Towards a morphological transducer and orthography converter for Western Tlacolula Valley Zapotec

Jonathan N. Washington
Swarthmore College
500 College Ave.
Swarthmore, PA 19081 USA
jonathan.washington@swarthmore.edu

Felipe H. Lopez
Pueblo of San Lucas Quiaviní &
Haverford College Libraries
370 Lancaster Ave.
Haverford, PA 19072
lieb@ucla.edu

Brook Danielle Lillehaugen
Haverford College
370 Lancaster Ave.
Haverford, PA 19072
blilleha@haverford.edu

Abstract

This paper presents work towards a morphological transducer and orthography converter for Dizhsa, or San Lucas Quiaviní Zapotec, an endangered Western Tlacolula Valley Zapotec language. The implementation of various aspects of the language's morphology is presented, as well as the transducer's ability to perform analysis in two orthographies and convert between them. Potential uses of the transducer for language maintenance and issues of licensing are also discussed. Evaluation of the transducer shows that it is fairly robust although incomplete, and evaluation of orthographic conversion shows that this method is strongly affected by the coverage of the transducer.

1 Introduction

In this paper, we present work towards a morphological transducer and orthography converter for Dizhsa, also known in the academic literature as San Lucas Quiaviní Zapotec (SLQZ), an endangered language variety of Western Tlacolula Valley Zapotec [zab].[1] To our knowledge, this is the first computational implementation of the morphology of a Zapotec language. (Throughout the paper we use the term "language variety" in place of "dialect" because of the pejorative force of the word *dialecto* in Spanish.)

A morphological transducer, implemented as a finite-state transducer (FST), is a tool that performs morphological analysis (converts between a word form and a morphological analysis) and morphological generation (the reverse). For example, a form like *gunydirëng* 'they won't do' can be quickly converted to an analysis like `uny<v><tv><irre><neg>+ëng<prn><pers><p3><prox><pl>` (read as the negative irrealis form of the transitive verb whose stem is "uny", followed by a 3rd person proximal plural personal pronominal enclitic); similarly, the analysis can be quickly converted to the form.

Not all speakers of SLQZ write their language, though more and more are doing so (Lillehaugen, 2016). There are published proposals for two orthographies, which we refer to as the phonemic orthography (Munro & Lopez et al., 1999) and the simple orthography (Munro et al., 2021). An orthography converter between these two orthographies based on the morphological transducer has been developed as part of this work.

Both tools have the potential to support language maintenance efforts. A morphological transducer can be used in various types of computer-assisted language learning software, such as for learning vocabulary (Katinskaia et al., 2018) and complex inflectional systems (Antonsen et al., 2013). FSTs are also used in electronic corpora (Saykhunov et al., 2019), paradigm generators,[2] text-reading tools,[3] and form-lookup dictionaries (Johnson et al., 2013). FSTs may be trivially converted to spell checkers (Washington et al., 2021) and can also be used in other types of text-proofing and language-learning tools (e.g., Antonsen, 2012); they can further serve as core elements of machine translation systems (Khanna et al., 2021).

Morphological transducers are being developed for languages globally (Khanna et al., 2021), including for entire language families, such as Turkic (Washington et al., 2021). Some of these FSTs are developed for languages with large corpora, such as the national languages of Western Europe (Khanna et al., 2021). One advantage of FSTs is that they can be created for a language without a large quantity of existing text. For example, a morphological trans-

[1] The tools presented in this paper are available publicly under a free/open-source license https://github.com/apertium/apertium-zab, and can be used online at https://beta.apertium.org.

The three authors recognise that we live and work on the homeland of the Lenape, and pay respect and honor to the caretakers of this land, from time immemorial until now, and into the future.

[2] Such as the prototype at https://apertium.github.io/apertium-paradigmatrix

[3] Such as https://sanit.oahpa.no/read/.

Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, pages 185–193

ducer has been developed for Zacatlán-Ahuacatlán-Tepetzintla Nahuatl (Pugh et al., 2021), a threatened language of Central Mexico with a relatively small corpus of texts. The fact that a morphological transducer can be developed with small corpora creates an entry point especially for threatened languages into the potential benefits of the types of language technology described above.

This paper is structured as follows. Section 2 situates SLQZ and overviews its socio-political context and basic morphological properties. Section 3 describes the morphological transducer and demonstrates several of the challenges which were overcome in its implementation. Section 4 presents a basic evaluation, including naïve coverage and accuracy of orthographic conversion. Section 5 overviews some issues related to licensing of the tools and section 6 concludes.

