Codex to corpus: Exploring annotation and processing for an open and extensible machine-readable edition of the Florentine Codex

Francis M. Tyers
Department of Linguistics
Indiana University
Bloomington, IN 47401
ftyers@iu.edu

Robert Pugh
Department of Linguistics
Indiana University
Bloomington, IN 47401
pughrob@iu.edu

Valery A. Berthoud F.
Department of Philosophy
Humboldt-Universität zu Berlin
Unter den Linden 6, Berlin 10099
valeryberthoud@gmail.com

Abstract

This paper describes an ongoing effort to create, from the original hand-written text, a machine-readable, linguistically-annotated, and easily-searchable corpus of the Nahuatl portion of the Florentine Codex, a 16th century Mesoamerican manuscript written in Nahuatl and Spanish. The Codex consists of 12 books and over 300,000 tokens. We describe the process of annotating 3 of these books, the steps of text preprocessing undertaken, our approach to efficient manual processing and annotation, and some of the challenges faced along the way. We also report on a set of experiments evaluating our ability to automate the text processing tasks to aid in the remaining annotation effort, and find the results promising despite the relatively low volume of training data. Finally, we briefly present a real use case from the humanities that would benefit from the searchable, linguistically annotated corpus we describe.

1 Introduction

The Nahuatl language, an agglutinating and polysynthetic member of the Uto-Aztecan family spoken throughout Mexico by about 1.5 million people today, has a rich literary tradition (Gingerich, 1975; León-Portilla, 1985). With a strong preconquest oral tradition and a hieroglyphic writing system, Nahuatl speakers quickly adopted the Latin alphabet for writing their language after its introduction almost immediately after the Spanish invasion. As a result, the volume of the colonial-era Nahuatl literary canon is unrivalled in Latin America (Olko and Sullivan, 2013). These texts are invaluable resources to scholars interested in the history, culture, and language of colonial and pre-invasion Nahua communities.

Perhaps the most notable Nahuatl text of the early colonial period, the *Historia General de las Cosas de Nueva España* "General History of the Things of New Spain" (Florentine Codex, FC) is an encyclopaedic work in Nahuatl and Spanish compiled by Indigenous scholars from the Colegio de Santa Cruz de Tlatelolco and Franciscan friar Bernardino de Sahagún.

The FC is undoubtedly one of the most valuable manuscripts of the early modern period. However, it was forgotten for centuries until Angelo Maria Bandini described it in 1793. He named it "Codice Fiorentino" after the Biblioteca Medicea Laurenziana in Florence, where it is still kept. But only at the beginning of the 20th century did Francisco del Paso y Troncoso bring it to a wider audience (Martínez, 1982). Charles Dibble and Arthur Anderson published a translation of the books into English throughout the second half of the 20th century. The original manuscript became available in the World Digital Library only ten years ago, thanks to the Library of Congress.

The impetus for the present project was the need of the third author, a humanities scholar, to search the text of the FC for specific linguistic constructions and terminology. This proposition is complicated by a number of factors:

First, there are few fully digitised versions of the FC, and those that do exist are under copyright, constraining the ability of a scholar to reproduce, annotate, and/or re-release any part of the text that results from a given research endeavour.

Second, the FC, having multiple authors and being written in the early years of Nahuatl alphabetic writing, contains numerous orthographic inconsistencies throughout the 12 books, with many words written in multiple distinct ways and decisions about word tokenisation not being standardised. Furthermore, due to constraints on column width in the original manuscript, words are frequently split by line breaks with no indication of whether the following line continues the word from the end of the previous one. Keyword searching this text is a seemingly-futile process involving determining all possible spellings for a given word and all possible tokenisations of a single syntactic

Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 19–29

fo. 1

Jc macujlli amuxtli, vncan mo
teneoa: intetzaujtl catca, inqujte
tzammatia, in iehoantin mexica.

Jc ce capitulo, vncan mitoa, inte
tzaujtl inqujtetzammatia: iniquac
ioaltica, aca quicaquja: in chocaia:
itla tequanj, inanoce, iuhquj aca
ilamachocaia: auh iequene, intlejn
qujtoaia, tonalpouhque: in mjtova
ia tlamatinjme catca.

I nievecauh, inoc tlateo
toco: injuh quitoa, ce
qujntin veuetque, ve
ue tlaca: inoc titixca
vitoque, inoc titomattoque, inoc to
tlalpan tiqujtoaia: inaia aque tech
nelotoque: inaiaque totlah onoque,
Yzqujtlamantli, y, in quj tetzamma
tia, inquitetzaujttaia, in tetzaujtl ipan
quimatia innetetzaujloia, in tlatetza
ujaia.

