="d1e64">búho</w>
  <w xml:id="d1e66">cornado</w>
</seg>
<linkGrp type="annotations">
  <link type="traduction" cible="#d1e53 #d1e60"/>
  <link type="traduction" cible="#d1e53 #d1e63"/>
</linkGrp>
```

Figure 152: Exemple d'encodage de <linkGrp> dans un contenu avec les traductions bilingues existantes

7. Dictionnaire TEI mixtèque de Mixtepec

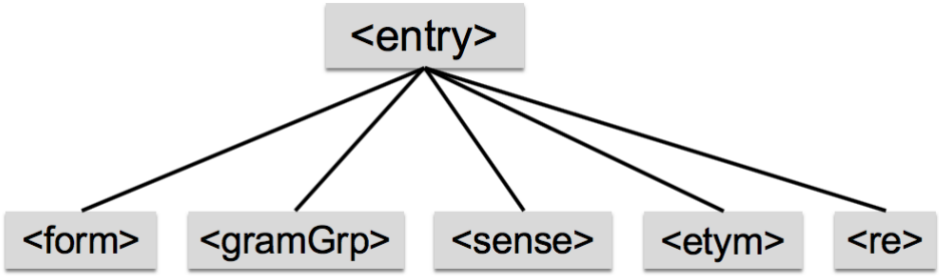
Outre le corpus annoté, les archives et les fichiers multimédias, le résultat principal de ce projet de documentation est le dictionnaire TEI trilingue dérivé des contenus présents dans le corpus et de tous modes d'observation autres. Les entrées contiennent en général les formes orthographiques des termes, les formes phonétiques (et leurs variantes), les informations grammaticales, sémantiques (sens), étymologiques et d'usage, ainsi que des exemples tirés du corpus. Dans les paragraphes qui suivent, chacune de ces caractéristiques et son encodage TEI[354]

[354] Il convient de noter que le dictionnaire est encore en cours d'élaboration, qu'au moment de la présentation de ces travaux, le formatage discuté dans la présente thèse n'est pas encore universellement appliqué, et que certains aspects des ensembles de données référencés subissent encore des modifications.

sont décrits en détail. Au moment de la présentation de ces travaux, le dictionnaire comprend 1 139 entrées. À cet ensemble s'ajoute la ressource additionnelle constituée par le Classical Mixtec Dictionary (Dictionnaire de mixtèque classique) (Alvarado, 1592) qui a été converti au format TEI à l'aide de l'outil logiciel GROBID-Dictionaries.

La méthodologie et la structure du dictionnaire TEI de mixtèque de Mixtepec ont été décrites dans la publication de Bowers et Romary (2018) qui présentait la ressource en cours de création. Ce chapitre reformule le contenu couvert dans ce dictionnaire, et apporte de nombreux détails complémentaires qui n'avaient pas été abordés, ainsi que des mises à jour nécessaires. Le dictionnaire TEI de mixtèque de Mixtepec a été initialement compilé et est de manière générale rédigé manuellement dans l'éditeur Oxygen XML Editor, même si des méthodes de script XSLT sont parfois utilisées selon les besoins, à la fois pour améliorer les entrées (c'est-à-dire avec des exemples d'items rencontrés dans le corpus), et pour créer de nouvelles entrées dès que du vocabulaire nouveau est collecté, annoté et identifié dans les données.

Le module Dictionnaire, qui définit les composants utilisés pour encoder des lexiques, constitue un composant lexicographique central du système TEI[355] . Les dictionnaires TEI sont utilisés pour encoder une grande variété de dictionnaires d'origine numérique et rétronumérisés, afin de produire des résultats orientés utilisateurs et outils. En raison, notamment, de la diversité des organisation structurelles, et donc des besoins d'encodage des dictionnaires d'origine numérique et rétronumérisés, la façon spécifique dont chaque dictionnaire TEI est structuré est très variable. Toutefois, les composants fondamentaux d'une entrée type, et de ses principaux blocs d'éléments enfants, sont ceux représentés à la figure 10 ci-dessous.