2 San Lucas Quiaviní Zapotec

San Lucas Quiaviní Zapotec is spoken by 98% of the population in San Lucas Quiaviní, Oaxaca, Mexico (DIEGPO, 2015) and by diaspora communities elsewhere in Mexico and the United States, especially the greater Los Angeles area (Lopez and Munro, 1999), with approximately 3500 total speakers. While children are still acquiring the variety as their first language, it should be considered endangered as the community is shifting to Spanish in more and more contexts (Munro, 2003; Pérez Báez, 2009).

Western Tlacolula Valley Zapotec encompasses a number of related varieties, with varying degrees of mutual intelligibility. In the present work we focus on the variety of San Lucas Quiaviní (SLQZ), but we also evaluate the transducer on the variety of San Juan Guelavía (SJGZ), also classified as Western Tlacolula Valley Zapotec. The two pueblos are separated by no more than 10km, but the two varieties of Zapotec differ in many relevant aspects of their grammar, including tone and phonation contrasts, verbal morphophonology, and pronominal systems.

Understanding the morphotactics of SLQZ is essential to developing a morphological transducer. A verb form in SLQZ includes at minimum an aspect marker followed by a verb stem, with very few exceptions. Additionally, a negative-marking enclitic, various other adverbial enclitics (Lee, 2006, 26–27), and pronominal enclitcs may follow. Nouns generally may be marked as possessed using a prefix—with some suppletive forms and a class of "essentially possessed" nouns which are always interpreted as possessed. Possessors follow possessed nouns, either as independent noun phrases or as pronominal enclitics. Pronominal enclitics also appear after predicate adjectives.

The morphophonology of verb forms in SLQZ is complex. Aspectual prefixes often have multiple realisations. Additionally, there is a large number of verbs whose stems alternate irregularly or are synchronically suppletive depending on aspect, subject, and any following enclitics. Some aspect markers have irregular realisations in these forms. There may also be changes in phonation type before certain enclitics.

San Lucas Quiaviní Zapotec has a very complicated system of tone and phonation with over 23 potential contrasts in a stressed syllable (although see Chávez Peón, 2010 for a different count). Representing all of these contrasts results in an orthography that is complicated. Members of the speech community have directly and indirectly expressed preference for a practical orthography that underrepresents these contrasts. Hence a phonemic orthography (described first in Munro & Lopez et al., 1999) is used in dictionaries and linguistic work, and a simplified orthography (described in Munro et al., 2021) which collapses many phonemic distinctions, is preferred by speakers of the language. Being able to convert the simplified practical orthography to the phonemic orthography would allow linguists and speech scientists to recover the phonemic contrasts from text written in the practical orthography.

3 Implementation

The transducer was implemented manually using the two-level approach (Koskenniemi, 1983) and is designed for use with HFST (Lindén et al., 2011), an open-source toolkit for finite-state morphology. In the two-level approach to morphology, the lexicon and morphotactics of a language are implemented in one finite-state transducer (FST), the morphophonology is implemented in another, and they two are intersected into a single FST with an analysis side and a form side. For the Dizsha transducer described here, both the morphotactics and morphophonology compile from hand-written patterns, lexicons, and rules. The `lexd` compiler (Swanson and Howell, 2021) was used to implement the morphotactics, and `twol` (part of HFST) was used to implement the morphophonology.

The grammatical patterns of SLQZ were implemented in these formalisms by the first author in part while receiving classroom instruction in the language from the second author and based largely on patterns observed in the first volume of a Dizhsa textbook (Munro et al., 2021). Later the transducer was expanded and revised in consultation with additional volumes of the textbook and other sources cited here, and under the guidance of the second and third authors, the former of whom is a native speaker and teacher of Dizhsa, and the latter of whom is a linguist with expertise in the language.

In section 3.1, the size and shape of the transducer's lexicon is presented. Section 3.2 discusses some design decisions and how some spelling variants were handled. We explain how some aspects of the language's morphotactics (section 3.3) and morphophonology (section 3.4) were implemented. Section 3.5 presents how orthography conversion was implemented.

3.1 Lexicon

The lexical entries of the transducer are divided by stem type based on morphological patterning. Table 1 shows the number of stems of various types in the transducer, and the overall number of stems.

Category	№ stems
Proper nouns	289
Nouns	133
Verbs	92
Pronouns	46
Complex verb elements	28
Adverbs	26
Punctuation	22
Numbers	31
Prepositions	17
Adjectives & determiners	10
Interjections & modal particles	10
Conjunctions	7
total	711

Table 1: The size of the transducer's entire lexicon, broken down by individual lexicons, corresponding to lexical category.