Jnic centlamantli, iehoatl, iniquac
aca quicaquia: in tequanj choca, te
quanj ipan choca: injuhquj tecciztli
qujpitza: tepetl qujnanqujlia: teh
cuiooa, tlaoalanja: iuhqujn ilama
pul choca. Ynjn, qujtovaia ie iaomj
qujz: anoço ie mjqujz, tlalmiqujz:
anoço , ie itla com monamjctiz, ic

Figure 1: On the left: first folio of Book 5 of the Florentine Codex "The Omens". The first paragraph translates as "Fifth book, where are told the omens, which the Mexicans believed". On the right: The transcription of the left-hand column of the folio. [Image credit: Library of Congress]

word into multiple orthographic words.

Finally, Nahuatl is a morphologically complex language with large amounts of inflection and derivation, making querying the surface/inflected form, instead of e.g., a lemma, particularly difficult.

The present project attempts to address these issues by creating an open-source, retokenised, and normalised corpus of the FC with queryable linguistic annotations following the Universal Dependencies framework (Nivre et al., 2020a). In the following sections, we describe the corpus, each component involved in its creation, and an investigation into automating the processing. We conclude by outlining a road map for the project's completion and a vision of future applications.

2 Related work

The FC has been the subject of a great deal of research in the humanities by scholars interested in the cultural beliefs and practices of the Nahua people during the early colonial period (Sullivan et al., 1966; Gingerich, 1988; Sigal, 2007; McDonough, 2020; Olivier, 2021). It has also served as a foundational component for work studying so-called "Classical Nahuatl," or Nahuatl spoken during the period (Launey, 1986; Lockhart, 1992, 2001). Both Olko et al. (2015) and Olko (2018) leverage corpus-based approaches using a multitude of historical Nahuatl documents, but it is unclear how much linguistic information was available in the corpus, and to our knowledge, this corpus has not been released to the public.

Gutierrez-Vasques et al. (2016) released *Axolotl*, a large, Spanish-Nahuatl parallel corpus with a focus on machine translation. It includes Nahuatl from multiple variants and time periods, including the early colonial period, but does not include text from the FC. Furthermore, the text in *Axolotl* is unprocessed and unannotated.

Other corpora that include Nahuatl texts include the Johns Hopkins University Bible Corpus (McCarthy et al., 2020), a parallel multilingual corpus that includes numerous contemporary Nahuatl variants. This corpus has been used to produce morphosyntactically-annotated resources for a large number of languages (Nicolai and Yarowsky, 2019; Nicolai et al., 2020).

The first open morphosyntactically-annotated corpus of Nahuatl was recently released by Pugh et al. (2022) and includes 10,000 tokens of the Western Sierra Puebla variety. Following this work, we also select UD as our annotation schema.

Marc Eisinger was the first to publish a computerised version of the FC, which is not freely available (Eisinger, 1977). The Universidad Autónoma de México (UNAM) hosts a website, *Temoa*, containing a large volume of digitised colonial-era Nahuatl texts, with minimal processing (at the very least, tokenisation problems in the FC appear to be corrected (Universidad Nacional Autónoma de México, 2023). However, the copyright and rights to use for annotation and re-release are retained by UNAM,[1] making it not possible to create derivative works, such as the annotated corpus described in this paper. Furthermore, the original text (before fixing tokenisation) is not available.

Related to the computational processing of colonial Mexican texts, The "Digging into colonial Mexico" project (Murrieta-Flores et al., 2022) involves the creation of a number of processed and machine-readable resources based on colonial Mexican documents, mostly written in colonial-era Mexican Spanish. As for colonial texts written in Mexican languages, the Ticha project (Broadwell et al., 2020), a collaboration between members of Zapotec-speaking communities and academics from universities in the United States of America, offers an "online digital text explorer" for colonial Zapotec texts and includes morphological analyses and translations.

3 Corpus

Our corpus comes from a typed transcription upholding the original layout, published in the open-access repository Zenodo[2] to allow the semantic and computational study of the text from the primary source (de Sahagún, 2022). In Figure 1 we present a folio from the manuscript where the text in Spanish (left) and Nahuatl (right) is seen in two columns, and an example of the transcription output in our corpus.

3.1 Orthography

There is a great deal of orthographic variation in the FC, in both the Nahuatl and Spanish sections, with multiple characters used inconsistently throughout. For example, the letter [v] can represent either /w/, e.g. *veue* /wewe/ 'big' (norm. *huehue*), or a long /o:/, e.g. *vmpa* /o:mpa/ 'there' (norm. *ompa*). [j] is used both for the vowel /i/ e.g., *jnpilhoan* /inpilwa:n/ 'their (pl) children' (norm. *inpilhuan*) and the glide /j/, e.g. *jollochicaoac* /jol:otʃika:wak/ 'brave' (norm. *yollochicahuac*). The letter [i] is also observed in both of these contexts.

