ComputEL 2025

Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages

Proceedings of the Workshop

March 4-5, 2025

The ComputEL organizers gratefully acknowledge the support from the following sponsors.

Gold

Google

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN None

Introduction

These proceedings contain the papers presented at the 8th Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-8), held on March 4–5, 2025 in Honolulu, Hawai‘i. The workshop is co-located with the 9th International Conference on Language Documentation & Conservation (ICLDC9) and offers hybrid attendance options, enabling participants to join either in-person or remotely.

As the name implies, this is the eighth workshop dedicated to the intersection of computational tools and endangered language research. The inaugural event took place at the Association for Computational Linguistics (ACL) main conference in Baltimore, Maryland in 2014. Subsequent workshops have been co-located with the International Conference on Language Documentation & Conservation at the University of Hawai‘i at Mānoa (2017, 2019, 2021, 2023) or ACL-related venues (2022 in Dublin, Ireland; 2024 in St. Julians, Malta). We are delighted to continue this tradition by returning to Honolulu, marking the fifth time the workshop has been held alongside ICLDC.

The primary aim of ComputEL-8 is to bring together computational researchers, documentary linguists, and community language practitioners. By uniting these diverse groups, the workshop fosters a collaborative environment for exchanging ideas, methods, and resources that support the documentation and revitalization of endangered languages. The organizers are gratified by the variety of contributions, reflecting the importance of collaborative efforts across different disciplines and communities.

This year, we received 45 submissions in the form of papers or extended abstracts. Following a thorough review process, 30 were accepted. In addition, 3 presentations formed our special session, titled “Building Tools Together,” which focused on strategies for joint development of language resources and technologies.

We extend our appreciation to all authors for their submissions and to the Program Committee for the thoughtful review of each proposal. We also thank the ICLDC9 organizers for their assistance in hosting this workshop. We hope that ComputEL-8 sparks discussions and partnerships that continue to enrich the field of endangered language research, ultimately contributing to more robust support for language communities worldwide.

Organizing Committee

General Chair

Jordan Lachler, University of Alberta

Deputy General Chair

Godfred Agyapong, University of Florida

Program Chairs

Antti Arppe, University of Alberta
Aditi Chaudhary, Google Research
Sarah Moeller, University of Florida
Shruti Rijhwani, Birla Institute of Technology and Science, Pilani
Daisy Rosenblum, University of British Columbia

Program Committee

Chairs

Jordan Lachler, University of Alberta
Godfred Agyapong, University of Florida
Antti Arppe, University of Alberta
Aditi Chaudhary, Google Research
Sarah Moeller, University of Florida
Shruti Rijhwani, Birla Institute of Technology and Science, Pilani
Daisy Rosenblum, University of British Columbia

Program Committee

Milind Agarwal, George Mason University
Felix Ameka, Leiden University Centre for Linguistics
Antonios Anastasopoulos, George Mason University
Gregory Anderson, Living Tongues Institute for Endangered Languages
Candy Angulo, SUNY at Buffalo
Helen Aristar-Dry, ICHL
Dorothee Beermann, Norwegian University of Science and Technology
Martin Benjamin, Kamusi Project International
Andrea Berez-Kroeker, University of Hawaii at Mānoa Department of Linguistics
Claire Bowern, Yale University
Katia Chirkova, CRLAO, CNRS
Rolando Coto-Solano, Dartmouth College
Christopher Cox, Carleton University
Anna Luisa Daigneault, Living Tongues Institute for Endangered Languages
Elizaveta Dorofeeva, Moscow State University
Suzanne Duncan, Te Hiku Media
Bill Dyer, University of Florida
Mengzhe Geng, National Research Council Canada
Luke Gessler, Indiana University
Jeff Good, University at Buffalo
Michael Wayne Goodman, University of Washington
Rachael Griffiths, EPHE
Njabulo Hadebe, Computational linguist
Christopher Hammerly, University of British Columbia
William N. Havard, LLL, Université d'Orléans and CNRS
Ryan Henke, University of Hawaii at Mānoa
Gary Holton, Univeristy of Hawai'i
David Huggins-Daines, Independent Researcher
Benjamin Hunt, George Mason University
Xin He Jiang, University of Victoria
Anagha Narasimha Joshi, University of Georgia, Athens, GA
Seth Katenkamp, Yale University
Anna Kazentseva, National Research Council of Canada
František Kratochvíl, Palacký University
Olga Kriukova, University of Saskatchewan
Roland Kuhn, unknown

Éric Le Ferrand, Boston College
Dylan Leddy, Boston College
Gianna Leoni, Te Reo Irirangi o Te Hiku o te Ika
Gina-Anne Levow, University of Washington
Patrick Littell, National Reseach Council Canada
Zoey Liu, University of Florida
Olga Lovick, University of Saskatchewan
Kavya Manohar, Digital University Kerala
Bradley McDonnell, University of Hawaii
Mel Mistica, Melbourne University
Timothy Montler, UNT
Emmanuel Ngue Um, University of Yaoundé 1
Ake Nicholas, University of Auckland
William O'Grady, University of Hawaii at Manoa
Maura O'Leary, Western Washington University
Shu Okabe, TUM
Aidan Pine, National Research Council Canada
Emily Prudhommeaux, Boston College
Robert Pugh, Indiana University
Manny Rayner, University of South Australia
Aleksandr Riaposov, Universität Hamburg
Enora Rice, University of Colorado
Elizabeth Salesky, Johns Hopkins University
Nay San, Stanford University
Emmanuel Schang, LLL, Univ. Orléans and CNRS
Yves Scherrer, University of Oslo
Katherine Schmirler, University of Lethbridge
Miikka Silfverberg, UBC
Mark Simmons, University of California San Diego
Gary Simons, SIL International
Sonal Sinha, Department of Linguistics, k.m. institute of hindi and linguistics, B.R.Ambedkar University
Conor Snoek, University of Lethbridge
Ngoc Tan Le, Industrial University of Ho Chi Minh, Université du Québec à Montréal
Nick Thieberger, The University of Melbourne
Maria Belen Ticona Oquendo, University of Buenos Aires
Jörg Tiedemann, University of Helsinki
Paul Trilsbeek, Max Planck Institute for Psycholinguistics
Trond Trosterud, UiT
Francis M. Tyers, Indiana University Bloomington
Daan van Esch, Google
Sahiinii Lemaina Veikho, University Of Bern, Switzerland
Nitin Venkateswaran, University of Florida
Sunny Walker, University of Hawaii at Hilo
Olivia Waring, University of Hawai'i Mānoa, Department of Linguistics
Jonathan Washington, Swarthmore College
Linda Wiechetek, UiT
Cheyenne Wing, University of Arizona
Winston Wu, University of Hawaii at Hilo
Fei Xia, University of Washington
Changbing Yang, University of British Columbia

Table of Contents

Formalizing the Morphology of Rromani Adjectives

Masako Watabe
Université de Franche-Comté
masako.watabe@univ-fcomte.fr

Max Silberztein
Université de Franche-Comté
max.silberztein@univ-fcomte.fr

Abstract

This paper presents a set of linguistic resources that formalizes the morphological behavior of simple Rromani adjectives. We describe the formalization of the adjectives' morphology and the implementation with the NooJ linguistic platform of an electronic dictionary associated with a formal morpho-syntactic grammar. We can then apply this set of resources to a corpus to evaluate the resources and automatically annotate adjectival forms in Rromani texts. The final set of resources can then be used to identify each Rromani dialectal variant and can be used as a pedagogical tool to teach Rromani as a second language.

1 Introduction

1.1 Rromani language

Rromani is the language of the Rromani people; it is an Indo-Aryan language. The number of Rromani speakers is estimated at 5.5 million (Gurbetovski, M. et al. 2010). UNESCO's "Atlas of the World's Languages in Danger" classifies Rromani as a "definitely endangered[1]" language (UNESCO. 2010). There are four Rromani dialects, formed by two isoglosses combining with each other (Courthiade, M. 2016):

- The first isoglossal criterion concerns the opposition between "o" and "e," *e.g.*, *phirdom vs. phirdem* [I walked], *o Rroma vs. e Rroma* [the Rroms].
- The second isoglossal criterion concerns the phonetic mutation of two consonants: the alveolar affricates "ʧʰ" and "ʤ" transform into alveolar-palatal fricatives [ɕ] and [ʑ], *e.g.*, "ʧʰavo" *vs*. "ɕavo" [Rromani boy, son], "ʤukel" *vs*. "ʑukel" [dog].

These four dialects are not areal: Rromani speakers living in nearby regions do not necessarily speak the same dialects, and the same dialect is used in distant countries.

The Rromani alphabet was standardized at the International Rromani Union Congress in 1990, see Figure 1.

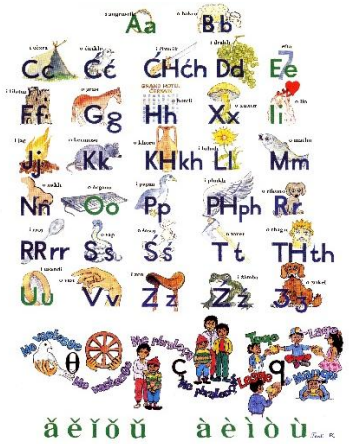


Figure 1: The Rromani standardized alphabet

If all Rromani speakers transcribe, for example, the word *ćhib* [language] using their local alphabets, there can be up to 60 different spellings. The written word *ćhib* is an underlying form including four possible pronunciations: [ʧʰb], [ʧʰp], [ɕib], and [ɕip], see Figure 2. The standardized alphabet enables speakers of different dialects to

[1] The UNESCO list has six categories of danger: Stable yet threatened, vulnerable, definitely endangered, severely endangered, critically endangered, extinct.

Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 1–10

understand each other in writing, giving them comfort in pronunciation.

No other standardization exists: neither lexical, nor grammatical, nor phonetic.

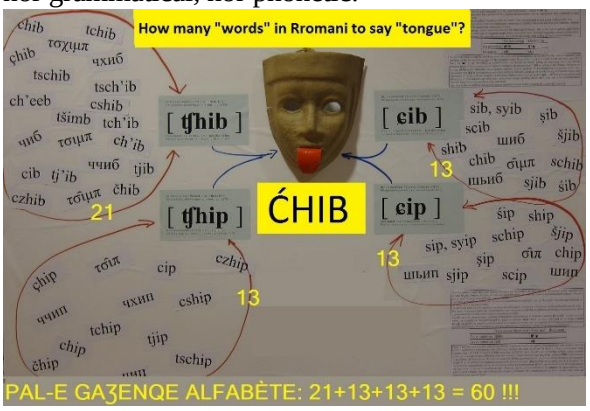


Figure 2: 60 different spellings of the word *ćhib* [language]

1.2 Natural Language Processing (NLP) applications in Rromani

Most of the NLP applications, such as Sketch Engine and DeepL, do not support Rromani. Very few NLP applications support Rromani, but always unsatisfactorily. For example, Rromani has been integrated into Google Translate in 2024. Translation quality is misleading at all levels: lexical, grammatical, orthographic, and dialectal. For example, translating "*thule bakră*" which means either "fat ewes" in the direct plural or "fat ewe" in the oblique singular, Google produces the single translation "fat goats[2]", see Figure 3.

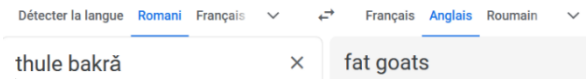


Figure 3: *thule bakră* translated by Google Translate (accessed September 12, 2024).

Conversely, Google Translate produces the translation "*thulo bakro*" for "fat sheep" correctly, whereas its translation for *fat ewe*: "*thuli bakro*" is incorrect at both lexical and grammatical levels, because the noun "ewe" should be translated as *bakri*, and there is a disagreement between the feminine adjective form *thuli* and the masculine noun *bakro*, see Figure 4.

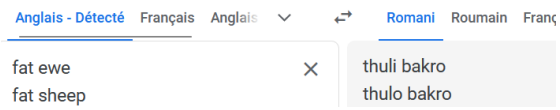


Figure 4: "fat ewe" and "fat sheep" translated by Google Translate (accessed October 2, 2024).

In addition, Google Translate incorrectly translates "my daughter, my daughters, the daughter, the daughters" as *miri ćhaj, mire čhaja, e shej, ćhaja* in Rromani. There are several problems at the grammatical, orthographic, and dialectal levels; the definite article in the plural (*e* or *le* depending on the dialect) is omitted, three different graphemes "ćh," "čh," and "sh" are confused, and two dialectal variants *ćhaj* and *ćhej*[3] are confused. Users, especially learners, would be lost.

Other NLP resources for Rromani are Russian Romani Corpus and ROMLEX. However, they do not adopt the standardized alphabet and do not clearly show the correspondences of dialectal variants, therefore, non-scientists and learners of Rromani cannot easily use them.

Facebook tries to process Rromani, but fails. When one posts a text in Rromani, Facebook will incorrectly recognize the source language (for example as French [4]), and then translate it incorrectly and partially (Watabe, M. 2024). For example, Facebook "translated" in French the Rromani text "Baxtalo 8 Aprilo savore Romenge, oven saste!" meaning [Happy April 8, for all the Rroms, may you be healthy!] but its translation "Baxtalo 8 avril Romenge au flaveur, saste de four!" is not a French text, see Figure 5.

Figure 5: A Rromani text incorrectly translated by Facebook (accessed April 8 2023).

[2] Rromani distinguishes lexically "ewe," "sheep," "she-goat," and "he-goat." Their equivalent words in Rromani are *bakri*, *bakro*, *buzni*, and *buzno*.

[3] This variant is transcribed as *shej* by Google Translate.

[4] French is the default language of the Facebook account of one of the authors, *i.e.*, Watabe, M.

2 Our project

We aim to describe the Rromani language by developing linguistic resources in the form of formalized dictionaries and grammar. Our initial dictionary, based on two small corpora; a two-page story written by a Rromani teacher (Duka, J. MS.) and a one-page poem (Đurić, R. 2006), contained only 747 lexical entries associated with a well-developed morphological grammar that includes 179 inflectional paradigms and 11 derivational paradigms [5] for nouns, verbs, adjectives, and grammatical words. A feature of these resources is that they take into account Rromani four dialects, as well as a few vernaculars.

An editorial dictionary (Courthiade, M. et al. 2009) including the four dialects of Rromani explains Rromani morphology in the grammar section. It is our principal lexical and grammatical resource.

In this paper, we are addressing the problem of describing adjectives and their inflection, which causes massive ambiguities.

3 Rromani adjectives

The inflectional morphology of Rromani adjectives is governed by two genders (masculine and feminine), two numbers (singular and plural), and two cases (direct and oblique[6]). Adjectival forms are according to noun genders, numbers, and cases. The basic form of adjectives is the masculine singular direct. Combining these three properties produces eight possibilities; in practice, however, most adjectives have no more than three forms (Courthiade, M. et al. 2009. Sarău, G. 2009). Consequently, there are many inflectional homonyms.

Most Rromani words are oxytonic; *i.e.*, the tonic stress is on the last syllable. One uses a grave accent to mark the stress when it is not on the last syllable. For example, *bakri* [ewe] and *thulo* [fat, thick, dense] are oxytonic, whereas *profesòri* [professor] and *sociàlo* [social] are non-oxytonic. The opposition oxytonic *vs.* non-oxytonic plays a role in inflectional morphology.

3.1 Oxytonic adjectives

Oxytonic adjectives are classified into four classes: large adjectives, narrow adjectives, plural adjectives, and invariable adjectives.

Large adjectives[7]: Large adjectives are vocalic and have three suffixes: "-o" in the masculine singular direct, "-i" in the feminine singular direct, and "-e" in the plural direct for both genders, as well as oblique for both genders and numbers, see an example of the adjective *thulo* [fat, thick, dense] in Table 1:

Form	Gender	Number	Case
thulo	masculine	singular	direct
thuli	feminine	singular	direct
thule	masculine	plural	direct
thule	feminine	plural	direct
thule	masculine	singular	oblique
thule	feminine	singular	oblique
thule	masculine	plural	oblique
thule	feminine	plural	oblique

Table 1: Inflected forms and properties of the adjective *thulo* [fat, thick, dense]

The form *thule* is therefore 6-time ambiguous. This high level of ambiguity is general in Rromani; as a matter of fact, we do not know any Rromani adjective that would inflect to eight different forms, each for each combination of properties.

Narrow adjectives [8] : Narrow adjectives are consonant, and the direct forms of both genders are identical in each number: "-Ø" in the direct singular and "-a" in the direct plural. The suffix of the oblique is "-e" for both genders and numbers, as in "large" adjectives, see an example of the adjective *godăver* [intelligent] in Table 2:

Form	Gender	Number	Case
godăver	masculine	singular	direct
godăver	feminine	singular	direct
godăvera	masculine	plural	direct
godăvera	feminine	plural	direct

[5] Only the diminutive and abstract nouns with the suffix "-pen" are described in the current Rromani module.

[6] In Rromani, the direct case of human and most animal nouns is used as a subject, while the oblique case is used as an object complement. The direct case of inanimate object nouns is used as a subject and an object complement.

[7] The Rromani adjective *buxlo* [large] is the origin of this designation.

[8] The Rromani adjective *tang* [narrow] is the origin of this designation.

godăvere	masculine	singular	oblique
godăvere	feminine	singular	oblique
godăvere	masculine	plural	oblique
godăvere	feminine	plural	oblique

Table 2: Inflected forms and properties of the adjective *godăver* [intelligent]

Plural adjectives: Plural adjectives are used specifically with plural nouns. In the direct case, they have the suffix "-∅" for both genders, and in the oblique "-e" for both genders, see an example of the adjective *but* [many, numerous] in Table 3:

Form	**Gender**	**Number**	**Case**
but	both	plural	direct
bute	both	plural	oblique

Table 3: Inflected forms and properties of the adjective *but* [many, numerous]

Invariable adjectives: So-called "international" adjectives have a tendency to be invariable. International oxytonic adjectives are completely invariable, see an example of the adjective *bordo* [Bordeaux-colored] in Table 4:

Form	**Gender**	**Number**	**Case**
bordo	both	both	both

Table 4: Invariable form and properties of the adjective *bordo* [Bordeaux-colored]

3.2 Non-oxytonic adjectives

Borrowed adjectives and suffixed adjectives: Inflectional paradigms of borrowed non-oxytonic adjectives and suffixed non-oxytonic adjectives are identical. Their suffix is "-o" in the direct singular for both genders, "-a" in the direct plural for both genders, and "-one" in the oblique for both genders and numbers. Oblique forms are oxytonic, therefore the stress is not marked, see an example of the adjective *veśìtko* [of the woods, wild] in Table 5:

Form	**Gender**	**Number**	**Case**
veśìtko	masculine	singular	direct
veśìtko	feminine	singular	direct
veśìtka	masculine	plural	direct
veśìtka	feminine	plural	direct
veśitkone	masculine	singular	oblique
veśitkone	feminine	singular	oblique
veśitkone	masculine	plural	oblique
veśitkone	feminine	plural	oblique

Table 5: Inflected forms and properties of the adjective *veśìtko* [of the woods, wild]

International adjectives: Compared to international oxytonic adjectives (*e.g., bordo* [Bordeaux-colored]), international non-oxytonic adjectives are not completely invariable. International non-oxytonic adjectives have two suffixes: "-o" in the direct for both genders and numbers and "-one" in the oblique for both genders and numbers. Oblique forms are oxytonic, therefore the stress is not marked, see an example of the adjective *sociàlo* [social] in Table 6:

Form	**Gender**	**Number**	**Case**
sociàlo	masculine	singular	direct
sociàlo	feminine	singular	direct
sociàlo	masculine	plural	direct
sociàlo	feminine	plural	direct
socialone	masculine	singular	oblique
socialone	feminine	singular	oblique
socialone	masculine	plural	oblique
socialone	feminine	plural	oblique

Table 6: Inflected forms and properties of the adjective *sociàlo* [social]

3.3 Conclusion

We have defined six classes of Rromani adjectives, according to their morphological properties:

- Oxytonic vocalic adjectives ending in "-o": *e.g., thulo* [fat, thick, dense], *buxlo* [large],
- Oxytonic consonant adjectives: *e.g., godăver* [intelligent], *tang* [narrow],
- Oxytonic consonant adjectives used only in the plural: *e.g., but* [many, numerous],
- Oxytonic vocalic international adjectives totally invariable: *e.g., bordo* [Bordeaux-colored], *pane* [breaded],
- Non-oxytonic borrowed adjectives and non-oxytonic suffixed adjectives: *e.g., zèleno* [green], *veśìtko* [of the woods, wild],
- Non-oxytonic vocalic international adjectives ending in "-o": *e.g., sociàlo* [social], *interesànto* [interesting].

4 Formalization of the Rromani vocabulary

4.1 The NooJ linguistic platform

NooJ is a linguistic development environment linguists use to describe natural languages, by constructing linguistic resources in the form of electronic dictionaries and formal grammars from the Chomsky-Schützenberger hierarchy: regular, context-free, context-sensitive, and unrestricted grammars. NooJ can formalize twelve levels of linguistic phenomena, from the typographical to the semantic level (Silberztein, M. 2016).

To formalize the Rromani adjectives vocabulary, one needs to construct the following linguistic resources:

- a dictionary containing the affixes, simple words, compound words, and discontinuous expressions that make up the Rromani vocabulary of adjectives
- a grammar containing the description of adjectives inflectional paradigms

One could describe Rromani's four dialectal variants in the dictionary and morphological grammar. The following sections present these levels of description.

4.2 Electronic dictionary

For example, the adjective *thulo* is represented by the following lexical entry in NooJ formalized (aka electronic) dictionary:

thulo,ADJ+EN="fat, thick, dense"+FLX=BUXLO +DRV=ĆAĆIPEN:SASTIPEN

In this extract, each lexical entry is composed of a lemma, its category "ADJ" (adjective), its English translation "+EN," and the name of its inflectional paradigm "+FLX".

The lexical entry *thulo* is associated with derivational paradigm ĆAĆIPEN and its derivative's inflectional paradigm SASTIPEN. ĆAĆIPEN describes the derivation of abstract nouns with the suffix "-pen," which applies to words of various categories, such as adjectives, nouns, and verbs.

The four main dialects are encoded using the following double codes:

- O-bi dialect: rro+rrbi
- O-mu dialect: rro+rrmu
- E-bi dialect: rre+rrbi
- E-mu dialect: rre+rrmu

In addition, we have added a third code to label specific language variants. For example, the northern speech used in Russia and Poland is defined by the extra label: "+rrn." This is the case of the entry *veśìtko* [of the woods, wild].

veśitko,ADJ+rro+rrbi+rrn+EN="of the woods, wild"+FLX=VEŚÌTKO+SYN= "veśutno"

The entry above shows that the adjective *veśìtko* belongs to the dialects O-bi (+rro+rrbi), and is used specifically in Russia and Poland (+rrn), its English translation is "of the woods, wild" (+EN="of the woods, wild"), it is inflected according to the paradigm named VEŚÌTKO (+FLX=VEŚÌTKO [9]), and it has the synonym "veśutno" (+SYN=veśutno) used in most Rromani dialects except the vernacular in Russia and Poland.

4.3 NooJ morphological grammar

In NooJ, inflectional paradigms are represented by regular or context-free grammars built over suffix/property factors: suffixes are added to the lexical entry to construct forms, which are associated with the corresponding properties (Silberztein, M. 2003-). For example, the following is the grammar rule that describes the inflectional paradigm RROM:

RROM = <E>/sg+dr | a/pl+dr | es/sg+ob | en/pl+ob;

This rule states that if one adds the empty string (<E>) to the lexical entry, one produces a singular (+sg) direct (+dr) form; if one adds an "a" to the lexical entry, one produces a plural (+pl) direct (+dr) form; if one adds "es" to the lexical entry, one produces a singular (+sg) oblique (+ob) form; if one adds "en" to the lexical entry, one produces a plural (+pl) oblique (+ob) form.

Suffixes may contain stack operators. For instance, operator <B> (for "Backspace") is used

[9] The lexical entry (*i.e., veśìtko*) is in lower case, whereas the paradigm name (*i.e., VEŚÌTKO*) is in upper case. It prevents confusion between two distinct values represented by identical writing.

to delete the current letter. In the following paradigm:

BUXLO = <E>/m+sg+dr | <B>i/f+sg+dr | <B>e/m+pl+dr | <B>e/f+pl+dr | <B>e/m+sg+ob | <B>e/f+sg+ob | <B>e/m+pl+ob | <B>e/f+pl+ob | <B>:CMP/comparative ;

The second term states that if one deletes the last letter of a lexical entry and then adds an "i" (suffix <B>i), one produces the feminine, singular, direct form (f+sg+dr) of the lexical entry.

For example, when the paradigm BUXLO is applied to the lexical entry *thulo* [fat, thick, dense], there will be no change (<E>) to the direct masculine singular, one final letter will be deleted (<B>) and an "i" will be added to the direct feminine singular, and one final letter will be deleted and an "e" will be added to produce the direct plural forms in both genders, and to the oblique forms in both genders and numbers. This paradigm then represents the three forms of *thulo*:

- *thulo*: masculine singular direct
- *thuli*: feminine singular direct
- *thule*: masculine plural direct, feminine plural direct, masculine singular oblique, feminine singular oblique, masculine plural oblique, or feminine plural oblique

It means that the wordform *thule* is associated with six potential linguistic analyses.
The last term of the BUXLO paradigm is used to produce the comparative forms of the lexical entries. :CMP refers to the name of the following rule (similarly to auxiliary symbols in generative context-free grammars):

CMP = eder/m+sg+dr | eder/f+sg+dr | edera/m+pl+dr | edera/f+pl+dr | edere/m+sg+ob | edere/f+sg+ob | edere/m+pl+ob | edere/f+pl+ob ;

The comparative suffix "-eder" is added in place of the final letter "o" in the *thulo* lexical entry (see "<B>:CMP" in the paradigm BUXLO above). The comparative is declined like the narrow adjectives: without suffix in the direct singular of both genders, "-a" in the direct plural of both genders, and "-e" in the oblique of both genders and numbers. This rule produces three forms: *thuleder*, *thuledera*, and *thuledere* for eight linguistic analyses.

Beside a dozen generic operators such as <B> that are available for any language, NooJ offers linguists the possibility of creating specific operators for each language. For instance, the Spanish operator <Á> is used to add an acute accent to the current vowel; the Hebrew operator <F> is used to de-finalize the last consonant of a word; the Tamazight operator <T> replaces letter "ḍ" with "ṭ", etc. For the Rromani language, we have implemented two specific operators:

- <A> deletes a grave accent, regardless the position and returns to the initial position,
- <À> adds a grave accent to the current letter.

For example, operator <A> is used in the following paradigms:

VEŚÌTKO = <E>/m+sg+dr | <E>/f+sg+dr | <B>a/m+pl+dr | <B>a/f+pl+dr | <A>ne/m+sg+ob | <A>ne/f+sg+ob | <A>ne/m+pl+ob | <A>ne/f+pl+ob ;

In the VEŚÌTKO paradigm, the fifth term states that if one removes the grave accent of the lexical entry, and then adds the suffix "ne" (<A>ne), one produces the masculine singular oblique form (+m+sg+ob) of the lexical entry. The operator <A> is typically used in paradigms associated with non-oxytonic words. For example, if the paradigm VEŚÌTKO is applied to the lexical entry *zèleno* [green], its oblique form will be *zelenone*, *i.e.*, without the grave accent.

By applying all inflectional NooJ paradigms to the dictionary, NooJ produces all the inflected forms for each lexical entry automatically. When applying these resources to a text, NooJ annotates all recognized word forms. For example, the wordform *veśitkone* will be annotated as an adjective (ADJ), its basic form is *veśìtko*, its inflectional value is the oblique (ob), its dialect value is a Northern vernacular belonging to the O-bi dialect (rro+rrbi+rrn), its synonym in other dialects is *veśutno*. That helps users pedagogically recognize the relationship between the basic form, an inflected form, and a dialectal variant. However, there are four sets of annotations because oblique forms are identical for both genders and numbers, see Figure 6. We need syntactic grammar to resolve ambiguities.

It is often better for pedagogical applications, to use NooJ graphical grammars to describe some phenomena. For instance, the paradigm BUXLO

shown above can equally be described with the following graph, see Figure 7.

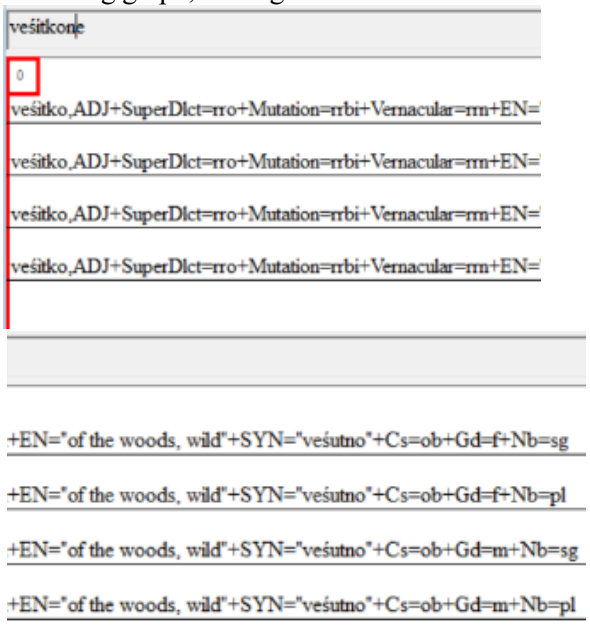


Figure 6: Inflected form *veśitkone* annotated automatically

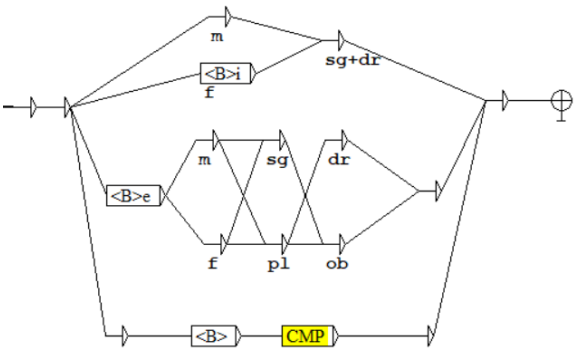


Figure 7: Inflectional grammar for adjective paradigm BUXLO.

The yellow node CMP refers to the embedded graph that represents the paradigm of the same name, see Figure 8.

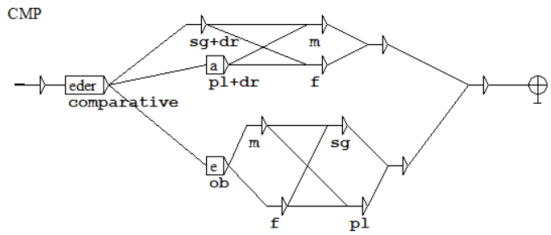


Figure 8: Inflectional grammar of the comparative suffix CMP.

The "DRV" property allows us to describe derivations. For example, the lexical entry *thulo* mentioned above:

thulo,ADJ+EN="fat, thick, dense"+FLX=BUXLO +DRV=ĆAĆIPEN:SASTIPEN

states that the word *thulo* can be derived according to paradigm ĆAĆIPEN, and the resulting derived form can be inflected according to inflectional paradigm SASTIPEN. Following is the definition of the derivational paradigm ĆAĆIPEN:

ĆAĆIPEN = <B><A>ipen/N+ina+m+abstract ;

From the lexical entry *thulo*, one deletes the final letter (<B>), removes the accent (<A>[10]), and then adds suffix "ipen." The resulting derivative *thulipen* [fatness, thickness, density], is a masculine inanimate noun (N+ina+m), that belongs to the semantic class "abstract."

The abstract noun *thulipen* is a generic masculine singular direct form, and has seven dialectal variants: *thulipe* used in the two dialects O-bi and E-bi belonging to a dialectal subgroup "without mutation (rrbi)," *thulipo* used in the O-mu dialect (rro+rrmu), *thulimos* and *thulimo* used in the E-mu dialect (rre+rrmu), *thuliben*, *thulibe*, and *thulibo* used in the Carpathian vernacular belonging to the O-bi dialect (rro+rrbi+rrc), all of which mean "fatness, thickness, density."

All these variants share the same inflectional morphology; *thulimàta* is the masculine plural direct form, *thulimas-* and *thulipnas-*[11] are the masculine singular oblique forms, and *thulimaten-* is the masculine plural oblique form. The following grammar rule describes the inflectional forms of the derivative as well as all its dialectal variants:

SASTIPEN = <E>/sg+dr | <B3>màta/pl+dr |
<B3>(mas|pnas)<E>/sg+ob |
<B3>maten<E>/pl+ob |
<B>/sg+dr+rrbi | <B2>o/sg+dr+rro+rrmu |
<B3>(mos|mo)<E>/sg+dr+rre+rrmu|
<B3>(ben|be|bo)<E>/sg+dr+rro+rrbi+rrc ;

NooJ automatically produces the derived forms for each lexical entry, and annotates all recognized derived forms in texts. For example,

[10] The <A> operator concerns only non-oxytonic words. Thus, it produces no change for the lexical entry *rromano*.

[11] This is an archaic form.

thulipe will be annotated as an abstract masculine inanimate noun (N+m+ina+abstract), its inflectional values are the singular direct (sg+dr), its dialect value is the "without mutation (rrbi)," and its initial category is the adjective meaning "fat, thick, dense." That helps users pedagogically recognize the relationship between the base word (*e.g.*, the adjective *thulo* [fat, thick, dense]) and its derivative (*e.g.*, the noun *thulipe* [fatness, thickness, density]), see Figure 9.

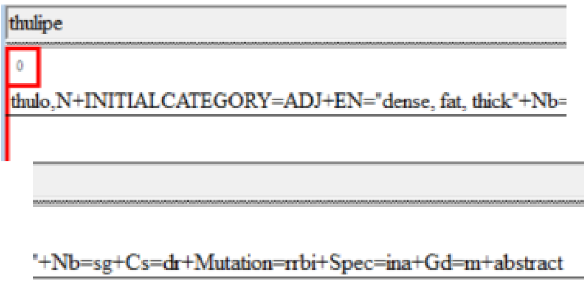


Figure 9: Derived form *thulipe* annotated automatically

4.4 Automatic Natural Language Processing

NooJ can use the same linguistic resources both to parse and generate texts. For example, one can apply a dictionary and its corresponding inflectional grammar to automatically produce all the forms associated with each lexical entry of the dictionary, see an extract in Figure 10.

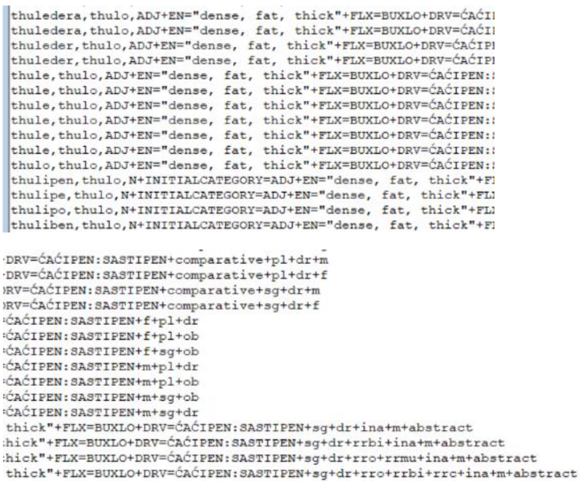


Figure 10: Inflected and derived forms of the adjective *thulo* [fat, thick, dense] generated automatically.

NooJ uses the same resources to parse texts, lemmatize and annotate their wordforms, to apply queries in the form of regular, context-free, context-sensitive, or unrestricted grammars, perform statistical analyses, compute semantic analyses in Predicative or XML format, translations (if accessing multilingual dictionaries), etc. For example, Figure 11 displays the query "<but>", which stands for "all inflected and derived forms of lexical entry *but* [many, numerous]" and the correspondence concordance computed by NooJ.

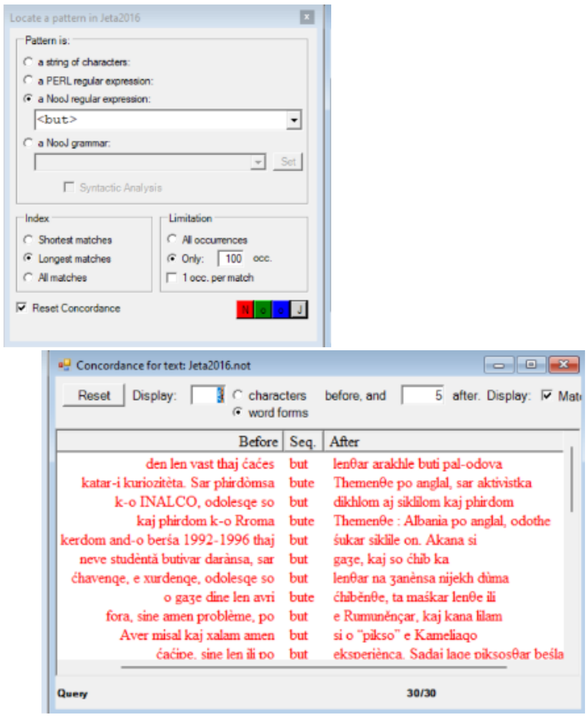


Figure 11: The query "<but>" and its resulting concordance.

The wordform *but* corresponds to three semantic values: the adjective "many, numerous," *e.g.*, *phirdòmsa bute Themenθe* [I traveled to many countries], the adverb "a lot, much," *e.g.*, *but dikhlom* [I have seen a lot], and another adverb "very," *e.g., but śukar siklile on* [They learned very well].

Syntactic grammar is underdeveloped in the current NooJ module for Rromani, this is why wordforms remain ambiguous in general.

However, the query "<but,ADV>" will not retrieve the adjectival inflected forms *bute* because adverbs are invariable, see Figure 12.

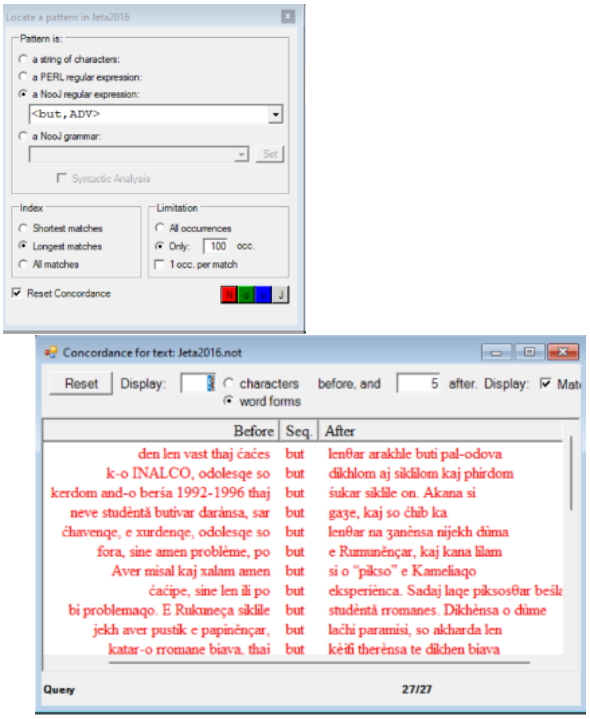


Figure 12: The query "<but,ADV>" and its resulting concordance.

If NooJ recognizes ambiguous forms, NooJ will show the annotation of all of them and not choose one randomly, as often happens in empirical applications. For example, as mentioned above, the adjectival inflected form *thule* is ambiguous because of six inflectional homonyms and the nominal inflected form *bakră* is ambiguous because of two inflectional homonyms, see Figure 13.