[355] https://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html

Figure 153: Composants les plus importants d'une entrée TEI

Comme le montre la figure 10, les composants principaux d'une entrée sont les éléments <forme> qui comprendront généralement un mot-clé <forme type="lemme">, des formes déclinées et variantes pouvant également être incluses en utilisant @type, respectivement en tant que « forme déclinée » et « variante ». Des formes orthographiques et phonétiques sont encodées séparément au sein de <forme> au moyen des champs <orth> et <pron>. Dans une entrée complexe (ou une entrée connexe) comme un mot composé ou une expression comprenant plusieurs mots, les contenus de <orth> ou <pron> peuvent encore être segmentés et reliés à des entrées séparées pour chaque élément en utilisant <seg @corresp>. En outre, des pointeurs vers des sources et/ou fichiers multimédias (notamment audio) peuvent être embarqués dans la forme ou ailleurs à l'aide de l'élément <média>. Des entrées connexes (<re>) sont incorporées dans une <entrée> et peuvent comporter les mêmes structures et éléments.

Des informations grammaticales peuvent être encodées pour l'entrée principale ou dans un sens spécifique d'un mot. Une entrée peut comporter autant de sous-sens séparés et/ou embarqués que nécessaire. Les sens peuvent comporter des traductions multilingues par le biais du champ <cit type="traduction">, des définitions <def>, des exemples <cit type="exemple">, un domaine et un registre en utilisant un champ typé <usg>. Des renvois peuvent être ajoutés à des endroits différents, habituellement dans <sens> ou <etym>. Dans <sens>, ils indiquent en général des relations sémantiques comme des synonymes, des antonymes, une méronymie, etc., qui sont toutes exprimées via l'attribut @type.

```
<entrée xml:id="marcher-MIX">
  <forme type="lemme">
    <orth xml:lang="mix">kaka</orth>
    <forme type="racine">
      <orth xml:lang="mix">tsika</orth>
      <gramGrp>
        <gram type="aspect">réel</gram>
      </gramGrp>
    </forme>
  </forme>
  <gramGrp>
    <pos>verbe</pos>
    <gram type="transitivité">trans</gram>
  </gramGrp>
  <sens>
```

```
<usg type="domaine">Mouvement</usg>
<cit type="exemple">
  <citation xml:lang="mix">Tsíka ra chi nuu inkaa yu.</citation>
  <cit type="traduction">
    <citation xml:lang="en">Il marche vers moi.</citation>
  </cit>
  <cit type="traduction">
    <citation xml:lang="es">Él camina hacia mi.</citation>
  </cit>
</cit>
<cit type="traduction">
  <forme><orth xml:lang="en">marcher</orth></forme>
</cit>
<cit type="traduction">
  <forme><orth xml:lang="es">caminar</orth></forme>
</cit>
</sens>
</entrée>
```

Figure 154: Exemple d'entrée pour le lemme *kaka* « marcher »

L'élément <etym> contient des informations étymologiques. Celles-ci peuvent être récursives, et peuvent apparaître au niveau de <entrée>, ou éventuellement à l'intérieur d'un sens, si ce sens a une étymologie particulière. Les processus étymologiques peuvent être typés, par exemple <etym type="emprunt">, ou <etym type="métaphore"> dans le cas d'une étymologie de sens. Les étymons, qui peuvent inclure des formes, des sens et des informations grammaticales, sont encodés à l'aide de <cit type="étymon">, et des termes apparentés issus de langues connexes peuvent être encodés soit dans le même esprit en tant qu'étymons (par exemple comme <cit type="cognat">) ou sous <xr>, en fonction du contexte. La prose peut être encodée par le biais du champ <seg type="desc"> (pour plus de détails, voir : Bowers et Romary, 2017 ; et Bowers et Romary, 2018b).

Outre la documentation générale de la langue, le dictionnaire est créé comme une base de données structurée d'informations étymologiques. L'éventail complet des processus étymologiques, ainsi que de nombreux cas de combinaison de ces processus, ont été observés dans les données, notamment : l'emprunt (la plupart du temps emprunt à l'espagnol, dans certains cas emprunt au nahuatl, voir la figure 12) ; l'héritage (du proto-mixtèque déduit en comparant des mots apparentés) ; les changements de forme comme la formation de mots composés, la dérivation, l'onomatopée, l'évolution phonologique ; les différents types de modification du sens comme la métaphore, la métonymie et la grammaticalisation. Le sujet de

d'encodage des informations étymologiques dans le système TEI, tel qu'appliqué à ce projet, a été discuté dans la publication Bowers et Romary (2018b), et les conventions utilisées sont conformes aux recommandations de la spécification TEI Lex-0 Etym (Bowers et al., 2018) et de Bowers et Romary (2016). En outre, l'intégralité du vocabulaire issu du Classical Mixtecan Dictionary (Dictionnaire de mixtèque classique) élaboré en 1593 par le frère dominicain Francisco Alvarado a été converti en format TEI (Bowers, Khemakhem, et Romary 2019) en utilisation l'outil GROBID-Dictionaries (Khemakhem et al., 2017), et ces contenus doivent être intégrés dans le dictionnaire en tant que référence historique importante[356].