There are also instances where a single sound, e.g. /ʃ/ can be represented by multiple letters, in this case [x] or [s]. For example, the word *axcan* /a:ʃka:n/ 'now, today' can appear as *ascan* or *axcan*. But [s] can also be the voiceless alveolar sibilant /s/ in loan words from Spanish *visorrej* /bisorei/ 'viceroy' (norm. *visorrey*).

4 Processing

A major theme of the processing of the FC is the use of initial detailed hand-annotation in order to bootstrap automated approaches for the remaining text. Crucially, the resulting corpus should be usable for academic research and, as such, must maintain the utmost quality. In this context, then, we consider automation a strategy to assist in human annotation, but still require manual auditing of the entirety of the annotated corpus.

4.1 Sentence segmentation

Full stops (or in dialogue, exclamation marks, and question marks) are used as sentence boundaries throughout the corpus, with the colon symbol often used to separate clauses, making sentence segmentation fairly straightforward. There are a number of abbreviations, such as *xpo.* for Christ and *p.* for Pedro. Table 5 presents the size of each book in terms of sentences, space-separated tokens, and words. Words are only given for the three books we have processed so far.

4.2 Retokenisation

There are a number of tokenisation inconsistencies in the original manuscript, resulting from (1) physical constraints, namely the author running out of room on one line and splitting a word across a line boundary (see Figure 1), (2) inconsistent tokenisation practices by the authors, such as sometimes writing the article subordinator *in* and an adjacent verb together as a single orthographic word, and (3) possible mistakes introduced during the process of manually typing up the manuscript.

[1] https://temoa.iib.unam.mx/creditos
[2] https://zenodo.org/

Y·njqc·oiuh·ipan·muchiuh,	Yn·jqc·oiuh·ipan·muchiuh,	In ihcuac oyouh ipan mochiuh,
·y:·njman·ic·iauh,¶qujttaz·	·y:·njman·ic·iauh, qujttaz·	y: niman ic yauh, quittaz
intonalpouhquj:·vmpa·quella¶	in·tonalpouhquj:·vmpa·	in tonalpouhqui: ompa
quaoa,·qujtlapaloa:·qujlvia.¶	quellaquaoa,·qujtlapaloa:·qujlvia.	quellacuahua, quitlapaloa: quilhuia.

Table 1: A sentence from Book 5 of the FC, the sentence reads "When it happened, he went to see the reader of the day signs, there he encouraged and greeted him and said:" Note that the original tokens *Y·njqc* have been retokenised into *Yn jqc* 'when', the token *intonalpouhquj* has been split into two tokens *in·tonalpouhquj* 'the reader of the day signs' and the tokens *quella¶quaoa* which have been split by a newline have been joined into *quellaquaoa* 'he encouraged him'.

Our first step in processing the codex, after obtaining text files transcribed from the original manuscript, involves "retokenisation": altering the word boundaries in the text to align them with canonical Nahuatl words.[3] An example of the input and output of this process is shown in Table 1, wherein a space is represented by the mid-dot character, ·, and newline is represented by the pilcrow character, ¶.

As with the rest of the processing steps, retokenisation starts as a manual process. For each identified case where retokenisation is necessary, we use the left and right contexts to write a rule for handling that case, ensuring that the contexts are large enough to avoid potential ambiguities (for instance, a minimal-context rule such as "n·c →nc" will likely produce many false positive matches). In the event that a rule produces false positives, we expand its contexts (e.g., "qujn·caoa →qujncaoa"). We use a left-to-right longest-match (LRLM) algorithm to apply the approximately 4,000 retokenisation rules.

4.3 Normalisation

Once the text is correctly tokenised, the next processing step is orthographic normalisation. We use the ACK (Andrews, Campbell, Karttunen) orthographic standard for the target orthography, since it is designed to reflect colonial-era Nahuatl writing (Campbell and Karttunen, 1989; Andrews, 2003; Karttunen, 1992).

For Spanish words we use contemporary orthography, so for example, *gouernadores* is normalised to *gobernadores* 'governors.'

For proper nouns, we also use modern orthographic conventions where available. For example, *tlatilulco* is normalised to *Tlatelolco*, and *motecuhcoma* is normalised to *Moctezuma*.

The process uses a hand-curated dictionary mapping original word forms to their normalised counterparts (e.g. the normalised form *yaoyotl* 'war' is written variably as *iaoiotl*, *iauiotl*, *iaviotl*, *iaujutl* and *iaujotl*. Thus, our dictionary has an entry for each of these forms mapping to the normalised form). To build the dictionary, we start with a naïve finite-state transducer (FST) model designed using general patterns of colonial-era Nahuatl writing. We then post-edit the output of the FST, adding all correct word pairs to the dictionary. We update the FST weights as we add forms to the dictionary to improve its performance. After processing three books, the dictionary contains 6,515 entries.