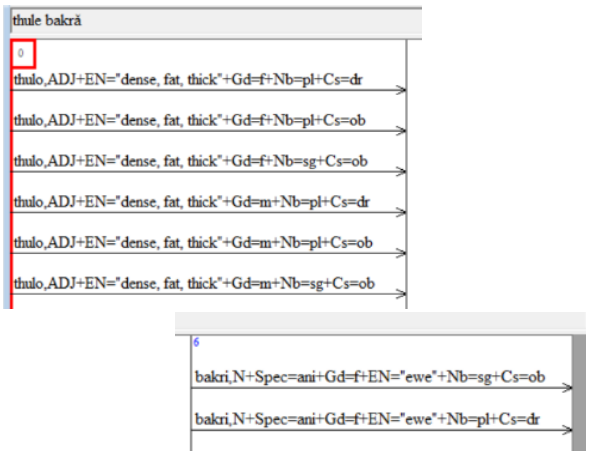


Figure 13: Inflected forms *thule* and *bakră* annotated automatically with ambiguity

4.5 Evaluation

There are 88 adjectives out of 747 lexical entries in the current NooJ dictionary for Rromani. Applying these lexical entries and their corresponding inflectional grammars generated 1,278 inflected and derivational forms.

In our corpus, all wordforms that might correspond to potential adjectives have been recognized and annotated correctly, *i.e.*, we have reached a 100% recall, which is expected as we are specifically constructing our linguistic resources from the corpus. However, without any syntactic grammar, wordforms that might function as adjectives or as adverbs (*e.g.,* the wordform *but*) and be associated with different properties (*e.g.*, the wordforms *thule* and *bakră*) remain ambiguous until we can apply a syntactic grammar.

5 Conclusion, perspective

The current Rromani module recognizes all 170 adjectival forms from a small corpus that contains 708 wordforms. We are currently importing around 4,500 lexical entries from an editorial dictionary (Courthiade, M. et al. 2009) into a formalized NooJ format.

Removing ambiguities is our current challenge. We are constructing syntactic local grammars to disambiguate frequent adjectives.

The resulting linguistic resources will be downloadable from the NooJ website. The NooJ dictionary for Rromani will use the standard Rromani alphabet and include dialectal variants at the lexical and morphological levels. It will be available as a new digital and linguistic tool for all speakers of Rromani: native speakers and learners, regardless of their dialects.

We believe this polylectal dictionary is valuable from a dialectological point of view. Furthermore, as the declaration of the first Congress of the International Rromani Union in 1971 stated that "no dialect is better than another," the dictionary will describe all dialects.

Unlike empirical methods, the NooJ platform produces analyses based on handcrafted linguistic resources, and thus offers linguists to describe and understand its properties. We believe that carefully and precisely handcrafting linguistic resources for Rromani is a worth scientific project, and will have many applications in Natural Language Processing, second-language teaching and corpus linguistics.

References

Gheorghe Sarău. 2009. *Struktùre rromane ćhibăqe*. Editura Universității din București, Bucharest.

Jeta Duka. Deś berś vaś-i rromani ćhib and-o INALCO. MS.

Marcel Courthiade. Structure dialectale de la langue rromani. Études tsiganes, 22-2005, pages 14-26. Le Centre de documentation, Paris.

Marcel Courthiade. 2016. The nominal flexion in Rromani. In Marcel Courthiade and Delia Grigore (eds.) *Professor Gheorghe Sarău: a life devoted to the Rromani language*. pages 157-211. Editura Universității din București, Bucharest.

Marcel Courthiade et al. 2009. *Morri angluni rromane ćhibăqi evroputni lavustik*. Romano Kher, Budapest.

Masako Watabe. 2024. A polylectal linguistic resource for Rromani. In Max Silberztein. (ed.) *Linguistic Resources for Natural Language Processing: On the Necessity of Using Linguistic Methods to Develop NLP Software*. pages 147-172. Springer, Cham.

Max Silberztein. 2003-. NooJ manual. https://nooj.univ-fcomte.fr

Max Silberztein. 2016. *Formalizing Natural Languages: the NooJ approach*. Wiley Ed.: Hoboken NJ.

Medo Gurbetovski, Mozes Heinschink, and Daniel Krasa. 2010. *Guide de conversation rromani de poche*. ASSIMIL, Paris.

Rajko Đurić. 2006. E rromani ćhib. In Marcel Courthiade (ed.) *La littérature des Rroms, Sintés et Kalés*. pages 67-68. INALCO, Paris.

2010. *Atlas of the World's Languages in Danger*. UNESCO, Paris.

Facebook. https://www.facebook.com/

Google Translate. https://translate.google.com/

NooJ platform. https://nooj.univ-fcomte.fr/

ROMLEX. http://romani.uni-graz.at/romlex/

Russian Romani Corpus. http://web-corpora.net/RomaniCorpus/search/

La langue romani - un atout pour l'éducation et la diversité (exhibition). 2014. Council of Europe, Strasbourg.

Bilingual Sentence Mining for Low-Resource Languages: a Case Study on Upper and Lower Sorbian

Shu Okabe[1,2] and **Alexander Fraser**[1,2,3]
[1]School of Computation, Information and Technology
Technische Universität München (TUM)
[2]Munich Center for Machine Learning
[3]Munich Data Science Institute
shu.okabe@tum.de, alexander.fraser@tum.de

Abstract

Parallel sentence mining is crucial for downstream tasks such as Machine Translation, especially for low-resource languages, where such resources are scarce. In this context, we apply a pipeline approach with contextual embeddings on two endangered Slavic languages spoken in Germany, Upper and Lower Sorbian, to evaluate mining quality. To this end, we compare off-the-shelf multilingual language models and word encoders pre-trained on Upper Sorbian to understand their impact on sentence mining. Moreover, to filter out irrelevant pairs, we experiment with a post-processing of mined sentences through an unsupervised word aligner based on word embeddings. We observe the usefulness of additional pre-training in Upper Sorbian, which leads to direct improvements when mining the same language but also its related language, Lower Sorbian.

1 Introduction

Machine Translation (MT) essentially relies on parallel corpora, which are widely available for 'winner' languages (Joshi et al., 2020). Yet, when it comes to lower-resourced languages, they become rarer, and such resources are more costly to obtain compared to monolingual corpora. This is why, to circumvent situations with too few or even no parallel sentences, parallel sentence mining is a task to find parallel sentences automatically in monolingual corpora. Research on parallel sentence mining is intertwined with MT since improving mining quality often leads to a better translation model.

The BUCC Shared Tasks (Zweigenbaum et al., 2017; Pierre Zweigenbaum and Rapp, 2018) notably focus on parallel sentence mining and acts as a benchmark. However, only four well-resourced language pairs are represented there. Hence, we try to fill this gap by evaluating sentence mining for low-resource languages.

In this work, we consider Upper Sorbian and Lower Sorbian, paired with German, which can be seen as a case study for low-resource sentence mining. We can effectively observe two data conditions (the former has more data than the latter) and also the impact of relatedness between the two languages.

We will try to answer the following questions: How well can we mine parallel sentences for a language with off-the-shelf word encoders? How useful is it to pre-train a model with the available monolingual data? How helpful is it to pre-train a model on a related language?

We consider two scenarios: (i) when computing resources are limited, we use already pre-trained models; (ii) otherwise, we fine-tune a language model on the available monolingual corpus in the low-resource language.

As such, we aim to foster further research on bilingual mining for low-resource languages and its challenges. We hope that this study provides important lessons useful even in a more data-restricted scenario.

To this end, we propose (a) two BUCC-style mining corpora, (b) a comparison of two state-of-the-art language models in mining Sorbian-German parallel sentences, (c) word encoders with different amounts of pre-training sentences in Upper Sorbian, and (d) an alignment post-processing to improve the mining quality. Thus, our work can serve as a benchmark for two low-resource languages in a realistic scenario. We release the corpora, the mining pipeline, and all related code material[1].

Section 2 will focus on the two languages and the creation of the corpora, while Section 3 compares the considered language models, the pre-training strategy and explains the mining method. Section 4 presents and analyses the mining results.

[1]At https://github.com/shuokabe/PaSeMiLL.

Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 11–19

2 Languages and datasets

2.1 On Upper and Lower Sorbian

Upper Sorbian (ISO code: hsb) and Lower Sorbian (dsb) are two West Slavic languages and constitute the Sorbian branch. Both are spoken in Germany (Saxony for the former and Brandenburg for the latter) and are currently classified as endangered according to Ethnologue (Eberhard et al., 2024). There are state-level laws that notably guarantee the use and teaching of both languages. For instance, the Witaj Sprachzentrum (Witaj Language Center) offers language courses in certain kindergartens and schools.

The NLP community has lately focused on the two Sorbian languages in cooperation with them and the Sorbian Institute. They both provided data for the successive WMT Shared Tasks in Unsupervised MT and Very Low Resource Supervised MT (Fraser, 2020; Libovický and Fraser, 2021; Weller-Di Marco and Fraser, 2022).

Hence, we focus on the Upper Sorbian-German and Lower Sorbian-German language pairs in this work. Previously, only Kvapilíková and Bojar (2023) focused on Upper Sorbian-German parallel sentence mining with a pre-training of XLM (Conneau and Lample, 2019), but the task has not been addressed on Lower Sorbian yet. It is the closest work, but their aim was to train a MT model, and their pre-trained encoder notably required 500K sentences in Upper Sorbian and German, which is already a large amount of available data and, hence, not a realistic scenario for most low-resource languages.

2.2 BUCC-style dataset creation

For our experiments, we apply the methodology of the BUCC 2017 Shared Task (Zweigenbaum et al., 2017) to Upper and Lower Sorbian by injecting known parallel sentences into their respective monolingual corpus.

This evaluation is an artificial approach, which can introduce some biases, such as parallel sentences that may stand out from the original monolingual sentences. However, this task is more difficult than the related sentence matching and gives a more realistic setting for bilingual mining.

We rely on the data provided for the above-mentioned WMT Shared Tasks and select its 2020 edition for Upper Sorbian and 2022 for Lower Sorbian for both monolingual and parallel sentences.

More precisely, for Upper Sorbian, we rely on the WMT 2020 Shared Task data and use the monolingual corpus provided by the Sorbian Institute (339,822 sentences). The monolingual German data comes from the Leipzig news corpora[2] (2020) (Goldhahn et al., 2012) and has 300K sentences. We chose to insert the development and development test data from the Shared Task (4,000 sentences) as parallel data.

For Lower Sorbian, we use the WMT 2022 Shared Task data for its monolingual corpus (66,408 sentences) and the parallel sentences from the development and development test datasets (1,353 sentences). The monolingual German data comes from the Leipzig news corpora of 2022 and contains 100K sentences.

Compared to the original BUCC methodology, presented in Zweigenbaum et al. (2017), we modified the following points. Instead of inserting a parallel sentence in a section of the monolingual corpus which deals with similar topics, we chose to shuffle all sentences. While we lose the context of each sentence, our mining pipeline does not take it into account. Besides, short sentences have been kept, while very long sentences of more than 40 words have been removed in the monolingual corpora, which explains the smaller datasets. Finally, we lower the possible amount of 'natural' parallel sentences (i.e., parallel sentences in the original monolingual corpora) by using the Leipzig news corpora, which is not directly related.

Table 1 presents the number of sentences in the Upper and Lower Sorbian datasets after inserting parallel sentences and shuffling. We also split the dataset into training and test subsets in a similar proportion of parallel sentences as in the German-English pair in the BUCC Shared Task.

	train	test
Upper Sorbian corpus	34,001	101,751
German corpus	32,915	98,747
of which parallel	1,000	3,000
Lower Sorbian corpus	22,303	44,616
German corpus	33,756	67,513
of which parallel	451	902

Table 1: Number of sentences in the Upper Sorbian (top) and Lower Sorbian (bottom) datasets.

[2] https://wortschatz.uni-leipzig.de/en/download/German.

2.3 Dataset difficulty

We verify whether the BUCC-style datasets created in Section 2.2 are suited to evaluate the mining task or not. If parallel sentences stand out from the other sentences which were originally in the unrelated monolingual corpus, the artificial dataset is deemed too easy.

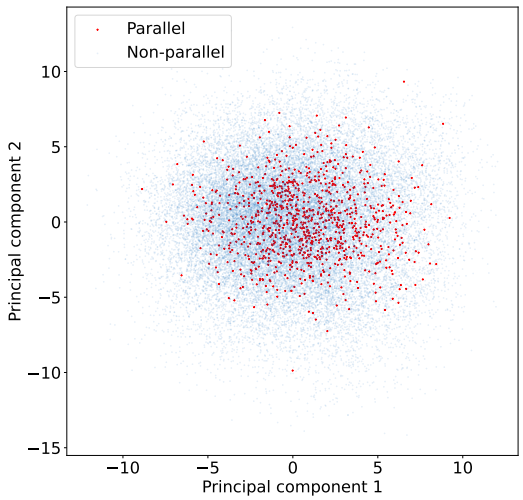


Figure 1: Distribution of embeddings of the sentences in the corpora according to the first two principal components for the created German dataset

We use the state-of-the-art sentence encoder LaBSE (Feng et al., 2022) to encode the well-resourced German dataset. We reduce the embedding dimension through a principal component analysis (PCA). Figure 1 displays each sentence embedding of the dataset according to the first two principal components. We can see that both parallel and non-parallel sentences are situated in similar regions with no clear cluster of sentences. Therefore, the task is not too easy.

3 Sentence mining methodology

3.1 Baseline models

We mainly study two multilingual pre-trained models to represent words. First, XLM-RoBERTa or XLM-R (Conneau et al., 2020) is a frequently used multilingual language model; we use its base version from the Transformers library. The other model is Glot500-m (Imani et al., 2023), which is an extension of XLM-R with additional pre-training for more than 500 low-resource languages.

It must be noted that XLM-R has seen German and two Slavic languages, namely Czech and Polish, during pre-training. Glot500-m has additionally been trained on 105K sentences in Upper Sorbian. From this perspective, both Sorbian languages are in a better situation than many other low-resource languages, which might not have as much available data or related well-resourced languages in the model pre-training.

3.2 Pre-training XLM-R in Upper Sorbian

Given that we have access to a monolingual corpus in Upper Sorbian, we also pre-train XLM-R on the available Upper Sorbian monolingual corpus. This model will also enable us to see how the additional pre-training in Upper Sorbian can indirectly help in the more closely related Lower Sorbian.

We replicate the pre-training strategy of Glot500-m (Imani et al., 2023). To gauge the amount of needed data to reach similar (or better) performance, we use different sizes and compositions of pre-training datasets.

In practice, we use the shuffled monolingual corpora presented in Section 2.2 for Upper Sorbian and German to pre-train XLM-R with a standard masked language modelling (MLM) objective. We obtain three models, named PT-HSB-3, PT-HSB-6, and PT-HSB-9, with different amounts of pre-training data: respectively, 30K, 60K, and 90K monolingual sentences in Upper Sorbian, coupled with and at least 30K sentences in German.

Providing bilingual cues Moreover, since Upper Sorbian is a Slavic language, we leverage additional data from the same language family. In our case, we choose to use parallel sentences in Czech and German, a better-resourced pair. Such a choice can be applied to other language pairs by considering neighbouring or related languages.

Hence, we carry out an additional pre-training on top of PT-HSB-9 with a MLM objective on a bilingual Europarl corpus in Czech and German from OPUS (Tiedemann, 2012), where we concatenate parallel sentences as one sentence for the model. We denote this model PT-HSB-9 + CS-DE. We restrict the training size to 220K sentences. The idea is twofold: give bilingual cues to the model, which is known to help the model, even when the language pair is different (Kvapilíková et al., 2020), and to indirectly improve the Upper Sorbian word representation thanks to Czech.

Experimental conditions To pre-train the models, we first relied on vocabulary extension, following the methodology of Imani et al. (2023). For

each pre-training setting, we extend the vocabulary to the used monolingual or bilingual corpora. Then, the pre-training itself uses the default parameters and approach given in the Transformers library.

All mining experiments have been carried out on 1 GPU (NVIDIA Tesla V100). The additional pre-training of XLM-R in Upper Sorbian or with the Czech-German corpus has been done on 1 to 4 of the same GPUs for 5 epochs. The longest pre-training is with the bilingual cues, due to a higher number of sentences and longer length; this took almost one week effectively. The other pre-trained models required a few days.

3.3 Mining and evaluation methods

The overall mining pipeline follows (Hangya and Fraser, 2019). First, we derive sentence representations by mean-pooling word embeddings with our encoders, which is a more effective approach than max-pooling (Kvapilíková et al., 2020). Then, we compute the similarity between a source (Sorbian) sentence and a target (German) sentence in the multilingual embedding space using the CSLS (Cross-Domain Similarity Local Scaling) score (Conneau et al., 2018)[3]. This metric is known to better deal with the hubness issue than the standard cosine similarity (Dinu et al., 2015). Formally, for two sentence vectors x and y, it is computed as in Equation (1):

$$\begin{aligned}\mathrm{CSLS}(x,y) &= 2\cos(x,y) \\ &- \sum_{z\in \mathrm{NN}_k(x)} \frac{\cos(x,z)}{k} - \sum_{z\in \mathrm{NN}_k(y)} \frac{\cos(y,z)}{k}, \end{aligned} \quad (1)$$

where $\mathrm{NN}_k(x)$ indicates the k-nearest neighbours of vector x. We choose $k = 20$.

Finally, we consider a source sentence and its most similar target sentence to be parallel according to a threshold that is chosen dynamically on the training dataset. Defined as in Equation (2) by Hangya et al. (2018), the threshold value depends on the mean and standard deviation (σ) from the found similarity values (S):

$$\text{threshold} = \text{mean}(S) + \lambda \times \sigma(S), \quad (2)$$

where λ is the tuneable hyper-parameter.

We evaluate the mining quality by computing the usual Precision, Recall, and F-score, following the BUCC Shared Task methodology. We also report the number of mined sentences (N_{sent}).

[3]This method is related to the margin-based methods presented by Artetxe and Schwenk (2019a); we observed comparable results on our dataset whether with CSLS or a ratio margin.

4 Experimental results

4.1 Mining results

embeddings	P (%)	R (%)	F (%)	N_{sent}
XLM-R	3.64	2.03	2.61	1,675
Glot500-m	32.82	20.63	25.34	1,886
PT-HSB-3	22.36	8.77	12.60	1,176
PT-HSB-6	34.54	17.23	22.99	1,497
PT-HSB-9	34.36	17.50	23.19	1,528
+ CS-DE	**36.96**	**26.30**	**30.73**	2,135

Table 2: Evaluation on the test dataset of the **Upper** Sorbian corpus.

Upper Sorbian Table 2 presents the quality of the mined parallel sentences with different word embeddings on the Upper Sorbian-German dataset. XLM-R's performance indicates that these embeddings are not suited for Upper Sorbian and cannot extend well to this language based on related pre-trained languages only. On the contrary, Glot500-m, which has seen a number of sentences in Upper Sorbian, has higher scores than XLM-R: the additional pre-training does indeed help to get a better word and, hence, sentence representation.

The bottom half of the table shows the performance of the different XLM-R models pre-trained on Upper Sorbian and German, as presented in Section 3.2. Not surprisingly, increasing amounts of Upper Sorbian data improve mining performance, reaching scores similar to Glot500-m, which was trained with roughly 100K Upper Sorbian sentences. Furthermore, using a bilingual cue from a related language pair (here Czech-German) enables us to go further, with an F-score of 31 for PT-HSB-9 + CS-DE. It is worth noting that this additional pre-training mainly helps with recall.

Lower Sorbian We use the same experimental methodology on the Lower Sorbian corpus and show the results in Table 3. XLM-R mines sentences of similar quality in both Sorbian languages: with no prior knowledge of the language, they equally struggle with F-scores of less than 3. Moreover, since Glot500-m has not seen any Lower Sorbian sentence, it also has a very low F-score compared to the Upper Sorbian case: pre-training in the language is indeed crucial, especially when mining

embeddings	P (%)	R (%)	F (%)	N_{sent}
XLM-R	5.88	1.88	2.85	289
Glot500-m	6.75	5.65	6.15	756
PT-HSB-3	5.99	5.21	5.57	785
PT-HSB-6	8.85	5.21	6.56	531
PT-HSB-9	10.06	6.87	8.17	616
+ CS-DE	**11.01**	**11.75**	**11.37**	963

Table 3: Evaluation on the test dataset of the **Lower** Sorbian corpus.

with averaged word embeddings. Here, the related languages help, with 3 points of F-score above the standard XLM-R, but only in a limited fashion.

Regarding pre-trained XLM-R models, they exhibit a comparable trend as for Upper Sorbian: more pre-training sentences improve the mining quality. The models see no Lower Sorbian during pre-training; the increase in performance is only due to the transfer between the related Slavic languages.

4.2 Precision-recall trade-off

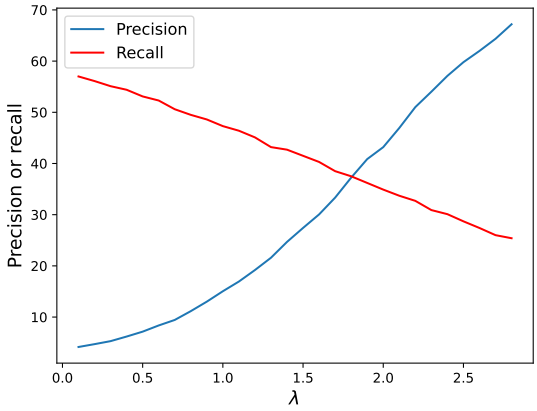


Figure 2: Evolution of the precision and recall for the best-performing PT-HSB-9 + CS-DE model on the *training* dataset in Upper Sorbian.

We defined the mining threshold by maximising the F-score on the training dataset in section 3.3. However, in real-life scenarios, the relevant criteria might differ depending on the use case. Another possible strategy would be to aim for a higher recall because, once mined, the precision can be increased through post-processing by filtering out wrong pairs.

Besides, as Figure 2 shows for the best model on Upper Sorbian, we notice that precision tends to rise faster than the decline in recall when increasing the threshold parameter λ. This means that voluntarily choosing a sub-optimal λ with a higher recall and post-processing could lead to higher F-scores.

Post-processing One approach is through manual annotation, which requires active involvement from the language community or speakers. This can be tedious, depending on the initial mining quality. A second method is to rely on unsupervised word aligners solely based on embeddings, such as SimAlign (Jalili Sabet et al., 2020). Given the lower amount of sentences, compared to the BUCC setting, for instance, this remains a reasonable option regarding the computing time and cost[4]. Moreover, since we focused on word encoders for Upper Sorbian in this work, embedding-based aligners can also benefit from the additional pre-training.

In this experiment, we select a very low threshold, $\lambda = 0.1$, to compromise between a scalable amount of sentences to align and a high enough recall. For all the kept pairs, we use SimAlign with our pre-trained embedding models[5]. Then, we compute a simple two-way alignment score by counting the found alignment links divided by the number of words in the sentence in both directions. Finally, using the dynamic threshold of Equation (2), we only consider sentence pairs above a threshold alignment score. This post-processing leads to a significant improvement when used on the best-performing PT-HSB-9 + CS-DE model in Upper Sorbian, for instance, with an F-score of 51.38 (to compare with 31, without alignment post-processing, in Table 2).

Since this approach relies on embeddings to correctly align words, it requires a decent modelling of the language. For instance, when we apply this post-processing method to Lower Sorbian (still with the PT-HSB-9 + CS-DE model), we only improve the F-score by 2 points, reaching 13.44.

4.3 Qualitative analysis

In Table 2, XLM-R obtained low metric scores on Upper Sorbian despite finding more than 1,000 sentences. This poor mining quality can be qualitatively seen in Figure 3, where the source Upper Sorbian and the found German sentences have nothing in common. Using the best model, in our case, PT-HSB-9 + CS-DE with alignment post-processing,

[4]In our experiments, the largest number of sentence pairs to align was 42,170, for the Upper-Sorbian test dataset, which took less than 10 minutes on one GPU.

[5]We use the 8th layer and align with the 'Argmax' strategy.

enables us to find the correct target German sentence.

HSB	Wón namjetuje moderěrowanu diskusiju wo tym.
XLM-R	Sie rechen das Laub der Laubbäume.
	They rake the leaves of the deciduous trees.
Best	Er schlägt eine moderierende Diskussion darüber an.
	He proposes a moderated discussion about this.

Figure 3: Example of mined sentences in Upper Sorbian. While XLM-R finds an unrelated sentence, PT-HSB-9 + CS-DE with alignment post-processing identifies the correct German sentence.

Figure 4 presents a pair of sentences wrongfully considered as parallel by the mining programme using PT-HSB-9 + CS-DE with alignment post-processing. One limitation of considering averaged word embeddings as sentence embedding is that nuances or details can get diluted in the final representation. A common issue is, hence, when two sentences have similar topics; even embedding-based word aligners will struggle in these cases. As such, the example sentences are incorrectly considered parallel because of a similar topic and structure. In the second half of the sentence, the dates and times do not correspond: Sunday, 15th of July ('njedźelu, 15. julija') at 17:00 ('17 hodź') in Upper Sorbian and Wednesday, 9th of December ('Mittwoch, den 9. Dezember') at 20:15 ('20.15 Uhr') in German. Nonetheless, this pair gets a high CSLC similarity score, and the computed align rate is 60%.

5 Related works

Parallel sentence mining has been extensively studied as an intermediate step geared towards Machine Translation, further stimulated by the BUCC Shared Tasks (Zweigenbaum et al., 2017; Pierre Zweigenbaum and Rapp, 2018). Previous works usually tackled parallel sentence mining with supervised bilingual and multilingual embeddings (e.g., Guo et al., 2018). When unsupervised, i.e. with no training parallel sentences, the embeddings stemmed from monolingual static embeddings such as fastText (Bojanowski et al., 2017) that were mapped to form bilingual or multilingual word embeddings (Hangya et al., 2018; Hangya and Fraser, 2018, 2019).

Then, static embeddings were replaced in the pipeline by multilingual contextual embeddings such as in (Kvapilíková et al., 2020; Kvapilíková and Bojar, 2023). The key point was to improve the bilingual (or multilingual) sentence representation, as proposed by Schwenk (2018).

Another reason to tackle sentence mining is to estimate the quality of embeddings; it is a simpler task computing-wise compared to the more resource-intensive machine translation. Similarly, an alternative method to assess the quality of multilingual word representations is sentence matching, where a parallel corpus is shuffled, and the true pairing must be found. It is more scalable to multiple languages due to the lower number of sentences to process.

An adjacent field of work worth mentioning here is on pre-trained multilingual embeddings, among which XLM-R (Conneau et al., 2020) and Glot500-m (Imani et al., 2023), that we consider here. The latter has notably been tested on the sentence-matching task to evaluate its word representation quality.

Finally, the task of parallel sentence mining itself is well-handled by multilingual sentence encoders, namely LASER (Artetxe and Schwenk, 2019b) and LaBSE (Feng et al., 2022), due to their massive training datasets and their specific objectives. More precisely, Costa-jussà et al. (2022) have actually striven to extend the initial LASER embeddings to more than 200 less-represented low-resource languages with LASER3 thanks to teacher-student distillation (Heffernan et al., 2022). By combining this approach with contrastive learning, Tan et al. (2023) get even further improvement on eight low-resource languages, with larger clean parallel data than Sorbian languages. These sentence embeddings are still unavailable for most low-resource languages, and extending them usually requires a significant amount of data or compute (if not both). The extension of our study to such embeddings goes beyond the scope of our current work but will be tackled in the future.

6 Conclusion

We studied the task of sentence mining, using averaged contextual word embeddings, by creating a benchmark for two Slavic low-resource languages: Upper and Lower Sorbian. We notably observed several advantages in carrying out additional pre-training on XLM-R for Upper Sorbian. The pre-trained model gets better word representations, which is reflected in a better mining capacity. Besides, the additional pre-training can improve the mining quality for related languages with even less data, in our case, for Lower Sorbian. Although

HSB	Kocorowy oratorij „Serbski kwas“ zaklinči po něhdźe dźesać lětach zaso, a to tutu njedźelu, 15. julija, w 17 hodź.
DE	Das große Finale von „Die Bachelorette“ läuft am Mittwoch, den 9. Dezember, um 20.15 Uhr bei RTL.

Figure 4: Example of a mining error in Upper Sorbian using PT-HSB-9 + CS-DE with alignment post-processing. Coloured parts respectively correspond to dates (in red) and times (in blue) in both languages but are not translations.

pre-training word encoders have a non-negligible computing cost, they open doors for other downstream tasks or parallel sentence post-processing with word aligners. Alternatively, if the language is already supported by Glot500-m, its word embeddings can be an off-the-shelf solution.

Our future work includes bringing the mining quality even higher by leveraging existing additional language resources (e.g., dictionaries). Besides, the natural downstream task would be machine translation, in a similar fashion to (Kvapilíková and Bojar, 2023) by using mined pseudo-parallel sentences during training.

More generally, we hope this work can foster further initiatives, namely real-life applications towards MT, for instance, together with language communities, in carrying out bilingual sentence mining for other low-resource languages. Our benchmark can also serve as a first place to evaluate upcoming tools before extending them to different languages. Indeed, it has yet to be confirmed whether our observations still hold true for other languages and language families.

Acknowledgments

We thank Viktor Hangya for his help and the anonymous reviewers for their comments. This work has received funding from the European Research Council (ERC) under grant agreement No. 101113091 - Data4ML, an ERC Proof of Concept Grant.

Limitations

We have focused on two low-resource languages, which might not be in the most challenging situation when it comes to pre-trained models: related Slavic languages such as Czech or Polish are commonly seen in the pre-training data, and both languages use the Latin alphabet. This is a favourable setting for an easier transfer between languages. The improvements we saw can hence be difficult to reproduce for languages with more different characteristics (grammar, morphology, language family, or script). Nonetheless, this work still represents an initial attempt towards parallel mining for low-resource languages, and we suggest that future researchers evaluate their tools initially on our benchmark.

Besides, since both Sorbian languages are close enough to two pre-training languages and German is also well-covered, some off-the-shelf *sentence* encoders, such as LASER or LaBSE, already have a high mining performance: with the latter model, the mining quality reaches a F-score of 73.17 on Upper Sorbian and 43.33 for Lower Sorbian. These results are tangential to our work, which focuses on improving *word* embeddings for Sorbian languages when mining sentences.

Finally, the task itself is only suitable for languages with a monolingual corpus large enough, which represents a subset of endangered languages; our work cannot handle left-behind or scraping-by languages (Joshi et al., 2020), where the essential challenge may indeed be to first create larger monolingual corpora in the first place (or to directly create parallel corpora so that sentence mining is not necessary).

References

Mikel Artetxe and Holger Schwenk. 2019a. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019b. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. *Cross-lingual language model pretraining*. Curran Associates Inc., Red Hook, NY, USA.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of the Workshop Track at ICLR*.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2024. *Ethnologue: Languages of the World*, twenty-seventh edition. SIL International, Dallas, Texas. Online version.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Alexander Fraser. 2020. Findings of the WMT 2020 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 765–771, Online. Association for Computational Linguistics.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Effective parallel corpus mining using bilingual sentence embeddings. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 165–176, Brussels, Belgium. Association for Computational Linguistics.

Viktor Hangya, Fabienne Braune, Yuliya Kalasouskaya, and Alexander Fraser. 2018. Unsupervised parallel sentence extraction from comparable corpora. In *Proceedings of the 15th International Conference on Spoken Language Translation*, pages 7–13, Brussels. International Conference on Spoken Language Translation.

Viktor Hangya and Alexander Fraser. 2018. An unsupervised system for parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 882–887, Belgium, Brussels. Association for Computational Linguistics.

Viktor Hangya and Alexander Fraser. 2019. Unsupervised parallel sentence extraction with parallel segment detection helps machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1224–1234, Florence, Italy. Association for Computational Linguistics.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Ivana Kvapilíková, Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Ondřej Bojar. 2020. Unsupervised multilingual sentence embeddings for parallel corpus

mining. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 255–262, Online. Association for Computational Linguistics.

Ivana Kvapilíková and Ondřej Bojar. 2023. Boosting unsupervised machine translation with pseudo-parallel data. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 135–147, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Jindřich Libovický and Alexander Fraser. 2021. Findings of the WMT 2021 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Sixth Conference on Machine Translation*, pages 726–732, Online. Association for Computational Linguistics.

Serge Sharoff Pierre Zweigenbaum and Reinhard Rapp. 2018. Overview of the third bucc shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234, Melbourne, Australia. Association for Computational Linguistics.

Weiting Tan, Kevin Heffernan, Holger Schwenk, and Philipp Koehn. 2023. Multilingual representation distillation with contrastive learning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1477–1490, Dubrovnik, Croatia. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Marion Weller-Di Marco and Alexander Fraser. 2022. Findings of the WMT 2022 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 801–805, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2017. Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67, Vancouver, Canada. Association for Computational Linguistics.

Citizen-linguists and Decolonial Lexicography: Co-creative Dictionary-building in Grassroots Digital Language Documentation

Anna Luisa Daigneault[#/^] **& Gregory D. S. Anderson**[^]
[#]Université de Montréal & [^]Living Tongues Institute for Endangered Languages
annaluisa@livingtongues.org, gdsa@livingtongues.org

ABSTRACT

Many endangered, under-represented, minority and Indigenous language communities around the world need access to multilingual online resources to survive in the digital age. The Living Dictionaries platform provides a collaborative online space for professional linguists and citizen-linguists alike to produce their own grassroots digital dictionaries that include multimedia such as audio recordings and images. These online lexica can play an important role in assisting present and future generations in combatting language loss and creating visibility for their languages and cultures on the Internet.

1 Introduction

While state-run language programs often serve as vectors of total assimilation to dominant languages and the abandonment of heritage ones (Skutnabb-Kangas, 2000; 2023), grassroots digital projects can serve as a counterbalance and bring visibility to lesser-known languages. Access to high-quality digital resources is essential for language communities in the modern age, as information is increasingly consumed and disseminated digitally, specifically through mobile platforms. Assisting communities in developing such accessible resources is a tangible contribution by linguists in response to the colonialist underpinnings of linguistics. Relying on institutional actors – state, academic, juridical – to act in the interests of linguistic minority communities and to enforce linguistic human rights (not just on paper) has proven to be largely ineffective to date, except in very few contexts where governments have successfully helped revitalize languages that are typically the sole or main minority Indigenous language of the nation (e.g., in Wales, Ireland, New Zealand).

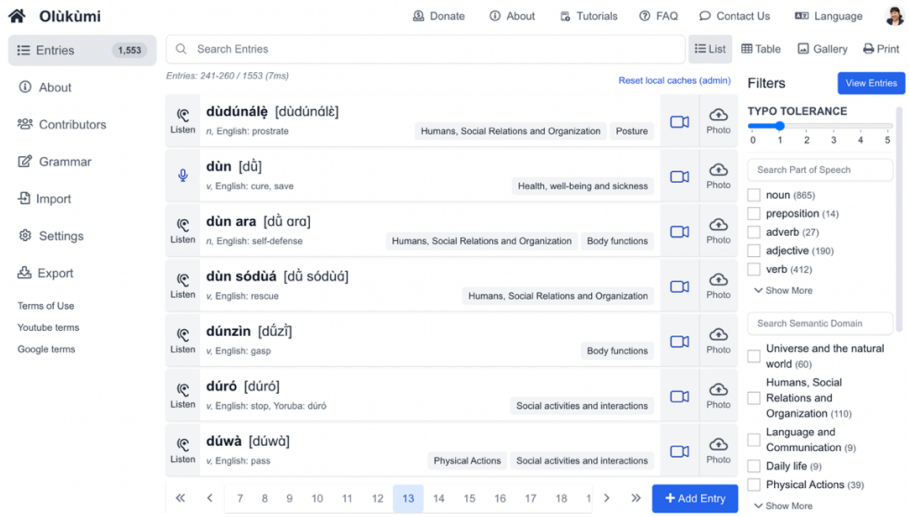


Figure 1: The Living Dictionary for Olùkùmi [ISO 639-3 code: ulb], an endangered Niger-Congo language of Nigeria, with glosses in English and some in Yoruba. It contains 1,553 entries tagged with semantic domains and parts of speech and includes multimedia. It was built by Dr. Bolanle Orokoyo (University of Ilorin) in close collaboration with scholars at Living Tongues Institute for Endangered Languages between 2012 and 2024. https://livingdictionaries.app/olukumi/entries

Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 20–28

As such, grassroots efforts that combine technology, collaboration and community stewardship provide a meaningful path to combat language loss. This paper shows the global applicability of the Living Dictionaries platform, which was engineered for straightforward co-creation of between grassroots collaborators, to create accessible digital resources for all underrepresented and endangered languages.

2 The Living Dictionaries platform

The Living Dictionaries platform is an online, source-available dictionary-builder that serves a wide range of underrepresented, endangered, Indigenous, creole and diaspora languages around the world, with the goal of providing communities with efficient online access to systematic language materials that benefit language learners as well as scholars. All languages, lects and regional varieties are welcome to be represented on the platform.[1] The Living Dictionaries website is an innovative tool for in-person as well as remote collaboration because it provides an accessible, interoperable[2] and user-friendly way for community language activists and linguists to work together to document, store and share large amounts of high-quality lexical data paired with digital images, audio, video and GPS coordinates.

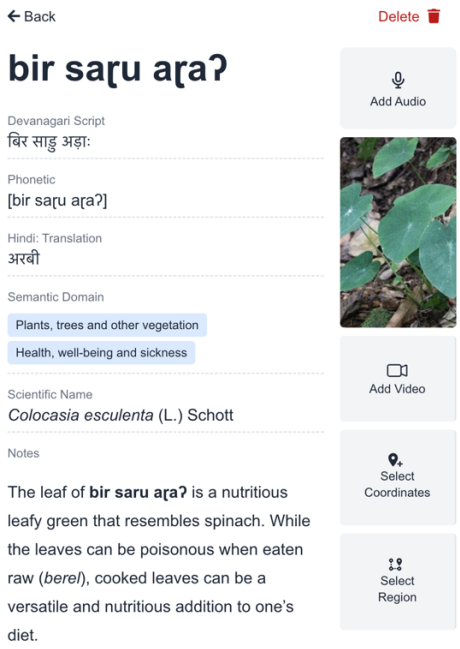


Figure 2: A detailed ethnobotanical entry from the Birhor Living Dictionary, with the headword represented in in IPA as well as the Devanagari script (locally dominant in India), the Hindi translation, two semantic domain tags, the scientific name in Latin, an image, and a "notes" section with relevant information about the plant's culinary use. Birhor [ISO 639-3 code: biy] is an endangered Munda language of India and this project was created by Living Tongues Institute for Endangered Languages in close collaboration with the Birhor tribal community of India. https://livingdictionaries.app/birhor/entries

The platform differs from other digital dictionary programs and online platforms in many distinct ways. For example, it functions in any web browser on any device, so there is no software to download, and any updates to the platform are visible instantly. While designed by linguists and usable by professional linguists, the platform's functionality is straightforward and easy to learn for citizen-linguists who may not have formal training in linguistics. To date, the Living Dictionaries platform houses

[1] When dictionaries are configured, language identifiers such as ISO 639-3 and Glottocodes can be added or updated "Settings page" of the dictionary. These identifiers also help index the dictionary online so that researchers can correlate the content in the dictionary with existing linguistic literature on the language.