Several of these categories span multiple lexicons. For example, under "verbs" are counted regular verb stems, irregular verb stems (currently spanning two lexd lexicons), and the copula. Additionally, verbs are subcategorised as intransitive (`<iv>`), transitive (`<tv>`), and ditransitive (`<dtv>`). "Pronouns" include both bound and free forms, which must be in separate lexicons due to their different morphological distribution.

3.2 Design decisions

Despite being the best studied variety of Western Tlacolula Valley Zapotec, many aspects of the grammar of SLQZ are not fully documented or described. Even when the patterns are understood, it is not clear whether particular phenomena are best accounted for through morphology or syntax.

For this reason, in many cases during the construction of the transducer, more than one implementation option seemed reasonable. For example, we chose to analyse verb stems followed by the negative marker ⟨di⟩~⟨dy⟩ as an inflected form of the verb stem, as in `uny<v><tv><irre><neg>+ëng<prn><pers><p3><prox><pl>` for *gunydirëng*. We could also have chosen to analyse it as a verb stem followed by an adverbial enclitic, e.g. `uny<v><tv><irre>+di<adv>+ëng<prn><pers><p3><prox><pl>`.

Another such decision is the choice to use verb stems as the lemma for all forms of a verb, and in the case of suppletive stems, the stem that patterns with the habitual aspect. Dictionaries for speakers and learners, such as the glossary in Munro et al. (2021), use the habitual form (prefix+stem) as the headword for entries. The transducer could just as easily use the habitual form as the lemma.

We made similar decisions regarding the lexicon. Some words in SLQZ have common variant pronunciations and corresponding spellings. For example, the word for 'fish' may be spelled *bel* or *beld*. In this case the lemma was chosen to be *beld*, but the generated form was chosen to be *bel*. The form *beld* is still analysed to the same lemma. This was implemented by adding the entry to the transducer with both spellings, and including a comment on the analyse-only variant that triggers the compiler to remove that line while creating the generator, but not the analyser. The lines corresponding to these entries are shown in Code Block 1.

```
beld:bel  behlld:behll   # "fish"
beld:beld behlld:behlld  # "fish" ! Dir/LR
```

Code Block 1: The entries in the `lexd` file for the word for 'fish'. All material after the # symbol is ignored by the compiler, but a preprocessing command strips all lines containing `Dir/LR` before compiling the generator transducer (but not the analyser transducer).

These analyses reflect our best current understanding of the grammar, but it would be trivial to change the implementation in the future.

3.3 Verbal morphotactics

A verb in SLQZ includes an obligatory prefix that signals aspect, optional endings that include a verbal extender (adding politeness) and a negative morpheme, and optional pronominal clitics. This was implemented fairly straightforwardly by defining a general pattern in lexd, shown in Code Block 2.

```
( :Aspect ( V-Stems(1) [<v>:] V-Stems(3):
) V-Extender(1)? Aspect: ) V-Neg(1)?
Prn-Bound(1)
```

Code Block 2: The pattern used for regular verbs in the SLQZ transducer. The numbers in parentheses after each element reference "components", described in section 3.5. The `:` character indicates separation of analysis and form. The `?` character represents optionality. The parentheses after lexicon names indicate column numbers within lexicons. The parentheses grouping parts of the pattern are not strictly necessary, but speed up compilation due to how matching works (described below).

The reason `lexd` was used instead of HFST's `lexc` or Lttoolbox's `dix` formats—the most common choices for implementing a transducer of this type—is because `dix` is not ideal for agglutinative patterns and `lexc` requires complicated tricks (flag diacritics or filter transducers) to implement prefixational morphology. The conventional structure of tag-based morphological analyses is a lemma followed by a part of speech tag, followed by any subcategory tags, followed by any grammatical tags. In an example like *runy* (form) `uny<v><tv><hab>` (analysis), the analysis presents that `uny` is the lemma (in this case a verb stem), `<v>` (verb) is the category of the word, `<tv>` (transitive) is the subcategory of the word, and `<hab>` (habitual) is a grammatical property of the form. Thus in a transducer we can define the form-analysis pairs `<hab>:r` and `uny<v><tv>:uny`, but if combined in that order, the result would be unconventional `<hab>uny<v><tv>:runy`.

The solution to this is lexicon matching, a feature unique to `lexd`. For SLQZ, we can create an `Aspect` lexicon (containing prefixes paired to their analyses, e.g. `<hab>:r`) and a `V-Stems` lexicon (which lists regular verbs). In the pattern that combines these lexicons shown in Code Block 2, the `lexd` compiler keeps track of multiple mentions of a lexicon and matches them. That is, instead of producing forms with all combinations of aspectual prefixes and tags, only the elements of pairs on the same line are used, despite the fact that the elements are referenced at different places in the pattern.