```
<etym type="emprunt">
  <seg type="desc" xml:lang="en">mot emprunté à :</seg>
  <cit type="etym">
    <lang>nahuatl</lang>
    <forme>
      <orth xml:lang="nah">tequitl</orth>
    </forme>
  </cit>
</etym>
```

Figure 155: Partie étymologique de l'entrée pour le mot *tekiu* empruntée au nahuatl *tequitl*

Le dictionnaire est converti en HTML (en utilisant le script XSLT) formaté avec le langage informatique CSS (feuilles de style en cascade), et un fichier PDF peut également être généré à partir du format HTML. Ces versions créées peuvent également contenir des images et lire des fichiers multimédias. Pour le moment, ces fichiers sont disponibles seulement dans le répertoire GitHub jusqu'à ce qu'un emplacement en ligne plus pérenne puisse être créé, permettant aux utilisateurs d'avoir accès à la fois aux contenus du dictionnaire et du corpus. Il est toutefois à noter que le formatage n'est pas finalisé. La figure 13 présente un exemple de l'état actuel.

[356] Le vocabulaire d'Alvarado (1593) a été converti dans GROBID-Dictionaries à partir d'une version PDF éditée des contenus réalisée par Jansen et Perez Jiménez (2009).

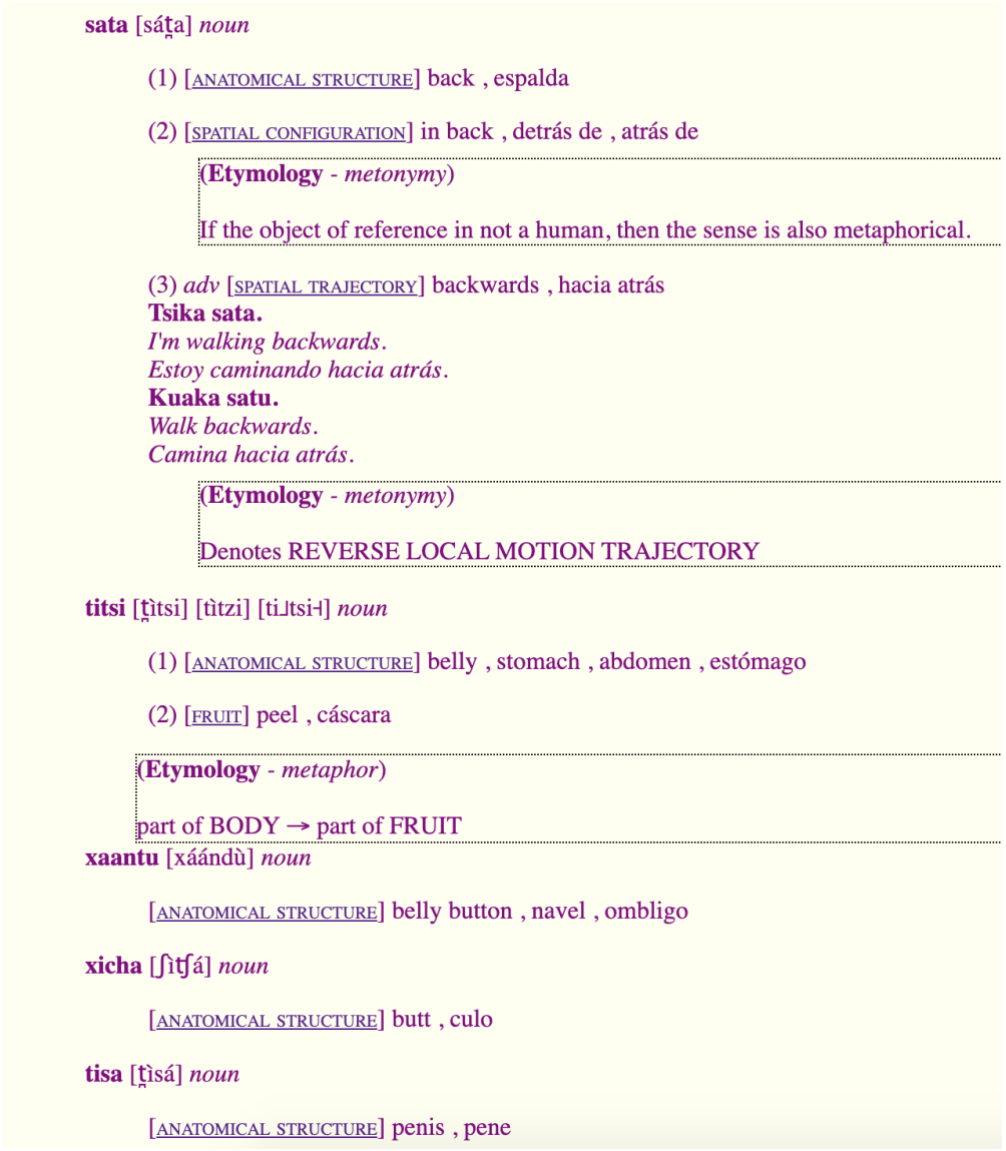


Figure 156: Capture d'écran de la version HTML du dictionnaire MIX TEI

En outre, l'utilisation d'un script XSLT permet maintenant d'exporter régulièrement le contenu aux formats CSV et Excel,[357] afin de rendre les données accessibles aux personnes qui ne travaillent pas en XML. Des travaux complémentaires devront être menés afin de développer un moyen de conversion entre les formats Flexfiles (FLEx), LIFT et TEI, ainsi qu'un script générant des formats de données interlinguistiques (CLDF-Cross-Linguistic Data Formats) à partir des contenus du dictionnaire TEI.

[357] Voir les contenus des jeux de données convertis en TSV et HTML à partir de XML dans le répertoire suivant : https://github.com/iljackb/Mixtepec_Mixtec/tree/master/exports

8. Conclusion