[2] The Living Dictionaries team is working to expand interoperability between data types, formats and software. They can currently import dictionary data from spreadsheets, .CSV files and FLEx files (Standard Format), and they aim to improve their existing import functionality by accommodating all legacy dictionary import types (such as formats coming from Toolbox, Shoebox, Lexique Pro, TshwaneLex and other dictionary programs) so that they can import any existing legacy data into the Living Dictionaries. Dictionaries may also be exported as .CSV files and professionally designed .PDFs for printing entire dictionaries. An "Offline mode" is also being developed in 2025.

dictionaries for over 400 languages and has a user interface that is available in fourteen languages, making it accessible to professional linguists and citizen-linguists in many parts of the world. The platform allows citizen-linguists to determine how they want their language to be named in the dictionary's title, and they can edit it at any time. The platform includes multimedia audio and video recording and uploading functionality, which is rare on other platforms. It includes a built-in list of tags for parts of speech and semantic domains, as well as customizable tags, which allow dictionary editors to tag, filter and sort according to their own categories. There is no paywall or fee of any kind to build digital dictionaries on the platform.

Unlike Wiktionary (where each dictionary has different user experience and layout features), every Living Dictionary has the same front-end layout and the same set of systematic linguistic features available to its editors (and they can choose what data fields they wish to fill out). Furthermore, the platform allows for citizen-linguists to configure and modify the "Settings" of their dictionary (including naming, language codes and locations) concerning the language(s) they are working on and decide what glossing languages they wish to include. Another notable feature of the Living Dictionary platform is that it includes geo-mapping for dictionary entries, allowing place names and other entries to be correlated to the dictionary's map.

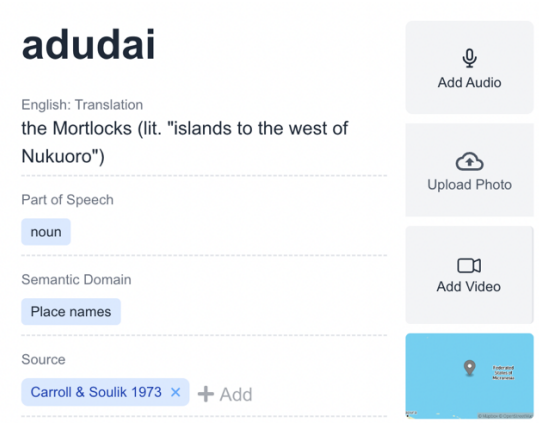


Figure 3: An entry for the place name "adudai" (a Nukuoro term that refers to the Mortlock Islands) with its map location in the Living Dictionary for Nukuoro [ISO 639-3 code: nkr], an endangered Austronesian language of the Federated States of Micronesia. It was led by linguist Emily Drummond (UC Berkeley) in close collaboration Nukuoro speakers, with assistance from Living Tongues Institute for Endangered Languages: https://livingdictionaries.app/nukuoro/entries

Living Dictionaries are unlimited in size, may contain as many glossing languages as one wants, and may represent entries in up to five local orthographies or scripts, which can be very useful in contexts where there are multiple competing orthographies and users may want to type in the search bar with their preferred script or writing system. Also, the platform allows users to generate and print a professionally designed .PDF of the dictionary directly within the browser. The number of columns, font size, data fields and optional inclusion of images and QR codes (linking directly to the Living Dictionary entry) can all be configured within the 'Print View' of each dictionary. With many 'sorting' filters on the right-hand side bar, it is easy to generate and print small and/or thematically targeted sets of lexical materials for pedagogical purposes.

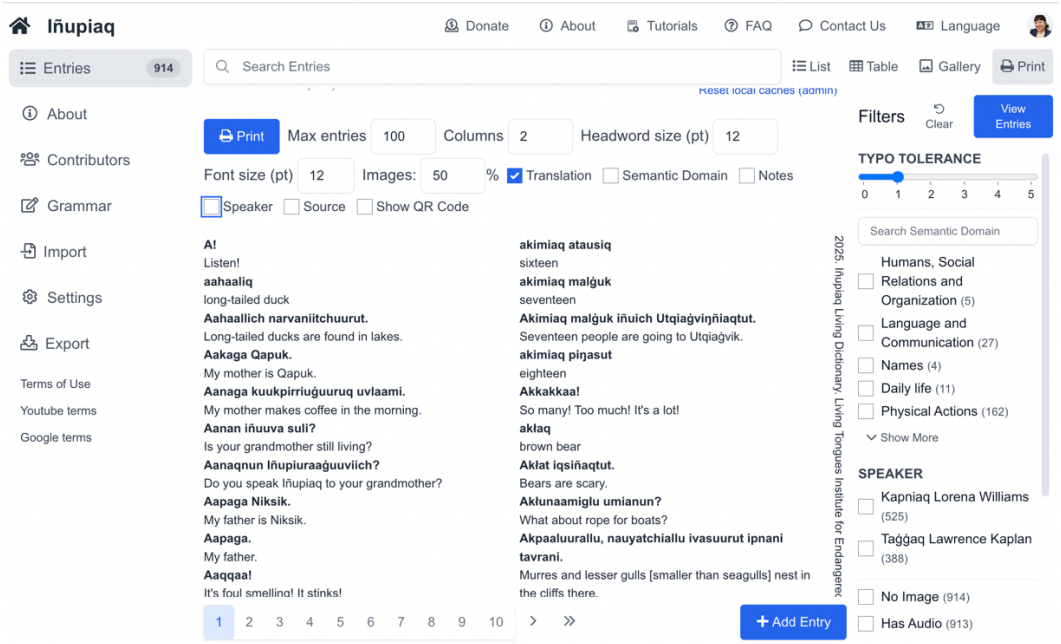


Figure 4: The 'Print View' of the Living Dictionary for Iñupiaq [ISO 639-3 code: ipk], an endangered Indigenous language of Alaska. It was created through a close collaboration between three nonprofit organizations: Doyon Foundation (Alaska), 7000 Languages and Living Tongues Institute for Endangered Languages. It is available here: https://livingdictionaries.app/inupiaq/entries

3 Co-creation on the Living Dictionaries platform

Collaboration is made easy since dictionary creators can send an automated email invitation to colleagues and assistants who wish to work on the dictionary. Once invited, collaborators join a project and they can add or edit content, record audio, and much more. These tools and features can also help facilitate collaboration and language exchange across generations, as younger, tech-savvy community members partner with experts from middle and elder generations in natural social contexts to record their voices on devices. The 'History' tab used for internal workflow makes it easy for various collaborators to edit a Living Dictionary at the same time and visualize what edits have been made and by whom. Through the availability of video tutorials[3] on the platform as well as scheduled Zoom trainings organized by the platform's leadership team, digital skills and best practices in citizen lexicography are made accessible to all users and indeed stand at the core of this initiative.

The platform's developers, linguists and curators work in close collaboration with citizen-linguists to create new Living Dictionaries; the co-creative workflow in the development of Living Dictionaries consists of five overarching phases: 1) planning and community consultation; 2) digital training, where citizen-linguists receive online or in-person training; 3) data collection and feature programming (if new features are being conceived in collaboration with a community); 4) data assessment and quality control, and lastly, 5) the import of large batches of linguistic and multimedia data to the Living Dictionaries platform and the deployment of new web features. Of course, these phases can be repeated as necessary over the course of a long project. All co-creators have editing access to the linguistic data in spreadsheets that are to be uploaded to the Living Dictionaries and editing access to the online dictionaries themselves. Imported dictionary entries may then be edited on Living Dictionaries manually (one by one) by any project participant directly to the platform or

[3] Free video tutorials are available on the platform in English and Spanish (with subtitles in Hindi, Russian, Chinese, Arabic and French) so that prospective dictionary creators can get started quickly without necessarily having to register for an upcoming Zoom training. The tutorials are available at: https://livingdictionaries.app/tutorials

may be uploaded in large batches from spreadsheets in collaboration with the lead digital curator. Once imported, any entry may be edited online at any time, and new entries can also be added. Entries also may include sample sentences with their translations into various languages, semantic domain tags and custom tags that allow for data organization and filtering, one or several audio recordings of the headword's pronunciation by local speakers, and one or multiple images related to the content of the dictionary entry. Entries may include a video that shows a speaker uttering the headword or using it in an example sentence, and some videos may include an explanation of the origin or etymology of the word, a demonstration of an event, etc. Each dictionary entry has its own unique URL, and is thus easily shareable via text messaging, email and social platforms.

Dictionary entries contain a headword represented in the language community's preferred orthography. The web development team works with community researchers to make sure the orthographies are displayed using the best current Unicode-compliant practices for web browsers. If community researchers eventually want to display alternate writing systems, that is not a problem. The system can accommodate up to five alternate orthographies within dictionary entries. The platform sends out regular community messages to all users on the platform several times a year, notifying them about upcoming Zoom trainings, feature updates and scheduled down time.

Living Dictionaries allow citizen-linguists and scholars to create, curate and expand digital lexicons that benefit present and future generations of speakers – locally and across their diasporas. The dictionaries can become large-scale, community-collaborative digital resources, incorporating extensively tagged multimedia materials that allow users to search, filter, sort, export and print data, thus providing language communities with free and shareable resources.

Speakers, regardless of their location, can use the multilingual interface[4] to record their voices directly to the cloud and store their audio recordings within dictionary entries for easy playback. Audio and video may be recorded directly into the web platform, using any device. The dictionary platform allows dictionary entries to be correlated with maps (helpful for place names and topographic features in the landscape). Users who are working on a language for which there already exists a Living Dictionary can click on the convenient "Contact Us" button located in the top menu toolbar (see the upper portion of Figure 4) to send an automated email message to the dictionary authors and ask for permission to join an existing project as a fellow editor.[5] Meetings with prospective editors may be convened to have a further discussion about collaboration.

4 Citizen-linguists and the broader impacts of decolonial lexicography

Linguistics is rightly critiqued as rooted in colonialist projects (Errington, 2001; Makoni, 2013; Zimmermann and Kellermeier-Rehbein (eds.), 2015; Hudley et al., (eds.) 2024). It is a moral imperative of the 21st century to address colonialist legacies and thus for linguists to make possible decentralized and decolonial approaches to language documentation, language resource development and linguistic analysis. Indeed, it is not enough to simply acknowledge and critique the colonialist underpinnings of the field, which, while an important step, if not tied to action, remains empty, self-defeating rhetoric. To be sure, documenting languages is not only important to the scientific field of linguistics, but also to speech communities who are urgently looking for tools to combat language loss. It is not an overstatement to assert that language documentation is crucial to conserving humanity's intangible heritage on Earth. Living Dictionaries offer a central point where different communities of stakeholders can contribute equally and respectfully. The platform was created to make

[4] At the time of writing, the platform's interface can be accessed in fourteen different languages, including Spanish, French, Hindi, Russian, KiSwahili, Bangla, Assamese, Portuguese, Malay, Bahasa Indonesia, Vietnamese, Hebrew, Chinese and English. More interface languages (Arabic, Thai, Italian) are under development.

[5] Each request is also forwarded to the platform administrators, who can assist the original authors in evaluating the inquiry and deciding if a new colleague should join the project or start a new Living Dictionary of their own instead.

building linguistic resources from the ground up easy and accessible to anyone who knows how to operate a smartphone or tablet. Equitable resource development is at the heart of this work. As platform administrators, we have sought to lessen the burden of colonial and capitalistic frameworks (such as bureaucracy and subscription models) so there are as few limitations as possible for newcomers to lexicography.

In particular, citizen-linguists play a huge role in the process of building these digital dictionaries. A *citizen-linguist*[6] is here understood as a person who is actively engaged in their speech community, believes in safeguarding their native (or heritage) language and works towards transmitting it to future generations. Citizen-linguists are motivated to create accessible language resources that are tailored to their community's needs. They are people who fulfill the multi-faceted roles of documentarian, language activist and digital content creator, whether they have formal training in linguistics or not. Some are educators, some are students; some are people with advanced training in other academic fields who see the value in protecting their language, whether they are fluent speakers or not.

In general, citizen-linguists take on the difficult challenge of recording and sharing their language, which may be under-studied by scholars and under-valued by their own community and the wider public. Many citizen-linguists are leaders in other areas of community life and undertake language work as volunteers in their spare time, because they understand that the act of preserving their language is connected to their cultural group's well-being, identity, and survival. Like other citizen scientists who study and celebrate local phenomena (e.g., flora or fauna, etc.), citizen-linguists are grassroots actors who may bridge divides between diverse groups of people and help bring local knowledge forward, while also feeling a sense of personal pride and connection to the language in question. Citizen-linguists are often excellent (co-) creators of language materials because they see the long-term value of the work they are doing and know that teamwork can benefit large, detailed undertakings like language documentation.

Living Dictionaries can serve as community-based access points to under-studied endangered traditional knowledge encoded in languages, in diverse domains such as flora and fauna, textile practices, spiritual traditions, food production, sacred sites and other elements of the landscape, and much more. For example, the Werikyana-Tiriyó-Portuguese-English Living Dictionary (see Figure 5) has over 3,000 entries with accompanying audio recordings and contains an impressive array of content regarding local culture, fauna and flora. This quadrilingual dictionary was created by the various Indigenous communities who speak the Werikyana language in the Brazilian Amazon, using solar-powered mobile devices and satellite Internet via Starlink. It is the first-ever digital resource of this kind for the Werikyana language and represents years of hard work led by many Werikyana speakers. This Living Dictionary was made publicly accessible to the world with the authorization of AIKATUK (Associação Indígena dos Kaxuyana-Tunayana-Kahyana).

[6] This term, which can also be used interchangeably with other broad designations like *language champion*, *language activist* and *community language activist*, expresses the notion that a member of a local speech community is contributing to science with their own resources and time, whether they have formal training or not.

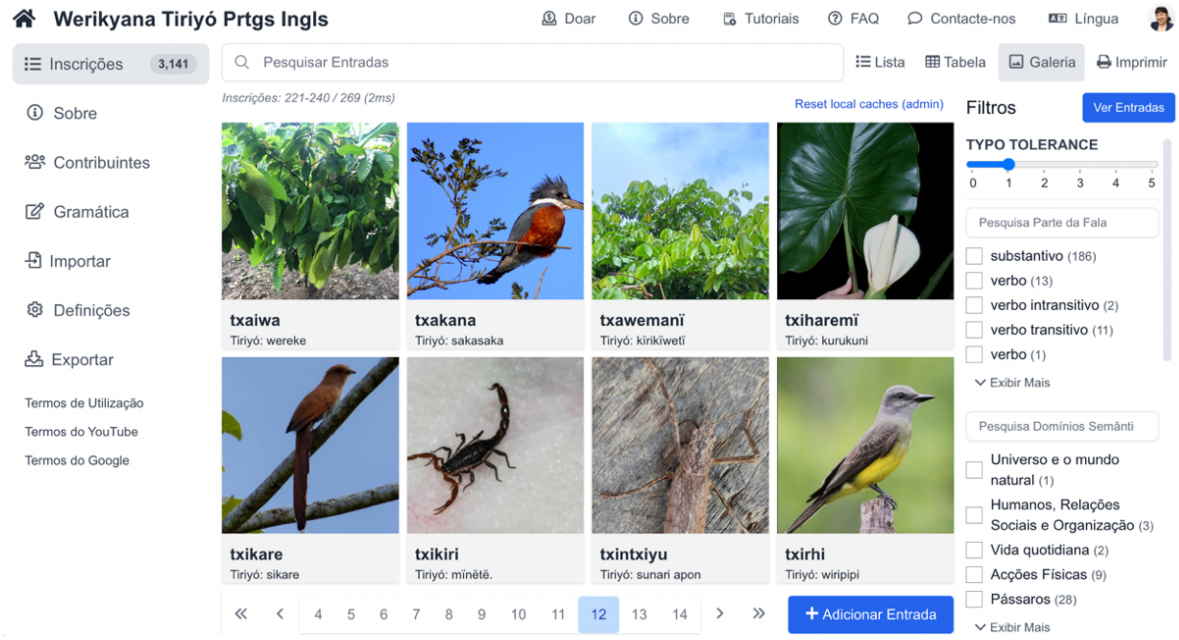


Figure 5: 'Gallery view' of the Werikyana-Tiriyó-Portuguese-English Living Dictionary, which was created by Werikyana speakers and Prof. Spike Gildea (University of Oregon) with technical assistance from Living Tongues Institute for Endangered Languages. Werikyana [ISO 639-3 code: kbb] is an endangered Indigenous language of the Brazilian Amazon of the Carib family. It is displayed in the Portuguese language interface in this screenshot (https://livingdictionaries.app/werikyana/entries)

The website offers the insights and systematicity of linguistic science in a user-friendly context, free of cost to citizen-linguists, with no institutional or other administrative roadblocks preventing access to this online tool.[7] It highlights and preserves essential ecological, social, and linguistic knowledge that lie at the foundation of cultural survival. All intellectual property rights associated with dictionary data remain with the language community from which the data originates.

This platform can help mitigate the global language extinction crisis by opening the door to linguistic documentation for all, thereby side-stepping colonial structures that seek to oppress minority languages. The platform supports citizen science by providing STEM opportunities to community activists to document their own languages and help them gain access to technical guidance from our team of professional linguists, anthropologists and web developers. The team also prioritizes racial equity by promoting access to (and awareness of) this platform to diverse communities of color worldwide, particularly in the Global South, with a focus on supporting academic colleagues and citizen-linguists long-term through digital literacy training online.

This platform provides an easy-to-use framework for systematically storing and sharing dictionary data for at-risk languages, thus increasing their viability for survival in the long-term. This comes with big implications: studies in North America and Australia show that weaving a connection to one's heritage language leads to better mental health, better performance in schools, and expanded economic opportunity (Zuckerman, 2020; Olko et al., 2022; van Beek, 2016; Olko and Andrason 2023). Therefore, pride in ethnolinguistic identity has numerous socio-economic and psycho-social-political benefits. Living Dictionaries, being multilingual tools at their core, also help promote bilingualism and multilingualism, which, in addition to social benefits, have positive biological outcomes such as improved cognition and protection against the onset of dementia (Bialystock et al., 2007; Perani and Abutalebi, 2015).

[7] Except for in nation-states that block access to certain websites for political/ideological reasons, like in China.

5 Code Accessibility on the Living Dictionaries platform

Accessible code is also important for future generations of developers building such tools. The Living Dictionaries code base is available on its GitHub page, and its license operates under a source-available, non-commercial license also listed in GitHub. Data in the Living Dictionaries are stored in a PostgreSQL database, backed up daily. Media is stored in a Google Cloud Storage bucket that is also backed up regularly. Dictionary managers can download their dictionaries as a .JSON file for their use in other applications. The .JSON is structured to make it easy for consuming applications to connect all relevant data points such as mapping speakers to dictionary entries. Dictionaries can also be exported as a .PDF file for easy printing. The site relies on popular, open-source technologies that make it easy to maintain and upgrade in the long-term. Two web developers who are specialized in mobile-friendly web applications help the platform's lead digital curator answer technical questions, improve the display and functioning of the platform on mobile devices and desktop computers, fix bugs on the platform, upload batches of content, manage the database, make necessary changes to the front-end and back-end of the website, plan and code new features and keep track of all technical issues on GitHub. They ensure that the code stays source-available at all times and stays up to date with the latest web technologies.

All edits to dictionary entries can be visualized online in real-time, without having to refresh the page, which facilitates instant remote collaboration between dictionary editors who are editing a dictionary at the same time (whether they are side by side or working at great distances). The design and engineering of the dictionary platform are created on an ongoing basis by an in-house web development team, guided by feedback from hundreds of scholars and citizen-linguists who attend our online workshops which provide training for community researchers involved in the editing, curation, recording process and construction of the dictionaries. The web developers working on Living Dictionaries have designed and implemented new features based on carefully assessing feedback from the user base and incorporating the needs of communities with limited digital literacy. Through engaging images, audio and video recordings, and the ability to add unlimited cultural information in the 'notes' section of each digital dictionary entry and grammatical information about the language in the 'Grammar' tab, the platform is able to showcase the unique features of each language and culture represented in the Living Dictionaries.

6 Summary

Living Dictionaries offer an inclusive, participatory citizen science approach to digital lexicography, thereby helping to decolonize and decentralize the process of language documentation. Grassroots digital language documentation is one of the only realistic approaches to combatting language loss in the long-term, and tools such as this one can benefit different communities of stakeholders. Living Dictionaries are playing an important role in helping under-represented, minority and Indigenous language communities worldwide to successfully claim space in the digital arena and thus safeguard their languages from extinction.

References

Bialystok, E., Craik, F. I. M., and Freedman, M. 2007. Bilingualism as a protection against the onset of symptoms of dementia. *Neuropsychologia*, 45(2):459–464. https://n.neurology.org/content/81/22/1938.

Errington, J. 2001. Colonial linguistics. *Annual Review of Anthropology*, 30(1):19–39.

Hudley, A. H. C., Mallinson, C., and Bucholtz, M., eds. 2024. *Decolonizing linguistics*. Oxford University Press.

Makoni, S. B. 2013. An integrationist perspective on colonial linguistics. *Language Sciences*, 35:87–96.

Olko, J., Lubiewska, K., Maryniak, J., Haimovich, G., de la Cruz, E., Cuahutle Bautista, B., Dexter-Sobkowiak, E., and Iglesias Tepec, H. 2022. The positive relationship between Indigenous language use and community-based well-being in four Nahua ethnic groups in Mexico. *Cultural Diversity and Ethnic Minority Psychology*, 28(1):132–143. https://doi.org/10.1037/cdp0000479.

Olko, J., and Andrason, A., eds. 2023. Introduction: Heritage languages and the well-being of speakers. In *Heritage languages and the well-being of speakers*, pages 5–16. Linguapax. https://www.linguapax.org/uploads/2024/01/Linguapax-2023-baixa.pdf.

Perani, D., and Abutalebi, J. 2015. Bilingualism, dementia, cognitive and neural reserve. *Current Opinion in Neurology*, 28(6):618–625. https://journals.lww.com/co-neurology/Abstract/2015/12000/Bilingualism,_dementia,_cognitive_and_neural.12.aspx.

Skutnabb-Kangas, T. 2000. *Linguistic genocide in education or worldwide diversity and human rights*. Mahwah, NJ/London: Lawrence Erlbaum Associates.

Skutnabb-Kangas, T. 2023. Preventing the implementation of linguistic human rights in education. In T. Skutnabb-Kangas and R. Phillipson, eds., *The handbook of linguistic human rights*, pages 109–126. Hoboken, NJ: Wiley-Blackwell.

van Beek, S. 2016. Intersections: Indigenous language, health and wellness. *First Peoples' Cultural Council*.

Zimmermann, K., and Kellermeier-Rehbein, B., eds. 2015. *Colonialism and missionary linguistics*. Colonial and Postcolonial Linguistics Vol. 5. Walter de Gruyter GmbH & Co KG.

Zuckerman, G. 2020. Our ancestors are happy. Language revival and mental health. In G. Zuckerman, ed., *Revivalistics: From the genesis of Israeli to language reclamation in Australia and beyond*, pages 9–36. Oxford University Press. https://doi.org/10.1093/oso/9780199812776.003.0009.

Supporting SENĆOŦEN Language Documentation Efforts with Automatic Speech Recognition

Mengzhe Geng[1*] **Patrick Littell**[1] **Aidan Pine**[1] **PENÁĆ**[2] **Marc Tessier**[1] **Roland Kuhn**[1]
[1]National Research Council Canada, Ottawa, ON, Canada
first.last@nrc-cnrc.gc.ca
[2]W̱SÁNEĆ School Board, Brentwood Bay, BC, Canada

Abstract

The SENĆOŦEN language, spoken on the Saanich peninsula of southern Vancouver Island, is in the midst of vigorous language revitalization efforts to turn the tide of language loss as a result of colonial language policies. To support these on-the-ground efforts, the community is turning to digital technology. Automatic Speech Recognition (ASR) technology holds great promise for accelerating language documentation and the creation of educational resources. However, developing ASR systems for SENĆOŦEN is challenging due to limited data and significant vocabulary variation from its polysynthetic structure and stress-driven metathesis. To address these challenges, we propose an ASR-driven documentation pipeline that leverages augmented speech data from a text-to-speech (TTS) system and cross-lingual transfer learning with Speech Foundation Models (SFMs). An n-gram language model is also incorporated via shallow fusion or n-best restoring to maximize the use of available data. Experiments on the SENĆOŦEN dataset show a word error rate (WER) of 19.34% and a character error rate (CER) of 5.09% on the test set with a 57.02% out-of-vocabulary (OOV) rate. After filtering minor cedilla-related errors, WER improves to 14.32% (26.48% on unseen words) and CER to 3.45%, demonstrating the potential of our ASR-driven pipeline to support SENĆOŦEN language documentation.

1 Introduction

Language documentation often plays an important role in the revitalization of Indigenous languages. Language revitalization is, in turn, crucially important for preserving the cultural heritage and identity of Indigenous communities. SENĆOŦEN (str), the language of the W̱SÁNEĆ people, has faced considerable challenges, largely due to the cumulative effects of historical marginalization and cultural suppression (Haque and Patrick, 2015; Pine and Turin, 2017). With a sharp reduction in fluent speakers, many Indigenous languages in Canada, including SENĆOŦEN, are at a critical juncture. Of the approximately 70 Indigenous languages in Canada, many urgently require revitalization efforts to prevent further loss (Littell et al., 2018). In this context, Automatic Speech Recognition (ASR) technology offers significant potential for language revitalization by supporting the transcription of spoken language, thereby potentially accelerating the development of educational curriculum developed from audio data (Jimerson and Prud'hommeaux, 2018; Foley et al., 2018; Littell et al., 2018; Gupta and Boulianne, 2020a,b; Liu et al., 2022; Rodríguez and Cox, 2023). While ASR technologies have made significant strides for widely spoken languages (Peddinti et al., 2015; Chan et al., 2016; Wang et al., 2020; Gulati et al., 2020; Hu et al., 2022; Li et al., 2023), research on ASR systems for Canadian Indigenous languages (Gupta and Boulianne, 2020a,b) remains limited.

SENĆOŦEN, also known as the Saanich language, is spoken around the Saanich peninsula in the southern region of Vancouver Island and on neighboring islands in the Strait of Georgia. The language is written with a distinct alphabet developed by the late Dave Elliott Sr. (FirstVoices, 2024). As of 2022, there are a reported 16 fluent SENĆOŦEN speakers and 165 semi-speakers (Gessner et al., 2022). While ongoing and vigorous revitalization efforts (Brand et al., 2002; Jim, 2016; Bird and Kell, 2017; Bird, 2020; Elliott Sr, 2024; Pine et al., 2025) are in place, there have been no prior efforts to leverage ASR techniques to support the documentation and revitalization of SENĆOŦEN.

This paper aims to address the gap by investigating cutting-edge ASR-based techniques that can support SENĆOŦEN language documentation

*Corresponding author.

Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 29–39

efforts. However, developing ASR systems for SENĆOŦEN presents two major challenges: **1) Limited data resources**: Compared with high-resource languages like English, there are very few digitized materials in SENĆOŦEN (Pine et al., 2022b), and even fewer audio recordings are available with aligned transcriptions; **2) Extensive vocabulary variation**: Beyond the relatively polysynthetic nature of SENĆOŦEN, metathesis driven by stress patterns further contributes to the vast number of possible word forms as illustrated below from Montler (1986, Section 2.3.5.4.3):

(1) *T̲QET*
ƛ'kʷ'ə́t
'Put it out (a fire).'

(2) *T̲EQT SEN*
ƛ'ə́kʷ't sən
'I'm putting it out.'

Such morphological and phonological complexity makes it impractical to construct a sufficiently large dictionary. As a result, many words to be transcribed are absent from the system's training data (i.e., out of vocabulary). These two challenges, taken together, significantly hinder the development of robust ASR systems for SENĆOŦEN.

To tackle the challenges associated with the development of ASR systems for SENĆOŦEN, this paper explores a range of state-of-the-art techniques, with an emphasis on end-to-end (E2E) models. E2E approaches offer a distinct advantage over traditional GMM-HMM or hybrid DNN-based systems, as they eliminate the need for a fixed lexicon. Given SENĆOŦEN's highly complex morphology, as well as the difficulty of building an exhaustive lexicon, E2E models are particularly well-suited to the task. However, a major drawback of E2E systems is their reliance on large datasets, which poses a significant obstacle for low-resource languages like SENĆOŦEN. To address this, we propose two strategies: **1) ASR data augmentation** through a carefully designed text-to-speech (TTS) synthesis pipeline, and **2) cross-lingual transfer learning** leveraging speech foundation models (SFMs). Additionally, we incorporate an external n-gram language model (LM) using either shallow fusion (Kannan et al., 2018) or n-best rescoring (Chow and Schwartz, 1989) to make the most of the available data.

Experiments were conducted using the SENĆOŦEN speech dataset, comprising 4 hours of recorded audio from the "Speech Generation for Indigenous Language Education project (Pine et al., 2025). The results show that systems employing cross-lingual transfer learning with speech foundation models significantly outperformed conventional hybrid time-delay neural networks (TDNNs), particularly when recognizing unseen words not present in the training set. Moreover, incorporating TTS-synthesized data in ASR training and an external n-gram LM further enhanced system performance.

This paper's key contributions are below:

1. **First comprehensive investigation of SFMs for documenting low-resource languages**: This study represents the first systematic investigation of speech foundation models for the development of ASR systems aimed at supporting the documentation of Canadian Indigenous languages. Prior research on languages such as Inuktitut (Gupta and Boulianne, 2020a), Cree (Gupta and Boulianne, 2020b), and other North American Indigenous languages, including Hupa (Liu et al., 2022), has predominantly employed hybrid ASR architectures. These approaches typically demand in-depth linguistic expertise, particularly for the careful design and selection of subword units. While models such as Wav2vec2, XLS-R and Whisper have been explored for language documentation tasks (Jimerson et al., 2023; Rodríguez and Cox, 2023), including for languages like Hupa (Jimerson et al., 2023; Venkateswaran and Liu, 2024), Seneca (Jimerson et al., 2023) and Oneida (Jimerson et al., 2023), our work is the first to conduct a comprehensive investigation of pre-trained SFMs in this context. By leveraging these models, we aim to widen the so-called "transcription bottleneck" and accelerate language documentation efforts.

2. **First ASR-driven documentation pipeline for the SENĆOŦEN language**: This work introduces the first ASR-driven documentation pipeline tailored for SENĆOŦEN. To address the challenges of limited data and high lexical variation, we adopt a two-pronged strategy: ASR data augmentation via TTS and cross-lingual transfer learning based on SFMs. Moreover, we perform a systematic analysis of ASR performance under more extreme conditions, reducing the available training data to as little as 10 minutes[1].

[1]Details can be found in the Appendix.

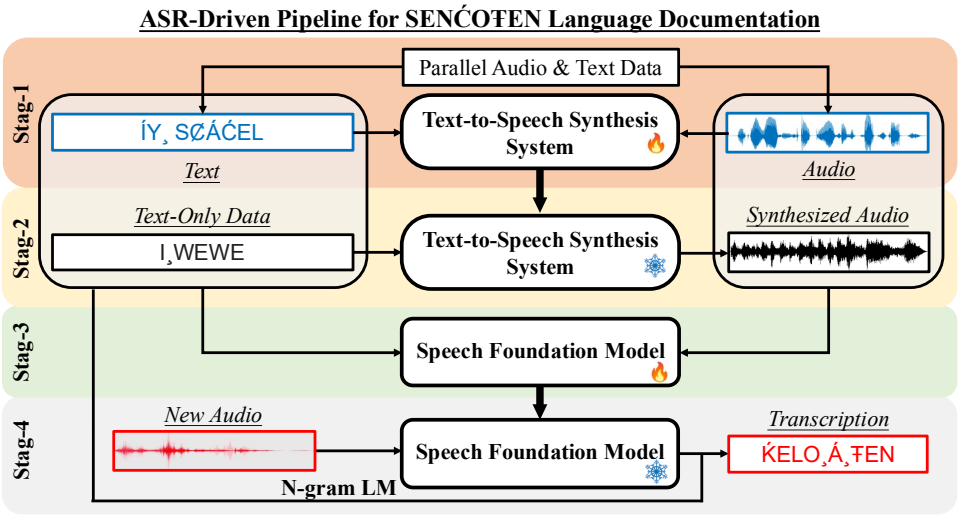


Figure 1: The proposed multi-stage ASR-driven pipeline to support SENĆOŦEN language documentation.

3. **Promising ASR performance with extended error analysis**: The top-performing system, integrating cross-lingual transfer learning, TTS-based data augmentation and language model fusion, achieves a word error rate (WER) of **19.34%** and a character error rate (CER) of **5.09%** on the test set with a **57.02% out-of-vocabulary (OOV) rate**. Furthermore, by filtering out minor errors involving missing or extraneous cedillas (¸), the WER and CER further improve to **14.32%** (26.48% on unseen words) and **3.45%**, respectively. These findings highlight the system's capability to significantly expedite the transcription process for SENĆOŦEN, providing valuable assistance in efforts to revitalize the language.

The rest of the paper is organized as follows. Section 2 outlines the proposed ASR-driven pipeline developed to support the documentation of the SENĆOŦEN language. Section 3 details the TTS system for synthesizing audio to augment ASR training data. Section 4 discusses the application of cross-lingual transfer learning based on speech foundation models. Experimental results and analysis on the SENĆOŦEN dataset are presented in Section 5. Section 6 provides conclusions and discusses potential directions for future research.

2 ASR-Driven Pipeline for SENĆOŦEN Language Documentation

As illustrated in Figure 1, our proposed ASR-driven pipeline for SENĆOŦEN language documentation consists of four stages, with carefully designed procedures to maximize the usage of the available audio and text data:

Stage 1: Train the TTS system: Parallel audio and text data in SENĆOŦEN are used to train a custom-designed text-to-speech (TTS) system, which will be described in detail in Section 3.

Stage 2: Generate synthesized audio via TTS: SENĆOŦEN text without accompanying audio is fed into the trained TTS system to generate the corresponding synthesized audio.

Stage 3: Perform cross-lingual transfer learning on the SFM: The original parallel audio and text data, combined with the synthesized audio from the text-only data, are utilized to perform cross-lingual transfer learning on the speech foundation model (SFM), which will be outlined in Section 4.

Stage 4: Transcribe new audio with the fine-tuned SFM: New audio in SENĆOŦEN is transcribed using the fine-tuned SFM, with the option to fuse an external language model (LM) trained on part or all of the available text data to further improve accuracy.

3 Text-to-Speech Synthesis

Text-to-speech (TTS) synthesis has emerged as a powerful technique for augmenting ASR training datasets (Gokay and Yalcin, 2019), particularly in scenarios where parallel audio and text resources are limited. As there exists written SENĆOŦEN text without corresponding audio recordings, TTS-generated audio can be used to augment the training data for developing SENĆOŦEN ASR systems.

To mitigate the scarcity of parallel audio & text data in SENĆOŦEN, a three-phase approach is

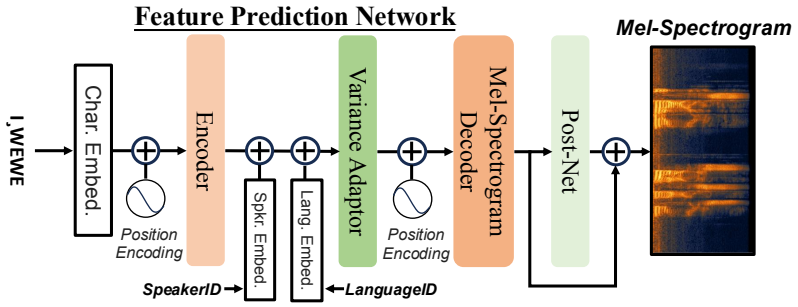


Figure 2: Architecture of the feature prediction work of our TTS system. "Char. Embed.", "Spkr. Embed." and "Lang. Embed." respectively denote character, speaker, and language embeddings.

adopted using the EveryVoice TTS Toolkit Pine et al. (2022b), including **1)** training a feature prediction network, **2)** developing a vocoder, and **3)** aligning vocoder outputs with mel-spectrograms generated by the prediction network (i.e., vocoder matching).

3.1 Feature Prediction Network

Building on the work of Pine et al. (2022b), the EveryVoice TTS toolkit uses a modified FastSpeech 2 (Ren et al., 2020) architecture as the feature prediction network. As illustrated in Figure 2, the key modifications are:

- Substitution of standard convolutions with depthwise separable convolutions in both the encoder and mel-spectrogram decoder (Pine et al., 2022b) to enhance parameter efficiency.
- Integration of learnable speaker embeddings (Figure 2, middle, in circled box).
- Incorporation of a decoder post-net (Figure 2, right, in light green).

Moreover, pre-generated forced alignments are replaced by a jointly-trained alignment module (Badlani et al., 2022), while pitch and energy are predicted at the phoneme level instead of the frame level to achieve smoother prosody.

3.2 Vocoder

Since speech foundation models directly process raw audio, ensuring high-quality waveform synthesis is crucial. The vocoder, which converts intermediate mel-spectrograms into waveforms, plays a key role in this process. To this end, we utilize HiFi-GAN (Kong et al., 2020), a widely adopted generative adversarial network recognized for generating natural and high-quality waveforms, as the vocoder in our TTS system.

3.3 Vocoder Matching

To mitigate the artifacts arising from limited training data, a vocoder matching strategy is employed after the initial training of the vocoder. This process fine-tunes the vocoder using mel-spectrograms generated by the feature prediction network as input, aligning it with the specific characteristics of these spectrograms to minimize discrepancies between training and inference conditions.

4 Cross-Lingual Transfer Learning

The limited availability of SENĆOŦEN data makes it impractical to train an end-to-end (E2E) ASR system from scratch. Alternatively, recent advances in speech foundation models (SFMs), which are pre-trained on large-scale datasets, offer a promising pathway for cross-lingual transfer learning in low-resource languages like SENĆOŦEN.

SFMs can be categorized into two main types:

Encoder-Based SFM: Encoder-based SFMs have gained widespread adoption due to their ability to convert raw audio into representations useful for various downstream tasks. Widely recognized models in this category include Wav2Vec 2.0 (Wav2Vec2) (Baevski et al., 2020), HuBERT (Hsu et al., 2021), WavLM (Chen et al., 2022), and Data2Vec (Baevski et al., 2022). These models employ a single encoder architecture to process audio, with a focus on self-supervised learning from unlabeled data. Both Wav2Vec2 and HuBERT excel at capturing rich speech representations, which are crucial for ASR in low-resource settings. WavLM further improves performance by effectively modeling not only speech but also environmental noise, making it particularly robust in challenging acoustic conditions. Data2Vec, on the other hand, expands the applicability of these models by generalizing the approach to multiple modalities.

Encoder-Decoder-Based SFM: In contrast, encoder-decoder-based SFMs integrate both an encoder to process the input audio and a decoder to generate transcriptions or other forms of output. Whisper (Radford et al., 2023) is among the most well-known models in this category. By combining these two components, Whisper is capable of end-to-end transcription, making it a powerful tool for ASR tasks. Its architecture is particularly useful for languages with limited resources, as the encoder-decoder framework allows for more sophisticated handling of complex linguistic structures through cross-lingual transfer learning.

5 Experiments

5.1 Task Description

Parallel Audio & Text Data: The SENĆOŦEN speech dataset, part of the "Speech Generation for Indigenous Language Education" project (Pine et al., 2025), consists of about 4 hours of single-speaker recordings. A Kaldi-based (Povey et al., 2011) GMM-HMM system is used to estimate emission probabilities for each utterance. 20% of the data is then allocated as the test set based on these estimates to ensure a balanced representation of difficulty. After silence stripping, the training set contains 1.7 hours and the test set 0.2 hours, with average utterance lengths of 2.06 and 2.04 seconds, respectively. The training set consists of 3k utterances with 3.6k distinct words, while the test set includes 0.8k utterances with 1.2k distinct words. The average word length is 3.6 characters in both sets. Due to SENĆOŦEN's polysynthetic nature and stress-driven metathesis, the test set shows a high out-of-vocabulary (OOV) rate of 57.02%.