Another lexd-specific feature employed is columns within lexicons. In the pattern, columns 1 and 3 of the `V-Stems` lexicon are referenced. These contain the simple-orthography form of verbs and the subcategory (transitivity) tag, respectively.

Some SLQZ verbs have irregular alternations in their stems when combined with perfective aspect prefixes or a first person plural (1PL) subject. This was implemented using filters in lexd, which allows for entries in a given lexicon which are tagged a certain way to be referenced from patterns, to the exclusion of other entries in that lexicon. In this way, separate patterns can be constructed that pull, e.g., (1) only the 1PL stems and pronoun forms, and (2) only the non-1PL stems and pronoun forms.

3.4 Verbal morphophonology

Many of the phonological alternations in SLQZ verb forms are regular. For example, the negative marker is written before a vowel as ⟨dy⟩, as in *queity runydyai / que'ity ruhnydya'ih* 'I don't do it' and elsewhere as ⟨di⟩ (simplified orthography) ⟨di'⟩ (phonemic orthography), as in *queity runydi Jwanyi / que'ity ruhnydi' Jwaanyih* 'Juan doesn't do it'.

This alternation is implemented by specifying the morpheme with a special character in the morphotactic transducer (`lexd`), as `<neg>:d{I}` and `<neg>:d{I}'` (depending on orthography), and then controlling the alternation of the `{I}` character using a morphophonology transducer (written in `twol`).

The `twol` formalism allows for symbol mappings to be restricted based on context. The mappings needed to condition the correct forms of the ⟨di(')⟩/⟨dy⟩ alternation are presented in Code Block 3. The compiled FST is intersected with the morphotactic transducer to produce correct forms.

```
"di' → dy before vowels: {I}"
%{I%}:y <=> _ (':) %>:* :0* :Vow ;

"di' → dy before vowels: '"
':0 <=> %{I%}:y _ ;
```

Code Block 3: Morphophonological mapping restrictions specified in the `twol` formalism to condition the alternation of `d{I}(')` as ⟨dy⟩ before vowels. In other contexts, `{I}` is realised as ⟨i⟩.[4]

In the transducer's `twol` file, there are currently 10 characters like `{I}` defined, and 20 mapping restrictions specified.

3.5 Orthography

This section outlines both the orthographic support of the transducer and how it is able to be used to convert between orthographies.

The morphological transducer is compiled into two generators: one for each of the simple and phonemic orthographies. A single analyser is compiled that supports both. This is made possible through a combination of the lexd features of lexicon matching and columns in lexicons, both discussed in section 3.3. For example the phonemic-orthography pattern for regular verbs is shown in Code Block 4, and can be compared to the pattern used for simple-orthography regular verbs shown in Code Block 2. The difference between these patterns lies in which column of the lexicons are referenced on the form side. For example, the verb stem lexicon is referenced using `V-Stems(1):V-Stems(2)` instead of `V-Stems(1)` (equivalent to `V-Stems(1):V-Stems(1)`). The second column of the `V-Stem` lexicon (and most lexicons in the transducer, cf. Code Block 1) is the phonemic-orthography form of each stem. The two sides of the lexicon are matched, as opposed to all elements of the first column being paired with all elements of the second column.

```
( :Aspect ( V-Stems(1):V-Stems(2) [<v>:]
V-Stems(3): ) V-Extender(2)? Aspect: )
V-Neg(2)? Prn-Bound(1):Prn-Bound(2)
```

Code Block 4: The pattern used for phonemic-orthography regular verbs used in the SLQZ transducer.

The other crucial part of this approach is control symbols in comments at the end of patterns for each orthography. Specifically, `Orth/Simp` is added to the end of lines containing simple-orthography patterns and `Orth/Dict` is added to the end of lines containing phonemic-orthography patterns. Then, as part of the compilation process for the transducer in each orthography, lines containing the control symbols for the other orthography are removed. This ensures that each transducer contains only forms in a single orthography. The respective analysers and generators are compiled from these pared-down `lexd` files, and the two analysers are unioned, resulting in an analyser that supports both orthographies.

The simple orthography, as discussed in section 2, collapses many of the distinctions made by the phonemic orthography. Because of this, it is mostly trivial to convert from the phonemic orthography to the simple orthography, but not vice versa.