The main motivation for performing the normalisation manually is to ensure a high-quality data set with which to train a model for automating the process. We discuss the evaluation of such an approach in §6.2.

4.4 Part-of-speech tagging

The part-of-speech tags are based on the Universal Part-of-Speech categories (UPOS) defined and used in the Universal Dependencies framework (Nivre et al., 2020b).

We accomplish part-of-speech tagging in three steps. We use a lexicon, a morphological analyser (see §4.5) and a set of ordered, regular-expression-based guessing rules applied to the normalised form, in sequence. We refer to this last component as 'the guesser.'

The lexicon is simply a list of normalised surface forms and their part of speech. Of the 10,959 types presently annotated for part-of-speech, 1,478 (6,916 tokens) received their POS from the lexicon.

In the event that a given surface form is not observed in the lexicon, we next run the word through the morphological analyser. This method accounts for 13,762 of the tokens thus far annotated (1,705 types).

[3] Following authoritative resources like Andrews (2003) and Campbell and Karttunen (1989) in identifying "canonical words", which should include subject, object, and aspectual affixes.

Finally, any word not identified in the previous two steps is passed to the guesser. The guesser consists of 36 rules which use regular expressions to look for particular prefixes and suffixes and assign part-of-speech tags with high precision. For example, words beginning with *nimitz-*, a combination of the first person subject marker and second person object marker are categorised as verbs, and words ending in *-tzitzin*, which is the plural reverential marker, are categorised as nouns. These rules are high precision, but low recall: a total of 986 forms out of 10,959 forms (1,471 tokens) in the three processed books receive guessed analyses.

We randomly sampled and manually checked 200 of these guesses and found that 198 were correct. In one case the mistake was due to a mistaken normalisation (*iehoatin* → **yehuatin* instead of *yehhuantin* 'they, them'), which resulted in the word being tagged as a noun due to the *-tin* 'PL' ending (plural). The second case was to do with the same plural rule, which resulted in the word *xixitin* 'it crumbled' (from the verb *xixintini* 'to crumble') being tagged as a noun.

4.5 Morphological analysis

Morphological analysis is the task of producing, for a given surface form, a lemma and a set of morphosyntactic tags describing that form. For example, given the form *tictlamacazque* /ti-c-tlamaca-z-que/ 'We will give something to him' (or 'We will make offerings to him') it would produce,

```
<s_pl1><i_sg3><o_nn3>maca<v><dv><fut>
```

Where `<s_pl1>` stands for 1st person plural subject, `<i_sg3>` stands for 3rd person singular secondary object, `<o_nn3>` stands for 3rd person inanimate indefinite object, `<v>` stands for verb, `<dv>` stands for ditransitive and `<fut>` stands for future. Note that there is a long distance dependency between the prefix *ti-*, which can be 2nd person singular or 1st person plural and the suffix *-que* which marks a plural subject.

A given token can produce more than one analysis, so for example, *quinchihua* 'They made them' or 'He made them' produces,

```
<s_pl3><o_pl3>chihua<v><tv><pres>
<s_sg3><o_pl3>chihua<v><tv><pres>
```

In this case, because of underspecification in the orthography, the plural subject-marking suffix *-h* is not written, resulting in an ambiguous analysis. The omission of this suffix is quite common in Nahuatl texts.

For implementing the morphological analyser we used the Helsinki Finite-State Toolkit (HFST) (Lindén et al., 2009). The analyser was implemented over the normalised forms. Morphotactics and the lexicon were implemented using `lexc`, while any morphographemic constraints were implemented with `twol`. A given surface form, for example, *omoyollochichili* 'He strove strongly' (lit. "he waited for himself on behalf of the heart"), consists of three parts, the surface form (1), the morphotactic form (2) and the lexical form/analysis (3).

1. `omoyollochichili`
2. `o>mo>«yollo»chichi>lia`
3. `<aug><s_sg3><o_ref>«yollotl<n>»chichilia<v><tv><past>`

The morphotactic form is the combination of the morphs before morphographemic rules are applied, it includes symbols to mark segment boundaries, such as '>' for an inflectional boundary, '«...»' for incorporated elements (in this case, the second object), ~ for reduplication and '·' for clitic boundaries. The symbols around the incorporated element allow that part of the surface form to be extracted for use in the representation of incorporation (see §5.1).

5 Representations

In this section we discuss a number of features of Nahuatl that require special attention in the Universal Dependencies framework.

5.1 Incorporation

Incorporation is the process by which a verb can incorporate, that is, be syntactically incorporated with one or more of its arguments or adjuncts. Incorporation has been understudied in the field of natural language processing, and there are few articles that describe annotation projects for languages exhibiting this feature.