Après ce résumé des points principaux de ma thèse, je récapitule ici toutes les questions discutées dans sa version intégrale dans laquelle je décris les travaux de documentation de la variété mixtèque de Mixtepec réalisés au cours de ces dix dernières années. Le principal résultat de cette documentation est la création d'un ensemble de ressources linguistiques multimédias réutilisables et évolutives en source ouverte (open source) comprenant : un dictionnaire TEI multilingue, une collection d'enregistrements audio publiés et archivés sur Harvard Dataverse ; un corpus de textes résultant de transcriptions de sources écrites et de langage parlé encodées et annotées dans le système TEI ; une description grammaticale préliminaire des aspects fondamentaux en matière de déclinaisons, de morphologie et de dérivation ; ainsi qu'une publication sur l'analyse linguistique de la sémantique des termes de parties du corps dans la langue mixtèque de Mixtepec (Bowers, sous presse).

Outre la création des ressources linguistiques et l'étude de la langue elle-même, une attention particulière a été portée, dans ces travaux, à l'articulation des différentes démarches fondamentales rencontrées dans un projet de documentation linguistique, qui portent sur la linguistique et sur d'autres domaines de recherche dont les humanités digitales, la linguistique descriptive, la lexicographie digitale, la linguistique computationnelle, et la plupart des autres sous-domaines de la linguistique. Le thème des données constitue le fil conducteur entre ces travaux et les disciplines précédemment mentionnées. Il inclut les métadonnées, tous les différents types de données linguistiques primaires, les standards de balisage, l'annotation, l'analyse et l'archivage, ainsi que les outils utilisés pour créer et gérer ces données. Dans la réalisation de ce projet et l'élaboration de cette thèse, la priorité a été d'identifier les limites actuelles au niveau des flux de travail liés à la création et à la gestion des ressources précitées, du fait de l'absence de capacités suffisantes en matière d'échange de données et d'interopérabilité entre les outils (à l'exception d'ELAN), et du manque de cartographies et de schémas de conversion bien établis entre les formats de données clés créés et les standards utilisés aux différents stades du processus de documentation linguistique.