For example, a word like *xyecwa* (simple) / *x:yèe'cwa'* (phonemic) 'my dog' can be converted from phonemic to simple orthography by simply removing the diacritics ⟨:⟩, ⟨ˋ⟩, and ⟨'⟩, and simplifying sequences of repeated vowels. The only other changes needed for most words is the simplification of doubled consonant letters ⟨ll⟩, ⟨mm⟩, and ⟨nn⟩, and the removal of ⟨h⟩ after vowels, e.g. *behlld* → *beld* 'fish'; *rille'eh* → *rile* 'knows how to'. However, as these examples show, conversion in the other direction is non-deterministic.

To convert between the orthographies, then, two transducers which share an interface are intersected along that interface. Specifically, the analyser in one orthography is intersected with the generator in the other orthography along the analysis side of each. This is possible because the analysis side is the same regardless of the orthography. An example of this method applied to one word is shown in Figure 1.

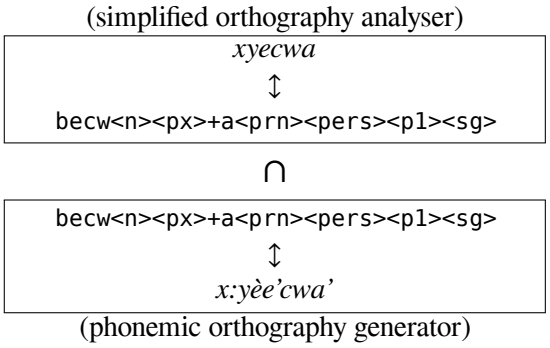

Figure 1: Demonstration of the intersection of two transducers to create an orthographic converter. In this example, an analysis in the simplified orthography analyser is matched to an analysis in the phonemic orthography generator, so that when a simplified orthography form is input to the resulting transducer, the corresponding phonemic orthography form is generated.

This approach provides fairly deterministic output, although as discussed in section 4.3, it does not solve the issue of simple-orthography homography.

One additional approach was used to handle orthographic variants, such as any of the apostrophe characters which might be used and the orthography of the Universal Declaration of Human Rights (UDHR) translation, which is like the phonemic or-

[4]For more on the `twol` formalism and its application, see https://github.com/hfst/hfst/wiki/HfstTwolc.

thography but uses a colon after a vowel to indicate creaky voice, represent by a grave accent over the vowel in the modern version of the phonemic orthography and in the `lexd` file. A "spellrelax" file, containing a series of regular expressions like those shown in Code Block 5, is compiled to an FST and intersected with an analyser. This allows it to accept forms with any of the specified variants used.

```
[ ?* [ ’ (->) [ %' | %’ | %` | %´ | %‘ | %‘
] ] ?* ] .o.
[ ?* ( à (->) [ a [ %: | : ] ] ) ?* ]
```

Code Block 5: Two of the regular expressions contained in the spellrelax file. The first one allows any number of apostrophe characters to be used in place of ⟨’⟩, and the second one allows for ⟨a⟩ followed by one of two colon characters to be used in place of ⟨à⟩. The `.o.` symbol conjoins the patterns.

4 Evaluation

The transducer was evaluated over available texts (4.1) for naïve coverage (4.2) and accuracy of orthographic conversion (4.3).

4.1 Texts used for evaluation

The transducer was evaluated against a number of available texts, including a number of genres in both the simple and phonemic orthographies.

The first two parts of the story *Blal xte Tiu Pamyël* (BxTP) are part of Munro et al. (2021), which is also the source for nearly all of the material in the transducer. A preliminary version of the transducer was evaluated using BxTP parts 1–2, whereafter the transducer was expanded to include unrecognised forms. Hence, BxTP parts 1–2 are treated as development data, and the remaining texts are treated as previously unseen data. Evaluating the transducer over BxTP parts 1–2 also allowed us to observe and correct mismatches between the phonemic and simplified orthographic versions.

A number of poems and stories were also used for evaluation. Those from Tlalocan are in individualised orthographies inspired by the phonemic orthography (Munro, 2014). There is also a blog post from the Ticha blog entirely in Dizhsa. The Universal Declaration of Human Rights (UDHR) is in an older version of the phonemic orthography, which is easily handled by the transducer due to the addition of some spellrelax mappings.

We also evaluated a translation of the New Testament in SJGZ, a language variety closely related to SLQZ which uses a distinct orthography.

The complete list of texts is presented in Table 2, along with naïve coverage results (see section 4.2). The sources for each set of texts are described in footnotes to the table.[5]

4.2 Naïve coverage

Naïve coverage was calculated as the percentage of tokens in a given corpus that received an analysis from the transducer, whether correct or not. Results are shown in Table 2.