In this project, we follow the proposal laid out by Tyers and Mishchenkova (2020) in which incorporated items are exposed in the enhanced dependency graph annotated with the relation of the slot that they fulfill in the argument structure.

# sent_id = Book_01_-_The_Gods.txt:87							
# text = [...] : qujlhuja, timotenoatzaz, titlacatlaquaz, timocujtlaxculcaoaz, naujlhujtl: [...]							
# text[norm] = [...] : quilhuia, timotenhuatzaz, titlacatlacuaz, timocuitlaxcolzahuaz, nahuilhuitl: [...]							
# text[orig] = [...] :·qujlhuja·,·timotenoa¶tzaz·,·titlacatlaquaz·,·timocujtlax¶culcaoaz·,·naujlhujtl·:·[...]							
[...]							
10	:	:	PUNCT	_	_	_	Norm=:
11	qujlhuja	ilhuia	VERB	_	_	_	Norm=quilhuia
12	,	,	PUNCT	_	_	_	Norm=,
13	timotenoatzaz	huatza	VERB	_	Subcat=Tran|Reflexive[iobj]=Yes†	_	Norm=timotenhuatzaz
13.1	ten	tentli	NOUN	_	_	_	Norm=ten
14	,	,	PUNCT	_	_	_	Norm=,
15	titlacatlaquaz	tlacatlacua	VERB	_	Subcat=Intr†	_	Norm=titlacatlacuaz
16	,	,	PUNCT	_		_	Norm=,
17	timocujtlaxculcaoaz	zahua	VERB	_	Subcat=Tran|Reflexive[iobj]=Yes†	_	Norm=timocuitlaxcolzahuaz
17.1	cujtlaxcul	cuitlaxcolli	NOUN	_	_		Norm=cuitlaxcol
18	,	,	PUNCT	_	_	_	Norm=,
19	naujlhujtl	nahuilhuitl	NOUN	_	_	_	Norm=nahuilhuitl
20	:	:	PUNCT	_	_	_	Norm=:
[...]							

Table 2: The second clause from the 87th sentence in Book 1. The sentence reads "He said to him: you will dry your mouth, you will fast, you will fast your entrails, four days". The underlined nouns are incorporated. † Feature=Value pairs Number[subj]=Sing|Person[subj]=2|Tense=Fut|VerbForm=Fin and repeated empty columns are left out for reasons of space.

Table 2 demonstrates this with the verb *moyollochichili*, where the verb *chichilia* 'enbitter' takes the incorporated object *yollo-* 'heart.'

5.2 Relational nouns

Relational nouns are nouns which express spatial and temporal relations when used with other noun phrases. These may be used as independent words in a possessive structure (1) or compounded to other words (2).

1. *inepantla in ilhuicatl* 'in the midst of the heavens' (lit. its-midst the heaven)
2. *ilhuicayollotitech* 'in the heart of the heavens' (lit. heavens-heart-on)

The first case is straightforward, each noun is analysed as a separate word, with the relational noun receiving a lexical feature NounType=Relat in addition to the necessary possessive morphology.

In the second, we take advantage of the multi-token word encoding in the CoNLL-U format and analyse the compound as consisting of two parts, the head and the compounded relative noun.

5.3 Lemmas

We also include the lemmas, or the stems, for each word. Lemmas ignore any of the inflectional morphology on the surface form of the word. Lemmatisation is performed first by looking up a surface form in the lexicon and, if the word is not in the lexicon, by the morphological analyser.

6 Automated processing

We experiment with the existing processed FC data to see to what extent we might be able to automate the retokenisation and normalisation steps. Following previous work showing that historical text normalisation can be modelled effectively as a character-based machine translation problem (Bollmann, 2019), we train an encoder-decoder Seq2Seq model with Attention on character sequences for both tasks. While a natural inclination would be to train both retokenisation and spelling normalisation jointly, we are interested in storing each intermediate step for potential future research, and so train a separate model for each task.

For the orthography normalisation model, we treat each word as a training instance, and map the unnormalised word (e.g. *qujchioa*) to its corresponding normalised form (e.g. *quichihua*).

For the retokenisation model, training on each word would not work since the phenomenon we are modelling spans word boundaries. Instead, we split the text on unambiguous punctuation ('.,:;?!'), creating numerous subsequences from each sentence.

Since the objective is to evaluate how well we could automate the text processing for future books, we used two of the three already-complete books (Books 1 and 8) for training, and held out

Book 5 for evaluation. The models used a bidirectional LSTM encoder, and training was done using OpenNMT (Klein et al., 2020). We trained both for 100 epochs.

Results of the experiments are listed in Table 3. They are generally favourable, though perhaps not quite to the point of being able to completely automate the low-level processing of the remaining books.