À ma connaissance, ce projet est le premier cas dans lequel les directives TEI ont été appliquées pour mettre en œuvre l'ensemble des composants centraux de la documentation

linguistique. Il permet ainsi de franchir une étape à la fois pour démontrer que ce standard a les capacités suffisantes pour encoder tous les contenus nécessaires, et pour établir un précédent concernant les pratiques utilisées pour ceux qui souhaiteraient l'adopter pour des projets similaires. Alors que le système TEI est tout à fait en capacité de traiter la majorité des nombreuses problématiques inhérentes à la documentation linguistique, notamment la représentation du langage parlé, la mise en relation de supports, le développement de dictionnaires/lexiques, l'annotation de corpus de textes et la gestion de différents types de métadonnées, quelques points mineurs nécessitant une amélioration ont été identifiés (sachant que j'ai déjà engagé des actions pour faire évoluer un certain nombre d'entre eux)[358]. L'un des plus importants (qui est discuté seulement dans la version intégrale de ma thèse) est en particulier que la TEI a grandement besoin d'une pratique établie pour traiter les textes avec gloses interlinéaires, qui constituent la méthode principale d'annotation de textes en documentation linguistique, mais comptent très peu d'exemples dans le système TEI. Alors que dans ce projet j'applique et présente une méthode de textes à gloses interlinéaires dans l'annotation du corpus, celle-ci est utilisée en association avec des annotations à distance, ce qui ne peut pas être affiché d'une manière orientée utilisateur sans transformation complémentaire.

L'annotation à distance constitue un point très important du projet qui n'a pour l'instant pas été complètement solutionné dans le système TEI. J'ai choisi d'appliquer une méthode d'annotation multi-niveaux à distance dans ce corpus, à la fois car le fait de séparer l'analyse du contenu source fait partie des bonnes pratiques en documentation linguistique, et parce que ce procédé est le meilleur pour exprimer des caractéristiques communes et un nombre potentiellement infini de caractéristiques distinctes qui ne doivent pas être appliquées uniformément dans l'ensemble du corpus. Malgré les avantages évoqués, il existe par contre toujours peu de cas ou d'exemples à l'appui concernant la recherche et l'extraction de données dans ce format particulier (réalisées ici avec des scripts XSLT et XQuery sur mesure), ou l'affichage de données dans ce format, qui nécessite des transformations et une programmation personnalisée. Pour l'instant, je n'ai pas pleinement atteint le niveau de récupération souhaité pour ce jeu de données en partie parce que l'annotation et, dans certains cas l'encodage, sont

[358] https://github.com/TEIC/TEI/issues?q=is%3Aissue+author%3Ailjackb

encore en cours. Enfin, le temps nécessaire à la réalisation des annotations à distance manuelles est conséquent, et cette tâche reste relativement lente même si elle est allégée par l'outil Oxygen XML Editor. Et s'il est possible pour moi d'effectuer un tel processus d'annotation et de créer les scripts personnalisés nécessaires pour rechercher, extraire et/ou transformer les données annotées dans un format de présentation facile à utiliser, ce ne serait pas possible pour une personne n'ayant pas de compétences en programmation.

Ainsi, même si ces travaux ont contribué significativement à faire progresser la capacité du système TEI à être utilisé dans le contexte de la documentation linguistique, et alimentent les cas de mise en œuvre du procédé d'annotation à distance dans un corpus TEI, outre la mise à disposition d'exemples détaillés d'encodage de chaque aspect des données de documentation linguistique, un travail substantiel reste à faire par la communauté TEI, en particulier dans le développement d'outils d'annotation et de gestion, et aussi en matière de schémas d'échange permettant une conversion entre les formats communément utilisés dans les outils de documentation linguistique tels que EAF (ELAN), LIFT et Flexfiles (FLEx).