The results show that the development text has good coverage, at over 90%—higher, not unexpectedly, than coverage over the remaining sources. Unseen texts vary, but average around two thirds coverage, as does the coverage over all available material. This indicates that the transducer has a solid base, but has many opportunities for expansion. It should also be noted that the development text, besides functioning as a graded reader in an introductory textbook for the language, is relatively short, and so lacks a wide range of vocabulary and morphological patterns.

The lower overall coverage on texts in the phonemic orthography is due primarily to the lack of phonological mappings accounting for all diacritic changes in verb forms, and the homography of the simple orthography. In the simple orthography many words are written the same that are written distinctly in the phonemic orthography. Words that are not in the transducer may receive an incorrect analysis, thus inflating the apparent coverage of texts in the simple orthography.

The individualised orthographies found in the Tlalocan texts are inspired by, but not the same as, the phonemic orthography, yielding much lower coverage results.

The translation of the New Testament in the related language variety of SJGZ, totalling 217K tokens, was also evaluated to test whether the SLQZ transducer could be applied to Western Tlacolula

[5] The entire set of texts is currently available at `https://github.com/jonorthwash/apertium-zab-corpus`. All testing was done on the contents of the transducer repository at revision `0866ec3` and the corpus repository at revision `85fda5c`.

[6] Munro et al. (2021)

[7] Drawn from Lopez and Lillehaugen (2018), Lopez and Lillehaugen (2017), and `https://felipehlopez.weebly.com/`.

[8] Chávez Peón and López Reyes (2009)

[9] Lopez (2018)

[10] `https://ticha.haverford.edu/updates/`

[11] `https://www.ohchr.org/EN/UDHR/Pages/Language.aspx?LangID=ztu1`

Use	Text	Orthography	Tokens	Coverage (%)
development	*Blal xte Tiu Pamyël* 1–2[6]	Simple	625	93.92
		Phonemic	628	91.40
testing	*Blal xte Tiu Pamyël* 3–7	Simple	1532	73.56
	Blal xte Tiu Pamyël 3–4	Phonemic	601	66.89
	Felipe H. Lopez poetry[7]	Simple	514	57.39
	Tlalocan poems & story[8]	Simple	635	57.95
		Individualised	788	47.72
	Niny Bac[6]	Simple	366	73.77
	Liaza Chaa[9]	Simple	963	58.67
	Ticha post 2020-07-17[10]	Simple	1026	60.04
	UDHR (9 articles)[6]	Simple	433	69.98
	UDHR (complete)[11]	Phonemic	1641	65.63
total	all	mixed	9934	67.47

Table 2: Naïve coverage results. BxTP 1–2 was used for development, and the remaining texts were used for testing. Tokens is the number of lexical units according to the transducer, and coverage is the percentage of tokens that received at least one analysis from the transducer.

Valley Zapotec more broadly. Even with a dedicated spellrelax transducer to account for a number of orthographic differences, the coverage was only a little over 34%. This suggests that perhaps a single transducer for Western Tlacolula Valley Zapotec may not be able to be applied to all varieties.

4.3 Orthographic conversion

The first four sections of *Blal xte Tiu Pamyël* are available in both the simple and phonemic orthography. To test orthographic conversion, we created two groups of texts, the first group consisting of sections 1 and 2 of BxTP and the second group consisting of sections 3 and 4.

The conversion of phonemic to simple orthography is almost entirely deterministic. We set up a simple regular expression (regex) replacement conversion system, which removed diacritics and ⟨h⟩ after vowels and also merged adjacent characters which were identical. The performance of this method provides a baseline measure of similarity between the two texts.

Performance was measured using Word Error Rate (WER), or the percentage of words that are different between the converted text and the "gold standard" of the text in the destination orthography. The results of both the regex-based method and the transducer-based method described in section 3.5 are presented in table 3.

The performance of the transducer-based approach has a ceiling defined by the level of coverage and the similarity of the two texts. For example, for phonemic→simple conversion of the first text, it would be impossible to get better (lower) than 8.6% WER, since the text has naïve coverage of 91.4%. None of the words which do not have an analysis in the transducer are able to be converted—although there is a possibility that some of those words would be "free rides", or words that are the same in both orthographies. The result of 11.78% WER should be taken in the context of this ceiling.

In the first group (BxTP 1–2), the simple-to-phonemic conversion performed worse than phonemic-to-simple, despite higher coverage of the source version. This is largely due to homography. While performing disambiguation between available analyses before orthography conversion might improve this result, there are some simple-orthography homographs that may never be possible to accurately decide between (without wider context), such as *re*, corresponding to both phonemic-orthography *rèe* 'that' and *rèe'* 'this'.