Text-Only Data: We have permission to access the SENĆOŦEN dictionary (Montler, 2018), the most comprehensive lexicographic resource for the language, containing over 30k words and example sentences. This data is text-only, with no corresponding audio. 27k words and sentences are retained after filtering out overly long entries exceeding 81 characters.

5.2 Experiment Setup

Data processing: We conduct silence stripping using SoX[2] and denoise the audio with an RNN-based denoiser[3]. The audio is then resampled to 16 kHz for ASR and 22.05 kHz for TTS development, while words are segmented into characters.

Text-to-Speech Synthesis: The TTS system outlined in Section 3 is built using the EveryVoice TTS Toolkit[4]. The train/test split mirrors that of the ASR system. The modified FastSpeech2 feature prediction network includes 4 encoder and 4 decoder blocks[5], while HiFi-GAN in its V1 configuration (Kong et al., 2020) is used as the vocoder. The synthesized audio is automatically evaluated using the TorchSquim (Kumar et al., 2023) model, which provides estimates for short-time objective intelligibility (STOI), perceptual evaluation of speech quality (PESQ), scale-invariant signal-to-noise ratio (SI-SNR), and mean opinion score (MOS).

Cross-lingual Transfer Learning: We utilize the Hugging Face platform to perform cross-lingual transfer learning with both encoder-based speech foundation models, including Wav2Vec2, HuBERT, WavLM, and Data2Vec, as well as encoder-decoder-based models[6] like Whisper. SFMs of varying model sizes and pretraining data serve as the starting point for this process.

Language Model Fusion: We construct two 4-gram language models (LMs) using the KenLM toolkit (Heafield, 2011):

- A smaller model ("small-4g") trained exclusively on the text from the training set.
- A larger model ("large-4g") that also incorporates the 27k text-only SENĆOŦEN sentences.

The "small-4g" LM covers 3.6k words, while the "large-4g" spans 14k. For encoder-based SFMs (e.g., Wav2Vec2), shallow fusion (Kannan et al., 2018) integrates the n-gram LM during decoding. For encoder-decoder SFMs (e.g., Whisper), the LM rescales the n-best hypothesis list.

5.3 Performance Analysis

Table 1: TTS evaluation on the SENĆOŦEN test set.

Vocoder Matching	STOI (↑)	PESQ (↑)	SI-SNR (↑)	MOS (↑)
✗	0.980	3.207	20.339	4.227
✓	0.985	3.324	20.809	4.336

[2] https://linux.die.net/man/1/sox
[3] https://github.com/xiph/rnnoise
[4] https://github.com/EveryVoiceTTS/EveryVoice
[5] Both encoder and decoder blocks have a 1024-dim feed-forward layer and two 128-dim attention heads.
[6] https://huggingface.co/blog/{fine-tune-wav2vec2-english,fine-tune-whisper}

Table 2: Performance of cross-lingual transfer learning using SFMs with different architectures. ✗ in the "Multilingual" column indicates that the SFM is pre-trained on English data only. "WER/CER" represents word/character error rate, while "seen" and "unseen" refer to whether the test words are included in the original training data.

Sys.	Model	Multi-Lingual	LM	CER% (↓)	WER% (↓)		
					seen	unseen	all
1	*Wav2Vec2-random-int*	-	-	*84.36*	*99.94*	*100.00*	*99.96*
2	Wav2Vec2-base	✗		10.68	36.18	65.99	49.10
3	Wav2Vec2-large			8.25	21.42	56.87	35.04
4	Wav2Vec2-xlsr-53	✓	-	11.23	33.48	69.13	51.44
5	Wav2Vec2-xls-r-300m			10.41	31.55	63.35	45.63
6	Wav2Vec2-xls-r-1b			**6.32**	14.87	**55.04**	**27.81**
7	Data2Vec-base	✗	-	14.89	40.09	78.56	60.61
8	Data2Vec-large			9.29	27.75	60.08	40.51
9	HuBERT-base	✗	-	13.34	45.15	72.04	58.61
10	HuBERT-large			12.29	44.66	69.78	56.06
11	WavLM-base	✗	-	11.70	36.18	71.27	50.71
12	WavLM-large			13.43	45.98	71.88	59.61
13	Whisper-medium-en	✗	-	7.36	15.38	57.30	28.13
14	Whisper-large-v2	✓		**7.11**	14.60	**58.01**	**27.66**
15	Wav2Vec2-xls-r-1b	✓	small-4g	6.05	12.18	57.92	25.13
16			large-4g	**5.63**	12.35	**49.09**	23.16
17	Whisper-large-v2	✓	small-4g	6.53	11.09	60.14	25.16
18			large-4g	6.12	10.95	50.35	**22.67**

The evaluation of the TTS system, cross-lingual transfer learning with SFMs, and the integration of TTS-based data augmentation and language models is conducted on the test set described in Section 5.1. In this context, "seen" and "unseen" words in terms of word error rate (WER) refer to whether the test words were present in the original training data.

Text-to-Speech Synthesis: We evaluate the TTS system on the SENĆOŦEN test set using four metrics: STOI, PESQ, SI-SNR, and MOS. As indicated in Table 1, performance improves across all four metrics with vocoder matching. Based on this, we use the vocoder-matched system to synthesize 27k SENĆOŦEN sentences outlined in Section 5.1, resulting in approximately 11.6 hours of generated speech for ASR data augmentation. Compared to the 13.3-hour augmented training set, the test set retains an OOV rate of 29.96%.

Cross-lingual Transfer Learning: Table 2 illustrates the results of cross-lingual transfer learning across various speech foundation models (SFMs). As part of an ablation study (Sys. 1), we also carry out an additional experiment where the weights of the Wav2Vec2-base model are randomly re-initialized to serve as the starting point. In addition, the top-performing encoder- and encoder-decoder-based SFMs are further integrated with the 4-gram LMs (Sys. 15-18).

Several insights can be drawn from Table 2: **1)** Larger SFMs do not consistently deliver better results than smaller models with similar architectures (Sys. 12 *vs.* 11). **2)** Although the top-performing SFMs are pre-trained on multilingual datasets (Sys. 6, 14), they do not always outperform monolingual models with similar structures trained solely on English (Sys. 4-5 *vs.* 3). **3)** Incorporating an external LM further boosts performance (Sys. 15-16 *vs.* 6 and Sys. 17-18 *vs.* 14), with larger LMs providing better outcomes (Sys. 16 *vs.* 15, Sys. 18 *vs.* 17). **4)** A substantial performance gap exists between words covered ("seen") in the training data and those that are not ("unseen"), while the top-performing systems (Sys. 16,18) correctly transcribe roughly half of the unseen words.

TTS-Based Data Augmentation: We progressively incorporate TTS-synthesized data into the cross-lingual transfer learning process of the top-performing SFM (Table 2, Sys. 18), i.e., Whisper[7].

[7] https://huggingface.co/openai/whisper-large-v2

Table 3: Performance of incorporating TTS-synthesized data in cross-lingual transfer learning with the **Whisper** model. The original training data is always included, while "all" denotes the full 11.6-h synthesized data.

Sys.	Aug. Data	LM	CER% (↓)	WER% (↓)		
				seen	unseen	all
1	1h	-	6.67	12.97	54.97	25.38
2	2h		6.02	12.67	51.42	23.99
3	4h		5.84	12.67	47.52	22.86
4	6h		5.72	11.34	49.79	22.56
5	8h		5.79	11.79	48.44	22.53
6	all		**5.63**	12.18	**44.81**	**22.01**
7	1h	small fg	6.80	10.55	56.31	23.74
8	2h		5.82	10.95	52.41	22.82
9	4h		5.50	9.71	49.79	21.21
10	6h		5.59	9.96	50.50	21.65
11	8h		5.42	9.57	48.73	20.70
12	all		**5.26**	9.91	**46.51**	**20.51**
13	1h	large fg	6.60	10.65	49.08	22.60
14	2h		5.96	10.06	49.36	22.42
15	4h		5.38	9.57	45.67	20.84
16	6h		5.18	9.22	46.60	20.44
17	8h		**5.09**	8.43	44.96	**19.34**
18	all		5.11	9.62	**43.53**	20.15

The augmentation begins with 1 hour of synthesized data and scales up to a total of 11.6 hours. As shown in Table 3, several trends can be observed: **1)** Incorporating TTS-synthesized data leads to ASR performance improvements both with or without an external LM (Sys. 1-6, 7-12, 13-18 in Table 3 *vs.* Sys. 14,17,18 in Table 2), with overall WER reductions of up to 5.65% abs. (20.43% rel.) and 13.20% abs. (22.75% rel.) on unseen words absent from the original training set (Sys. 6 in Table 3 *vs.* Sys. 14 in Table 2). **2)** For systems fused with the large 4-gram LM covering all text used for TTS, the inclusion of synthesized audio further improves ASR performance (Sys. 13-18 in Table 3 *vs.* Sys. 18 in Table 2). **3)** There is a general trend of performance convergence when 8 hours of TTS-synthesized data are added (Sys. 5,11,17 in Table 3).

5.4 Language Documentation Support

The motivation for this project stemmed from discussions between the authors in the context of regular meetings related to a multi-year TTS research project, described in detail in Pine et al. (2025). ASR was not an explicit goal of the project that brought us together, but the first author of this paper has expertise in speech recognition and realized that we had some of the requisite pieces to develop a proof-of-concept ASR system, namely, an established relationship with the language community in question, pre-trained TTS models, and some modest amounts of parallel text-audio data. The first author proposed the idea, along with possible benefits and risks, to members of the W̱SÁNEĆ school Board at a meeting for the TTS project, which was met with enthusiasm and support leading to this initial effort. Despite the strong results from these initial experiments, many more steps and protocols will be required to connect this technology with on-the-ground language efforts.

To help us demonstrate the capabilities of this technology, we developed an intuitive, web-based user interface and API using the Gradio framework (Abid et al., 2019) in Python. The interface, as illustrated in Figure 3, allows users to interact with the model by either speaking directly into the microphone or uploading a pre-recorded audio file. After providing the input, users can click the "Submit" button (Figure 3, bottom, in orange) to generate an automatic transcription, displayed in the text box labeled "output" (Figure 3, right).

Additionally, users can select specific segments of the audio (Figure 3, left) to view their corresponding transcriptions, enabling precise analysis of smaller portions of the recording. The "Flag" button, located on the right, allows users to mark an audio-transcription pair for further review, annotation, or reference. This functionality is particularly useful for collaborative workflows, where flagged segments may require validation or additional context from language experts. By simplifying transcription workflows, the interface streamlines language documentation, enabling linguists, community members, and researchers to efficiently process spoken language data with decreased manual effort. Features like audio segmentation and flagging support iterative transcription processes, while the web-based design ensures accessibility across a wide range of devices, making it well-suited for teams working in different locations.

To safeguard the model's privacy, the interface is currently accessible exclusively through a private web gateway. Future developments aim to facilitate closer alignment with documentation and language revitalization workflows (e.g., Cox, 2019; Adams et al., 2021).

A closer examination of the decoded outputs from the SENĆOŦEN ASR systems reveals that a notable portion of the errors involve missing or extraneous cedillas (¸) which indicate either glottalization when following resonant or glottal stops otherwise. Given that these errors are relatively

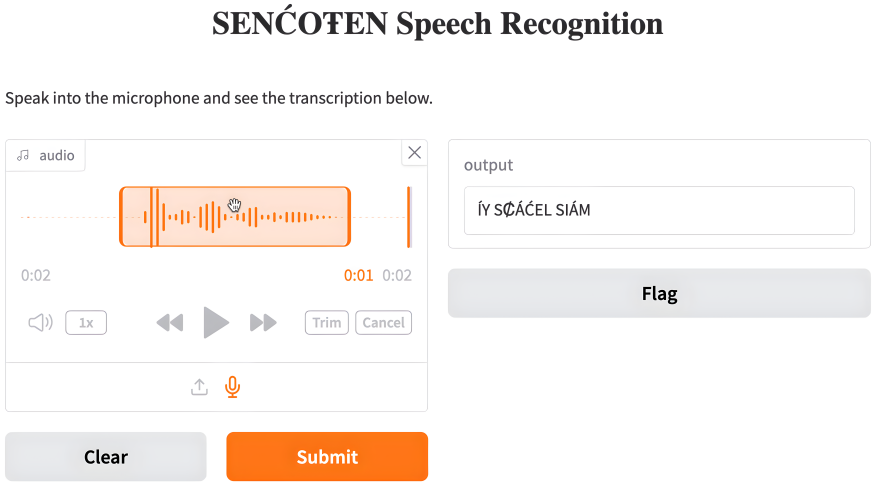


Figure 3: Demonstration of the user interface designed to support SENĆOŦEN language documentation.

easily correctable by a SENĆOŦEN speaker, and that the consistency of their use varies, we reassess the ASR performance with these errors excluded. As shown in Table 4, the system achieves an overall WER of **14.32%**, a CER of **3.45%**, and a WER of **26.48%** on unseen words. This demonstrates the potential of our proposed ASR-driven pipeline to support the documentation for SENĆOŦEN.

Table 4: Performance of the top-performing SFM (Sys. 17 in Table 3), excluding errors related to cedilla (¸).

Model	CER% (↓)	WER% (↓)		
		seen	unseen	all
Whisper	**3.45**	6.88	**26.48**	**14.32**

6 Conclusion

In this paper, we proposed an ASR-driven pipeline designed to tackle the unique challenges of documenting the SENĆOŦEN language, which is hindered by data scarcity, substantial vocabulary variation, and phonological complexity. By incorporating augmented speech data from a TTS system, cross-lingual transfer learning using speech foundation models (SFMs), and an n-gram language model via shallow fusion, we demonstrated the effectiveness of our approach in improving ASR performance for low-resource languages. Our experiments on the SENĆOŦEN dataset yielded a WER of 19.34% and a CER of 5.09%, with further improvements to a WER of 14.32% (26.48% on unseen words) and a CER of 3.45% after mitigating minor cedilla-related errors. These results highlight the potential of the proposed pipeline to enhance SENĆOŦEN language documentation, offering a valuable tool for ongoing language revitalization efforts. Future work will focus on more linguistically oriented techniques, for example, modeling stress-driven metathesis in SENĆOŦEN.

References

Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*.

Oliver Adams, Benjamin Galliot, Guillaume Wisniewski, Nicholas Lambourne, Ben Foley, Rahasya Sanders-Dwyer, Janet Wiles, Alexis Michaud, Séverine Guillaume, Laurent Besacier, Christopher Cox, Katya Aplonova, Guillaume Jacques, and Nathan Hill. 2021. User-friendly automatic transcription of low-resource languages: Plugging ESPnet into elpis. In *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-4)*, pages 51–62.

Rohan Badlani, Adrian Łańcucki, Kevin J. Shih, Rafael Valle, Wei Ping, and Bryan Catanzaro. 2022. One TTS Alignment to Rule Them All. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*.

Sonya Bird. 2020. Pronunciation among adult Indigenous language learners: The case of SENĆOŦEN. *Journal of Second Language Pronunciation*, 6(2):148–179.

Sonya Bird and Sarah Kell. 2017. The role of pronunciation in SENĆOŦEN language revitalization. *Canadian Modern Language Review*, 73(4):538–569.

Peter Brand, John Elliot, and Ken Foster. 2002. Language revitalization using multimedia. *Indigenous languages across the community*, pages 245–247.

William Chan, Navdeep Jaitly, Quoc Le, et al. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*.

Yen-Lu Chow and Richard Schwartz. 1989. The n-best algorithm: Efficient procedure for finding top n sentence hypotheses. In *Speech and Natural Language: Proceedings of a Workshop Held at Cape Cod, Massachusetts, October 15-18, 1989*.

Christopher Cox. 2019. Persephone-ELAN. Available at: https://github.com/coxchristopher/persephone-elan.

Dave Elliott Sr. 2024. *Saltwater People*. SD 63, Victoria, BC, Canada. Available at: https://www.uvicbookstore.ca/general/browse/abor/9780000105356.

FirstVoices. 2024. SENĆOŦEN. https://www.firstvoices.com/sencoten. Accessed: 2024-10-01.

Ben Foley, Josh Arnold, Rolando Coto-Solano, Gautier Durantin, T. Mark Ellison, Daan van Esch, Scott Heath, František Kratochvíl, Zara Maxwell-Smith, David Nash, Ola Olsson, Mark Richards, Nay San, Hywel Stoakes, Nick Thieberger, and Janet Wiles. 2018. Building Speech Recognition Systems for Language Documentation: The CoEDL Endangered Language Pipeline and Inference System (ELPIS). In *6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, pages 205–209.

Suzanne Gessner, Tracey Herbert, and Aliana Parker. 2022. The Report on the Status of B.C. First Nations Languages 2022. Accessed: 2024-10-01.

Ramazan Gokay and Hulya Yalcin. 2019. Improving Low Resource Turkish Speech Recognition with Data Augmentation and TTS. In *2019 16th International Multi-Conference on Systems, Signals & Devices (SSD)*.

Anmol Gulati, James Qin, Chung-Cheng Chiu, et al. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

Vishwa Gupta and Gilles Boulianne. 2020a. Automatic transcription challenges for Inuktitut, a low-resource polysynthetic language. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Vishwa Gupta and Gilles Boulianne. 2020b. Speech transcription challenges for resource constrained indigenous language Cree. In *Proceedings of the Joint Workshop on Spoken Language Technologies for Under-resourced Languages and Collaboration and Computing for Under-Resourced Languages (SLTU-CCURL)*.

Eve Haque and Donna Patrick. 2015. Indigenous languages and the racial hierarchisation of language policy in Canada. *Journal of Multilingual and Multicultural Development*.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*.

Shoukang Hu, Xurong Xie, Mingyu Cui, Jiajun Deng, Shansong Liu, Jianwei Yu, Mengzhe Geng, Xunying Liu, and Helen Meng. 2022. Neural architecture search for LF-MMI trained time delay neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*.

Jacqueline Jim. 2016. W̱SÁNEĆ SEN: I am emerging: An Auto-Ethnographic study of life long SENĆOŦEN language learning. Master's thesis, University of Victoria, Victoria, British Columbia, Canada.

Robbie Jimerson and Emily Prud'hommeaux. 2018. ASR for documenting acutely under-resourced indigenous languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Robert Jimerson, Zoey Liu, and Emily Prud'Hommeaux. 2023. An (unhelpful) guide to selecting the best asr architecture for your under-resourced language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1008–1016.

Anjuli Kannan, Yonghui Wu, Patrick Nguyen, Tara N Sainath, Zhijeng Chen, and Rohit Prabhavalkar. 2018. An analysis of incorporating an external language model into a sequence-to-sequence model. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*.

Anurag Kumar, Ke Tan, Zhaoheng Ni, Pranay Manocha, Xiaohui Zhang, Ethan Henderson, and Buye Xu. 2023. Torchaudio-Squim: Reference-Less Speech Quality and Intelligibility Measures in Torchaudio. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Guinan Li, Jiajun Deng, Mengzhe Geng, Zengrui Jin, Tianzi Wang, Shujie Hu, Mingyu Cui, Helen Meng, and Xunying Liu. 2023. Audio-visual end-to-end multi-channel speech separation, dereverberation and recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*.

Patrick Littell, Anna Kazantseva, Roland Kuhn, Aidan Pine, Antti Arppe, Christopher Cox, and Marie-Odile Junker. 2018. Indigenous language technologies in Canada: Assessment, challenges, and successes. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Zoey Liu, Justin Spence, and Emily Prud'hommeaux. 2022. Enhancing documentation of Hupa with automatic speech recognition. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 187–192.

Timothy Montler. 1986. *An Outline of the Morphology and Phonology of Saanich, North Straits Salish*. Number 4 in Occasional Papers in Linguistics. University of Montana, Linguistics Laboratory, Missoula, Montana.

Timothy Montler. 2018. *SENĆOŦEN: A Dictionary of the Saanich Language*. University of Washington Press, Seattle, WA, USA.

Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. A time delay neural network architecture for efficient modeling of long temporal contexts. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

Aidan Pine, Erica Cooper, David Guzmán, Eric Joanis, Anna Kazantseva, Ross Krekoski, Roland Kuhn, Samuel Larkin, Patrick Littell, Delaney Lothian, Akwiratékha' Martin, Korin Richmond, Marc Tessier, Cassia Valentini-Botinhao, Dan Wells, and Junichi Yamagishi. 2025. Speech generation for Indigenous language education. *Computer Speech & Language*, 90.

Aidan Pine, Patrick William Littell, Eric Joanis, David Huggins Daines, Christopher Cox, Fineen Davis, Eddie Antonio Santos, Shankhalika Srikanth, Delasie Torkornoo, and Sabrina Yu. 2022a. Gi2Pi Rule-based, index-preserving grapheme-to-phoneme transformations. In *ComputEL*.

Aidan Pine and Mark Turin. 2017. Language revitalization. *Oxford Research Encyclopedia of Linguistics*.

Aidan Pine, Dan Wells, Nathan Brinklow, Patrick Littell, and Korin Richmond. 2022b. Requirements and motivations of low-resource speech synthesis for language revitalization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7346–7359.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The Kaldi speech recognition toolkit. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Lorena Martín Rodríguez and Christopher Cox. 2023. Speech-to-text recognition for multilingual spoken data in language documentation. In *Proceedings of the 7th Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-7)*.

George Saon, Hagen Soltau, David Nahamoo, and Michael Picheny. 2013. Speaker adaptation of neural network acoustic models using i-vectors. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

Nitin Venkateswaran and Zoey Liu. 2024. Looking within the self: Investigating the impact of data augmentation with self-training on automatic speech recognition for hupa. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 58–66.

Yongqiang Wang, Abdelrahman Mohamed, Due Le, et al. 2020. Transformer-based acoustic modeling for hybrid speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Appendix A Further Ablation Studies

To get an in-depth analysis of our proposed ASR-driven documentation pipeline (Figure 1), two sets of ablation studies are further conducted: **1)** replacing the end-to-end speech foundation model with a conventional hybrid TDNN ASR system, and **2)** Reducing the training data to as little as 10 mins to simulate ultra-low resource settings.

Hybrid TDNN Systems: The hybrid TDNN system is constructed following the Kaldi Chain recipe[8] but with a more compact architecture featuring 7 context-splicing layers with time strides of $\{1, 1, 0, 3, 3, 6\}$. I-Vectors (Saon et al., 2013) are incorporated while speed perturbation is omitted. The text is transcribed into International Phonetic Alphabet (IPA) representations using the g2p library (Pine et al., 2022a).

Table 5 reveals the following trends: **1)** Using a larger language model (LM) leads to noticeable performance degradation when the additional words in the LM lack corresponding audio in the training set (Sys. 2 *vs.* Sys. 1). **2)** Using the small 4-gram, expanding the training set's word coverage with TTS-synthesized data leads to marginal performance improvement (Sys. 3 *vs.* Sys. 1). However, a substantial gain is achieved when the text used for TTS is also included in LM training (Sys. 4 *vs.* Sys. 3). **3)** The top-performing SFMs (Sys. 17-18 in Table 3) largely outperforms the best hybrid TDNN system (Sys. 4 in Table 5 across all metrics, showing the effectiveness of using SFMs in the proposed ASR-driven documentation pipeline.

Table 5: Performance of hybrid TDNN system. "Data Aug." refers to TTS-based data augmentation. "# Hrs" denotes the duration of the training set.

Sys.	Data Aug.	# Hrs	LM	CER% (↓)	WER% (↓)		
					seen	unseen	all
1	✗	1.7	small-4g	19.68	18.93	100.00	46.92
2			large-4g	19.59	20.46	100.00	49.34
3	✓	13.3	small-4g	19.72	16.67	100.00	46.23
4			large-4g	**9.65**	16.62	**58.92**	**36.63**

Ultra-Low Resource Scenarios: We simulate ultra-low-resource conditions by utilizing just 10 minutes of parallel audio and text data to perform cross-lingual transfer learning with SFMs, excluding the external LM, and assuming this limited training data is the only available resource. As shown in Table 6, Whisper (Sys. 5-8) is more sensitive to the amount of data available for transfer learning compared to Wav2Vec2 (Sys. 1-4). This may be attributed to differences in their architectures, with Wav2Vec2 being encoder-based, while Whisper follows an encoder-decoder structure.

Table 6: Performance of cross-lingual transfer learning on SFMs in ultra-low resource scenarios with as little as 10 min training data. No LM fusion is incorporated.

Sys.	Model	Train Data	CER% (↓)	WER% (↓)		
				seen	unseen	all
1	Wav2Vec2	all	6.32	14.87	55.04	27.81
2		1h	7.54	20.15	54.99	32.92
3		30min	8.78	23.79	59.77	37.74
4		**10min**	**12.11**	34.42	**67.94**	**50.13**
5	Whisper	all	7.11	14.60	58.01	27.66
6		1h	9.13	17.36	59.94	31.17
7		30min	10.95	26.53	69.22	41.17
8		10min	21.28	43.34	82.58	63.00

[8] https://github.com/kaldi-asr/kaldi/tree/master/egs/librispeech/s5

Speech Technologies with Fieldwork Recordings: the Case of Haitian Creole

William N. Havard[1,2], **Renauld Govain**[3], **Benjamin Lecouteux**[2], **Emmanuel Schang**[1]

[1] LLL, Université d'Orléans, CNRS, 45000 Orléans, France
[2] LIG, Université Grenoble Alpes, CNRS, Grenoble INP, 38000 Grenoble, France
[3] LangSé, Université d'État d'Haïti, Port-au-Prince, Haïti

william.havard@univ-orleans.fr

Abstract

We use 40-year-old digitalised tape-recorded fieldwork data in Haitian Creole to train a *native* self-supervised learning (SSL) model of speech representation (WAV2VEC2). We also use a continued pre-training approach on pre-trained SSL models of two *foreign* languages: the lexifier language – French – and an unrelated language – English. We compare the performances of these three SSL models, and of two other foreign SSL models directly fine-tuned, on an ASR task, where all five models are fine-tuned on transcribed fieldwork recordings in Haitian Creole. Our results show the best-performing model is the one trained using a continued pre-training approach on the lexifier language, followed by the native model. We conclude that the 'mobilising the archive'-approach advocated by (Bird, 2020) is a promising way forward to design speech technologies for new languages.

1 Introduction

Most of the so-called low-resourced languages are often low-resourced from the perspective of computer scientists only: they often have many resources that were collected over the years by linguists, missionaries, and generally by the community of speakers itself (Bird, 2020). The data is often not readily accessible (i.e. in a digitalised format), but existent nonetheless. The question we aim to answer in this paper is the following: how far can we go with state-of-the-art speech processing models using *only* fieldwork data? By 'fieldwork data', we mean data that was *not* originally collected to serve as training data for computational applications (e.g. Automatic Speech Recognition, ASR), but was collected for linguistic purposes (e.g. to study dialectal variation). In this paper, we focus on spoken data in Haitian Creole, consisting of recorded interviews between linguists and their collaborators. Haitian Creole is a French-based Creole (i.e. French is called its lexifier language, the language that gave Haitian Creole most of its vocabulary (Hazael-Massieux, 2012)), spoken by 13M speakers (Simons and Fennig, 2023) in Haiti and by the Haitian diaspora.

Most of the data we use in this paper (see Section 2) was collected 40 years ago with tape recorders to study dialectal variation in Haitian, with a focus on lexical variations. Contrary to the clean audiobooks commonly used to train neural models (e.g. Librispeech, (Panayotov et al., 2015)), the data we used is inherently noisy: reverberated, echo-y, full of environmental noise (e.g. chickens, cars, passers-by, etc.). Yet, this type of data represents the majority of the data available for most of the world languages. As collecting and transcribing data is a costly process (Himmelmann, 2018), is it possible to make use — as advocated by (Bird, 2020) in the 'mobilising the archive'-approach — of already existing (and potentially old) fieldwork data and re-purpose them for computational applications?

1.1 Related Works

The field of speech processing for Creole languages is relatively sparse, except for the work of (Breiter, 2014) for Haitian Creole, that of (Macaire et al., 2022; Le Ferrand et al., 2023; Le Ferrand and Prud'hommeaux, 2024) for Guadeloupean and Mauritian Creole, and (Gooda Sahib-Kaudeer et al., 2019) for Mauritian Creole (with a focus on the medical domain). Hence, speech processing for Creole languages — whatever the lexifier language, be it French, English, Portuguese, etc. — remains largely unexplored.

Unrelated to Creole languages — but related to our experimental settings — (Nowakowski et al., 2023) explored continuous pre-training (CPT) approaches, followed by an ASR fine-tuning task for Ainu speech recognition using old fieldwork data. In short, CPT is a form of transfer learning which

Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 40–46

consists in using large quantities of unlabelled data (i.e. raw speech) to continue to pre-train models that were already pre-trained on another language. However (Nowakowski et al., 2023) do not train their models 'on a budget' as (i) they use 4 GPUs and (ii) use the XLSR-53 model (Conneau et al., 2021) which is based on WAV2VEC2-LARGE and pre-trained on 56k hours of data, and (iii) use multilingual fine-tuning by which the ASR model is not only trained on the target language (Ainu), but on several languages at once (English, Japanese, alongside Ainu). We aim for a stricter approach that only uses fieldwork data at all steps.

1.2 Research Questions

In this work, **we *only* assume the existence of (potentially old) fieldwork data** to train the models, which corresponds to several real-world use cases: that of field linguists documenting a language and that have gathered a certain amount of both untranscribed and transcribed recordings (our case), or that of a community of speakers that uses archival material to build models for their language.

More precisely, the questions we tackle in this paper are the following: **(a)** Would noisy, but ecologically valid, fieldwork data be usable to train self-supervised learning (SSL) models of speech (e.g. WAV2VEC2, (Baevski et al., 2020))? **(b)** Should said models be trained from scratch or should continued pre-training (CPT) (Gururangan et al., 2020; Nowakowski et al., 2023) approaches be used? **(c)** How much training data is necessary to fine-tune the models on an ASR task? And finally, **(d)** given our use-cases, is it possible to train such models 'on a budget'? (i.e. only 1 GPU, as having more – e.g. 64 as (Baevski et al., 2020) – is generally impossible for laypeople).

Additionally, as we work in the context of Creole languages, we also aim to explore the influence of the lexifier language (as a clear case of related languages) and **(e)** whether CPT be done on SSL models of the lexifier language (e.g. French in the case of Haitian Creole), or do models trained on an unrelated language (e.g. say English in the case of Haitian Creole) also work?

2 Data

ALH. We used the *Atlas Linguistique d'Haïti* (Fattier, 1998), which consists of a collection of 499 audio recordings in Haitian Creole (Kreyòl ayisyen) collected in Haiti between 1978 and 1987 for the purpose of creating a linguistic atlas. The recordings were originally done on audio cassettes with tape recorders which were then digitalised by the French National Library (*Bibliothèque Nationale de France*, BNF) in 2010. Each recording is on average 45 minutes long and features one or several interviewers eliciting words or phrases from their native collaborators. This data has been made publically available by the BNF and is accessible on the COCOON[1] platform. Although the recordings are associated with field notebooks containing partial handwritten transcriptions (phonetic transcription at word level), these have not been digitised (nor aligned with the recordings). As a result, this corpus consists entirely of raw speech. We partitioned the data set (356.3 hours) into 3 splits (train/val/test). The data was partitioned so that the validation set would contain at least 5 hours of data and a minimum of 2 unseen speakers, and the test set at least 5 hours of data and a minimum of 3 unseen speakers. We reached the following distribution which fulfilled our constraints: train = 345.6 hours; val = 5.3 hours, 5 unseen speakers; and test = 5.4 hours, 8 unseen speakers, the latter being left for future work.

CNCH. The *Corpus of Northern Haitian Creole*[2] (Valdman et al., 2015) consists of 10 recorded interviews, conducted in Cap-Haïtien (Northern Haiti) to study dialectal variation with regard to standard Haitian. This corpus was entirely transcribed by the linguist who collected it. However, the transcriptions used are non-standard and impressionistic, in the sense that spelling variations deviating from the norm are used to transcribe the speaker's pronunciation more faithfully: "*Powoprens*"/"*Potoprens*", Port-au-Prince, the capital city of Haiti; "*eskeu*"/"*eske*", question words; "*deu*"/"*de*", two; etc.). We partitioned the data set (9.0 hours) into 3 splits (train/val/test). The data was partitioned so that the val set would contain at least 1 hour of data and a minimum of 1 unseen speaker, and the test set at least 1 hour of data and a minimum of 1 unseen speaker. We reached the following distribution which fulfilled our constraints: train = 6.9 hours; val = 1.1 hours, 1 unseen speaker; test = 1.0 hours, 2 unseen speakers.

Other data sets. The two previous data sets are the ***only*** publicly available data sets of fieldwork recordings in Haitian Creole. We however wish to

[1] https://cocoon.huma-num.fr/

[2] https://archive.org/details/interview-8-ujf-107-a-ujm-107-a

acknowledge the existence of other data sets featuring speech in Haitian Creole, which we purposefully excluded during the training phase as they do not consist of fieldwork data: the freely accessible Haiti-CMU data set[3] which features read speech ($\sim$ 20 hours), mainly from sections of the Bible, which do not reflect everyday language use; and the proprietary IARPA-Babel data set consisting of "203 hours of Haitian Creole conversational and scripted telephone speech" (Andrus et al., 2017). We use both data sets to test our models on out-of-domain data and compare them with Facebook's MMS model (Pratap et al., 2023). For Haiti-CMU, we generated a test set that consists of 2 hours of data by randomly sampling recordings; and for IARPA-Babel we used the development set as a test set (as it is commonly done with IARPA-Babel data sets, as the evaluation set was kept private), which consists of 20 hours of data.

3 Experimental Settings

Given our low-budget setting, we focus on the WAV2VEC2-BASE architecture, thus excluding fine-tuning a multilingual model such as XLSR-53 which is based on the LARGE architecture.

3.1 Native and Foreign-SSL Pre-Training

We use the ALH corpus to pre-train our SSL models. A voice activity detection model (Pyannote (Bredin et al., 2020)) was used to isolate sections corresponding to speech from surrounding noises, resulting in 229h of spoken sections. The resulting segments were rather short ($\sim$ 2.3s) and unsuited to pre-train SSL models as-is. Thus, we merged them until the resulting concatenated segments reached 19s on average (19.4s$\pm$5.8). The WAV2VEC2 models were trained on a single GPU[4] using gradient accumulation for 16 steps (to simulate 16 GPUs) with 16-bits floats and a maximum batch size of 5.2 minutes. All the models were implemented using fairseq's standard WAV2VEC2-BASE implementation and training pipeline (Ott et al., 2019). Three models were trained:

- One model pre-trained from scratch (i.e. not based on any existing pre-trained model):
  - NATIVE -HAT-SSL+Ø : this model was pre-trained on the ALH data and has never been exposed to any other language other than Haitian (HAT) throughout pre-training;
- Two models pre-trained using a continued pre-training approach:
  - FOREIGN-FRA-SSL+CPT: the base model was pre-trained on a French (FRA) (wav2vec2-FR-7K-base, pre-trained on 7k hours in French (Parcollet et al., 2023)) and was continued pre-trained on the ALH data;
  - FOREIGN-ENG-SSL+CPT: the base model was pre-trained on English (ENG) (wav2vec2-base pre-trained on Librispeech 960 (Baevski et al., 2020)) and was continued pre-trained on the ALH data.

3.2 ASR fine-tuning

We fine-tuned our pre-trained and continued pre-trained models on the CNCH data set. 5 models were fine-tuned:

- Three models from models that had seen Haitian speech in a (continued) pre-training phase:
  - NATIVE -HAT-SSL+Ø +FT: where NATIVE -HAT-SSL+Ø was fine-tuned after the pre-training phase on ALH;
  - FOREIGN-FRA-SSL+CPT+FT: where FOREIGN-FRA-SSL+CPT was fine-tuned after the continued pre-training phase on ALH;
  - FOREIGN-ENG-SSL+CPT+FT: where FOREIGN-ENG-SSL+CPT was fine-tuned after the continued pre-training phase on ALH;
- Two models from models that hadn't seen any Haitian speech before being fine-tuned:
  - FOREIGN-FRA-SSL+Ø +FT: where the French (wav2vec2-FR-7K-base) was directly fine-tuned.
  - FOREIGN-ENG-SSL+Ø +FT: where the English (wav2vec2-base) was directly fine-tuned.

In order to understand the impact of the training size on the final performance of the models, we use different train sizes: max (360 minutes), 320,

[3] http://www.speech.cs.cmu.edu/haitian/

[4] 32Gb Nvidia Tesla V100 or 45Gb Nvidia A40 depending on availability.

Table 1: Configurations that yield the best performances in terms of WER (left) and CER (right) for each type of fine-tuned model. *Rank* shows the models' rank(from 1/best to 200/worst) when WER/CER is used as sorting key.

Model Type	**WER** ↓	CER ↓	Train Size	Decoding	Rank
FOREIGN-FRA-SSL+CPT+FT	**36.8**	21.6	320	4-gram	**1**
NATIVE -HAT-SSL+Ø +FT	37.4	**21.5**	360 (max)	3-gram	5
FOREIGN-ENG-SSL+CPT+FT	37.5	22.4	320	4-gram	6
FOREIGN-FRA-SSL+Ø +FT	42.5	24.5	360 (max)	3-gram	27
FOREIGN-ENG-SSL+Ø +FT	50.4	29.0	320	3-gram	49

Model Type	WER ↓	**CER** ↓	Train Size	Decoding	Rank
FOREIGN-FRA-SSL+CPT+FT	**38.2**	**17.1**	320	Viterbi	**1**
NATIVE -HAT-SSL+Ø +FT	39.8	17.8	360 (max)	Viterbi	3
FOREIGN-ENG-SSL+CPT+FT	40.3	18.6	360 (max)	Viterbi	6
FOREIGN-FRA-SSL+Ø +FT	46.2	21.7	360 (max)	Viterbi	12
FOREIGN-ENG-SSL+Ø +FT	57.1	26.6	360 (max)	Viterbi	38

Table 2: Comparison between Facebook's MMS (Pratap et al., 2023), our best-performing model (FOREIGN-FRA-SSL+Ø +FT) and a native model (NATIVE -HAT-SSL+Ø +FT). For a fair comparison between models, only Viterbi decoding was used. Note that MMS was pre-trained on the IARPA-Babel data.

Corpus	Model	CER↓
CNCH	FOREIGN-FRA-SSL+CPT+FT	**17.1**
	NATIVE -HAT-SSL+Ø +FT	17.8
	MMS (Pratap et al., 2023)	28.4
Haiti-CMU	FOREIGN-FRA-SSL+CPT+FT	09.5
	NATIVE -HAT-SSL+Ø +FT	11.6
	MMS (Pratap et al., 2023)	**07.9**
IARPA-Babel	FOREIGN-FRA-SSL+CPT+FT	36.6
	NATIVE -HAT-SSL+Ø +FT	38.5
	MMS (Pratap et al., 2023)	**34.6**

160, 80, 40, 20, 10, and 5 minutes. Each train size including the previous sizes (e.g. max $\supseteq$ ... $\supseteq$ 10 $\supseteq$ 5). Each model is fine-tuned for 20k steps[5] with a CTC loss and the best model is selected on the lowest WER on the validation set. To prevent overfitting, the parameters were frozen for the first 10k steps. The text was lower-cased and diacritics removed (due to inconsistent use).