Malgré toutes les avancées significatives, un travail important reste à accomplir pour que ce projet exprime tout son potentiel pour produire un résultat convivial pour l'utilisateur, réutilisable, évolutif et librement accessible pour la communauté mixtèque de Mixtepec, les apprenants et les utilisateurs non spécialistes de la technologie. Il s'agit notamment :

- de transcrire plusieurs dizaines d'heures d'enregistrements restantes ;
- de continuer l'annotation du corpus en appliquant à tous les fichiers l'ensemble des caractéristiques essentielles décrites ici ;
- de créer un site web stable comprenant une interface de recherche pour les contenus du dictionnaire et du corpus avec une capacité multimédias ;
- d'élaborer une infrastructure pour l'affichage parallèle d'éditions numériques de ressources de mixtèque historique encodées ;
- de créer des schémas complémentaires pour reformater les documents de corpus annotés en documents plus conviviaux pour les utilisateurs, en déplaçant, dans l'idéal, le contenu des traductions dans une annexe susceptible de servir de référence aux apprenants ;

- d’obtenir un financement pour engager/employer un ou des locuteur(s) natif(s) pour co-créer le dictionnaire et aider aux transcriptions ;
- de collaborer avec un spécialiste en phonologie computationnelle pour tester et mettre en œuvre des méthodes d’apprentissage machine pour les transcriptions en retard et la classification des tons ;
- de déposer le dictionnaire TEI avec Mesolex ;
- d’établir des relations, incluant le partage de données et d’analyses, avec des organismes communautaires qui apportent leur soutien à la communauté mixtèque et à la langue au Mexique et aux diverses communautés issues de la diaspora ;
- de réaliser davantage d’analyses linguistiques et de descriptions de base de la langue à partir de données.

Dans le cadre de ce projet, afin de convertir les transcriptions du langage parlé du système Praat à la structure de corpus utilisée dans la TEI, des scripts de conversion XSLT ont été créés, ce qui constitue probablement le premier schéma de conversion entre Praat et TEI. Il s’agit là seulement de l’une des nombreuses étapes qui permettront d’assurer le niveau d’échange de données qui est absolument requis tant dans le domaine des humanités digitales que dans celui de la documentation linguistique. Même si ce point n’est pas spécifique au projet MIX, un écosystème de données plus interopérable est nécessaire pour progresser dans les domaines de la documentation linguistique et des humanités digitales, et doit permettre d’assurer la compatibilité entre le système TEI et les standards et formats de données les plus couramment utilisés aux différents niveaux de la documentation linguistique, notamment :

- *Pour les métadonnées :* IMDI, CMDI, OLAC
- *Pour la transcription de la langue parlée (y compris textes à gloses interlinéaires) :* EAF, EXMARaLDA
- *Pour le corpus (et autres textes à gloses interlinéaires)* : FLEx, Toolbox
- *Pour les dictionnaires et lexiques :* FLEx.

OxGarage, l'outil de conversion en ligne existant,[359] qui permet de réaliser des conversions entre le système TEI et un certain nombre de formats de données différents, serait un candidat potentiel évident dans lequel les schémas de conversion complémentaires pourraient être ajoutés.

Pour aller plus avant, un travail substantiel reste à faire dans la création d'un ensemble de ressources mixtèques librement disponible, accessible et interopérable susceptible d'être utilisé tant pour les travaux relatifs au mixtèque de Mixtepec Mixtec que pour des variétés apparentées, ce travail comprenant :

- l'intégration de l'ensemble du vocabulaire contenu dans le dictionnaire de mixtèque de Vera Cruz (Vera Cruz Mixtec dictionary, Galindo Sánchez, 2009) dans le dictionnaire MIX ;
- la création d'éditions numériques des manuscrits mixtèques, idéalement avec des descriptions dans une ou plusieurs variétés de mixtèque ;
- la création de documents encodés TEI pour les données contenues dans les travaux fondateurs sur le proto-mixtèque et le proto-mixtèque-amazugo, issues en particulier de :
    - Longacker (1957)
    - Josserand (1983)
    - Dürr (1987)
    - Mak et Longacre (1960)
    - Longacre et Millon (1961).

Alors que, comme discuté tout au long de cette thèse, des travaux complémentaires sont nécessaires pour rendre l'édition et la recherche de certains aspects (en particulier les annotations à distance dans le corpus) des données TEI plus accessibles aux non-experts (idéalement à des membres du projet dans la communauté), le modèle utilisé dans le dictionnaire digital TEI pour la langue MIX pourrait facilement être étendu pour créer un corpus digital pan-mixtèque qui

[359] https://oxgarage.tei-c.org/#

trouverait une utilisation immédiate pour le milieu universitaire, le gouvernement et la communauté.