The second group of texts (BxTP 3–4) has much lower correspondence between the two orthographies than the first group due to slight differences between the texts, such as words or sentences that seem to be present in one version but absent in the other. That together with the lower coverage over the second group to start with compound for much worse performance.

While phonemic→simple orthography conversion is deterministic (and hence possible to perform

Text	Direction	Method	Tokens	Coverage (%)	WER (%)
Blal xte Tiu Pamyël 1–2	Simple→Phonemic	transducer	625	93.92	20.10
	Phonemic→Simple	transducer	628	91.40	11.78
	Phonemic→Simple	regex	”	”	1.63
Blal xte Tiu Pamyël 3–4	Simple→Phonemic	transducer	574	77.53	46.75
	Phonemic→Simple	transducer	601	66.89	46.79
	Phonemic→Simple	regex	”	”	10.35

Table 3: Orthographic conversion accuracy. Tokens is the number of lexical units according to the transducer, coverage is the percentage of tokens that received at least one analysis from the transducer, and WER is word error rate, or the percentage of tokens after orthography conversion that do not correspond to the text in the other orthography.

accurately with a series of regular expressions), simple→phonemic conversion is not, and hence must be done in some other way. These initial experiments in using a lexical approach show that it is a viable method, although it currently suffers from the low overall coverage of the transducer.

5 Licensing

We have chosen to license this work under the GNU Affero General Public License (AGPL) because we want it to be available for others to use and build on. This work is also part of a long-term commitment to collaboration with Zapotec communities and community members. The AGPL license allows for uses of our work that would be inconsistent with our commitment to the community.

Reciprocity is a defining Zapotec cultural value and practice. Zapotec speakers have shared their knowledge and language in the creation of these resources. Others are allowed to use the tools and in doing so enter into a reciprocal commitment with the Zapotec community that we define in what we call the Guelaguetza clause, shown below:

> While licensed under a free/open-source license that permits commercial uses, it is expected that anything created using this resource be made available to the community of San Lucas Quiaviní free of charge. This is consistent with the community's practice of guelaguetza, a complex system of reciprocity and exchange of goods and labor.

This context reminds us that the more broadly available licenses could use refinements in particular cultural contexts, particularly Indigenous contexts, and that the field should be open to discussions of how culturally specific practices may interact with open source licensing.

6 Conclusion

This paper has overviewed the development of a morphological transducer and orthography converter for San Lucas Quiaviní Zapotec.

An evaluation of the analyser over available texts demonstrates that despite being incomplete, it is fairly robust. Future work to improve the transducer will focus on expanding the lexicon, adding missing morphological patterns, refining the morphophonological patterns, and finding better ways to deal with the nuances of SLQZ verb morphology.

Text in another variety of Western Tlacolula Valley Zapotec was evaluated using the morphological transducer, and the results suggest that a separate transducer might be needed.

An evaluation of the orthography converter shows that this method of orthography conversion has potential, but is affected heavily by the coverage of the transducer.

It is our hope that this resource will be useful to the SLQZ community. In particular, we are excited about the many roles it could play in language maintenance efforts. This work also impacts conversations on language technology for under-resourced languages and open licensing in Indigenous contexts.

Acknowledgements

The authors are grateful to the anonymous reviewers and editors of this volume. We also thank Daniel Swanson for his assistance with lexd. Felipe and Brook express their gratitude to Pamela Munro, who was integral in their understanding of Zapotec grammar. Runybia ra xauzanën, ra Bunyza ni za Quiabni, bsan Dizhsa ni zezaneën nazhi.

References

Lene Antonsen. 2012. Improving feedback on L2 misspellings – an FST approach. In *Proceedings of the SLTC 2012 workshop on NLP for CALL*, pages 1–10.

Lene Antonsen, Ryan Johnson, Trond Trosterud, and Heli Uibo. 2013. Generating modular grammar exercises with finite-state transducers. In *Proceedings of the second workshop on NLP for computer-assisted language learning at NODALIDA 2013*, pages 27–38.

Mario E. Chávez Peón. 2010. *The interaction of metrical structure, tone, and phonation types in Quiaviní Zapotec*. Ph.D. thesis, The University of British Columbia.

Mario E. Chávez Peón and Román López Reyes. 2009. Zidgyni zyala rnalaza liu 'Vengo de la luz del amanecer, recordándote'. Cuatro poemas y un cuento del zapoteco del Valle. *Tlalocan*, 16:17–49.

DIEGPO. 2015. San Lucas Quiaviní. libro demográfico.