Finally, we also train 2-to-5-gram LMs using KenLM (Heafield, 2011), with default Kneser-Ney discounting parameters. LMs were trained on the transcriptions of the CNCH data set only (hence, preserving our 'fieldwork data'-only setting). We trained a separate LM for each size of the training data set (e.g. a LM trained on train-10 only uses the text corresponding to the transcription of 10 minutes of speech), resulting in 32 different LMs (4 n-gram sizes $\times$ 8 train sizes) that will be used to compare raw (i.e. Viterbi) decoding and LM-rescored decodings.

4 Results & Discussion

Results. We used the SCTK toolkit[6] to compute standard Word Error Rate (WER) and Character Error Rate (CER). Standard Viterbi decoding, and LM rescoring with 2-to-5-gram LMs was used. This resulted in 5 fine-tuned models $\times$ 8 training sizes $\times$ (1 Viterbi + 4 ngram) decoding = 200 decoding strategies. A general overview of our results is shown in Fig.1 (for clarity, only Viterbi, and 5-gram LM rescoring is shown) and the best configuration for each of the 5 model types is shown in Tab. 1. A performance comparison between Facebook's MMS and our models is shown in Tab. 2.

Using Fieldwork Data. Turning back to our original research questions, our results show that **(d)** it is possible to train competitive models on a budget using a single GPU and that **(a)** using fieldwork data to train SSL models of speech is effective. Despite such data being inherently noisy — as opposed to audiobooks or broadcast speech commonly used to train SSL models — the NATIVE -HAT-SSL+Ø Haitian model we trained remained very competitive compared to other approaches. This is particularly interesting in the case of low-resourced languages, such as are most of the French-based Creole languages spoken in the Caribbean (Haitian, Guadeloupean, Saint Lucian, etc.) or in South America (Guianan). This means that *no new data needs to be collected*, but that old tape-recorded fieldwork data, once digitalised, can be repurposed for this matter. This opens an avenue for many languages of the world to have cutting-edge speech processing models at their disposal.

Train From Scratch or Use CPT. Now, turning to whether we should fine-tune SSL models that have been pre-trained from scratch or models pre-trained using a CPT approach **(b)**, our results show that the CPT models show a slight advantage over native models trained from scratch (-1.6 WER points, and -0.7 CER points, Viterbi decoding, using lowest CER as sorting key). However, our result show that **(e)** this advantage is only true when the model used for continued finetuning is *that of the lexifier language* (here, French). This advantage seems to disappear when it is not the case, as

[5]Given how little data we have, the models quickly converge and remain stable and do not evolve after 20k steps, hence this cutoff value.

[6]https://github.com/usnistgov/SCTK

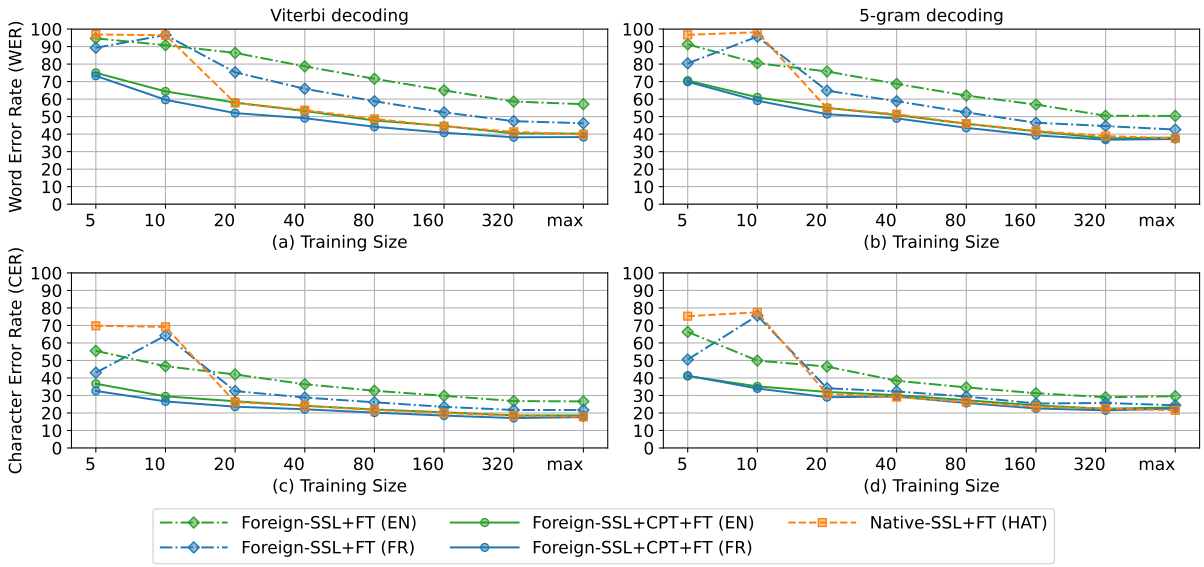


Figure 1: (a, b) Word Error Rate (WER) and (c, d) Character Error Rate (CER, at the character level) of fine-tuned models on an ASR task with Viterbi decoding (left) and with 5-gram LM (right) as a function of the amount of CNCH data used for training (in minutes, from 5 to *max*, where $max = 6.9$ hours, ~ 360 minutes).

the model fine-tuned from another language (here, English) has generally worse performances than either a model fine-tuned from the lexifier language ($+2.1$ WER, $+1.5$ CER, *id.*) or from the native language ($+0.5$ WER, $+0.8$ CER, *id.*). However, what seems most critical is the CPT approach. The ASR models directly fine-tuned from SSL models (FOREIGN-SSL+Ø +FT) that have *not* seen any Haitian speech in a CPT setting lag far behind ($+8$ WER, $+4.6$ CER for the French-based models, *id.*) or very far behind the best model ($+18.9$ WER, $+9.5$ CER for the English-based models, *id.*).

Amount of Fine-tuning Data. Turning to **(c)** and the amount of data necessary to fine-tune SSL models on an ASR task, our results show a marked difference between three groups of models: (i) FOREIGN-SSL+CPT+FT very robust to a reduced amount of training data, (ii) FOREIGN-SSL+Ø +FT not very robust to a reduced amount of data, and (iii) NATIVE -HAT-SSL+Ø showing in between results. Using 20 minutes of data closes the gap between (i) and (iii) while models in group (ii) required approximately 4 times this amount of data (80 minutes) to reach similar performances. We hypothesise that models in group (i) benefit from having seen more speech altogether, as they were pre-trained in their respective language (French or English), have seen Haitian data in the CPT phase, and were further fine-tuned, which could explain why they are more robust.

Viterbi or LM Decoding. Finally, we observed mixed results with the use of LMs for decoding. While they do not significantly improve (nor hurt) the NATIVE -HAT-SSL+Ø +FT or FOREIGN-SSL+CPT+FT models, they significantly improved the WER scores of the FOREIGN-SSL+Ø +FT (Fig. 1a and 1b): e.g. -10 WER with a 5-gram LM for FOREIGN-ENG-SSL+Ø +FT model fine-tuned with 40 minutes of data. Hence, when no pre-training data is available and that foreign models can only be directly fine-tuned, using LM-rescoring is indispensable. However, it seems that using LMs, while improving WER scores, comes at the expense of higher CERs (Fig. 1c and 1d); which hints at the fact that while there are more words accurately transcribed, the others are less well transcribed resulting in a higher CERs.

Comparison with MMS. Tab. 2 shows a comparison of the performances between our models and Facebook's MMS (Pratap et al., 2023) model with the Haitian adapter. To ensure a fair comparison, only Viterbi decoding was used. MMS obtains better scores (-1.6 CER for Haiti-CMU, and -2 CER for IARPA-Babel) compared to our best-performing model FOREIGN-FRA-SSL+CPT+FT (though, the comparison is not entirely fair, as MMS was pre-trained on the IARPA-Babel data). However, both FOREIGN-FRA-SSL+CPT+FT and NATIVE -HAT-SSL+Ø +FT obtain better CERs than MMS on fieldwork data (-11.3 and -10.6 respectively). This shows that our models are very competitive compared to MMS, particularly given

the fact that MMS was pre-trained on 491k hours of data and fine-tuned 44.7k hours of labelled data (including roughly 20 hours of Haitian). In contrast, our models are pre-trained on 340 hours of data and fined-tune on less than 6 hours of data. It also shows that using fieldwork recordings does not hinder zero-shot adaptation to out-of-domain (i.e. non-fieldwork) data, contrary to MMS which performs much worse on out-of-domain fieldwork data.

5 Limitations and Future Work

In this paper, we focused on exploring the validity of using fieldwork data to pre-train self-supervised models to ultimately fine-tune ASR models from them (*ex*trinsic evaluation), but have left aside the study of the pre-trained models and representations themselves (*in*trinsic evaluation). In future works, we wish to use an ABX task (Schatz et al., 2013) to compare the latent representations and their transfer at the phoneme level. This would help us gain more insight into the performances of our models. The data we use for continued pre-trained was collected 40 years ago, and the language between that time and now has evolved (e.g. its phonology, etc.). Hence, the question of the impact of the diachronic shift and how to measure it is open. Finally, our results show that 350 hours of fieldwork recordings is enough to pre-train a native SSL model and obtain competitive results when fine-tuned on an ASR task. Yet, such a treasure trove with as many recording hours might not exist for all languages: the question of the minimal amount of fieldwork data to use is open.

6 Conclusion

In this work, we used 40-years old digitalised tape-recorded fieldwork data in Haitian to train SSL models. We trained a native SSL model, and also used a CPT approach on pre-trained SSL models of the lexifier language (French) and of an unrelated language (English), which we fine-tuned on another data set of fieldwork recordings on an ASR task. We obtained competitive results and showed that the best model is the pre-trained model of the lexifier language with CPT on Haitian fieldwork recordings, followed by the native SSL model, obtaining close results. Hence, when no model of the lexifier language is available, it is still worth training a native model with fieldwork data. Being able to train a native model is all the most important, as a native model might be a matter of self-pride to the speaker community, as opposed to a model derived from the lexifier language, generally that of the former colonising power.

Contrary to the work of (Nowakowski et al., 2023), ours is the first that demonstrates the feasibility of training SSL models using ***only*** fieldwork recordings, and their usability on downstream tasks, such as ASR. This methodology opens an avenue for many languages of the world to have cutting-edge speech-processing models at their disposal, by digitalising recordings collected decades ago. Hence, the 'mobilising the archive'-approach advocated by (Bird, 2020) constitutes a promising way forward.

The best-performing foreign & native models will be made public, along with the scripts used to format the data.

References

Tony Andrus, Aric Bills, Thomas Conners, Erin Smith Crabb, Eyal Dubinski, Jonathan G. Fiscus, Breanna Gillies, Mary Harper, T. J. Hazen, Brook Hefright, Amy Jarrett, Hanh Le, Jessica Ray, Anton Rytting, Wade Shen, Ronnie Silber, and Evelyne Tzoukermann. 2017. Iarpa babel haitian creole language pack iarpa-babel201b-v0.2b.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. pyannote.audio: neural building blocks for speaker diarization. In *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain.

Wojtek Breiter. 2014. Rapid bootstrapping of haitian creole large vocabulary continuous speech recognition.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proc. Interspeech 2021*, pages 2426–2430.

Dominique Fattier. 1998. *Contribution à l'étude de la genèse d'un créole: l'Atlas linguistique d'Haïti, cartes et commentaires, 6 vol.* Bibliographical record, Presses Universitaires du Septentrion, Villeneuve d'Ascq. Ph.D. Dissertation, Université de Provence.

Nuzhah Gooda Sahib-Kaudeer, Baby Gobin-Rahimbux, Bibi Saamiyah Bahsu, and Maryam Farheen Aasiyah Maghoo. 2019. Automatic speech recognition for kreol morisien: A case study for the health domain. In *Speech and Computer*, pages 414–422, Cham. Springer International Publishing.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Marie-Christine Hazael-Massieux. 2012. *Les Créoles à base française*. Editions Ophrys, Gap, France.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Nikolaus P. Himmelmann. 2018. *Meeting the transcription challenge*. University of Hawai'i Press.

Eric Le Ferrand, Claudel Pierre-Louis, Ruoran Dong, Benjamin Lecouteux, Daphné Gonçalves-Teixeira, William N Havard, and Emmanuel Schang. 2023. Outiller la documentation des langues créoles. In *LIFT 2023 : journées scientifiques du GdR Linguistique Informatique, Formelle et de Terrain*, Vandoeuvre-Lès-Nancy, France.

Éric Le Ferrand and Emily Prud'hommeaux. 2024. Automatic transcription of grammaticality judgements for language documentation. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 33–38, St. Julians, Malta. Association for Computational Linguistics.

Cécile Macaire, Didier Schwab, Benjamin Lecouteux, and Emmanuel Schang. 2022. Automatic speech recognition and query by example for creole languages documentation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2512–2520, Dublin, Ireland. Association for Computational Linguistics.

Karol Nowakowski, Michal Ptaszynski, Kyoko Murasaki, and Jagna Nieuważny. 2023. Adaptation of a multilingual speech representation model for a new, underresourced language via multilingual fine-tuning and continued pretraining. *Science Talks*, 8:100249.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Titouan Parcollet, Ha Nguyen, Solene Evain, Marcely Zanon Boito, Adrien Pupier, Salima Mdhaffar, Hang Le, Sina Alisamir, Natalia Tomashenko, Marco Dinarelli, Shucong Zhang, Alexandre Allauzen, Maximin Coavoux, Yannick Esteve, Mickael Rouvier, Jerome Goulian, Benjamin Lecouteux, Francois Portet, Solange Rossato, Fabien Ringeval, Didier Schwab, and Laurent Besacier. 2023. Lebenchmark 2.0: a standardized, replicable and enhanced framework for self-supervised representations of french speech. *Preprint*, arXiv:2309.05472.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages. *arXiv*.

Thomas Schatz, Vijayaditya Peddinti, Francis Bach, Aren Jansen, Hynek Hermansky, and Emmanuel Dupoux. 2013. Evaluating speech features with the minimal-pair ABX task: analysis of the classical MFC/PLP pipeline. In *Proc. Interspeech 2013*, pages 1781–1785.

Gary F Simons and Charles D Fennig, editors. 2023. *Ethnologue: Languages of the world*. Summer Institute of Linguistics, Academic Pub.

Albert Valdman, Anne-José Villeneuve, and Jason F. Siegel. 2015. On the influence of the standard norm of haitian creole on the cap haïtien dialect: Evidence from sociolinguistic variation in the third person singular pronoun. *Journal of Pidgin and Creole Languages*, 30(1):1–43.

Evaluating Indigenous language speech synthesis for education: A participatory design workshop on Ojibwe text-to-speech

Viann Sum Yat Chan, Christopher Hammerly
Department of Linguistics
University of British Columbia
Vancouver, Canada
chan.sum@northeastern.edu
chris.hammerly@ubc.ca

Abstract

This paper reports methods and results from a participatory design workshop aimed at evaluating the use of speech synthesis and text-to-speech for Ojibwe language education. Using an existing text-to-speech feature as a starting point, we worked with two groups of Ojibwe language instructors using a guided trial of the speech synthesis system and a two hour semi-structured workshop with the aim of creating a lesson plan that utilizes text-to-speech. We highlight the insights from this work, both in how to design and deliver speech synthesis systems for Indigenous language education, but also how to approach and design such a workshop to ensure a fruitful discourse.

1 Introduction

Ojibwe is a North American Indigenous language in the Algonquian family known in different regions as Anishinaabemowin, Nishnaabemwin and Ojibwemowin. It is spoken in both the US and Canada, with 25,440 speakers recorded in the 2021 Canadian Census (Statistics Canada, 2023). Colonial policies like the residential school system aimed to force assimilation through means such as reduced use of the language and separation of children from their families (Truth and Reconciliation Commission of Canada, 2015). Because of this, the Ojibwe speaker population is characterized by a high average age of L1 speakers and a parent generation who may understand the language but do not primarily speak it to their children (UNESCO, 2010).

In addition to its effects on language use within families, the lack of L1 speakers in the current parent generation also means many instructors of Ojibwe are as much learners of the language as they are teachers (Engman and Hermes, 2021). Because not all families are able to support students' language learning at home, students rely heavily on their teachers and peers in the classroom to practice the language, thus limiting their exposure to the language in other contexts and environments. Combined with the unique position of teachers as teacher-learners, the task of teaching Ojibwe poses challenges beyond what is typical of second-language learning.

One way to address this issue is through the development of synthetic text-to-speech (TTS) systems which can act as an audio supplement to existing text-based tools like verb conjugators, dictionaries, and phrasebooks (Pine et al., 2024). Currently, there are 70 Indigenous languages spoken throughout Canada, but only a handful of existing TTS systems (e.g. Harrigan et al., 2019; Pine et al., 2022; Conrad, 2020; Hammerly et al., 2023). Low-resource languages face challenges in the development of TTS due to a limited number of fluent speakers and these speakers having limited time to record data for training. Pine et al. (2024) also identifies challenges in the evaluation of Indigenous TTS systems—a small L1 population means there might not be a large enough sample to contribute to a meaningful and generalizable quantitative evaluation of the synthetic speech system. While efforts to create TTS systems have been successful, not much work has been done to investigate how language communities are using these TTS systems, and whether the intended benefits can be enjoyed.

The goal of the current study is therefore to answer the following research questions:

1. What are the strengths and limitations of our existing Ojibwe TTS feature?

2. What are teachers' priorities when approaching new tools in educational technology like TTS?

These questions address the present and the future of developing TTS for Ojibwe and other Indigenous languages. Exploring the strengths and limitations of TTS can help us troubleshoot existing problems, while understanding teachers' priorities when

Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 47–64

using TTS in their teaching can help researchers and developers focus their improvements on the needs of the community. Observing how teachers interact with unfamiliar technology and understanding the strengths of TTS can give researchers and developers insight into what the barriers to usage are currently, and how usage of new technology can be encouraged in the future.

2 The current Ojibwe Text-to-Speech Feature

Hammerly et al. (2023) describes the development of a TTS synthesis system for Border Lakes Ojibwe that is being deployed on the *Anishinaabemodaa* web-based language learning platform produced by teams at the Seven Generations Education Institute, SayItFirst, CultureFoundry, and the University of British Columbia. Only users with "teacher" profiles are given access to the TTS feature, delivered as a standalone webpage independent of the other learning materials on the platform. The webpage (Figure 1) includes a text box for users to input text, a button to generate speech labelled "Speak!", and an audio clip once the "Speak!" button is clicked. Users can play the audio clip on the webpage or download the clip to use in different learning materials by clicking the three dots next to the audio clip to reveal a drop down menu.

As detailed in the Hammerly et al. (2023) paper, this standalone TTS feature was intended for teachers to use to generate audio files that can be sped up, slowed down or downloaded for offline use. It was also planned for teachers to be able generate their own materials and integrate the audio into games like a flashcard activity. Despite this resource being available to teachers, surveys and consultation conducted by CultureFoundry found that teachers were not using this resource, nor were they aware of it. We aim to understand why this feature has not yet seen widespread use on the platform.

3 Participatory Design and Indigenous Research Methods

Pine et al. (2024) highlighted the need for synthetic speech systems to be developed through collaboration with their respective language communities to avoid ethical issues in consent, data collection and usage. To ensure adequate community engagement and consultation in the development of our TTS tool, we seek to use participant-centred research methods to facilitate collaboration between teachers of Ojibwe, researchers and developers.

Participant-centred research methods position participants as the subject matter expert, a role traditionally held by the researcher (Zelenko et al., 2021; Flaskerud and Anderson, 1999). Participatory design (PD) or co-design is most commonly used in human-computer interaction (HCI) research as a way for users of computer technology to participate in its development, with the goal of aligning these tools with the practice and beliefs of the users (Hansen et al., 2019). It is often used to develop educational technology, inviting students, parents and teachers to contribute to the design process (Roschelle et al., 2006; Lin and Van Brummelen, 2021). While co-design focuses on creating and reporting on a tangible finished product, PD often requires the reinterpretation of the design outcomes to understand users' needs and values (Lim et al., 2008). Outcomes of PD can include intangible products like knowledge of current practice, new practices and visions for the future on top of the tangible product or prototype (Hansen et al., 2019).

PD research involving Indigenous communities place a strong emphasis on establishing a warm and welcoming environment for participants, giving participants time to build rapport, begin friendly dialogue, and get to know each other on a personal level (Parsons et al., 2016). This emphasis can be seen in researchers designating a separate workshop session for this purpose (Barcham, 2023), or spending considerable preparation time on building trust before formal data collection begins (Woodward and Marrfurra McTaggart, 2016). While participant-centred research methods have always put the spotlight on participants' voices with minimal input from the researcher (Zelenko et al., 2021), Indigenous co-design practices appear to be characterised by a disproportionately long duration of time dedicated solely to rapport-building, relative to formal data collection. Additionally, Parsons et al. (2016) recommend Indigenous co-design workshops conform to culturally appropriate ways of interacting, incorporate traditional practices in the workshop, and tie research to relevant cultural priorities.

This study is concerned with the evaluation and improvement of an EdTech tool, typical of HCI research, while also understanding the need to be respectful and sensitive of the cultural context surrounding the development of the tool. We aim to combine practices from both HCI and Indigenous

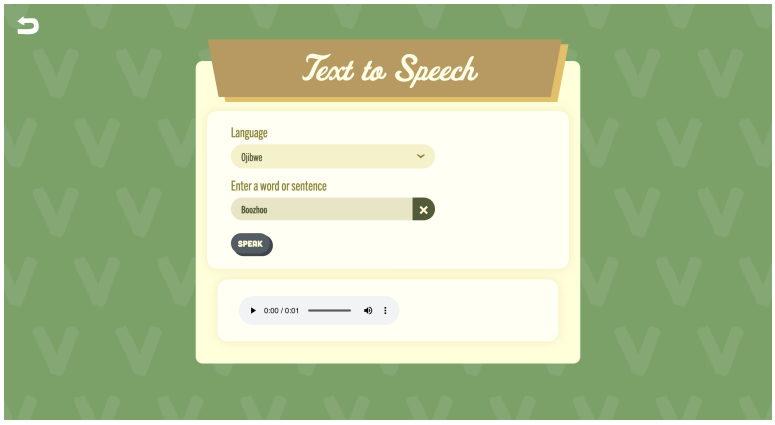


Figure 1: Screenshot of current Ojibwe text-to-speech feature on the *Anishinaabemodaa* platform

PD research to approach the design process as a vehicle for inquiry rather than simply a means to create a tangible end product.

4 Method

A single-session participatory design workshop was conducted with two groups of teachers who use the *Anishinaabemodaa* language learning platform as part of their instruction. Participants first completed a pre-workshop questionnaire and guided trial of the TTS feature on the *Anishinaabemodaa* language learning platform. The workshop involved the creation of a lesson plan that includes the use of the TTS feature and aimed to explore the capabilities and limitations of the feature through active engagement with it. Each workshop session lasted approximately two hours. All procedures were approved by the UBC Office of Research Ethics.

4.1 Participants

All five teachers who participated in the study have had contact with CultureFoundry due to their involvement with the *Anishinaabemodaa* language learning platform and were recruited through CultureFoundry's mailing list. Two workshop sessions were run, first with a group of three, then a group of two. The participants were between the ages 25 and 55 and all participants were female. Their years of experience with Ojibwe ranged from two to 51, while their years of experience teaching Ojibwe ranged from one to eight. Participants were located in Northwestern Ontario, the Greater Toronto Area and Wisconsin. All participants considered themselves learners as well as teachers of the language, and were encouraged to draw from their unique teacher-learner perspective throughout the workshop. The participants were grouped by their availability to participate in the workshop, and the group size was limited to three participants to ensure enough opportunity for everyone's ideas to be heard. Participants were paid CAD $50 per hour for their time.

4.2 Materials

4.2.1 Pre-Workshop Materials

The pre-workshop questionnaire (see Appendix A for the full set of questions) was hosted on Qualtrics and consisted of five parts: (1) a consent form, (2) a demographic questionnaire, (3) a guided trial of the TTS feature, (4) a general user-experience questionnaire and (5) a brainstorm area for participants to write down initial ideas they might have for the lesson planning activity in the workshop. The questionnaire was sent to participants prior to the workshop.

The guided trial of the TTS feature consisted of six tasks, each task prompted participants to enter a different type of text input into the TTS feature and share their first impressions. The types of text input include: (1) one word, (2) one sentence, (3) one paragraph, (4) one question, (5) sentences that convey different emotions, and (6) any other text input they would like to try.

Each task was structured in the same way: the participants were first prompted to try entering one type of text into the TTS feature, then, they were to type out the text input they chose, listen to the synthetic speech output generated by the TTS feature, and rate how accurate, natural and contextually appropriate the speech sounded on four-point Likert scales. In the final two tasks, participants had the option to try out additional phrases, sentences or paragraphs and report their thoughts in more detail

in an open-ended text box. The purpose of this guided trial was to ensure participants had interacted with the TTS feature in a meaningful way, and had an opportunity to discover the capabilities and limitations of the TTS feature on their own terms prior to involvement from other participants and the researchers.

4.2.2 Workshop Materials

The workshop was hosted on Zoom and the Whiteboard feature was used for collaboration between participants. The Whiteboard was set up prior to the session with four different areas (Figure 2), moving from one area to the next as the workshop progresses. The first is a brainstorm area framed by a white square where participants can add "sticky notes" with their ideas on how to incorporate the TTS feature into a lesson or resource. Prior to the workshop, sticky notes with ideas that were previously suggested in the brainstorm portion of the pre-workshop activity were placed onto the brainstorm area on the Whiteboard. The second is the sorting area which included three rectangles labelled "Let's discuss this!", "Maybe discuss these if we have time" and "Save for another day". Participants were expected to move their sticky notes and sort each idea into one of these three boxes. The third area included three examples of built-in templates that can be used for lesson planning. There are many templates to choose from on Zoom Whiteboard, this sample template area was meant to give suggestions but not limit what participants eventually chose to use in lesson planning. The fourth area was the lesson planning area, used to create the lesson plan or resource together to reach a final product.

To guide participants through the introduction and discussions, a PowerPoint presentation with a progress bar was created. The same progress bar was included on the Zoom Whiteboard.

4.3 Workshop Design

The two-hour long workshop sessions were planned as described below, but we were flexible with our approach and did not follow it strictly. Changes in plans are addressed in Section 8, and full details of workshop plans, design, goals and time management are included in Appendix B.

Each workshop started with a Welcome presentation and self-introduction activity to help participants warm up and build rapport. This was followed by a brainstorm task for participants to share their ideas on how to incorporate the TTS feature into a lesson plan. Sorting tasks were planned for participants to parse through these ideas but these tasks were skipped, and participants directly identified one idea to develop further. This led into the design of a full lesson plan from the idea that was chosen and wrapped up with a workshop debrief and reflection.

5 Workshop Products

Through workshop discussions and activities, participants in the two workshop sessions created the following lesson plans to incorporate the TTS feature into their teaching.

Group 1 designed a make-your-own phrasebook activity where students would create their own customizable digital phrasebook. Teachers would model to students how to add new phrases they come across in daily life to the digital phrasebook in text and audio form and encourage usage of this phrasebook outside the classroom. The audio clip would be created with the TTS feature. The full lesson plan and additional ideas from Group 1 are included in Appendix C.

Group 2 designed a make-your-own flashcards activity. Teachers would model to students how to create digital flashcards. Students are meant to listen to audio clips of target words or phrases repeatedly and practice their pronunciation at home. When ready, they can record themselves saying these words and phrases, and embed the audio clips onto the flashcards. This activity can double as an oral assessment. The full lesson plan and additional ideas from Group 2 are included in Appendix D.

6 Synthetic Voice Quality

The pre-workshop questionnaire revealed a number of interesting results. There is a consensus between participants that while the TTS feature does not produce speech that sounds contextually appropriate the pronunciation of specific words and phrases are accurate. Pre-workshop questions on whether participants believed the synthetic speech sounded accurate received 11 responses rated "Strongly Agree", 8 rated "Agree" and 1 rated "Disagree" (See Appendix A for full results). Participant 4 further highlighted in a questionnaire response: "I tried the glottal sounds and a few other different sounds we have that are unique (different from English) [...] and all were pronounced correctly." In regards to the TTS system's ability to differ-

Figure 2: Zoom Whiteboard set up including a progress bar, brainstorm area, sorting area and lesson planning area

entiate between similar sounds, Participant 4 also suggested in a questionnaire response that the TTS feature would be a good tool to demonstrate how misspelling leads to a change in morphemes and results in words that look similar but are different in meaning. They give the example of the first person suffix *-yaan* versus the second person suffix *-yan*, which differ only in vowel length, so are frequently confused. This accuracy makes it possible for students to use the TTS feature as a secondary resource for speaking and listening practice. Students need as many reference points for the language as they can get and it is important for them to "hear a voice other than [their teachers']" (Participant 2). However, because the synthetic speech lacks natural rhythm and tone modulation (Participant 5 on questionnaire), the feature is better used for pronunciation practice than conversation practice.

7 Teachers' Priorities

We identify four priorities based on direct feedback on the TTS feature and language learning platform provided by participants on the pre-workshop Qualtrics questionnaire, workshop discussions, participants' approach to the lesson planning task and additional responses to personal reflection questions.

7.1 Representation

Participants appreciated that the TTS feature and the synthetic speech used across the online learning platform 'can allow students to hear the language from a voice other than [theirs]' (Participant 2) because a lot of their students come from families who do not speak the language at home, remarking that 'even though it is synthetic it does sound spot on' (Participant 4). However, when asked how to make the feature more culturally relevant to its potential users, participants across both workshops suggested the inclusion of different voice options, as there is currently only one voice of a middle-aged male behind the synthetic speech output. Participants highlighted the importance of having a female voice on the feature:

- 'It is important for kids to hear female voices and know that men aren't the only speakers [of Ojibwe], there are great female speakers out there as well.' (P4)
- 'There might be some trauma with men, so if they have a voice they felt comfortable with, that might be [a good] option as well.' (P5)

Choice and autonomy are key to recovery from trauma related to gender-based violence (Elliott et al., 2005), and having the option of a female synthetic voice would support that.

The importance of having a younger voice on the feature was also highlighted:

- 'It would be amazing for young folks to hear the language spoken accurately by a young sounding speaker, not necessarily culturally relevant but definitely more relevant to young people.' (P3)
- 'My kids know on fun days I play TV shows dubbed over in Anishinaabemowin like Spongebob or Scooby Doo, and they always think it's hilarious that the voices are much older than the characters they are portraying. In high school we talk about why that is, and it's obviously a serious concern that so many of our fluent speakers are getting so old.' (P2)
- 'A kid voice would be more engaging, especially since there is only one voice on the language learning platform' (P2)

Finally, participants suggested synthetic speech as a means to preserve the voices of elders, saying, 'We're losing our elders and we will lose their

voices as well' (P5). Participant 5 gave the example of a feature on the Ojibwe People's Dictionary which allows you to choose between voices of different elders when playing recordings of words by clicking on the elder's initials, as a great way to add more voice options and pay tribute to important members of the community.

7.2 Accessibility

Participants' concern with access was three-fold: the TTS feature should be more accessible on the app, the interface should include accessible language and user-friendly buttons, and there was general concern for access to technology in rural areas.

Currently, the TTS feature is only made available to teachers and it takes four clicks to reach the interface from the home page. Furthermore, awareness of the feature among teachers is limited. Participants expressed that it was through this workshop that they first heard of the feature. Just knowing that the feature is available and understanding what it is for would be huge steps in increasing access and usage. Additionally, specific parts of the TTS feature like the download, slow down and speed up functions are hard to locate. Participants appreciated these functions when told about them, but crucially needed to be told explicitly about their existence and where to access them.

When asked to complete a guided TTS trial in their own time prior to the workshop, participants reflected that there was a learning curve and using the TTS feature was not an intuitive experience. There is a button on the TTS interface labelled "Speak!" under the text box to indicate that the user is telling the TTS program to speak, but several participants thought this was an instruction for them to speak to the TTS feature and record their own voice. Participant 2 suggested this label should be replaced with the phrase "Generate Speech" which is more straightforward and tells users exactly what the button does. In trying to avoid technical language or jargon to make tools more user-friendly, the actual meaning of the instruction might be lost and have the opposite effect to the accessibility that word choice was intended to achieve.

Brief interactions with the TTS feature before the workshop already revealed several barriers to access. Participants who worked in rural school districts brought up barriers to access in terms of internet connection and access to a device at home as an additional hurdle. This makes it difficult for students to access the benefits of using the TTS feature at home, such as aiding in independent study and practicing the language in private. Along with the lack of exposure to devices at home comes with an unfamiliarity towards educational technology in general, meaning the learning curve for these students would be steeper than those who have been using all sorts of technology in their learning across different subjects. Certain rural school districts limit access to the online platform to only high school students because the technology is too hard to use, thus widening the gap in access to Ojibwe language learning resources between students in rural and urban school districts.

Because access to internet is an issue, Participants 4 and 5 particularly expressed their appreciation for the download function of the TTS feature, as it can be used to create offline multimodal resources.

7.3 Encouraging Language Usage

Encouraging usage of the Ojibwe language itself as well as the resources for language learning emerged as a priority for teachers. Participant 3 approached her teaching based on the idea that "The only wrong way to speak your language is to not speak it at all." This means getting students to engage with the language as much as possible regardless of how accurate or "good" they are. Participants liked that the TTS feature offers students a chance to practice their pronunciation independently at home by listening to the audio clips and copying the sounds. This is especially key as some students' families do not speak the language, and they rely on their teacher and lesson time to practice interactive language-use.

Another barrier to increasing language-use is students getting self-conscious. Participant 5 offers students the option to take their oral assessments or activities to a private room to complete independently, which does help students feel more comfortable, but might not be conducive to the maximized language exposure needed for effective language acquisition (Matusevych et al., 2017). The TTS feature can help these students gain exposure to the language without opening themselves up to the social anxiety of speaking to a figure of authority like a teacher, elder, or older family member.

In the lesson plans created by both groups, the first part of the lessons involved the teacher directly modelling how to use the technology. This suggests

the first barrier teachers and students face when being introduced to new tools is always the simple question of "How do I use this?" Following initial instruction, participants across both workshops had a plan for encouraging habitual usage of the resource built into their lesson plans. The participants in the first workshop session included a plan to add their phrasebook to the class' daily routine. Participant 3 suggested incorporating this phrasebook into her class' existing word-of-the-day routine—asking students to record these phrases and words in their phrasebook, while also reminding them to use the phrasebook throughout the day. Both the phrasebook and flashcard activities were designed in a way that allows students to continuously add to the resources created, with the goal of helping students build the habit of language learning in their day-to-day lives, outside of school, creating "a living document of [the students'] learning" (P2). These considerations are in line with Indigenous views that learning is "a life-long, self-directed process of experiencing, processing and reshaping existing knowledge," (English, 2008), without the distinction between adult-learning and K-12 education typical of Western conventions. The priority of encouraging language-use is perhaps a reflection of cultural values held by teachers of Ojibwe, as well as a desire to document and revitalize the language.

Encouraging language usage means involving families and community members so students can practice the language in different contexts. Both workshop products included an element of allowing students to take their work home and show their parents as a way to help parents learn the language alongside their children. The phrasebook or flashcards created can be as much a resource for parents as it is for students, and students are encouraged to continually add to these resources outside of school, which can be a bonding activity for families.

Participant 2 also mentioned how other teachers in her school who do not speak or teach Ojibwe have expressed the desire to learn a few words in Ojibwe to use with the students so they can hear the language from more people and in more contexts. Participant 2 suggested that the TTS feature would be a great resource for these teachers to practice and look up the pronunciation of certain words they had forgotten, making it easier for them to be a part of the community. This benefit can also be extended to teacher-learners of Ojibwe who are not completely fluent in the language.

7.4 Inclusion

The inclusive education framework Universal Design for Learning (UDL; CAST 2024) encourages teachers to create multimodal resources that offer multiple means of representation so students with a range of needs can access the same lesson in different ways. For instance, an audio clip next to a chunk of text would help students who have difficulty reading understand the content and having both modalities would be helpful for all L2 learners regardless of their needs. Participant 1 said teachers are "constantly recording [themselves] to create materials" for their classes, and Participant 5 was delighted to find out about the download button, commenting, "I know what I'll be playing with this evening!" The download function of the TTS feature makes something that teachers were already doing more convenient, so it is easy for them to integrate this standalone feature into their existing teaching practices.

Multimodality was heavily considered in the design of the first group's lesson plan, not only in the inclusion of both audio and text, but also in adding cross-curricular elements like having students create a customized background for their phrasebook so the phrasebook feels like their own or a themed background to match the content. Participant 2 suggested that a student interested in basketball phrases can decorate their page with basketball drawings. Participants prioritized offering students a comprehensive learning experience that does not stop at the text and the language itself, and can benefit a range of students who might prefer to learn in different ways.

UDL (CAST, 2024) also calls for multiple means of expression, meaning teachers should offer different assessment pathways for the same content to cater to diverse needs. Participants in the second workshop session highlighted challenges faced by teachers in providing accommodations and modifications for students in a school subject lacking in standard resources and practices, especially since creating custom materials adds to teachers' workload. Efforts to differentiate are often covert, designed so students are unaware of it. For one module, Participant 4 offered three different modes of assessment, one of which was a Kahoot quiz that appeared to be a lighthearted and interactive activity for the whole class, but actually assessed students who struggled with plain text. Such considerations were apparent in this group's lesson

plan, which involved students creating multimodal flashcards with text and audio clips of students' own voice recordings made after practicing pronunciation with the TTS feature. This activity offers teachers the opportunity to assess students orally, while also being a hands-on activity students can enjoy without feeling like they are being assessed.

There is a need to differentiate because a number of students struggle with language learning, even with English. Participants 4 and 5 raised concern about Ojibwe being the harder language, and having to learn it as L2 when students' L1 English abilities are not up to grade level is particularly challenging. For these students, the greatest barrier to using the TTS feature is in the feature's adherence to the standard Double Vowel orthography, which they report is not taught in certain school districts. This indicates a broader problem of literacy in Ojibwe and English, rather than an issue with the TTS feature design per se. However, the TTS itself can be a useful tool for those struggling to read a given text, since students could use the TTS feature to listen while reading along to a passage to aid in their comprehension. Participant 5 wrote in a questionnaire response, "I would think many students who are not strong in English language will have difficulty as they would also not have a strong grasp of Ojibwe words. The words need to be in front of them to be able to type it in properly and be able to identify the word. Without having the properly spelled words in front of you, if you misspell the word, [the TTS feature] does not correct it." Moving forward, one direction for our work could be to integrate a spell-checking mechanism into the TTS input, which could correct deviations from the standard orthography. We could also explore the possibility of expanding to other writing systems in the language such as syllabics.

8 Lessons Learned

8.1 Building trust and rapport is as much a priority as meeting the aims of the study

The emphasis on rapport and trust building in Indigenous participatory design research is reflected in our flexible approach to the workshop design—not intentionally allotting too much time to unstructured chatting in our plans, but allowing conversations to run as long as participants felt comfortable doing so. In line with practices in other EdTech participatory design workshops (Lin and Van Brummelen, 2021), the first workshop session we held started with a formal welcome presentation where the goal of our research was discussed, so participants and researchers were on the same page. Participants listened with their microphones muted throughout the presentation, until they were prompted to introduce themselves. There were questions on a PowerPoint presentation slide providing suggestions to guide their introductions, and the facilitator introduced herself with those same questions first to give participants time to prepare. These questions included a mix of personal, professional and lighthearted, fun questions. For a more reserved group like this one, having these questions might help participants warm up without overwhelming them.