6.1 Retokenisation

A number of the mistakes we see from the retokenisation model involve a type of 'hallucinations,' where the output contains characters not in the input. This is an effect of treating this problem as one of translation with a relatively low volume of training data. To remedy this problem, we may try adding an additional auto-encoding or "copying" auxiliary task as discussed in Mager et al. (2019), wherein we add training examples that are already correctly tokenised in order to provide more examples of correct outputs.

Alternatively, the task of retokenisation can be straightforwardly modelled as a one-to-one sequence tagging problem, where for each input character the model must assign one of three "retokenisation actions": (1) merge, or remove a token boundary that follows the current character, (2) split, or add a token boundary after the current character, or (3) do nothing. For comparison, we also evaluate this approach, using a bidirectional LSTM also trained for 100 epochs.[4] This approach has a slightly worse word error rate compared to the MT-based approach, but has a lower character error rate. The advantage to this approach is that we don't risk transforming characters or inserting substrings during the tokenisation step.

6.2 Orthographic normalisation

The orthographic normalisation model correctly normalises 87% of the words in the held out book. The errors suggest a similar issue seen in the retokenisation model, namely the insertion of multiple additional characters not corresponding to the input (e.g. converting input *ie*, to **yeyecye* instead of *ye*). This issue, as mentioned above, would likely be alleviated with some data augmentation and/or multi-task training to ensure the model sees enough examples of properly formed output strings. We plan to leverage this model as a backup in the case where we are not able to identify a normalisation via our dictionary-lookup approach. For example, by first checking if we have seen a given word in the training data and, if so, using the corresponding output from training and using the model's prediction on unseen words only, the word error rate drops to 8.3.

[4] Given our limited data volume and the interest to simulate testing on an unseen book, the results we report here do not include a hyper-parameter tuning step using a heldout development set. With an additional held out book we could tune these models' hyperparameters and improve performance.

7 Use cases

In this section, we provide descriptions of a few research questions that could be informed by our corpus. The use cases are based on information that is available in the corpus and is not found in other editions of the manuscript.

The first use case concerns the status of the *tlamatinimeh* 'sages, wise men' (lit. those who know things). It is widely claimed that there is no philosophy outside Western philosophy (Maffie, 2014), but this claim has been contested by scholars, starting from Ángel María Garibay and his student Miguel León-Portilla who identify the *tlamatinimeh* with philosophers and argue that the pre-contact Mexicans had long philosophical traditions (León-Portilla, 1956). Analysing individual words has since this work been the basis of understanding Nahua thought. However, to date this process is difficult and error-prone as it involves carefully reading through unannotated concordances of surface forms and it is easy to miss examples that appear in forms that are unknown or unfamiliar to the researcher.

Our corpus will be able to help by allowing scholars to extract examples that are morphologically and syntactically related. For example, by allowing queries based on lemmas (encompassing for example *tlamatini* 'sage', *tlamatinimeh* 'sages', etc.) It will also allow for searching for specific syntactic constructions, such as those where a *tlamatini* is the subject of a speech verb.

The second use case concerns concepts of time. For instance, consider Maffie (2014)'s statement about the Mexica conceiving time and space as a single unit. He argues that the Mexica did not separate time and space but had a perception of a "time-place" . This argument is based on the word *cahuitl*, which means 'time', and the intransitive verb *cahui*, which means "to stay or end". However, Maffie argues that *cahuitl* is also related to

Task	Model	Train	Test	CER	WER
Retokenisation:	NMT	5,245	1,264	3.6	23.1
	Sequence labelling	5,245	1,264	2.8	23.9
Orthographic normalisation:	NMT	15,208	3,209	4.3	13.3
	NMT + Dictionary	15,208	3,209	1.7	8.3

Table 3: Results from experiments on automating text processing tasks (retokenisation and normalisation). The training set includes books 1 and 8, while the test set includes book 5.

Ret. Action	Precision	Recall	F1
Merge	0.856	0.952	0.902
Split	0.978	0.926	0.952
Nothing	0.997	0.995	0.996

Table 4: Results on predicting the "retokenisation actions" to correctly tokenise each original sequence. The corresponding word and character error rates are listed in Table 3.

the transitive verb *cahua*, which means "to leave or abandon", among other senses.

The morphological annotations (including lemmas) in our corpus will allow for searching on lemma (to be able to distinguish forms of *cahui* from forms of *cahua*). And the syntactic annotations will allow for the extraction of time and place obliques that are dependents of those two verbs.