Ryan Johnson, Lene Antonsen, and Trond Trosterud. 2013. Using finite state transducers for making efficient reading comprehension dictionaries. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 59–71. Linköping University Electronic Press, Sweden.

Anisia Katinskaia, Javad Nouri, and Roman Yangarber. 2018. Revita: a language-learning platform at the intersection of ITS and CALL. In *Proceedings of LREC: 11th International Conference on Language Resources and Evaluation*, Miyazaki, Japan.

Tanmai Khanna, Jonathan N. Washington, Francis M. Tyers, Sevilay Bayatlı, Daniel G. Swanson, Tommi A. Pirinen, Irene Tang, and Hèctor Alòs i Font. 2021. Recent advances in Apertium, a free / open-source rule-based machine translation platform for low-resource languages. *Machine Translation*.

Kimmo Koskenniemi. 1983. *Two-level Morphology: A General Computational Model for Word-Form Recognition and Production*. Ph.D. thesis, University of Helsinki.

Felicia Lee. 2006. *Remnant Raising and VSO Clausal Architecture: A Case Study of San Lucas Quiaviní Zapotec*. Springer, Dordrecht.

Brook Danielle Lillehaugen. 2016. Why write in a language that (almost) no one can read? twitter and the development of written literature. *Language Documentation and Conservation*, 10:356–392.

Krister Lindén, Erik Axelson, Sam Hardwick, Tommi A. Pirinen, and Miikka Silfverberg. 2011. HFST—framework for compiling and applying morphologies. *Communications in Computer and Information Science*, 100:67–85.

Felipe H. Lopez. 2018. Liaza chaa / I'm going home. *Latin American Literature Today*, 1(7). With Brook Danielle Lillehaugen, translator.

Felipe H. Lopez and Brook Danielle Lillehaugen. 2017. Mam and Guepy: Two Valley Zapotec poems. *Latin America Literary Review*, 44(88):83–84.

Felipe H. Lopez and Brook Danielle Lillehaugen, translator. 2018. Seven poems. *Latin American Literature Today*, 1(7).

Felipe H. Lopez and Pamela Munro. 1999. Zapotec immigration: The San Lucas Quiaviní experience. *Aztlán*, 24:129–49.

Pamela Munro. 2003. Preserving the language of the Valley Zapotecs: The orthography question. Presented at Conference on Language and Immigration in France and the United States: Sociolinguistic Perspectives.

Pamela Munro. 2014. Breaking rules for orthography development. In Michael Cahill and Keren Rice, editors, *Developing orthographies for unwritten languages*, pages 169–189. SIL International Publications, Dallas.

Pamela Munro, Brook Danielle Lillehaugen, and Felipe H. Lopez with Benjamin Paul. 2021. *Cali Chiu? A Course in Valley Zapotec*, 2nd edition. Haverford College Libraries Open Educational Resources.

Pamela Munro and Felipe H. Lopez with Rodrigo Garcia & Olivia Mendez. 1999. *Di'csyonaary X:tèe'n Dìi'zh Sah Sann Luu'c (San Lucas Quiaviní Zapotec Dictionary / Diccionario Zapoteco de San Lucas Quiaviní)*. UCLA Chicano Studies Research Center.

Robert Pugh, Francis M. Tyers, and Marivel Huerta Mendez. 2021. Towards an open source finite-state morphological analyzer for Zacatlán-Ahuacatlán-Tepetzintla Nahuatl. In *Proceedings of the 4th Workshop on Computational Methods for Endangered Languages*, volume 1, pages 80–85.

Gabriela Pérez Báez. 2009. *Endangerment of a transnational language: the case of San Lucas Quiaviní Zapotec*. Ph.D. thesis, State University of New York at Buffalo.

M.R. Saykhunov, R.R. Khusainov, and T.I. Ibragimov. 2019. Сложности при создании текстового корпуса объемом более 400 млн токенов. In *Финно-угорский мир в полиэтничном пространстве России: культурное наследие и новые вызовы*, pages 548–554. UdmFITS UrO RAN.

Daniel Swanson and Nick Howell. 2021. Lexd: A finite-state lexicon compiler for non-suffixational morphologies. In Mika Hämäläinen, Niko Partanen, and Khalid Alnajjar, editors, *Multilingual Facilitation*, pages 133–146. Helsingin yliopisto.

Jonathan N. Washington, Ilnar Salimzianov, Francis M. Tyers, Memduh Gökırmak, Sardana Ivanova, and Oğuzhan Kuyrukçu. 2021. Free/open-source technologies for Turkic languages developed in the Apertium project. In *Proceedings of the Seventh International Conference on Computer Processing of Turkic Languages (TurkLang 2019)*.