In the second workshop session, despite never having met, the participants dove into an open discussion on the challenges faced by teachers in their communities before the formal welcome presentation. Insights shared in this unstructured time were incredibly valuable, and we believe that unprompted comments gave the best representation teachers' priorities. This went on for over 30 minutes before the workshop started as planned.

The rapport-building portions of these two sessions of the workshop went differently, yet both were beneficial for their respective groups and fit the personalities of the participants. It is important for researchers to anticipate and hold space for both possibilities. Helping participants balance open-endedness and freedom to speak with the pressure of having to come up with new ideas on the spot is the facilitator's job during the workshop and should be heavily considered in workshop plans as well.

In addition to prioritizing rapport-building, Participant 1 reflected that simple yet explicit mention that "This is a safe space," was already helpful in making her feel more comfortable. Encouragement and positive feedback throughout the workshop can also contribute to this welcoming environment, but feedback should be kept non-specific so as to not influence participants' opinions.

The benefits of rapport-building and interactive workshops can be seen in our study. CultureFoundry regularly solicits feedback on the *Anishinaabemodaa* language-learning platform, but this workshop process has helped teachers generate new ideas on how the platform can be improved. The participants in this TTS-focused study had lots of additional ideas for the language-learning platform in general, which suggests the collaborative inquiry

done in this study can be further extended to other EdTech tools and platforms in the future.

8.2 Role reversal in researcher-participant dynamics

The role reversal between researchers and participants in participant-centered research design means researchers can afford to be more flexible in their study design.

Researchers are trained to be precise in their methods, focusing on sensitivity and validity in their experimental design in order to elicit a meaningful outcome in data analysis (Lipsey and Hurley, 2009). Along with this mindset comes research anxiety, referring to how researchers can feel pressured throughout the research process to design methods, collect data and analyze results in a way that is publishable (Cooper et al., 2023).

In the case of the current study, we had discussion questions prepared for the workshop debrief and reflection as a way to guide the conversation towards being relevant to our research questions. However, in the second workshop, the participants started speaking freely, independent of any input from the facilitator, and had already addressed the research questions before the workshop formally started. In addition, we had designed a sorting task to prompt participants to consider and explain their decision-making in greater depth. Both groups opted not to participate in the sorting task and moved straight to choosing an idea to further develop. Because of how willing participants were to share their thoughts and expertise without prompting, the sorting task likely would not have added any more depth to the conversation. In retrospect, the inclusion of this task was intended as a way for researchers to feel confident in the richness of the data collected. Unlike most scientific research, changing the methods and being flexible did not impact the quality of the data in this study. Teachers who are passionate about their work will tell you what their priorities are without prompting.

In research where the participants are consulted for their expertise, researchers should approach the design in an exploratory manner, which might go against their training but will ultimately be rewarding. While helping participants feel as though they can trust the researchers is key to effective collaboration, here, we see how that trust can go both ways.

9 Conclusion

The goal of this study was to understand the priorities of teacher-learners of Ojibwe when approaching new tools in EdTech like this text-to-speech feature, and this was achieved through the pre-workshop activity, the workshop itself, lesson plans generated, and post-workshop interview. Participants cared about how well the synthetic speech represented their community, how easily the feature can be accessed, how they can encourage their students to use the language and the tools available, as well as how the TTS feature can be used to aid in inclusion.

We were also interested in the strengths and limitations of our existing Ojibwe TTS feature for the purpose of improving it for its users. Feedback provided by teacher-learner participants exposed gaps between what developers of the online platform and teachers understand as "accessible". This highlights where more work needs to be done in consultation of teachers or users of new digital tools in order to better serve the community. Additionally, an area in which representation is lacking on the online platform was revealed in our use of a single synthetic voice. However, there is great potential in how the TTS feature can be improved and used. It is possible to use synthetic speech as a document of the different voices in the community. The feature can also be useful in its ability to help teachers create multimodal resources conveniently and involve more of the community in supporting students who are developing their language skills.

Limitations

As is common for research on low-resource languages, there was a limited pool of possible participants who were available to participate in this research and this was reflected in our small sample size of five. This workshop was lengthy and required a time commitment of at least three hours as well as access to a computer and internet connection. This proved difficult for some of our participants but our workshop plans were kept open and flexible in anticipation of these challenges.

Similar workshops in the past often involved multiple sessions, with one independent rapport-building session. For the purpose of this limited project focusing on a simple TTS feature, the choice to do one session was appropriate as we were mindful of challenges in recruitment, but participants did have much to share and the session

could have run for longer if not for time constraints.

The choice to conduct the workshop on Zoom was due to logistical reasons, with participants spread out across Ontario and one participant in Wisconsin, while the researchers were based in Vancouver, BC. As much as possible, this kind of research should be conducted in-person as it would be beneficial for rapport-building and communication. It would also be much easier to run a lesson plan creation workshop with sticky notes, pen, paper and other physical materials than a blank digital space like the Zoom Whiteboard which participants were unfamiliar with and found difficult to use.

Acknowledgements

This work would not be possible without the teachers who so generously shared their time, insight, professional expertise and personal experiences with us, and continue to do so in their work every day. We thank Nozomi Nagashima at CultureFoundry for helping with recruitment and liaising with teachers. We also thank Hope and Danielle for their participation in the pilot session of the workshop. Special thanks to Dongwook Yoon for help in the ideation process for our research methods, for comments and feedback throughout, and to Jian Zhu for helpful feedback as well. Finally, we thank Jason Jones for lending his voice to the Ojibwe text-to-speech feature and developers at CultureFoundry for creating and maintaining this tool.

References

Manuhuia Barcham. 2023. Towards a radically inclusive design–indigenous story-telling as codesign methodology. *CoDesign*, 19(1):1–13.

CAST. 2024. Universal Design for Learning guidelines version 3.0.

Michael Conrad. 2020. Tacotron2 and Cherokee TTS.

Katelyn M Cooper, Sarah L Eddy, and Sara E Brownell. 2023. Research anxiety predicts undergraduates' intentions to pursue scientific research careers. *CBE—Life Sciences Education*, 22(1):ar11.

Denise E Elliott, Paula Bjelajac, Roger D Fallot, Laurie S Markoff, and Beth Glover Reed. 2005. Trauma-informed or trauma-denied: Principles and implementation of trauma-informed services for women. *Journal of community psychology*, 33(4):461–477.

Mel M Engman and Mary Hermes. 2021. Land as interlocutor: A study of Ojibwe learner language in interaction on and with naturally occurring 'materials'. *The Modern Language Journal*, 105(S1):86–105.

Jacquelyn H Flaskerud and Nancy Anderson. 1999. Disseminating the results of participant-focused research. *Journal of Transcultural Nursing*, 10(4):340–349.

Christopher Hammerly, Sonja Fougère, Giancarlo Sierra, Scott Parkhill, Harrison Porteous, and Chad Quinn. 2023. A text-to-speech synthesis system for Border Lakes Ojibwe. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 60–65.

Nicolai Brodersen Hansen, Christian Dindler, Kim Halskov, Ole Sejer Iversen, Claus Bossen, Ditte Amund Basballe, and Ben Schouten. 2019. How participatory design works: mechanisms and effects. In *Proceedings of the 31st Australian conference on human-computer-interaction*, pages 30–41.

Atticus Harrigan, Timothy Mills, and Antti Arppe. 2019. A preliminary Plains Cree speech synthesizer. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1.

Youn-Kyung Lim, Erik Stolterman, and Josh Tenenberg. 2008. The anatomy of prototypes: Prototypes as filters, prototypes as manifestations of design ideas. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 15(2):1–27.

Phoebe Lin and Jessica Van Brummelen. 2021. Engaging teachers to co-design integrated ai curriculum for K-12 classrooms. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–12.

Mark W Lipsey and Sean M Hurley. 2009. Design sensitivity: statistical power for applied experimental research. In *The SAGE handbook of applied social research methods*, pages 44–76. SAGE Publications, Inc.

Yevgen Matusevych, Afra Alishahi, and Ad Backus. 2017. The impact of first and second language exposure on learning second language constructions. *Bilingualism: Language and Cognition*, 20(1):128–149.

Meg Parsons, Karen Fisher, and Johanna Nalau. 2016. Alternative approaches to co-design: insights from indigenous/academic research collaborations. *Current Opinion in Environmental Sustainability*, 20:99–105.

Aidan Pine, Erica Cooper, David Guzmán, Eric Joanis, Anna Kazantseva, Ross Krekoski, Roland Kuhn, Samuel Larkin, Patrick Littell, Delaney Lothian, Akwiratékha Martin, Koren Richmond, Marc Tessier, Cassia Valentini-Botinhao, Dan Wells, and Junichi Yamagishi. 2024. Speech generation for indigenous language education. *Computer Speech Language*.

Aidan Pine, Dan Wells, Nathan Brinklow, Patrick Littell, and Korin Richmond. 2022. Requirements and

motivations of low-resource speech synthesis for language revitalization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7346–7359.

Jeremy Roschelle, William Penuel, and Nicole Shechtman. 2006. Co-design of innovations with teachers: Definition and dynamics. In *Proceedings of ICLS 2006*, volume 2. International Society of the Learning Sciences.

Statistics Canada. 2023. Indigenous languages in Canada, 2021.

Truth and Reconciliation Commission of Canada. 2015. *Honouring the Truth, Reconciling for the Future: Summary of the Final Report of the Truth and Reconciliation Commission of Canada*. House of Commons.

UNESCO. 2010. Atlas of the world's languages in danger. United Nations Education, Scientific and Cultural Organization.

Emma Woodward and Patricia Marrfurra McTaggart. 2016. Transforming cross-cultural water research through trust, participation and place. *Geographical Research*, 54(2):129–142.

Oksana Zelenko, Rafael Gomez, and Nick Kelly. 2021. Research co-design: meaningful collaboration in research. In *How to Be a design academic*, pages 227–244. CRC Press.

A Qualtrics Questionnaire

Welcome to our research study evaluating the quality of an Ojibwe text-to-speech feature. This questionnaire will include some demographic questions, followed by a guided trial of the text-to-speech feature on the *Anishinaabemodaa - Waking Up Ojibwe* language learning platform. Be sure have this questionnaire and the text-to-speech feature open on your screen at the same time so you can follow along. At the end of the questionnaire, you will be asked to give some ideas on how to incorporate the text-to-speech feature into your instruction, or some ways you can use it as a learner. The questionnaire should not take more than an hour. Thank you for your time!

A.1 Demographic Questions

1. What is your current age?
2. At what age did you start learning Ojibwe?
3. How long have you been teaching Ojibwe?
4. Would you consider yourself a learner of the language alongside being an instructor? (Yes/No)

Read each "I can..." statement and think about which answer best describes where you are in your usage of Ojibwe (Likert Scale: "Not Yet", "Rarely", "Sometimes", "Mostly", "Always")

1. I can sound out individual words
2. I can accurately spell individual words
3. I can initiate a conversation and stay on topic
4. I can recognize individual words when listening to elders speak
5. I can understand whole sentences when listening to elders speak
6. I can understand what elders say and I am able to identify the main idea

A.2 Text-to-Speech Guided Trial

The following guided trial of the text-to-speech feature will involve entering five different kinds of text into the text-to-speech system, and evaluating the synthetic speech output. You will be asked to record what you entered into the system and share your impressions of the output. The questions will ask you to input one of each type of text, but you are encouraged to experiment with more than one word, phrase or sentence; be sure to record all of them in the text box. You can use words, phrases and paragraphs from textbooks or any existing media, but feel free to come up with your own ideas and other kinds of text input we have not listed. There will be an opportunity for you to record anything else you have tried at the end of the guided trial.

Question 1a: Try entering one word into the text-to-speech system, write down what you entered in the text box below: [text box]
Question 1b: To what extent do you agree with the following (Likert Scale: Strongly Disagree, Disagree, Agree, Strongly Agree, N/A):

- The word was sounded out accurately
- The tone of voice was contextually appropriate

Question 2a: Try entering one sentence with at least three words into the text-to-speech system, write down what you entered in the text box below: [text box]

Question 2b: To what extent do you agree with the following (Likert Scale: Strongly Disagree, Disagree, Agree, Strongly Agree, N/A):

- The words were sounded out accurately
- The tone of voice was contextually appropriate
- The transitions between words sounded natural

Question 3a: Try entering one paragraph with at least three sentences into the text-to-speech system, write down what you entered in the text box below: [text box]
Question 3b: To what extent do you agree with the following (Likert Scale: Strongly Disagree, Disagree, Agree, Strongly Agree, N/A):

- The words were sounded out accurately
- The tone of voice was contextually appropriate
- The transitions between words sounded natural
- The transitions between sentences sounded natural

Question 4a: Try entering one question into the text-to-speech system, include a question mark in your input, write down what you entered in the text box below: [text box]
Question 4b: To what extent do you agree with the following (Likert Scale: Strongly Disagree, Disagree, Agree, Strongly Agree, N/A):

- The words were sounded out accurately
- The tone of voice was contextually appropriate
- The transitions between words sounded natural

Question 5a: Try entering sentences that convey different emotions into the text-to-speech system, write down all sentences you entered in the text box below: [text box]
Question 5b: To what extent do you agree with the following (Likert Scale: Strongly Disagree, Disagree, Agree, Strongly Agree, N/A):

- The words were sounded out accurately
- The tone of voice was contextually appropriate
- The transitions between words sounded natural

Question 6: Please share anything you found interesting from trying out the different sentences. Did the results meet your expectations? Was there anything you found surprising? [text box]
Question 7a: Feel free to experiment with the text-to-speech feature and come up with new ideas to enter into the system. Write down what you entered in the text box below: [text box]
Question 7b: Please share any interesting observations or insights from your additional experiments: [text box]

A.3 User Experience Questions

To what extent do you agree with the following statements (Likert Scale: Strongly Disagree, Somewhat Disagree, Somewhat Agree, Strongly Agree)?

1. The text-to-speech feature is easily accessible on the platform
2. The text-to-speech feature is easy for a new user to navigate with no prior knowledge of the feature

3. The text-to-speech feature is easy to use
4. The text-to-speech feature is able to generate synthetic speech output in a timely manner
5. I can hear the synthetic speech output clearly
6. I can understand the synthetic speech output clearly
7. Organization of information on the screen is clear and easy to follow
8. The text-to-speech feature can be useful for individuals seeking to improve their general fluency in Ojibwe

Please share any other first impressions from interacting with the text-to-speech feature that you would like to highlight. Were there any results that were unexpected or surprising? [text box]

A.4 Lesson Plan Ideas

We are interested in new and innovative ways to use the text-to-speech feature. Please use the following space to write down between three and five ideas you have on how to incorporate the text-to-speech feature into an existing or new lesson activity, OR how you might use this feature as a learner of the language. You will be expected to share these ideas with other teacher participants during the co-design workshop. [text box]

A.5 Guided Trial Response Summary Table

	Question	Strongly Agree	Agree	Disagree	Strongly Disagree
Enter one word	The word was sounded out accurately	3	1	0	0
	The tone of voice was contextually appropriate	2	2	0	0
Enter one sentence	The words were sounded out accurately	1	3	0	0
	The tone of voice was contextually appropriate	1	2	1	0
	The transitions between words sounded natural	0	4	0	0
Enter one paragraph	The words were sounded out accurately	1	2	1	0
	The tone of voice was contextually appropriate	2	2	0	0
	The transitions between words sounded natural	1	2	1	0
	The transitions between sentences sounded natural	1	3	0	0
Enter one question	The words were sounded out accurately	3	1	0	0
	The tone of voice was contextually appropriate	1	1	2	0
	The transitions between words sounded natural	2	2	0	0
Enter sentences that convey different emotions	The words were sounded out accurately	3	1	0	0
	The tone of voice was contextually appropriate	0	1	3	0
	The transitions between words sounded natural	2	2	0	0
Total	The words were sounded out accurately	11	8	1	0
	The tone of voice was contextually appropriate	6	8	6	0
	The transitions between words sounded natural	5	10	1	0

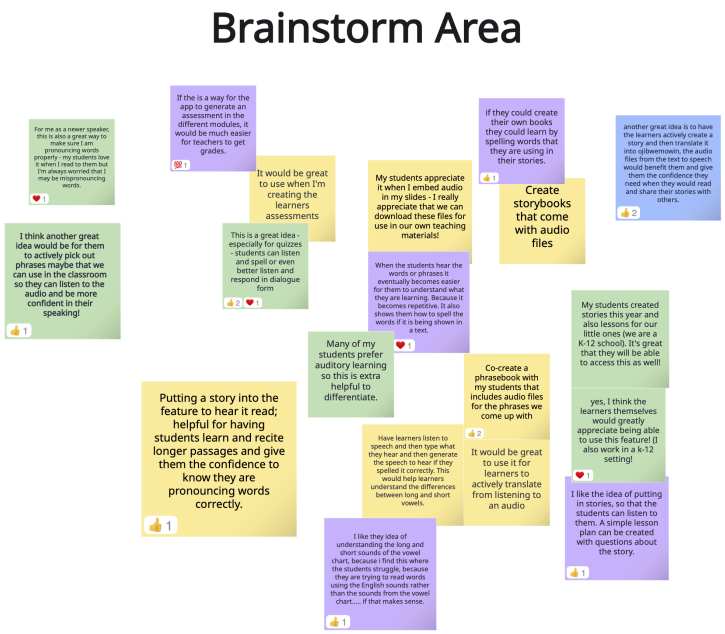
A.6 User Experience Questions Response Summary Table

Question	Strongly Agree	Somewhat Agree	Somewhat Disagree	Strongly Disagree
The text-to-speech feature is easily accessible on the platform	2	0	3	0
The TTS feature is easy for users to navigate without prior knowledge	2	0	3	0
The TTS feature is easy to use	3	0	2	0
The text-to-speech feature is able to generate synthetic speech output in a timely manner	4	0	1	0
I can hear the synthetic speech output clearly	4	0	0	1
I can understand the synthetic speech output clearly	4	0	1	0
Organization of information on the screen is clear and easy to follow	3	0	1	1
The text-to-speech feature can be useful for individuals seeking to improve their general fluency in Ojibwe	5	0	0	0

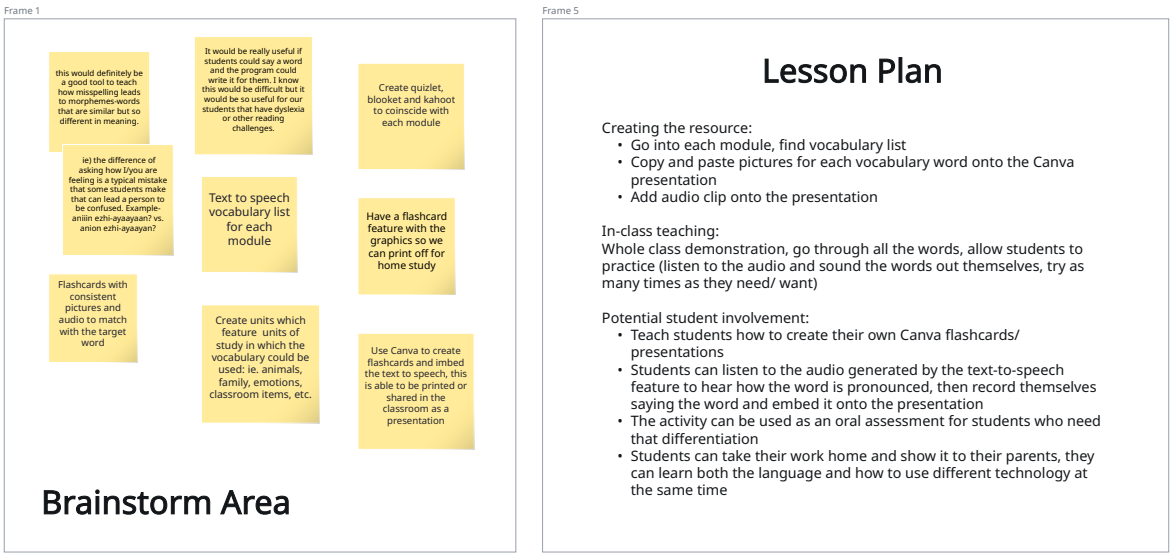
B Workshop Design Details

Activity	Duration	Goal	Additional details
Welcome and introductions	15-20 minutes	Warm up and build rapport	Land acknowledgment; explanation of the research and what to expect during workshop; encourage participants to think out loud and talk through decision making. Participants introduced themselves in Ojibwe, sharing spirit name and connection with the language.
Secondary brainstorm	10-15 minutes	Additional opportunity to share ideas, perhaps ones inspired by others	Facilitator modelled how to add sticky notes to Whiteboard. Participants encouraged to look through existing ideas to further develop or combine them, and add brand new ideas to populate the area with more sticky notes. Participants also interacted with each other through sticky notes
Sorting	5-10 minutes	Consider the factors that inform their decision making when using new technology in teaching	Participants prompted to sort sticky notes into three different areas on the Whiteboard labelled "Let's discuss this!", "Maybe discuss if we have time" and "Save for another day". Participants in both workshop sessions did not participate in this activity.
Choosing an idea from "Let's discuss this!"	5-10 minutes	Consider factors that inform decision making when using new technology in teaching	Participants asked to choose an idea out of the ones sorted into "Let's discuss this!" to create a lesson plan out of. Participants in both workshop sessions chose an idea directly from the brainstorm area.
Develop lesson plan	15-20 minutes	Reveal teachers' priorities in applying TTS to pedagogical contexts	A blank space was set up on Zoom Whiteboard for participants to take notes and create lesson plan. They were given as much time as needed to collaborate. The facilitator supported participants in using the technology when needed.
Workshop debrief	10-15 minutes	Gain extra feedback on the TTS feature in an applied context	Questions: (1) How feasible is it to run this lesson/ activity? (2) What are some possible logistical barriers you might run into? (3) What do you hope students will gain from this lesson/ activity? (4) Do you think students would enjoy the lesson?
Personal reflection	10-15 minutes	Reflect on workshop, bring specific personal and professional experiences into the conversation, additional opportunity for feedback on the TTS feature	Questions: (1) How did you find the workshop? What did you learn? (2) What are some unexpected challenges you came across during the planning process? (3) Is there anything you would do differently if you participated in this workshop again? (4) Does your perspective on the learning platform and TTS feature change when you consider different parts of your identity? (5) How can TTS be made culturally relevant to you? (6) What would you like to see us change, improve or build for the app and the TTS feature?

C Group 1 Workshop Product



D Group 2 Workshop Product



Zero-Shot Query Generation for Approximate Search Algorithm Evaluation

Aidan Pine[1], David Huggins-Daines[2], Carmen Leeming[2], Patrick Littell[1], Timothy Montler[3], Heather Souter[4], Mark Turin[5]

[1]National Research Council Canada, [2]Independent Researcher, [3]University of North Texas, [4]University of Winnipeg, [5]University of British Columbia

Correspondence: aidan.pine@nrc-cnrc.gc.ca

Abstract

Approximate search is a valuable component of online dictionaries for learners, allowing them to find words even when they have not fully mastered the orthography or cannot reliably perceive phonemic differences in the language. However, evaluating the performance of different approximate search algorithms remains difficult in the absence of real user queries. We detail several methods for generating synthetic queries representing various user personas. We then compare the performance of several search algorithms on both real and synthetic queries in two Indigenous languages, SENĆOŦEN and Michif, that are phonologically and morphologically very different from English.

1 Introduction

Online dictionaries are one of the most commonly used and important tools in language revitalization and reclamation programs (Anderson, 2020; Leavitt, 2023; Lyon et al., 2023). For under-resourced languages, online dictionaries are very often the only lexical resource available to learners in a community where no print dictionary has ever been compiled or published. For authoritative monolingual dictionaries, such as the Oxford English Dictionary, users are assumed to be fluent and literate in the language of the dictionary. The same expectations of users of bilingual dictionaries and phrasebooks in language revitalization contexts cannot be made. Users of bilingual dictionaries are often learners, and trying to harness the power of an online dictionary can present learners with an unwelcome paradox: they may wish to look up a word in the dictionary in order to learn it and/or verify the spelling, but in order to look it up in a dictionary with only an exact-match search algorithm, they already need to know exactly how to spell it. This can lead to a Catch-22, particularly with complex writing systems for which keyboard input systems are less standardized or easily available. For these reasons, it is extremely important in a language learning context that users can benefit from fuzzy search algorithms that accommodate anticipated errors or idiosyncratic spellings.

Despite the importance of online dictionaries in language revitalization, they remain resource-intensive endeavours that are often the first project that communities and scholars start and the last one to be completed (Sear and Turin, 2021; Schreyer and Turin, 2023). Compiling lexicographic data, let alone managing and maintaining software, websites and mobile apps all present significant technical hurdles (Trotter et al., 2023). On top of these requirements, building a language-specific approximate search algorithm is also a significant challenge. In some cases, language models already exist for the language in question and can be applied to provide morphologically-aware search results (Johnson et al., 2013; Arppe et al., 2021). Alternatively, Littell et al. (2017) describe software that allows users to define language specific phonologically-aware approximate search algorithms. Originally published under the name Waldayu, the software was generalized and renamed 'Mother Tongues Dictionaries' (MTD) in 2018. MTD is a Python library and collection of visualization frameworks that, given the MTD data specification, allows users to create online dictionaries (web, Android, iOS) from a potentially heterogeneous set of data (i.e., a spreadsheet, JSON file, and XML file). In addition to the data wrangling and visualization capabilities of the library, it also allows users to customize an approximate search algorithm based on weighted or unweighted edit distances. MTD has been used to develop online dictionaries for dozens of Indigenous languages around the world including in Canada, the US, and Japan.

Since 2017, the MTD search algorithm has been updated to allow multi-word, multi-field search,

Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 65–73

and to also include a multi-field variant of the BM25 ranking algorithm (Zaragoza et al., 2004) as a secondary score in addition to edit distance. Impressionistically, and through gathering informal user feedback, we believe that the changes to the MTD search algorithm have led to improved search results, although this has not been formally investigated. Part of the difficulty in evaluating approximate search is that a corpus of common misspellings or otherwise plausible queries does not exist for the Indigenous languages that we are working with. In this paper, we demonstrate a variety of techniques for generating plausible queries that can be applied to other written and unwritten languages. We then apply these techniques to dictionary data in the SENĆOŦEN and Michif languages and show that the recent updates to the MTD search algorithm provide improvements for each type of query generation strategy that we test. We believe the query generation techniques that we describe could be applied to other languages and used in other contexts to help evaluate approximate search algorithms.

2 Methodology

2.1 Data

To investigate a variety of approximate search algorithms, we apply our proposed query generation techniques to two lexical resources from two different languages; SENĆOŦEN and Michif.

2.1.1 SENĆOŦEN Dictionary

The SENĆOŦEN language is a Salish language spoken traditionally in the territories of the W̱SÁNEĆ people. Contemporary revitalization efforts were catalyzed by the late Dave Elliott Sr., who developed the SENĆOŦEN orthography which is still the standard used by the community. The SENĆOŦEN dictionary (Montler, 2018) is the largest lexicographic resource available for the SENĆOŦEN language, and contains over 30 000 words and example sentences. From an approximate search perspective, the language is particularly challenging given its rich and complex phoneme inventory. The language has 36 different consonants with phonemic contrasts between velar and uvular consonants as well as rounding and glottalization. These contrasts are all represented in the orthography, often only with small diacritical changes to indicate them, as illustrated in Table 1.

K̲ /q/	Ḱ /q^{w}/	K /q'/
K̓ /q^{w}'/	Q /k^{w}'/	Ȼ /k^{w}/

Table 1: A subset of the SENĆOŦEN consonant inventory illustrating how uvular/velar, rounding, and glottalization contrasts are encoded in the orthography. This phonological richness presents a challenge for learners when searching in the dictionary.

2.1.2 Michif Dictionary

Southern Michif is one of three language varieties spoken by the Métis (Bakker, 1997; Sammons, 2019). It is a contact language combining elements from Algonquian languages—Plains Cree and the Saulteaux dialect of Ojibwe—with Métis French. Traditionally, it has been written with a mixture of English and French spelling conventions, notably as seen in the Turtle Mountain Dictionary (Laverdure et al., 1983) and its recent digital version (Souter et al., 2024b). More recently there has been an effort to further develop and use the Southern Michif Learners Orthography which is based on a double-vowel system similar to that used for Ojibwe. It has its roots in the work done initially by the late Rita Flamand of Camperville, Manitoba with later input from Robert Papen. Further refinement was carried out by a number of learners. And, after reflection on the early work of Ida Rose Allard, a decision was made to use one special symbol, ñ, to mark nasalization of preceding vowels in order to help support accurate pronunciation.

The Southern Michif for Learners website (Souter et al., 2024a) includes an extensive set of illustrated phrases and words with audio recordings, used with permission here.

2.2 Evaluation

To evaluate the performance of our search strategies we employ the use of mean reciprocal rank (MRR). MRR is a measure for evaluating the order of results given a query. Concretely for our use case, if we type a query and the dictionary entry we intended to find is ranked first in the search results then it has a reciprocal rank of 1 (which denotes the best possible score); however if the expected entry does not rank at all in the search results then the reciprocal rank is 0 (which denotes the worst possible score). If the expected dictionary entry appears as the second result, it would have a recip-

rocal rank of $\frac{1}{2}$ (and $\frac{1}{3}$ if it appeared as the third result, etc). The mean reciprocal rank is then the mean of each reciprocal rank for each query that we evaluate.

More formally, we calculate MRR as $\frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{r_i}$ where Q is a set of queries and r_i is the rank of the expected entry from the dictionary for the *i*-th query. So, in order to calculate this metric, we need a set of queries Q and a corresponding set of expected dictionary entries E.

One can imagine obtaining Q from actual queries logged from an online dictionary - the expected dictionary entries E could also then be obtained by asking users to select the entry they were looking for if it appeared in their search results. This approach is somewhat noisy though, and more importantly, is incompatible with the privacy terms of the dictionaries we use. As a rule, we do not log open user input since even when this information is anonymized, there is no guarantee that users will not input de-anonymized or sensitive search terms (e.g., searching for one's own name or other identifying features). Furthermore, these dictionaries often operate entirely offline which would complicate our ability to record these results.

Instead, this paper presents eight methods for *approximating* user queries given a set of dictionary entries. Since some of our methods are time-consuming, we limit our evaluation to a randomly-sampled 50-word subset from each dictionary. That is, we apply each of the eight methods discussed in the following section to a randomly sampled subset of the dictionaries. We then use mean reciprocal rank to evaluate how robust the search algorithms discussed in §3 are with respect to the approximated user queries.

Our approach here has the simplifying assumption that there is only one expected entry for any given generated query. In reality, there are often multiple relevant entries given a query, for example morphologically related words or matches found in other fields related to the main entry (e.g., an example sentence). If we were able to accurately identify all relevant entries in the dictionary for a particular query, we might instead have considered evaluating using mean average precision.

2.3 Generating queries

To help guide the creation of our query generation functions, we borrow descriptions of likely users from Littell et al. (2017). The sections 2.3.1 to 2.3.4 describe a variety of different types of users and our corresponding query generation techniques. To further approximate the types of queries made by a learner, we consider additional approaches to query generation in sections 2.3.5 to 2.3.8.

2.3.1 Users who can distinguish phonemes but do not always know the orthographic conventions

In order to generate queries for this category of user, the first and second authors hand-transcribed words in the target language by listening to audio of those words. Both transcribers were familiar with the sound systems of the languages they were transcribing, but were not speakers, and had not had any instruction in the language or its writing system. They simply listened to the audio with headphones and transcribed what they heard using the (non-IPA) keyboard available to them. Both transcribers are first-language English speakers who also speak French and have formal training in linguistics.

2.3.2 Users who know the orthography, but cannot reliably discern certain phonemes

We approach query generation for this category of user as a data corruption task. We target specific classes of graphemes and phonemes that we expect to be challenging for our users to distinguish (i.e., the velar/uvular contrast in SENĆOŦEN). We then randomly corrupt up to $N = 3$ of these phonemes' related graphemes with another confusable from the same class, for example swapping out Ḱ (/qʷ/) for Ḳ (/qʷ'/) in SENĆOŦEN or ñ (indicating nasalization on the preceding vowel) for n (/n/) in Michif.

2.3.3 Users without access to a keyboard

We assume this category of user to be able to accurately identify and discern phonemes in the target language, and to also be familiar with the target language's writing system, but to be unable to type due to the unavailability of a Unicode input system on their device. This is less of an issue for Michif, but for SENĆOŦEN there are many specialized Unicode characters and diacritics used in the writing system, and typing in the language requires installing a language-specific keyboard (Chase and Borland, 2022).

To approximate the type of user queries expected when such a keyboard is not installed, we transform each non-ASCII character to its closest ASCII

equivalent. We do this by performing NFD Unicode normalization, removing any diacritics in the range U+0300 to U+036F, and then applying the Unidecode[1] library to the resulting text.

2.3.4 Users who know an alternative orthography

For this class of user queries, we generate queries for Michif in the Turtle Mountain Dictionary (TMD) orthography, an alternate orthography to the one used in the Michif dictionary in this study. SENĆOTEN has historically had multiple orthographies, including an Americanist phonetic representation and the Bouchard practical orthography (see Turner and Hebda (2012, p. 155)). These orthographies are not in standard use by the community and while they might still be used in some queries of the SENĆOTEN dictionary, it would be uncommon, and we do not include them.

2.3.5 IPA-based query generation

The vast majority of users of the SENĆOTEN dictionary are first-language English speakers. This is similar for Michif except in some cases users might speak French as a first language. This query generation technique seeks to approximate a query by mapping it through the International Phonetic Alphabet (IPA) to a query language such as English.

First, we use a rule-based grapheme-to-phoneme (G2P) library (Pine et al., 2022) to derive the IPA pronunciation form of a given word in the dictionary. We then use PanPhon (Mortensen et al., 2016) to map from the IPA symbols in the target language, to the closest English IPA equivalents. For Michif, we also map to the closest French IPA equivalents for comparison, since speakers of that language are more likely to speak French as a first language.

Finally, we train two sequence-to-sequence Transformer based models using the DeepPhonemizer[2] software. We train an English system using a reversed IPA version of the CMU pronunciation dictionary[3] to predict IPA from English graphemes, and we train a French system in the same way using the WikiPronunciation dictionary[4]. In both cases we keep the default hyperparameter settings and train until convergence (140k steps for English with a 12% character error rate (CER), and 1700k steps for French with a 10% CER).

We release our English[5] and French[6] phoneme-to-grapheme models publicly. For generating plausible English or French queries from another language, a method of turning graphemes into phonemes in the target language would be required. For generating plausible queries in a language other than English or French, a similar phoneme-to-grapheme model in that language would also have to be trained.

2.3.6 LLM-based query generation

We also consider the use of Large Language Models (LLM) as naive transcribers of the languages in question. We use ollama and the publicly available 'llama3' model. For the Michif prompts, we ask the question 'The following is a list transcriptions of words in the Michif language. How would you write these words using only the English or French orthography?' and for SENĆOTEN we prompt it to only write the words using the English orthography. We then provide the list of IPA transcriptions of Michif or SENĆOTEN words and record the results. Like the previously described method, adapting this approach for other languages would also require providing the LLM with a pronunciation form of the words in question.

2.3.7 Audio-based query approximation (ASR)

Instead of approximating user queries through a transformation of the original text, we also consider approximating user queries by decoding the original audio using automatic speech recognition (ASR) models. To mimic how a user of the dictionary 'with English ears' might transcribe a word, we decode audio corresponding to a given query with the pre-trained wav2vec2-base-960h model[7]. Importantly, we use a greedy decoder that is not constrained by a language model, so the model will decode the audio into characters in the English orthography, but will allow for non-English words to be decoded. While the Michif dictionary had audio available for each of the 50 words that were sampled, the SENĆOTEN did not, so we synthesized the audio using the SENĆOTEN speech synthesis model described in Pine et al. (2025).

[1] https://pypi.org/project/Unidecode/
[2] https://github.com/as-ideas/DeepPhonemizer
[3] https://github.com/open-dict-data/ipa-dict
[4] https://github.com/DanielSWolf/wiki-pronunciation-dict
[5] https://bit.ly/eng-p2g-model
[6] https://bit.ly/fra-p2g-model
[7] https://huggingface.co/facebook/wav2vec2-base-960h

2.3.8 Teacher-curated queries

It is difficult to draw any conclusions from artificially generated queries. Part of the problem is that for each word in the dictionary, there are many ways to misspell it. So, we cannot know if the ways our query generation techniques have misspelled these terms are similar to the way the target audience of these dictionaries will misspell them.

To help corroborate the results seen among our query generation techniques, the third, fifth and sixth authors, who have experience in teaching Michif and SENĆOŦEN and are familiar with common misspellings from students, compiled a list of common misspellings for each of the words in the 50-word subsets of the Michif and SENĆOŦEN dictionaries.

2.4 Examples and CERs of each technique

In Table 2 we show the result of applying each method to one of the words in each dictionary.

Query Type	Michif	SENĆOŦEN
Original	pashikook	NEW̱SPEŦTENEK̲
IPA	pʌʃɪkoːk	nəxʷspəs̟tə́nəq
Human (§2.3.1)	peshkop	nuhuhspahstanak
Phon. (§2.3.2)	pawshiihkok	NEW̱SPEŦTENEK̶
ASCII (§2.3.3)	pashikook	NEWSPETTENEK
P2Eng (§2.3.5)	puchikouk	neckspothtinick
P2Fra (§2.3.5)	péchecauque	nekspestenek
LLM (§2.3.6)	Pashikotak	Nexwspetheniq
ASR (§2.3.7)	PUSHCOG	NOSPASTANA
Teacher (§2.3.8)	pashikohk	NEW̱SPESTENEK̲

Table 2: An example of how a sample word in each dictionary is transformed by each query generation method. Each example here is the raw output from each query generation method (i.e., prior to case normalization).

As mentioned in §2.2, to generate our test set, we randomly sample a 50-word subset from each dictionary. We then apply each proposed query generation method to the 50-word test set. In Table 3 we report the character error rate (CER) between the generated queries and the original terms. Note that this is not an evaluation of the query generation technique (which we cannot do without data of actual misspelled words and a model of their distribution), rather it is just meant to be an indication of how much the generated queries deviate from the original terms. A higher CER indicates an increased difficulty for the task of approximate search, but not necessarily a less valid or less plausible query.

Across the board, our query generation techniques incurred higher CERs in SENĆOŦEN than they did for Michif. For example, the ASCII query generation technique (§2.3.3) incurs a 30% CER for SENĆOŦEN but only a 1% CER for Michif. In other words, for SENĆOŦEN, non-ASCII characters make up 30% of the characters in our 50-word set, whereas they only make up 1% of the characters in our set for Michif.

Query Type	Michif	SENĆOŦEN
Human (§2.3.1)	0.37	1.38
Phon. (§2.3.2)	0.52	0.27
ASCII (§2.3.3)	0.01	0.30
P2Eng (§2.3.5)	0.43	1.18
LLM (§2.3.6)	0.34	1.01
ASR (§2.3.7)	0.59	0.81
Teacher (§2.3.8)	0.40	0.29

Table 3: Query generation Methods and their Character Error Rates (CER). CERs in terms of the character edit distance between the words generated by the query generation method, and the terms in the dictionary they are meant to approximate.

2.5 Adapting to other languages

Beyond evaluating the recent changes to the MTD search algorithm, part of the goal of this paper is to provide query generation techniques that can be applied to languages other than SENĆOŦEN and Michif. Table 4 shows the data or models required to implement each technique, since some techniques require only audio and some techniques require text, or an available grapheme-to-phoneme library for the language in question.