8 Concluding remarks

We have outlined the strategies and approaches involved in creating a free and open, linguistically-annotated corpus of the FC. Having nearly completed retokenisation, orthographic normalisation, lemmatisation, part-of-speech tagging, and morphological analysis for 3 of the 12 books, we have established the key linguistic information to include in the corpus, and have engineered the foundations of the annotation process. Results of our preliminary experiments into automatic annotation suggest that some tasks, like orthographic normalisation, can largely be automated with the existing data, whereas others, e.g., retokenisation, likely still require more labelled data and/or a more powerful architecture.

8.1 Future work

Our first priority for the future is to continue the annotation process, automating some of the text normalisation, expanding the lexica, and enhancing the morphological analyser. We are optimistic that with each subsequent book, the additional amount of available annotated data will enable faster future annotation via automation. Finally, adding dependency syntax annotations will enable quantitative analysis of colonial Nahuatl syntax, a field with relatively little prior work.

The study described in §7 is one of many potential uses of an annotated corpus as described here. We expect that the release of this corpus with complete morphosyntactic annotations and an unambiguous free licence will promote future research from scholars in a variety of fields.

Additionally, the tools for automatic processing of the FC will likely be applicable to the numerous additional texts written in Nahuatl during the colonial period, contributing to the advancement of language technology development for Nahuatl.

Finally, another important project related to the development of this corpus involves the translation of the FC into contemporary Nahuatl variants, making the rich cultural heritage of the Nahuatl language more accessible to Nahuatl-speaking communities. It is our hope that the production of this corpus can aid in the translation process.

Acknowledgements

We would like to thank Maira Cayetano Nemecio and Stephanie Berthoud Frías for their valuable contributions. We are grateful to Mitsuya Sasaki and Joe Campbell for fielding numerous questions about language use in the Florentine Codex, and to the anonymous reviewers for their helpful feedback. Finally, a special thanks to Daniel Swanson, Andrew Davis, Zack Leech, and Maria Lucero Guillen Puon, for stimulating discussions about Nahuatl and the Florentine Codex.

References

James Richard Andrews. 2003. *Introduction to classical Nahuatl*, volume 1. University of Oklahoma Press.

Marcel Bollmann. 2019. A large-scale comparison of historical text normalization systems. *arXiv preprint arXiv:1904.02036*.

George Aaron Broadwell, Moisés García Guzmán, Brook Danielle Lillehaugen, Felipe H Lopez, May Helena Plumb, and Mike Zarafonetis. 2020. Ticha: Collaboration with indigenous communities to build digital resources on zapotec language and history. *DHQ: Digital Humanities Quarterly*, (4).

Joe R. Campbell and Frances Karttunen. 1989. Foundation course in Nahuatl grammar.

Bernardino de Sahagún. 2022. Transcript of the florentine codex (nahuatl). In *The General/Universal History of the things of New Spain*. Zenodo.

Marc Eisinger. 1977. *Codex de Florence et informatique: propositions pour l'étude systématique des textes nahua*. Ecole des Hautes Etudes en Sciences Sociales.

Willard Gingerich. 1988. Chipahuacanemiliztli: The purified life, in the discourses of book vi, florentine codex. In *Smoke and Mist: Mesoamerican Studies in Memory of Thelma D. Sullivan*, 402. British Archaeological Reports.

Willard P Gingerich. 1975. A bibliographic introduction to twenty manuscripts of classical nahuatl literature. *Latin American Research Review*, 10(1):105–125.

Ximena Gutierrez-Vasques, Gerardo Sierra, and Isaac Hernandez Pompa. 2016. Axolotl: a web accessible parallel corpus for spanish-nahuatl. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4210–4214.

Frances E Karttunen. 1992. *An analytical dictionary of Nahuatl*. University of Oklahoma Press.

Guillaume Klein, François Hernandez, Vincent Nguyen, and Jean Senellart. 2020. The opennmt neural machine translation toolkit: 2020 edition. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 102–109.

Michel Launey. 1986. *Catégories et opérations dans la grammaire nahuatl*. Ph.D. thesis, Paris 4.

Miguel León-Portilla. 1956. *La filosofía náhuatl estudiada en sus fuentes*. Instituto Indigenista Interamericano, Mexico City.

Miguel León-Portilla. 1985. Nahuatl literature. In *Supplement to the Handbook of Middle American Indians, Volume 3: Literatures*, pages 7–43. University of Texas Press.

Krister Lindén, Miikka Silfverberg, and Tommi Pirinen. 2009. Hfst tools for morphology–an efficient open-source package for construction of morphological analyzers. In *State of the Art in Computational Morphology: Workshop on Systems and Frameworks for Computational Morphology, SFCM 2009, Zurich, Switzerland, September 4, 2009. Proceedings*, pages 28–47. Springer.

James Lockhart. 1992. *The Nahuas after the conquest: A social and cultural history of the Indians of Central Mexico, sixteenth through eighteenth centuries*. Stanford University Press.