Query Type	G2P	Audio	Text
Human (§2.3.1)	✗	✓	✗
Phon. (§2.3.2)	✗	✗	✓
ASCII (§2.3.3)	✗	✗	✓
P2Eng (§2.3.5)	✓	✗	✓
LLM (§2.3.6)	✓	✗	✓
ASR (§2.3.7)	✗	✓	✗

Table 4: Query generation Methods and their requirements. ‘G2P’ indicates that the method requires a grapheme-to-phoneme engine to be adapted to a new language.

3 Search Algorithms

Following Littell et al. (2017) we compare results using both an unweighted Levenshtein edit distance $ULev$ and a weighted Levenshtein edit distance $WLev$. The unweighted Levenshtein edit distance

between two strings X and Y is equal to the number of single-character edits (additions, deletions, substitutions) required to change X into Y. By comparison, the weighted Levenshtein edit distance allows edits to be weighted differently, for example allowing substitutions involving commonly confused characters to accrue a lesser penalty. We used the hand-written substitution weights that have been in use for the dictionaries already.

In addition to ranking results based on edit distance, the most recent version of the MTD search engine also applies a secondary ranking based on a weighted multi-field variant of BM25 (Zaragoza et al., 2004); a language agnostic ranking function based on the inverse document frequency of the query. Therefore, in addition to evaluating the difference between weighted and unweighted edit distance, we also report the effect of including BM25 as a secondary score. Although MTD is capable of handling multi-word queries and indexing multiple fields, for the purposes of this evaluation we limit ourselves to single word queries and only search based on a single field in the dictionary entries. The MTD search engine also allows for optional stemming when creating the inverted index used in searching, as well as some basic normalization functions including case and Unicode normalization and the removal of punctuation. These configurations result in the same normalization processes being applied to each term in the inverted index and to each query. For the purposes of this paper we do not configure a stemmer, but we do apply both case and Unicode normalization to all of the queries and to each term in the inverted indices built by MTD.

4 Results

To evaluate the approximate search algorithms described in §3, we randomly sample 50 words from each of the dictionaries. We then apply each of our query generation techniques to the random 50-word sample sets for both languages and compute the mean reciprocal rank (MRR) for the queries generated by each technique. We present our results in Table 6 on the following page.

As expected, given the wide range of CERs for our various query generation techniques, there is also a wide range of results and the relationship between CER and MRR appears roughly inverse. For example the P2Eng (§2.3.5) technique, which had a CER of 1.18, only receives a MRR of 0.07 in the best system for SENĆOŦEN while the ASCII system for Michif had a 0.01 CER and resulted in a MRR of 0.96 in the best systems.

The addition of BM25 results in MRR improvements across all query generation strategies for both weighted and unweighted edit distance. We also see improvements to the MRR when BM25 is included for unmodified queries. That is, when we pass the original word unchanged as the query, we see improvements of +0.09 MRR for SENĆOŦEN and +0.15 MRR for Michif as well as improvements among all query generation techniques. We believe that this is sufficient for justifying the use of an approximate search strategy that is combined with BM25, like the one found in MTD.

To weight or not to weight The difference between weighted and unweighted edit distance is less clear than the improvements seen with the addition of BM25. In Table 5 we compare the results when prompting the LLM to produce either English or French outputs, as well as mapping through English or French pronunciation forms for the phoneme-to-grapheme based technique (§2.3.5). Unexpectedly, the results from the English LLM and P2Eng methods do not seem to show a strong difference between weighted and unweighted edit distances whereas we see a stronger improvement for the generated 'French' queries using an unweighted edit distance. Since the Michif dictionary substitution weights were written by a first-language English speaker who works primarily with English-speaking students, the pattern that we see here could be the result of linguistic bias in the substitution weights, which could be either desirable or undesirable depending on the target audience for the dictionary. In this case, it is possible that the weights are resulting in worse performance for French-influenced queries, since the weights were created with an English speaking audience

	MRR ↑	
Query Type	MTD_w	MTD_u
P2Eng	0.39	0.41
P2Fra	0.18	**0.38**
LLM Eng	0.66	0.65
LLM Fra	0.30	**0.39**

Table 5: Mean Reciprocal Ranks (MRR) for Michif IPA-based (§2.3.5) and LLM (§2.3.6) query generation with both English and French outputs. CER denotes the Character Error Rate for the 50 word set for each particular query generation technique.

Query Type	Language	CER	MRR ↑ $ULev$	$WLev$	MTD_w	MTD_u
Original Text	SENĆOŦEN	0.0	0.91	0.91	1.0	1.0
Phon. (§2.3.2)	SENĆOŦEN	0.27	0.63	0.54	0.58	**0.68**
Teacher (§2.3.8)	SENĆOŦEN	0.29	0.09	0.09	0.11	0.12
ASCII (§2.3.3)	SENĆOŦEN	0.30	0.35	0.36	**0.45**	0.42
ASR (§2.3.7)	SENĆOŦEN	0.81	0.01	0.03	0.04	0.03
LLM (§2.3.6)	SENĆOŦEN	1.01	0.06	0.10	**0.14**	0.11
P2Eng (§2.3.5)	SENĆOŦEN	1.18	0.02	0.04	0.06	0.07
Human (§2.3.1)	SENĆOŦEN	1.38	0.0	0.01	0.02	0.0
Original Text	Michif	0.0	0.80	0.81	0.96	0.96
Phon. (§2.3.2)	Michif	0.52	0.33	0.33	0.41	**0.44**
Teacher (§2.3.8)	Michif	0.40	0.47	0.48	0.61	0.60
ASCII (§2.3.3)	Michif	0.01	0.79	0.79	0.96	0.96
ASR (§2.3.7)	Michif	0.59	0.12	0.12	0.15	**0.20**
LLM (§2.3.6)	Michif	0.34	0.46	0.52	0.66	0.65
P2Eng (§2.3.5)	Michif	0.43	0.25	0.26	0.39	0.41
Human (§2.3.1)	Michif	0.37	0.43	0.42	0.50	**0.55**
TMD Queries (§2.3.4)	Michif	0.79	0.18	0.26	**0.30**	0.25

Table 6: Mean Reciprocal Ranks (MRR) for different query generation techniques given 50 randomly sampled words from the SENĆOŦEN and Michif dictionaries. CER denotes the Character Error Rate for the 50 word set for each particular query generation technique. MTD indicates the search strategy used by Mother Tongues Dictionaries ranks results based on edit distance and a secondary BM25 score. A higher MRR for a particular search strategy indicates that it is more robust to that type of query.

in mind. Ultimately, we believe that the decision of whether to use substitution weights should depend on how well the target audience is known in advance. In most cases though, given the time and expertise required to create custom substitution weights for each language, unweighted edit distance is likely sufficient.

5 Conclusion

In this paper, we have proposed and developed a variety of methods for approximating user queries. We provide guidance and release models so that they might be adapted to other languages. Using the described query generation techniques, we compared the effectiveness of a variety of approximate search algorithms in both SENĆOŦEN and Michif dictionaries. We showed that fuzzy search can be improved by combining BM25 as a secondary score with Levenshtein edit distance. Despite these improvements, and the relatively successful results for Michif, approximate search remains a difficult problem, particularly for languages with large phoneme inventories like SENĆOŦEN.

Future work should compare the queries generated using our described techniques with actual misspellings, for example using corpora like the ones described in Max and Wisniewski (2010) and Flor et al. (2019). Additional future work could also more thoroughly explore the difference between weighted and unweighted edit distances for example by applying the methods described here to more languages, or by devising techniques for learning optimal edit distance weights from data. The latter approach would require a corpus of real or artificial misspelled data, as well as careful evaluation to avoid over-fitting to the training data.

Additional future work might also consider morphologically-aware query generation and approximate search algorithms, for example comparing the FST-based morphologically-aware approaches of Johnson et al. (2013) with the phonologically motivated techniques described here. We expect that languages with higher degrees of polysynthesis might in turn require search algorithms with greater morphological awareness, but it is not clear at what point the benefits of morphologically aware search would be large enough to motivate the additional effort compared to, for example, a simple unweighted edit distance in combination with a secondary BM25 score.

Acknowledgments

This work would not have been possible without the support from our collaborators at the W̱SÁNEĆ School Board, PENÁĆ, SX̱EDŦELISIYE, and Tye Swallow. We would also like to thank Delaney Lothian, Maria Ryskina, and Roland Kuhn for proof-reading and assistance formatting and typesetting this document.

References

Patricia Anderson. 2020. *Revitalization Lexicography: The Making of the New Tunica Dictionary*. University of Arizona Press, Tuscon.

Antti Arppe, Jolene Poulin, Eddie Antonio Santos, Andrew Neitsch, Atticus Harrigan, Katherine Schmirler, Daniel Hieber, Ansh Dubey, and Arok Wolvengrey. 2021. Towards a morphologically intelligent and user-friendly on-line dictionary of Plains Cree—next next round.

Peter Bakker. 1997. *A Language of Our Own : The Genesis of Michif, the Mixed Cree-French Language of the Canadian Metis*. Oxford University Press, Oxford & New York.

Bridget Chase and Kyra Borland. 2022. Networks of support: How online resources are built, maintained, and adapted for community language revitalization needs at firstvoices. *Language Documentation & Conservation*, pages 209–227.

Michael Flor, Michael Fried, and Alla Rozovskaya. 2019. A benchmark corpus of English misspellings and a minimally-supervised model for spelling correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 76–86, Florence, Italy. Association for Computational Linguistics.

Ryan Johnson, Lene Antonsen, and Trond Trosterud. 2013. Using Finite State Transducers for Making Efficient Reading Comprehension Dictionaries. *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 59–71.

Patline Laverdure, Ida Rose Allard, and John C. Crawford. 1983. *The Michif Dictionary: Turtle Mountain Chippewa Cree*. Pemmican Publications, Winnipeg.

Robert Leavitt. 2023. Creating the Passamaquoddy-Wolastoqey dictionary: A personal reflection on fifty years of lexicography. *Dictionaries: Journal of the Dictionary Society of North America*, 44:187–206.

Patrick Littell, Aidan Pine, and Henry Davis. 2017. Waldayu and Waldayu Mobile: Modern digital dictionary interfaces for endangered languages. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 141–150, Honolulu. Association for Computational Linguistics.

John Lyon, Justine Manuel, and Kathleen Michel. 2023. The Upper Nicola Nsyilxcn talking dictionary project: Community-driven revitalization lexicography within an academic context. *Dictionaries: Journal of the Dictionary Society of North America*, 44:107–126.

Aurélien Max and Guillaume Wisniewski. 2010. Mining naturally-occurring corrections and paraphrases from Wikipedia's revision history. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Timothy Montler. 2018. *SENĆOŦEN: A Dictionary of the Saanich Language*. University of Washington Press, Seattle, WA, USA.

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. PanPhon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484, Osaka, Japan. The COLING 2016 Organizing Committee.

Aidan Pine, Erica Cooper, David Guzmán, Eric Joanis, Anna Kazantseva, Ross Krekoski, Roland Kuhn, Samuel Larkin, Patrick Littell, Delaney Lothian, Akwiratékha' Martin, Korin Richmond, Marc Tessier, Cassia Valentini-Botinhao, Dan Wells, and Junichi Yamagishi. 2025. Speech generation for Indigenous language education. *Computer Speech & Language*, 90.

Aidan Pine, Patrick Littell, Eric Joanis, David Huggins-Daines, Christopher Cox, Fineen Davis, Eddie Antonio Santos, Shankhalika Srikanth, Delasie Torkornoo, and Sabrina Yu. 2022. G_i2P_i rule-based, index-preserving grapheme-to-phoneme transformations. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 52–60, Dublin, Ireland. Association for Computational Linguistics.

Olivia Sammons. 2019. *Nominal Classification in Michif*. Ph.D. thesis, University of Alberta.

Christine Schreyer and Mark Turin. 2023. Indigenous lexicography: An introduction. *Dictionaries: Journal of the Dictionary Society of North America*, 44:1–5.

Victoria Sear and Mark Turin. 2021. Locating criticality in the lexicography of historically marginalized languages. *History of Humanities*, 6:237–259.

Heather Souter, Carmen Leeming, Marlee Paterson, and Terry Ireland. 2024a. *Southern Michif for Learners*. Prairies to Woodlands Indigenous Language Revitalization Circle.

Heather Souter, Olivia Sammons, and David Huggins Daines. 2024b. Creating digital learning and

reference resources for Southern Michif. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 67–75, St. Julians, Malta. Association for Computational Linguistics.

Bailey Trotter, Christine Schreyer, and Mark Turin. 2023. An open-access toolkit for collaborative, community-informed dictionaries. *Dictionaries: Journal of the Dictionary Society of North America*, 44:161–185.

Nancy Turner and Richard Hebda. 2012. *Saanich Ethnobotany: Culturally Important Plants of the W̱SÁNEĆ People*. Royal British Columbia Museum.

Hugo Zaragoza, Nick Craswell, Michael J. Taylor, Suchi Saria, and Stephen E. Robertson. 2004. Microsoft Cambridge at TREC 13: Web and hard tracks. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004*, volume 500-261 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Exploring Limitations and Risks of LLM-Based Grammatical Error Correction for Indigenous Languages

Flammie A Pirinen
Divvun
UiT Norgga árktalaš universitehta
Tromsø, Norway
flammie.pirinen@uit.no

Linda Wiechetek
Divvun
UiT Norgga árktalaš universitehta
Tromsø, Norway
linda.wiechetek@uit.no

Abstract

Rule-based grammatical error correction has long been seen as the most effective way to create user-friendly end-user systems for grammatical error correction (GEC). However, in the recent years the large language models and generative AI systems based on that technology have been progressed fast to challenge the traditional GEC approach. In this article we show which possibilities and limitations this approach bears for Indigenous languages that have more limited digital presence in the large language model data and a different literacy background than English. We show experiments in North Sámi, an Indigenous language of Northern Europe.

1 Introduction

Grammatical error correction (GEC) is a crucial for supporting writers in their writing process, especially new writers and those who do a large workload in production and translation of administrative texts, educational material, news articles, fiction.

For writers of Indigenous languages proofing tools have an even higher significance which is due to literacy in these languages. Indigenous and minority languages that compete with an official majority language typically stand much stronger orally than written, and competent speakers are not necessarily competent in writing to the same degree as in speech. However, a feeling of competence is an important factor in text production, and writers typically feel more confident when they can verify grammar and spelling. An increase in high-quality text production (representative of the language we want as an output) again is an important factor in developing large language models. In other words, we need a sufficient amount of the type of language we want to be produced as an input to the models, and in order to build a text corpus, we need someone to writing skills and motivate native speakers to write. Behind that is usually the work of highly motivated native language experts who actively push forward a language revitalization process (Olthuis et al., 2013). As a part of this process, language technology can provide the necessary tools like spell- and grammarcheckers.

Up until late it has been obvious that linguistically demanding tasks like grammatical error correction require a component of expert-built, rule-steered grammar, not only to be accurate enough, but also to have the legitimacy of an expert controlling the language norms and ongoing standardisation. However, in the few recent years it has raised into a question if more data-driven approaches can also work for this problem. In this article we perform some experiments to find out to which extent this is plausible and what kind of limitations there are.

The *research question* we solve in this article is to evaluate how efficient the contemporary large language models are in the actual task of grammatical error correction—specifically in endangered language context with North Sámi as example study. We set to find out the effort needed to use them and develop existing systems. We also consider how much work it might take to fix problems in the large language models versus rule-based models when it comes to, e.g. bad suggestions and mistakes in the error detection (i.e. false positives). Regardless of the paradigm, the improvement of the system is driven by developers with language skills or developers with linguists co-operating, a resource that is very sparse. Another dimension is how time-critical the system is; a high quality GEC is a time critical resource for Indigenous language maintenance and revitalisation in digital era and leaving a low quality or disfunctional GEC with a promise of potential better version in the future is unacceptable.

Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 74–81

2 Background

Grammatical error correction system for Indigenous languages in the Sámi region have existed for over a decade. (Wiechetek, 2012) These systems use rule-based approaches to natural language processing: Finite-State Morphology (Beesley and Karttunen, 2003) for modelling lexica and morphology and Constraint Grammar (Karlsson, 1990) for modelling linguistic grammars including syntax. Rule-based approaches have historically been considered as an ideal fit for grammatical error correction, since it directly concerns writing grammatical rules. In a rule-based approach it is possible to target exact grammatical phenomena and also provide user feedback precise to the situation: "if there is a first singular personal pronoun and verb in third singular form, mark an error and tell user about the mismatch, suggest using first singular form of the verb instead" would be a typical grammatical order of action in a GEC tool build on rule-based natural language processing system. Historically, statistical and data-driven approaches have been limited to flagging unlikely word-forms and suggesting more likely forms, without addressing complex grammatical constructions or the logical error that leads to the error and eventually helps the writer to understand what has gone wrong. The missing link between error and cause in these approaches deprives the user of understanding their mistakes and improving their grammar. However, it has been suggested that the LLM-based approaches may be able to overcome this limitation, and, at least the most popular chatbot-driven user interface to large language models does indeed generate explanations alongside corrections when requested. Inspired by these innovations we decided to test the current capabilities of large language models and compare them to the rule-based approach.

The open source LanguageTool (Naber et al., 2003) and the closed-source browser plugin / webapp Grammarly (Alikaniotis and Raheja, 2019) are two of the most widely used GEC tools. On top of that popular office applications like Google Docs provide writers aids, which seem to mainly focus on spelling errors, but may contain grammatical error correction features as well. We have not found suitable scientific documentation of these to give a fair comparison.

2.1 Languages and literacy

We are experimenting with North Sámi, an Indigenous and low-resource language of Northern Europe. With approx. 20,000 speakers according to Ethnologue Campbell and Grondona (2008) it is the biggest of the 9 Sámi languages. It is spoken in Norway, Finland and Sweden and competing with the three national majority languages Norwegian, Finnish and Swedish as (nearly) all speakers of North Sámi are bilingual. While Finnish and Sámi are related (Finno-Ugric) languages, Sámi and Norwegian/Swedish are on opposite branches of the language tree. Bilingualism and loss of language domains are the cause of a higher frequency of grammatical errors among North Sámi writers. On the other hand, the widespread use of technologies requires us to express ourselves in writing in all domains. If the Sámi languages are to have a digital future, written Sámi needs to be strengthened and correction tools need to be available for everybody.

2.2 Risks related to quality of GEC

When we think of spell- and grammar checkers as tools that somewhat enforce a language standard in the same way as a teacher and educational books, we rely on a high level of knowledge/accuracy from these language authorities. Proofing tools that do not comply with these high standards will eventually have a negative effect on their user communities. In the case of Indigenous user communitites this effect can be even stronger if proofing tools are used in the absence of daily language arenas and language experts. Related research in the field of spell-checking and correction for L2 learners in schools suggests (Högström et al., 2024) that there are patterns of usage of automatic language correcting tools that can be detrimental to the end-users. Some of the problems of this sort can be avoided by ensuring the quality of GEC.

One trend of data-driven language technology products for (minority) languages has been providing inferior products to the existing rule-based ones, promising that they will be improved eventually. However, for example as it is in the case of spellchecking for Finnish, in the past 10 or so years, the so-called autocomplete set of spelling checkers have not improved to be able to handle rare morphologically complex word-forms at all, which is a clear downgrade from earlier rule-based spellcheckers. For example, a recent version of gboard Finnish keyboard for android does not think

ovikoodit ‘door codes, i.e. keycodes for a door’ is a word and marks it as an error, while it is a normal and quite lexicalised compound already. On another example, it erroneously suggest that the correct *suukoistamme* ‘about our kisses’ is replaced with the more likely *puukoistamme* ‘about our knives’, which apparently exists in the training data. In order for future GEC to be useful and not destructive, problems leading to this sort of downgrades ought to be fixed before pushing them to end users.

3 Methods

We are using two existing systems out of the box for comparison: one rule-based and one based on the large language model technology. We use the systems as black boxes, without rewriting the existing rules of the rule-based systems and without finetuning or re-training the large language model. In the rule-based system, we use command-line tooling to get suggestions for the grammatical error corrections with explanations and with the LLM we use the available chat interface to get corrections and explanations. The LLM is prompted in English and advised to do Grammatical Error Correction.

Here are two pictures showing the end-user experience of grammar checking with the two systems as it is now, see *ChatGPT* in Figure 1 and *GramDivvun* in Figure 2. *ChatGPT* helpfully provides a translation of the sentence with its error correcting explanations. It translates *jáhkkán* with the gerund *waiting* instead of past participle *thought* and by that introduces two errors– one semantic and one syntactic. The correction *jáhkán* is not a gerund as promised, but a first person singular. The list of flaws for this short sentence goes on and on. *Guorosnaga* does not mean *suddenly*, but *empty-handed* and the spelling correction to *guorusnaga* is incorrect. *[G]ii livčči* does not mean *who was*, but *who would have (thought)*. *[J]a dál diekkár* means *and now that*, which is entirely correct and should not be corrected to *ja dál diehtá*.

We have gathered an error corpus by harvesting sentences with several types of grammatical errors, where we focused on 1. frequent errors in the error corpus and 2. errors of different main types and complexity. The main grammatical error types are categorized in the error corpus are lexical errors (misuse and non-idiomatic use of a word), real-word errors (forms that are likely caused by a typo, but result in existing words), morpho-syntactic errors (errors that have a syntactic impact, where the difference between error and correction can be described by means of morphology), syntactic errors (errors that have a syntactic impact, where the difference between error and correction requires adding/taking away or moving one or several word forms). In addition, the error data contains punctuation and style errors. We have not evaluated the punctuation and style errors in this article.

Error type	Instances
Adjective inflection errors	6
Global agreement errors (subject-verb)	7
Nominal case errors	8
Compound errors (2>1)	6

Table 1: Morpho-syntactic and syntactic error types

The Figure 3 shows the raw corpus data for a morpho-syntactic error. The third person plural verb *eai* does not agree with the singular subject *dihtor* ‘computer’, cf. ex. (1).

(1) Mus ***eai** leat
I.LOC NEG.3PL be.CONNEG
dihtor dahje TV.
computer.NOM.SG or TV.NOM.SG
‘I don’t have a computer or a TV’.

In this investigation, we focused only on the two categories of morpho-syntactic and syntactic errors, specifically the error types represented in Table 1. Since we are using real-world texts as test data, some of the sentences do contain further error types; this is common and unavoidable in the realistic use cases for Indigenous corpora and GEC.

4 Results and discussion

To evaluate the grammatical error correction systems, we have collected and hand-annotated 101 sentences, some of which are error-free and some which have one or more errors. We have done both quantitative and qualitative analysis of the error corrections performed by both *GramDivvun* and *ChatGPT*.

To get a rough idea of the quality, we measured the precision and recall using the usual formulas, on a per error basis, counting a correction as a true positive only when the detected error and the correction are exactly the correct substrings, true negative, when no errors were expected and none were marked, false positive, when system marked a non-error substring as an error and false negative,

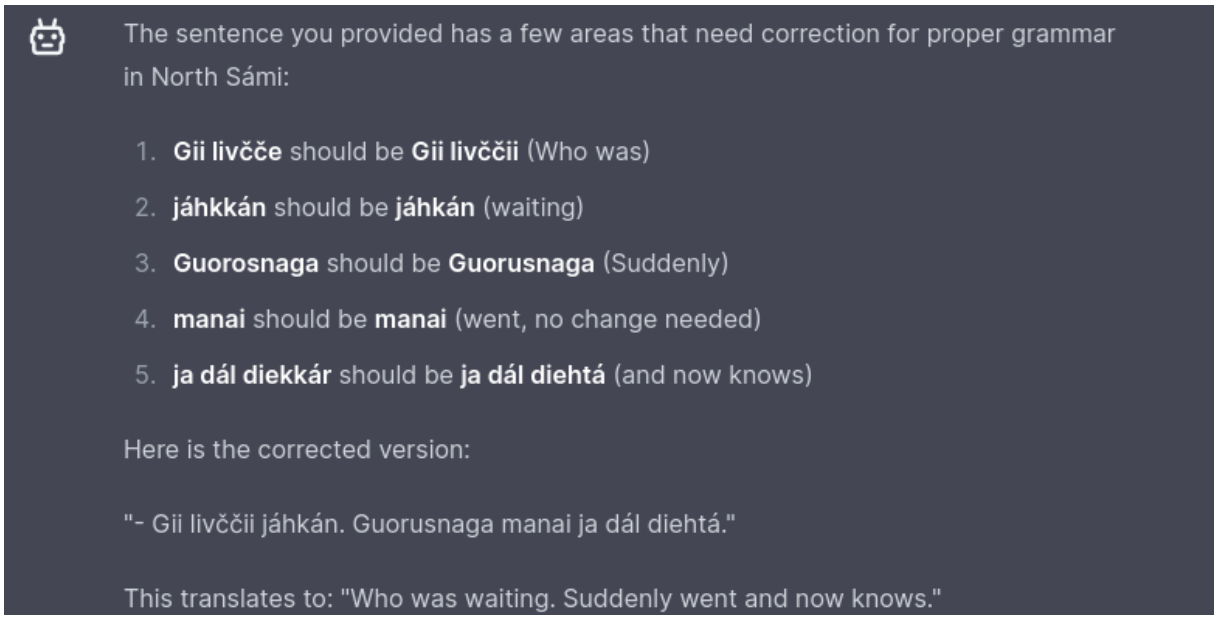


Figure 1: ChatGPT correcting a sentence in North Sámi.

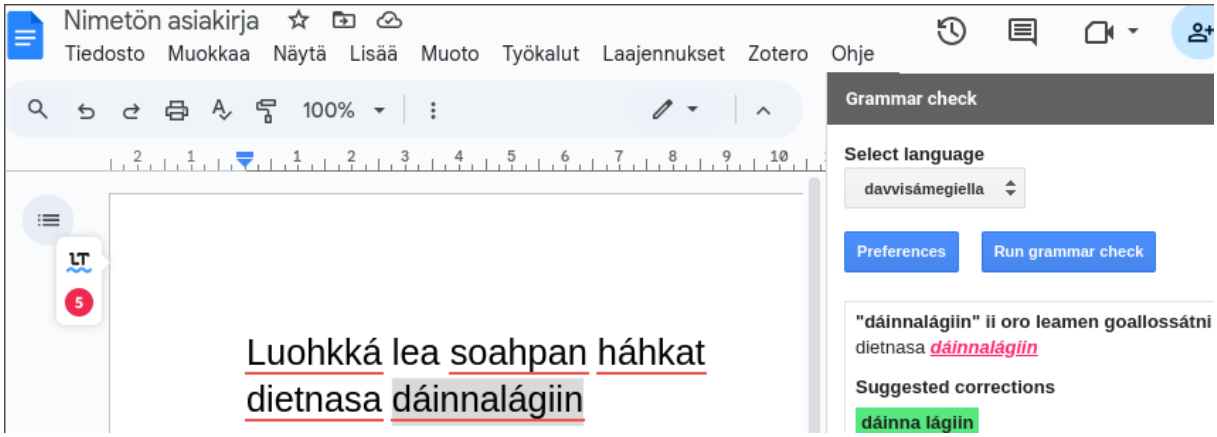


Figure 2: *GramDivvun* in Google Docs correcting a sentence with default red lines for corrections of English language

```
Mus {eai}£{verb,fin,sg3prs,
pl3prs,kongr|ii}
leat dihtor dahje TV.
```

Figure 3: Example of marked up error in the hand-annotated corpus.

System	Precision	Recall	$F_{0.5}$
GramDivvun	58 %	60 %	0.58
ChatGPT	17 %	13 %	0.16

Table 2: Precision, recall and $F_{0.5}$ scores of the systems we tested.

when system did not flag an error substring, or also when all corrections were incorrect. The statistics resulting are shown in the figure 2. We included an $F_{0.5}$ score to underscore our preference for high precision over high recall.

We have performed a linguistic error analysis on both correction sets and summarised the results in the following subsection 4.1.

4.1 Error Analysis

We first analyse common North Sámi grammatical error types linguistically keeping in mind the probable cause of the error from the end-user perspective. We then show how both GEC tools do the grammatical error analysis (successfully/unsuccessfully) and summarise our findings. The source texts are presented in examples (2), (5), (8), (11), and (14) followed by *ChatGPT*'s corrections in examples (3), (6), (9), (12), and (15), and *GramDivvun*'s (4), (7), (10), (13), and (16) respectively.

Example (2) has a compound error in *dáinnalágiin* 'in this way'. It is perceived as a semantic unit, therefore many people write it as one word. However, the official spelling requires it to be two words. In example (5) there aren't any grammatical errors. Example (8) has a common verb error in *livčče*, which is due to dialectal forms that are not used in the written standard. In order to satisfy subject-verb agreement third person plural *livčče* should be third person singular *livčči*. Example (11) has another subject-verb agreement error. The first person plural verb *guorahallat* should be third person plural *guorahallet* in agreement with the nominal subject in plural. For uneven-syllable verbs, first and third person plural forms in present tense are homonymous. However, for uneven syllable verbs that is not the case. In example (14), there is an adjective error. In North Sámi, adjective forms differ with regard to their position in the sentence. In this sentence it is used before a noun, i.e. attributively. However, the form used here is predicative *álkis*. The correct form is *álkes*.

A common type of error with *ChatGPT* is that it comes up with non-words as a corrections, as in example (3).

(2) Luohkká lea soahpan
class have.PRS.3SG agree.PTCP
háhkat dietnasa dáinnalágiin
provide income in.this.way
'The class has agreed to provide the income this way'.

(3) Luohkká lea soahpan háhkat dietnasa dáidnalaččat

(4) Luohkká lea soahpan háhkat dietnasa dáinna lágiin

In this example *ChatGPT* came up with the nonsense word *dáidnalaččat*. *GramDivvun*, on the other hand, splits the compound as it should.

In the following example (5) *ChatGPT* corrects the predicative adjective form *boaris* to adjective *boares*, where it should not be corrected. *GramDivvun*, correctly, does not give us this false correction. However, it does not recognize the foreign name *Hijiri* and corrects it to *Hiiri*.

(5) Sin namat
They.GEN.PL name.PL.NOM
leat Jola ja Hijiri. Jola
be.PRS.3PL Jola and Hijiri. Jola
lea gávcci
be.PRS.3SG eight.SGG.GEN
jagi boaris. Hijiri
year.SG.GEN old. Hijiri
lea logi jagi
be.PRS.3SG ten.SG.GEN year.SG.GEN
boaris.
old.
'Their names are Jola and Hijiri. Jola is eight years old. Hijiri is ten years old'.

(6) Sii namat leat Jola ja Hijiri. Jola lea gávcci jagi boares. Hijiri lea logi jagi boares.

(7) Sin namat leat Jola ja Hiiri. Jola lea gávcci jagi boaris. Hiiri lea logi jagi boaris.

In the following example (8) *ChatGPT* corrects the third person verbal form *livčče* to third person singular *livčči* correctly. However, it introduces two realword errors which were not there beforehand, *jáhkkán*>*jáhkán* (changing past participle to first person singular present tense and the demonstrative pronoun), *diekkár* to a similar sounding third person singular verb *diehtá*. *GramDivvun* corrects the agreement error correctly, and does not introduce any false positives.

(8) Gii livčče jáhkkán.
Who be.COND.3PL think.PTCP.
Guorosnaga manai ja dál
Empty-handed go.PAST.3SG and then
diekkár.
that.
'Who would have thought. S/he went there empty-handed and then that.'

(9) Gii livčči jáhkán. Guorusnaga manai ja dál diehtá.

(10) Gii livčči jáhkkán. Guorosnaga manai ja dál diekkár.

In example (11), *ChatGPT* erroneously introduces a new lexeme, i.e. *mánáid* 'children', instead of *mánusiid* 'manuscript'. Neither *ChatGPT*, nor *GramDivvun* manage to correct the subject-verb agreement error and change first person plural *guorahallat* to third person plural *guorahallet* in agreement with the plural subject *konsuleanttat*.

(11) Ovdalgo ášši joavdá dán
Before thing move.PRS.3SG this
muddui, de konsuleanttat
way, then consultants
*guorahallat
deal.with.PRS.1PL
mánusiid.
manuscript.ACC.PL
'Before the matter goes this way, the consultants deal with the manuscripts.'

(12) Ovdalgo ášši joavddá dán muddui, de konsuleanttat guorahallat mánáid.

(13) Ovdalgo ášši joavdá dán muddui, de konsuleanttat guorahallat mánusiid.

(14) Árbejuohku lei duššiid dihte
Árbejuohku lei duššiid dihte
maŋŋonan, ja gárvvisin
maŋŋonan, ja gárvvisin
jurddašuvvon álkis kuohku lei
jurddašuvvon álkis kuohku lei
šaddan váddáseabbon.
šaddan váddáseabbon.
'The inheritance settlement had been delayed due to trivial matters, and the planned simple settlement had become more complicated.'

(15) Árbejuohku lei dušše dihte maŋŋonan, ja gárvvisin jurddašuvvon álkis juohku lei šaddan váddáseappot.

(16) Árbejuohku lei duššiid dihte maŋŋonan, ja gárvvisin jurddašuvvon álkes juohku lei šaddan váddáseabbon.

GramDivvun corrects the adjective error *álkis* to attributive *álkes*. *ChatGpt*, on the other hand, firstly, does not find the adjective error and secondly, corrects several forms that are correct in the original sentence, *duššiid* 'nonsense' to *dušše* 'only' and *váddáseabbon* to *váddáseappot*. Both,

ChatGpt and *GramDivvun* find the spelling error *kuohku* and correct it to *juohku*.

4.2 Discussion

As an overall conclusion of *ChatGPT*'s performance, the overwhelming problem is the false positive rate, which can be rather bothersome for end-users and, more importantly, contradicts the authorative nature of a spell- and grammarchecker in language expertise. The rule-based grammar checker, which performs significantly better in our experiment, can have issues with recall at the expense of not alerting the user with false positives. An open question for the LLM-based approach is, what kind of effort it would take to get the false positive rate down, or if the correct way forward is to use a hybrid control where rule-based grammar can identify actual errors with more precision, and possibly validate or guide correcting as well.

One recent trend in LLM-based NLP applications, especially in low-resourced contexts, is to bring specific examples of the target language in the context of the prompts, e.g. in RAG or in-context training. This would be an interesting future experiment. However, in order to fully retain the authoritative, norm-building grammar correction, we would envision an ideal hybrid LLM-application that would be able to interact with the linguistic resources of rule-based implementation in the same way as they do with calculators, python scripting and web browsers to overcome hallucinations in the LLM-based math-answering, programming, and smart agent applications respectively.

One noteworthy thing about the current experience with *ChatGPT*-driven grammatical error correction is that the generated helpful descriptions (c.f. Figure 1 for reference on what they look like in the chat interface) do not always properly keep track of the actual corrections that the system makes, so it provides an itemized list of corrections and the corrected text snippet, which do not necessarily match with each other. Furthermore, some of the explanations provided are not corrections at all, generally formulated such as: "*heivehit* should be *heivehit* (to develop)" even if, in this case, the suggestion and original word form are the same. *ChatGPT* also does the opposite, saying "jus beatnaga should be jos beatnaga (if the dog, no change needed)" when the suggestions actually does include a change. If this was used in an end-user product, it would be very confusing for the users. Given that this type of problem has been a known issue of the generative LLMs for a while now, it might be a risk if pivoting to fully LLM-guided grammatical error correction.

Furthermore, when reading the explanations provided by *ChatGPT*, the interpretation of the sentence and its correction, e.g. in Figure 1, contain serious flaws that, instead of helping the user, present them with the additional workload of deciding when to trust the tool and when not. The errors that are made by the tool are completely random and do not follow a certain pattern, which makes it impossible for the user to trust it.

To loop back to our initial research question and quite concretely the setup that we have: if we have available one North Sámi computational linguist, what is the most reasonable use of time for them to improve the North Sámi grammar checker; writing the grammar rules, collecting and annotating error corpora or giving human feedback to a chatbot; all of which can be tedious at times and not very exciting? At the time it still seems that the former is more beneficial, but it is an open question and possibly changing in near future?

5 Conclusion

We have tested LLM and traditional grammatical error correction for North Sámi. LLMs a few correct frequent forms of grammatical errors correctly (like the copula form *livčči*, which is more common than the form *livčče*). At the same time it introduces an uncontrollable amount of unsystematic false positives that the tool becomes useless for any user that seeks linguistic help from a grammar checker. It also tends to replace a lot of forms with a completely different lexeme.

An expert of the language with above-average language intuitions may be able to evaluate the correctness of the grammar checker suggestions. However, when false alarms outnumber the correct suggestions as in this case, the tool does not reduce, but add to the workload of the writer. More importantly, as we have pointed out, grammar checking for the Sámi writer is meant predominantly as an active help where spelling and grammar skills may be incomplete. In this case, the writer is left with an unreliable tool that does not provide linguistic stability, but instead increases the insecurity of the writer. Language confidence is an important factor in revitalization and feeling comfortable to use the language even if it is not one's first language.

Limitations

The LLM testing was made with a closed source, commercial LLM and the results cannot be easily reproduced. However, this method of impressionistic exploratory testing seems to be a de facto standard in contemporary natural language engineering.

Ethical Concerns

The LLM-based experiment has consumed an estimated hundreds of liters of drinking water[1] and a not insignificant amount of energy (Strubell et al., 2019). With this background it seems almost irresponsible to conduct more LLM experiments, however, given the strong hype in the scientific discourse at the moment, debunking some of the hype may prove invaluable also in putting a cap for wasted experimentation.

We have used no crowd-sourcing or underpaid external workers for this article, all the linguistic and computational work has been done by authors and colleagues who are fully paid for their work.

References

Dimitris Alikaniotis and Vipul Raheja. 2019. The unreasonable effectiveness of transformer language models in grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 127–133, Florence, Italy. Association for Computational Linguistics.

Kenneth R Beesley and Lauri Karttunen. 2003. Finite-state morphology: Xerox tools and techniques. *CSLI, Stanford*.

Lyle Campbell and Verónica Grondona. 2008. Ethnologue: Languages of the world. *Language*, 84(3):636–641.

Jenny Högström, Marie Nilsberth, and Anna Slotte. 2024. "it is not always very cooperative" distributed agency in the use of spell check software in a lower secondary classroom. *Nordic Journal of Digital Literacy*, (1):8–24.

Fred Karlsson. 1990. Constraint grammar as a framework for parsing unrestricted text. In *Proceedings of the 13th International Conference of Computational Linguistics*, volume 3, pages 168–173, Helsinki.

Daniel Naber et al. 2003. A rule-based style and grammar checker.

[1] https://www.thetimes.com/article/9167a8a8-96d1-4a68-9a13-824d862f627a?shareToken=ee1797a1e9992a631f79c82dd49c3a6b

```
Please correct the grammar in the following North Sámi texts:
```

Figure 4: Example of ChatGPT prompt for grammatical error correction.

Marja-Liisa Olthuis, Suvi Kivelä, and Tove Skutnabb-Kangas. 2013. *Revitalising indigenous languages: How to recreate a lost generation*, volume 10. Multilingual matters.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Linda Wiechetek. 2012. Constraint Grammar based correction of grammatical errors for North Sámi. In *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL 8/AFLAT 2012)*, pages 35–40, Istanbul, Turkey. European Language Resources Association (ELRA).

A ChatGPT Example Prompt

We thank the anonymous reviewers for the suggestions of inclusion of example prompt. Since this test is performed on a commercial black box solution, there is no scientific reproducibility of any kind, and indeed the whole underlying system has changed drastically between time of our writing the article and writing this camera ready. An example prompt is provided in the Figure 4.