James Lockhart. 2001. *Nahuatl as written: Lessons in older written Nahuatl, with copious examples and texts*, volume 6. Stanford University Press.

James Maffie. 2014. *Aztec Philosophy*. University Press of Colorado, Boulder.

Manuel Mager, Monica Jasso Rosales, Özlem Çetinoğlu, and Ivan Meza. 2019. Low-resource neural character-based noisy text normalization. *Journal of Intelligent & Fuzzy Systems*, 36(5):4921–4929.

José Luis Martínez. 1982. *El "Códice Florentino" y la "Historia General" de Sahagún*, 1. edition. Archivo General de la Nación, Mexico City.

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2884–2892, Marseille, France. European Language Resources Association.

Kelly McDonough. 2020. Intercultural (mis) translations: Colonial static and "authorship" in the florentine codex and the relaciones geográficas of new spain. In *The Routledge Hispanic Studies Companion to Colonial Latin America and the Caribbean (1492–1898)*, pages 393–405. Routledge.

Patricia Murrieta-Flores, Diego Jiménez-Badillo, and Bruno Martins. 2022. Digital resources: Artificial intelligence, computational approaches, and geographical text analysis to investigate early colonial mexico. In *Oxford Research Encyclopedia of Latin American History*.

Garrett Nicolai, Dylan Lewis, Arya D. McCarthy, Aaron Mueller, Winston Wu, and David Yarowsky. 2020. Fine-grained morphosyntactic analysis and generation tools for more than one thousand languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3963–3972, Marseille, France. European Language Resources Association.

Garrett Nicolai and David Yarowsky. 2019. Learning morphosyntactic analyzers from the Bible via iterative annotation projection across 26 languages. In

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1765–1774, Florence, Italy. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020a. Universal dependencies v2: An evergrowing multilingual treebank collection. *arXiv preprint arXiv:2004.10643*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020b. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Guilhem Olivier. 2021. Teotl and diablo: Indigenous and christian conceptions of gods and devils in the florentine codex. In *The Florentine Codex*, pages 110–122. University of Texas Press.

Justyna Olko. 2018. Unbalanced language contact and the struggle for survival: Bridging diachronic and synchronic perspectives on nahuatl. *European Review*, 26(1):207–228.

Justyna Olko and John Sullivan. 2013. Empire, colony, and globalization. a brief history of the nahuatl language. In *Colloquia humanistica*, 2. Instytut Slawistyki Polskiej Akademii Nauk.

Justyna Olko et al. 2015. Language encounters: Toward a better comprehension of contact-induced lexical change in colonial nahuatl. *Politeja-Pismo Wydziału Studiów Międzynarodowych i Politycznych Uniwersytetu Jagiellońskiego*, 12(38):35–52.

Robert Pugh, Marivel Huerta Mendez, Mitsuya Sasaki, and Francis M. Tyers. 2022. Universal Dependencies for Western Sierra Puebla Nahuatl. In *Proceedings of the 13th Language Resources and Evaluation Conference*.

Pete Sigal. 2007. Queer nahuatl: Sahagún's faggots and sodomites, lesbians and hermaphrodites. *Ethnohistory*, 54(1):9–34.

Thelma D Sullivan et al. 1966. Pregnancy, childbirth, and the deification of the women who died in childbirth: texts from the florentine codex, book vi, folos 128v-143v. *Estudios de cultura Nahuatl*, 6:63–95.

Francis Tyers and Karina Mishchenkova. 2020. Dependency annotation of noun incorporation in polysynthetic languages. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 195–204.

Universidad Nacional Autónoma de México. 2023. Temoa [en línea]. `http://temoa.iib.unam.mx`; Accessed 4th April, 2023.

Book	Title	Sentences	Tokens	Words
01	The Gods	178	6,066	6,481
02	Ceremonies	664	29,209	–
03	The Origins of the Gods	186	5,794	–
04	The Art of Divination	341	24,283	–
05	The Omens	111	3,546	4,470
06	Rhetoric and Moral Philosophy	1,450	57,021	–
07	The Sun, Moon, Stars, and the Binding of the Years	229	5,189	–
08	Kings and Lords	348	13,711	13,970
09	The Merchants	506	21,022	–
10	The People	1,217	35,196	–
11	Earthly Things	3,074	78,066	–
12	The Conquest of Mexico	667	27,099	–
		8,971	306,202	24,921

Table 5: A breakdown of the FC by book. "Tokens" refers to raw whitespace-separated tokens, prior to the retokenisation process described in §4.2. At present, we have processed approximately 637 sentences containing a total of 25,000 words. We strategically started with the shorter books for manual processing with the idea that we can leverage this data to mostly automate the processing of the longest books.

A Books of the Florentine Codex

Table 5 presents some statistics about the books of the Florentine Codex.