Speech Technologies Datasets for African Under-Served Languages

Emmanuel Ngue Um
Institute of African DH
ngueum@gmail.com

Francis Tyers Morton
University of Indiana
ftyers@iu.edu

Eliette-Caroline E. Ngo Tjomb
University of Yaounde 1
eliette1404@gmail.com

Landry Dibengue
University of Yaounde 1
landrydibengue@gmail.com

Blaise-Mathieu Banoum Manguele
University of Yaounde 1
banoummanguelleblaisemathieu@gmail.com

Blaise Abbo Djoulde
Institute of African DH
tblaiso@gmail.com

Mathilde Nyambe A
Institute of African DDH
mathildenyambe@gmail.com

Brice M. Atangana Eloundou
Institute of African DH
martialatangana@gmail.com

Jeff S. Ngami Kamagoua
University of Yaounde 1
jeffsterlingngami@gmail.com

José Mpouda Avom
Institute of African DH
joseavom@gmail.com

Zacharie Nyobe
Institute of African DH
znyobe@gmail.com

Emmanuel Giovanni. Eloundou Eyenga
Institute of African DH
eloundoueyenga13@icloud.com

André Pascal Likwa
Institute of African DH
likwaiandre66@gmail.com

Abstract

The expansion of the speech technology sector has given rise to a novel economic model in language research, with the objective of developing speech datasets. This model is expanding to under-served African languages through collaborative efforts between industries, organisations, and the active participation of communities. This collaboration is yielding new datasets for machine learning, while also disclosing vulnerabilities and sociolinguistic discrepancies between industrialised and non-industrialised societies. A case study of a speech data collection camp that took place in September 2024 in Cameroon, involving representatives of 31 languages throughout the continent, illustrates both the prospects of the new economic model for research on under-served languages and the challenges of fair, effective, and responsible participation.

Introduction

There is a growing momentum in industry and academia to develop speech technologies on a massive scale. In the industrial domain, one of the most emblematic moves in this regard is the Massively Multilingual Speech (MMS) project initiated by Meta (Pratap et al., 2024), which aims to extend the coverage of speech technology across the global linguistic landscape. There are currently 336 African languages for which the MMS project has developed automatic speech recognition (ASR) and text-to-speech (TTS) models. MMS uses multilingual datasets to pre-train wav2vec 2.0 models, and the labelled dataset used for this pre-training consists of aligned New Testament recordings. This has enabled coverage of many of Africa's under-served languages, for which the Bible is often the only substantial textual resource. At an institutional level, academics and organisations are working together to build language datasets for machine learning in African languages. This is evidenced by initiatives such as The Lacuna fund[1], which has enabled the creation of a diverse range of language datasets, including speech datasets in more than 20 African languages over the past three to four years (Babirye et al., 2022).

Despite this progress, significant limitations remain, particularly in the dominant crowdsourced data collection model employed by platforms such as Mozilla Common Voice (MCV)[2] (Ardila et al., 2020). While MCV is widely recognised for enabling community participation in the creation of speech datasets, several critical flaws undermine its effectiveness for under-served languages. A significant challenge pertains to the dearth of publicly accessible text sources that can be collated for utilisation as reading prompts, compelling the reliance on religious texts such as the Bible, which are frequently the sole non-licensed text data sources.

[1] https://lacunafund.org/datasets/language/
[2] https://commonvoice.mozilla.org/en/about

Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 82–90

While the Bible may not be the predominant text source in most of the MCV's collecting interfaces for African languages, the absence of text diversity in under-resourced languages leads to a limited representation of language use, significantly differing from the fluid and varied nature of daily language usage. Additionally, the platform's framework tends to impose a single orthography model for each language, disregarding the linguistic diversity and orthography multiplicity found within many African communities. This rigid approach has the potential to marginalise certain dialects or writing traditions. Another challenge stems from the dependency on literacy participation, which excludes individuals who are fluent speakers but not proficient readers. Finally, the incentivisation of participation, while effective in the short term, raises questions about the sustainability of community engagement and the quality of collected data over time. The speech data collection camp organised by the Institute of African Digital Humanities (INHUNUM-A)[3]– in partnership with MCV, which constitutes a use case in this discussion – highlights these challenges. This experience has underscored the necessity for a more inclusive and adaptable approach to the development of speech technologies for African languages.

The initiative had two main goals. First, it sought to expand the reach of the MCV ecosystem in Africa by engaging community representatives to lead responsible, long-term crowdsourced speech data collection efforts. These efforts would be critical to the future development of speech technologies. Secondly, the initiative aimed to collect a 310 hour benchmark labelled speech dataset for 31 under-served African languages[4]. This paper reports on the key areas of the project and the challenges encountered during its implementation. These are grouped under (1) methodological, (2) technological, (3) sociolinguistic, (4) quality control, (5) incentivisation, (6) ethical aspects, and (7) discussion, and (8) recommendations.

1 Methodological aspects

In this section we discuss the approach to 1) the selection of languages and team members and 2) the collection and pre-processing of sentences.

1.1 Selection of languages and teams

The Institute of African Digital Humanities is a newly established organisation that aims to provide capacity building and networking in the use of digital methods and tools in the humanities and social sciences on the continent. Its outreach includes affiliated members, but more broadly any African-based institutional or individual stakeholder with an interest in digital humanities. In order to promote greater inclusivity across the regions and linguistic communities of the continent, an open call was launched to select teams, ideally consisting of two representatives of different genders and dialects within the same linguistic community. Candidates were also required to be fluent and literate in the language they were representing. In a sense, the selection was aimed at grassroots language enthusiasts who were not necessarily trained in linguistic research. In the same vein, the selection mechanism was designed to ensure, as far as possible, an equitable representation of linguistic diversity, to the extent that a given language was endowed with at least a standard orthography and a basic body of literature. Less emphasis was placed on criteria used in similar initiatives, such as regional representation, number of speakers or degree of standardisation (Butryna et al., 2020; Agirre et al., 2021). Languages with existing ASR or TTS models, including those developed in the MMS project, were excluded from the selection, even if they were more under-served. While this selection process was consistent with the principles of equity and representativeness that underpin the philosophy of our initiative, it did introduce some biases and inequalities. In terms of bias, the current ASR and TTS models developed within MMS, which are largely trained on biblical recordings, have not been sufficiently evaluated for performance, inclusivity and representativeness, raising concerns about the reliability of these technologies for the wider language community. In terms of inequality, the selection excluded de facto languages for which there was no existing orthography and/or a minimal body of literature.

Overall, The number of languages launched on MCV increased from 137 to 166, with the addition of 29 new languages[5], after the language data col-

[3] https://inhunumaf.hypotheses.org/

[4] https://github.com/Ngue-Um/INHUNUMA2024/blob/main/Inhunuma2024.md

[5] Setswana, one of the 31 languages involved, was already launched prior to the data collection event. Representatives of the Setswana languages attended the event with the objective of expanding the existing collection of sentence prompts to include the Kgatla dialect. At the time of this writing, Tunen,

Figure 1: MCV ecosystem in Africa before the data collection camp

Figure 2: MCV ecosystem in Africa after the data collection camp

lection camp held on September 9-14, 2024. This represents a growth of approximately 21.17%. The camp's contribution to expanding speech data collection for under-served African languages resulted in a significant increase in the platform's language offering, as represented on figures 1[6] and 2[7].

1.2 Sentence collection and preprocessing

There are two approaches to designing speech datasets using MCV. The first approach is Spontaneous Speech, whereby speakers are provided with prompts in their language, e.g. "What is the history of the origins of your community?", and are asked to respond in a few sentences, resulting in voice clip recordings. Subsequently, the recordings are listened to and transcribed, resulting in the alignment of voice and script labels. The second approach is called Read Speech, and consists of speakers reading sentence prompts. The resulting voice clips are then listened to by two different speakers who validate or invalidate the voice clip, assigning labels to the voice clip in the validation process. The second approach was used in our data collection camp. A prerequisite for the Read Speech approach is the provision of sentence prompts, which in the case of this project had to be provided by language teams. Each language teach was required to provide a minimum of 1000 sentences, the sources of which had to be licensed under Creative Commons (CCO). The majority of these sentences were either elicited by the team representatives or derived from their personal manuscripts, with some requiring digitisation and preliminary processing. Digitisation entailed the deployment of OCR (Optical Character Recognition) or manual typesetting by team members or project staff. In numerous instances, both processes resulted in inadequate rendering of characters, necessitating re-encoding or character conversion, and posing technological challenges. To address these challenges, language teams received support from language technologists and data scientists who are part of the MCV staff.

2 Technological aspects

In this section we discuss 1) the technological challenges of navigating competing writing norms and 2) the localisation of MCV interfaces.

2.1 The "ortho-graphy" challenge

The term 'orthography' has its roots in the Greek word *orthos*, meaning 'straight', 'correct' or 'right'. The emphasis on correctness in writing is based on the idea that languages are realities that can be reduced to coherent parts that reflect the range of possible uses within a linguistic community. The very notion of 'linguistic community' (Gumperz, 1968) is based on the assumption of the unity of the members of a given language group. While 'correctness' in orthography and 'unity' within the linguistic community are relatively easy to achieve in societies with a long history of political organisation and centralisation, with the exception of

a second language of the 31, is awaiting its launch.

[6] https://tinyurl.com/mcv-languages-before

[7] https://tinyurl.com/mcv-languages-after

societies such as Luxembourgish (Bellamy, 2021), many African societies in the post-colonial era have yet to achieve such ideals, if they have to at all. In the context of this study, there were regular instances where the materials submitted by the language teams revealed issues of competing orthographic norms. This was particularly pronounced in languages with a history of early missionary literacy before independence. Literature produced in the pre-independence missionary alphabet tended to contrast with post-independence orthographic standards. The latter were promoted by the second generation of missionaries, led by the Summer Institute of Linguistic (SIL) and Evangelical Missions, and operationalised by the first generations of linguists of African descent.

The coexistence of different, sometimes divergent, orthographic norms was difficult to resolve in the context of this initiative. In any case, the project leadership did not have the legitimacy and responsibility to make decisions regarding the choice of a particular orthographic norm. At the same time, the technological interface of linguistic infrastructures such as MCV is designed in accordance with the dominant, monolithic view that there should be one and only one orthographic norm for a given language. Final decisions about the choice of orthography were left to the team members. In such circumstances, an agreement was reached with the project leadership to give priority to the orthography standard that is widely used in the community.

2.2 Localisation of MCV Interfaces

Incidentally, decisions on the choice of spelling standard for the sentence collection did not always coincide with the choices made by the translators responsible for localising the interfaces in the various languages. For reasons related to the project schedule and the scarcity of competent human resources in the selected languages, the task of translating for localisation was sometimes entrusted to actors other than those involved in providing the sentence collections. The ideal situation would have been to reach a compromise between the translators and the sentence contributors. However, such arrangements were not always feasible, given the remote nature of the workflow between translators, sentence collectors, project management and MCV, and the critical impact of any delay on the project schedule. As a result, there are interfaces, such as that for Eton[8], where the localisation follows a different orthography standard from the sentence collection.

3 Sociolinguistic aspects

For want of a better option, the project managers had to force language representatives to pool their sentence samples. Initially, teams were asked to provide unified sentence collections for their languages. However, in cases such as Tupuri and Batanga, the two members of the team, each representing a particular dialect, provided a sample for their dialect. While in the case of Batanga the two samples used the same orthography, in the case of Tupuri the orthography used in the sentence sample from Tupuri Banwere, spoken on the border between Chad and Cameroon, differed slightly from the orthography used for Tupuri Bango, spoken in the area of Kaele in Cameroon. The two orthographies seemed to reflect the sociolinguistic configuration of the Tupuri linguistic community, and there did not seem to be any socio-political contestation of this reality. At the same time, MCV allows only one unique locale for each specific language, where the locale is represented by a two- or three-letter code, e.g. 'tui' (for Tupuri), 'bnm' (for Batanga), 'tn' (for Setswana). Technically, therefore, the MCV infrastructure does not appear to be configured to accommodate the sociolinguistic reality of Tupuri, which is manifested in the fluidity of usage in both spoken and written form. The example of Tupuri is not uncommon in accounts of applied language work in Africa. Roberts et al. (2021) refer to a similar situation among the Yambasa community in Cameroon, where groups of arguably distinct dialects have reclaimed orthographic autonomy and developed separate writing norms and practices.

4 Quality control

The quality control process was divided into seven stages and was subject to oversight from the MCV staff and a pool of local experts, as illustrated in Table 1.

5 Incentivisation

Incentivisation through cash and in-kind rewards is common practice in language work in general, for example in language documentation research involving community contributors (Ngue Um, 2019;

[8] https://commonvoice.mozilla.org/eto

Levels of control	Oversight
Localisation (sheets)	Local team
Sentences (Sheets)	Local team
Localised (Pontoon)	Local team
Approved (Pontoon)	MCV staff
Sentences (Checked)	MCV staff
Sentences (MCV)	MCV staff
Launched	MCV Staff

Table 1: Levels of quality control and oversight involved in the project

Akumbu, 2024). It has also been implemented in the creation of language datasets for machine learning as part of the Lacuna Fund initiative (Babirye et al., 2022). The benefits of paid labour can be measured in terms of the level of mobilisation of the actors involved and the extent to which they have contributed to the achievement of the project's objectives. In the specific case of the speech data collection camp organised by INHUNUM-A in September 2024, the impact of the incentives can be seen in the mobilisation of the participants before, during and after the data meeting, which enabled the recording and validation of more than 300 hours of voice data over a period of 30 days. In terms of diversity and linguistic representativeness, this represents a significant growth in the ecosystem of both MCV and speech datasets for machine learning.

However, there are a couple of side effects of incentivisation. One is the sustainability of community mobilisation beyond the scope of a particular project, such as the one undertaken. Withholding a portion of the monetary compensation for teams that did not meet the goal of 10 hours of voice recording and validation during the camp timeline, and paying it only after the goals were met, proved effective for continued mobilisation after the camp. However, for almost all the languages involved, once the incentives are fully paid, the tendency to contribute decreases significantly and sometimes stops altogether. This raises questions about the long-term sustainability of a crowdsourced approach to speech data collection and, by extension, the voluntary, informed and qualitative participation of under-served communities in the development of speech technologies in their languages.

A notable dimension of this language data collection event is the under-representation of professional linguists, which contradicts the initial assumptions of the project leadership about a possible over-representation of linguists. In fact, of the 70 or so people who attended the meeting, only 3 professional linguists were listed. In comparison, there were three computer scientists. The majority of participants were grassroots language workers, either indigenous language teachers, translators, community literacy experts or language enthusiasts.

6 Ethical considerations and copyright

One of the major challenges in developing language datasets is the ethical considerations around data sources and community participation. For many under-served languages, existing text resources are sparse, and those that do exist are often limited to biblical texts. As a result, many existing ASR and TTS models in African under-served languages have been developed using these sources. This is the case with the MMS project, but also with the Building African Voices (Perez Ogayo, 2022) and Google Crowdsourced Speech Corpora for Low-Resource Languages and Dialects (Butryna et al., 2020) projects. This reliance on a religious text raises questions about the representativeness of the data, as it may not reflect everyday language use or cultural diversity within the community. In order to avoid expanding the inclusion of biblical texts in the language technologies of Africa's under-served languages, our project management reached an agreement with MCV to exclude such texts from the sentence collections. Although this provision was made explicit in the Call for Participation, a number of teams submitted sentence collections that were either entirely biblical or contained large swathes of religious texts taken from the Bible. In such cases, team representatives were asked to submit new collections. This has resulted in some of the initially selected teams dropping out of the project, or in long delays in the provision of the MCV interfaces for these languages.

In addition, the project had to deal with copyright issues, especially for languages such as Tunen, where the sentence sources were licensed under Creative Commons Attribution-ShareAlike (CC BY-SA), but needed to be licensed under Creative Commons (CCO) according to MCV standards. Community representatives were generally not well informed about copyright, and although the Call for Participation was explicit about these issues, the project leadership had not provided adequate

guidance and resources to help community representatives navigate and resolve these issues as they arose.

7 Discussion

Crowdsourcing is a mode of participation that is becoming increasingly prevalent in social, behavioral, and educational research (Bagherzadeh et al., 2023; Kwek, 2020). Bagherzadeh et al. (2023) have identified two distinct approaches to the recruitment of participants in crowdsourced routines, which they have metaphorically designated as "fishing" and "hunting." The "fishing" routine targets a wide range of external knowledge on a specific domain, with the assumption that the diversity of the participants' input will enhance the robustness of the solution that is being engineered. In contrast, the "hunting" approach targets specific individuals with expert knowledge in the domain under investigation, seeking to elicit solutions from those with the greatest expertise.

In the domain of linguistic research, an analogy can be drawn with language documentation, a form of crowdsourced perspective of linguistic research in which data collection leverages the involvement of diverse contributions, profiles, and situations (Ajo et al., 2010; Grenoble, 2010; Maxwell, 2010; Himmelmann, 2006). While MCV's crowdsourcing perspective is generally of the "fishing" type, language documentation predominantly employs the "hunting" technique, with various accounts of success stories (Dwyer, 2010), as well as shortcomings (Akumbu, 2024; Ngue Um, 2019).

One aspect of crowdsourcing for speech data that appears to be overlooked in the "fishing" approach employed by MCV is the distinction between the literacy rate in WEIRD (Western, Educated, Industrialized, Rich, and Democratic) populations and that in non-WEIRD ones (Brice et al., 2024). The implication of the literacy rate is that it indicates the degree of exposure of the average population to written text in the language for which speech datasets are collected. It is commonly assumed that a vast array of literacy expertise is readily available for crowdsourcing speech by reading sentence prompts, as well as for evaluating pre-recorded sentences. This is undoubtedly the case in literate societies and in WEIRD settings, but it is not the case in non-WEIRD, African under-served linguistic communities. Despite the fact that these communities have developed a considerable literacy rate through education, the reading and writing skills of individuals are still largely confined to the former colonial languages that serve as the medium of instruction in the majority of educational institutions across Africa. The implementation of the "fishing" approach in such circumstances thus renders crowdsourcing vulnerable.

Languages	Hours	Speakers	Validation
Duala	11	13	91%
Borgu Fulfulce	10	9	100%
Mbo	11	12	91%
Mokpwe	8	9	75%
Yoruba	7	123	72%
Hausa	13	50	39%
Ahmaric	3	34	67%

Table 2: Status of voice data contribution on MCV for 6 African languages (Language = "language name"; Hours = "total hours of speech recording, updated: 13th Oct. 2024 10:42am"); Speakers = "total number of contributors of recordings and validation"; Validation = "total number of labelled hours of speech data recording".)

As previously noted in Section 5, in the context of the project described in this paper, 100% of the contributions for the 30 languages included in the collection have either ceased or decreased significantly after the final payment of incentives. This may be in alignment with the analysis presented by Bagherzadeh et al. (2023), which suggests that the "fishing" approach attracts a significant number of non-domain experts, primarily driven by financial incentives. This hypothesis can be further substantiated by examining the trends in speech data contributions for African languages that were launched on MCV but not included in our data camp, as illustrated in Table 2.

This analysis does not imply that participants who are primarily attracted by financial incentives lack domain expertise. In the context of this study, domain expertise is defined as literacy skills in the language in which speech data is crowdsourced. The argument, therefore, is that the motivation of those who are attracted primarily by financial motives is more likely to decrease drastically in the absence of incentivisation. Conversely, Bagherzadeh et al. (2023) suggest that elite experts, that is to say, the category of participants in crowdsourcing who are recruited using the "hunting" approach, do not engage out of the prospect of financial gain in the first place.

With respect to the number of contributing speakers and the total population of the linguistic community, the three languages indicated in the shaded section of Table 2 exhibit a comparatively larger population. This may justify why their contributing population is more significant than the number of the contributing population of the languages in the unshaded area. Thus, the "fishing" approach to crowdsourcing that represents MCV's standard contribution "doctrine" would result in a higher level of contribution from the languages in the shaded area compared to those in the unshaded area. As the data in Table 2 show, this is not the case. In particular, a greater number of contributors does not necessarily result in a proportional increase in hours of recorded speech and validation. The discrepancy in the contribution rate observed in this case can be attributed to at least two factors. First, the influence of incentives, which is reflected in the higher contribution rate of the languages in the upper part of Table 2. Second, in the context of under-served linguistic communities, the standard "fishing" approach of MCV does not attract elite experts, who are likely to spend more time recording and validating voices, even in the absence of financial reward. It is also noteworthy that the timing of the contribution rate in the languages at the top of Table 2 indicates that participation in the "fishing" approach is primarily driven by financial incentives.

8 Recommendations

The participation of individuals in crowdsourced linguistic datasets in exchange for financial compensation highlights the economic vulnerability of those engaged in such activities. In the specific context of African under-served linguistic communities, where literacy in indigenous languages is often low, this raises further questions about the quality of participation. In light of the above, there is an urgent need to develop robust protocols for crowdsourcing data for speech technologies such as ASR and TTS that aim for inclusivity and efficiency. This is especially true for crowdsourced participation aimed at collecting and labelling speech data. Similarly, the evaluation of the performance of ASR and TTS models trained on crowdsourced speech data in under-served linguistic communities should include an assessment of the crowdsourcing methods used, as well as an investigation of the potential influence of the socio-economic vulnerability of the contributors on the quality of the technological solutions developed. The success of the experience of the Speech Data Camp reported in this study, which we describe in terms of the achievement of the objectives initially stated, owes much to 3 main factors. The first is the incitement through cash payment of the contributors, which has attracted a critical mass of candidates to the speech contribution, and has enabled the management side to define selection criteria that could guarantee a reasonable level of literacy expertise of the selected participants, as well as the diversity of voices, in terms of representativeness of coexisting dialects and gender. Here it is important to emphasize that the design of the data camp model is an important step for the success of such an initiative. The second factor is the timing of data collection. In our model, most language teams achieved the best contribution scores in terms of number of hours and rate of progress during the camp. In other words, on-site mobilisation and emulation among peer groups is critical for the onboarding and self-motivation of contributors, even with the promise of financial reward. In comparison, the rate of contribution within one month after the data camp was significantly lower compared to the 6 days of contribution during the camp, despite the incentives. Reasons for this are related to the lack of focus when participants are in their normal social environment, as well as access to internet and electricity. The third factor is the quality of supervision and monitoring of the contributions. Once again, the examples of Yoruba, Hausa and Amharic in Table 2 show that in the absence of leadership to create a momentum of voice-data contributions, the growth of contributions may remain uncertain. The status of the Kinyarwanda[9] contribution illustrates this state of affairs. Namely, under the leadership of a speech data collection startup, Digital Umuganda[10], Kinyarwanda is currently the third most contributing language on MCV, just behind English and Catalan, and surpassing better endowed languages such as Spanish, French, and Chinese.

Conclusion

The initiative to enhance speech technologies for under-served African languages has highlighted both challenges and opportunities in language data collection. This paper details the methodological,

[9] https://commonvoice.mozilla.org/rw
[10] https://digitalumuganda.com/

technological, sociolinguistic, ethical, and incentive aspects of the project, while highlighting the significant progress made in collecting over 300 hours of speech data for 30 languages[11]. However, critical issues remain, such as uneven language representation, barriers to community engagement, and the biases introduced by reliance on pre-existing automatic speech recognition (ASR) and text-to-speech (TTS) models, many of which are rooted in religious texts.

The project also grappled with competing orthographic norms, issues of copyrights applicable to the sources of the sentence prompts, and the long-term sustainability of crowdsourced data collection efforts. Despite the tangible results achieved, ensuring continued community participation beyond financial incentives remains a challenge. Going forward, a deeper commitment to fostering authentic collaboration between language communities, linguists and industry is essential to ensuring the equity and efficiency of the new economy model brought by voice technologies.

In addition, expert linguists specialising in underserved African languages need to develop a critical awareness of the solution-oriented approaches driven by industry that are increasingly influencing applied linguistic work. Without a deep understanding of industrial and commercial practices in product and service design, linguists cannot critically and productively engage with industrial actors who own many of the technological solutions and financial resources. These industrial actors often lack key insights into which approaches are most appropriate for specific languages and contexts. Productive collaboration between linguists, communities and industry is essential to ensure that the technologies developed are not only linguistically sound, but also socially and culturally relevant to the communities they are intended to serve.

Acknowledgments

The study reported in this paper is the result of a data collection workshop (which we refer to here as the " Speech Data Camp") funded by Mozilla through its Common Voice initiative. The authors of this paper are grateful for the generous support that made it possible to achieve the goals of the project. More importantly, the organisation of this workshop has led to what one Mozilla staff member has called a "sea change" in the Mozilla Common Voice ecosystem. The authors of this paper are grateful to Mozilla for their trust and excellent oversight at all stages of the event.

[11] As of the date of submission of this paper, one language, Tunen, is awaiting clearance for copyright issues regarding the collection of sentence prompts submitted by representatives before its launch.

References

Eneko Agirre, Izaskun Aldezabal, Iñaki Alegria, Xabier Arregi, Jose Mari Arriola, Xabier Artola, Arantza Díaz de Ilarraza, Nerea Ezeiza, Koldo Gojenola, Kepa Sarasola, and Aitor Soroa. 2021. Developing language technology for a minority language: Progress and strategy. *ELSNews. 10.1*.

Frances Ajo, Valérie Guérin, Ryoko Hattori, and Laura C. Robinson. 2010. *Native speakers as documenters A student initiative at the University of Hawai'i at Manoa*, chapter 19. John Benjamins, Berlin.

Pius W. Akumbu. 2024. A community approach to language documentation in africa. In *ACAL in SoCAL: Selected papers from the 53rd Annual Conference on African Linguistics*, page 1–25.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, page 4211—4215.

Claire Babirye, Joyce Nakatumba-Nabende, Andrew Katumba, Ronald Ogwang, Jeremy Tusubira F., Jonathan Mukiibi, Medadi Ssentanda, Lilian D. Wanzare, and Davis David. 2022. Building text and speech datasets for low resourced languages: A case of languages in east africa. *AfricaNLP 2022*.

Mehdi Bagherzadeh, Andrei Gurca, and Rezvan Velayati. 2023. Crowdsourcing routines: the behavioral and motivational underpinnings of expert participation. *Industrial and Corporate Change*, 32(6):1393–1409.

John Bellamy. 2021. *Contemporary Perspectives on Language Standardization*, chapter 26. Cambridge University Press, Cambridge, UK.

Henry Brice, Benjamin Zinszer, Danielle Kablan, abrice Tanoh, Konan N. N Nana, and Kaja K Jasińska. 2024. Individual differences in leveraging regularity in emergent l2 readers in rural côte d'ivoire. *Scientific Studies of Reading*, 28(4):391–410.

Alena Butryna, Shan-Hui C. Chu, Isin Demirsahin, Alexander Gutkin, Linne Ha, Fei He, Martin Jansche, Cibu Johny, Anna Katanova, Oddur Kjartansson, Chenfang Li, Tatiana Merkulova, Yin M. Oo, Knot Pipatsrisawat, Clara Rivera, Supheakmungkol

Sarin, Pasindu de Silva, Keshan Sodimana, Richard Sproat, Theeraphol Wattanavekin, and Jaka A. Eko Wibawa. 2020. Google crowdsourced speech corpora and related open-source resources for low-resource languages and dialects: An overview. *arXiv preprint*.

Arienne Dwyer. 2010. *Models of successful collaboration*, chapter 13. John Benjamins, Berlin.

Lenore A. Grenoble. 2010. *Language documentation and field linguistics: The state of the field*, chapter Conclusion. John Benjamins, Berlin.

John Gumperz. 1968. The speech community. *International Encyclopedia of the Social Sciences*, pages 381–386.

Nikolaus P Himmelmann. 2006. *Language documentation: What is it and what is it good for?*, page 1–30. Mouton de Gruyter, Berlin.

Adrian Kwek. 2020. Crowdsourced research: Vulnerability, autonomy, and exploitation. *Ethics & Human Research*, 42(1).

Judith M. Maxwell. 2010. *Training graduate students and community members for native language documentation*, chapter 18. John Benjamins, Berlin.

Emmanuel Ngue Um. 2019. *Achieving sustainable language preservation through economic empowerment in endangered language settings in West Africa*, chapter 20. Rüdiger Köpper Verlag, Cologne.

Alan W Black Perez Ogayo, Graham Neubig. 2022. Building african voices. *arXiv preprint*.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling speech technology to 1,000+ language. *Journal of Machine Learning Research*, pages 1–52.

David Roberts, Ginger Boyd, and JeDene Reeder. 2021. *Elip, Mmala and Yangben*, chapter 9. John Benjamins Publishing Company, Amsterdam.

Towards a Hän morphological transducer

Maura O'Leary[1], Joseph Lukner[2], Finn Verdonk[2], Willem de Reuse[3], Jonathan Washington[2]
[1]Western Washington University
[2]Swarthmore College
[3]The Language Conservancy
Correspondence: Maura.O'Leary@wwu.edu, Jonathan.Washington@swarthmore.edu

Abstract

This paper presents work towards a morphological transducer for Hän, a Dene language spoken in Alaska and the Yukon Territory. We present the implementation of several complex morphological features of Dene languages into a morphological transducer, an evaluation of the transducer on corpus data, and a discussion of the future uses of such a transducer towards Hän revitalization efforts.

1 Introduction

In this paper, we present work towards a morphological transducer for the Dene language Hän. The paper provides background on Hän, data collection, and morphological transducers (§2); overviews decisions made during implementation as well as our approaches to various challenges presented by Hän morphology (§3); and offers a preliminary evaluation (§4), directions for future work (§5), and some concluding thoughts (§6). The eventual goal is for this transducer to complement ongoing revitalization efforts.

2 Background

2.1 Hän

Hän (ISO 639-3: `haa`) is a Dene (more specifically, Northern Athabaskan) language spoken in the Native Village of Eagle in Alaska, USA, and in Moosehide, Yukon Territory, Canada. Hän is a critically endangered language, with only five remaining native speakers. While the number of native speakers is low, the communities in both Eagle and Moosehide are both engaged in significant revitalization efforts, including locally taught introductory language courses, language teacher training, and the creation of learning materials (lessons, textbooks, flashcards, etc.).

The primary complication in the process of learning (and thereby also in the process of revitalizing) an Athabaskan language is the rather complex verbal morphology. Verbs often surface with a string of both derivational and inflectional prefixes, which can be difficult for speakers of less-inflecting languages such as English. The complexity of Hän verbs stands in stark contrast to every other lexical category, which are at most bimorphemic.

In order to progress the community's revitalization efforts, there is a clear need for an understanding of Hän's verbal morphology. Understanding the inner workings of verbs has been a long-standing battle for many Athabaskan languages (see Rice, 2000, for an overview of many of the relevant works), and Hän is no exception. We intend for this transducer, and the resources which stem from it, to clarify the inner workings of the Hän verb as an aid to future Hän language learners.

Table 1 presents the structure of verbs in Hän, with some example verb forms broken down accordingly in Table 2. Each cell represents a distinct morpheme "slot". Many of the slots are optional—a valid verb form must contain at a minimum a stem marked for aspect and a subject marker. However, some verbs additionally require other elements, such as a theme or disjunctive prefix. Additionally, several slots interact with one another; for example, generally subject morphology is indicated in the slot before the stem, but plurality is indicated by a morpheme's occurrence in the "plural subject" slot for 3rd person plural and another morpheme's occurrence in the "deictic subject" slot for 1st person plural. Object marking and the presence of a reflexive morpheme appear to be mutually exclusive. 3rd person singular object markers vary depending on the person features of the subject (Lehman and O'Leary, 2019). Object marking is used only if an overt object DP is not present in situ in the verb phrase (Manker, 2014). Additionally, the specific form of subject marking depends on the classifier (l, ɬ, 0, or d), the aspect (imperfective, perfective, etc.), and the conjugation marking (0, dh, gh) associated with the verb stem (de Reuse and Las, 2014). Verb stems alternate irregularly for a given lexeme based on aspect and sometimes number of the subject (de Reuse, 2015b,a).

Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 91–99

10	9	8	7	6	5	4	3	2	1	0
(disjunctive prefix)	(pl. subj.)	(object)	(deictic subj.)	(reflexive)	(directive)	(future/ inceptive)	(gender/ qualifier)	(theme)	conjugation marker, subject, classifier	stem

Table 1: The structure of verbs in Hän, with numbers assigned to each prefix slot. The stem occurs at the end of a verb form, with prefixes stacking before it. Prefix slots that are not used in every verb form are described in ()s.

	10	9	8	7	6	5	4	3	2	1	0
a.	nä- ITER-		n- OBJ.2SG-			u- DIR-			n- THM-	ök- (ɬ/0) SUBJ.3SG-	gòt punch.IMPF
b.		hë- SUBJ.PL-						jë- GENDER-	n- THM-	èh- (ɬ/dh) SUBJ.3-	tlot boil.PERF

Table 2: Morphological breakdown of two example verb forms: (a) *nänunökgòt* 'I keep hitting you (sg) over and over again' and (b) *hëjënèhtlot* 'they boiled (a liquid)'. Numbers corespond to those for prefix slots in Table 1. Classifier and conjugation marker are specified in the gloss of the subject prefix (slot 1).

2.2 Language data and elicitation

The data used in this project comes primarily from in-person elicitation done by the fourth author between 2006 and 2012 (de Reuse, 2015b) and, to a lesser extent, from in-person elicitation done by the first author between 2016 and 2022. (As is discussed in §4.1, short stories written by one of the speakers are also used to test coverage and build the lexicon.) In addition to descriptive fieldwork, both the first and fourth authors have also been involved directly in Hän revitalization efforts since 2017, with projects yielding in-person language workshops and physical language learning materials (flashcards, language games, a phrasebook, and a short textbook). As revitalization efforts continue, the first author remains in close contact with the Eagle Village Chief—who is also daughter to one of the remaining speakers and niece to two others—so that all efforts can be made to fit the desires and needs of the language learning community. §5 discusses the potential future uses of a Hän morphological transducer in the revitalization process, which the community has shown great excitement for.

2.3 Finite-state transducers

A morphological transducer is a finite-state model of a language's morphology such that valid forms of a language receive one or more analyses (morphological analysis), and a valid form of a language is output when an analysis is input (morphological generation), as illustrated in Figure 1.

A finite-state transducer can be a useful tool for a marginalized language for a number of reasons. Most directly, it can be used for linguistic research and analysis of texts. It can also expand access to language

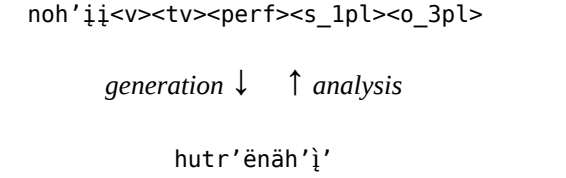


Figure 1: An example of morphological analysis and generation, as different directions in the mapping between an analysis (`noh'ɨɨ<v><tv><perf><s_1pl><o_3pl>`) and a form (`hutr'ënäh'ì'`). The example translates roughly as 'We saw them.'

technology, a crucial element for vitality of a language in the 21st century (Kornai, 2013), including as a core component of tools such as machine translation systems and spell checkers (Khanna et al., 2021). Additionally, a finite-state transducer can be useful for language revitalization as a component of pedagogical tools, such as Computer-Assisted Language Learning tools (Snoek et al., 2014; Katinskaia et al., 2018; Ivanova et al., 2019), word-form creators (Fernald et al., 2016; Kazantseva et al., 2018), and paradigm generators.[1] It is our intention to move to integrating the present transducer into any of these tools that the Hän community might find useful once the transducer is mature enough.

3 Implementation

One major challenge presented by Dene languages for development of a morphological transducer is the fact that the verb morphology is complex (§2.1) and almost entirely prefixational. Morphological analy-

[1] A prototype paradigm generator using transducers is available at https://apertium.github.io/apertium-paradigmatrix/ with source code at https://github.com/apertium/apertium-paradigmatrix.

ses of the type returned by transducers are usually organised in a suffixational order: `lemma`, `POS tag`, `subcategory tag(s)`, `grammatical tag(s)`, e.g., `noh'ɨɨ<v><tv><perf><s_1pl><o_3pl>`, where `noh'ɨɨ` is the lemma, `<v>` represents the part of speech (verb), `<tv>` represents the subcategory (transitive verb), and `<perf><s_1pl><o_3pl>` constitute grammatical tags (perfective aspect, first-person plural subject, third-person plural object). This order is much easier to implement when subsequent grammatical tags match the order of added [suffixational] morphology and occur after the stem; formalisms that rely on continuation lexicons, such as `lexc`, fail to offer a straightforward solution for non-suffixational morphology. For such languages, including Dene languages, a combination of several approaches is used to circumvent these limitations: the use of flag diacritics, intricate continuation lexicons, and collapsing intricate verbal morphology into simplified "zones" (Harrigan et al., 2017; Arppe et al., 2017; Holden et al., 2022). The main disadvantages of these approaches seem to be cleanliness of code (and hence maintainability) as well as transducer size and compilation and runtime speed.

To get around these limitations of previous approaches, the `lexd` formalism and compiler (Swanson and Howell, 2021) was used to implement a model of Hän morphology. The `lexd` formalism was designed to handle non-suffixational morphology efficiently, and has proven effective for other languages which make use of non-suffixational morphology (Washington et al., 2021; Christopherson, 2023).

We use the Apertium framework (Forcada et al., 2011; Khanna et al., 2021) for compilation scripts and other features and HFST format and tools (Linden et al., 2011) for storing and working with the compiled transducer, and adhere closely to the Apertium tagset standards.[2]

The remainder of this section reports on the implementation of the lexicon (§3.1), aspectual verb stem alternations (§3.2), and distributed morphology (§3.3), as well as how we deal with spelling variation and tone spreading (§3.4), and an initial foray into implementing a guesser (§3.5).[3]

[2]Described at `https://wiki.apertium.org/wiki/List_of_symbols`.

[3]Source code is available under a free/open license at `https://github.com/SwatLangTech/apertium-haa/`. All reports of code and performance are based on the latest code at time of submission: revision `b334130`, dated 2025-01-17.

3.1 Lexicon

We have mostly focused our efforts on implementing verbal morphology. Other parts of speech have been included in the lexicon to "clear out" the list of top unanalysed forms over corpora so that verb forms become more visible for additional morphology work. A first stab at non-verbal morphology, which is limited in Hän to pronominal possessor prefixes on nouns and pronominal prefixes on prepositions denoting indirect objects, has been implemented. The number of stems of various types are listed in Table 3.

part of speech	unique	total
nouns	167	183
verbs	15	64
adjectives	18	20
prepositions	15	17
adverbs	6	8
conjunctions	3	4
modal words, determiners, pronouns, numerals, anthroponyms, etc.	22	23
total	246	319

Table 3: The number of stems of various parts of speech: unique excludes spelling variants or context-dependent stems; total is the total number of entries in each lexicon.

Uninflected verb stems in Hän are never uttered in isolation, and verbs have different stems depending on their patterning with aspectual morphology, so verb lemmas must inherently be inflected for subject and aspect. We originally selected the 3rd person singular imperfective form of a verb as its lemma for morphological reasons—primarily that this form is also used as a base on which the 1st person plural and 3rd person plural forms are built, and thus is present in three of six person/number combinations. However, recent speaker judgments suggest that the 1st person singular imperfective form feels like a more appropriate label for the verb, so we will be transitioning to a 1st person singular imperfective lemma system.

3.2 Aspectual verb stem alternations

Verb stems in Hän take different forms depending on the aspect marker they pattern with, as well as (in some cases) whether the subject is singular or plural. Since these alternations are unpredictable, they could not easily be encoded as phonological alternations. Instead, we implemented these alternations using filter tags, a feature of `lexd`. An example is provided in

Code Block 1.

Additionally, subject markers take different forms based on the classifier and conjugation marker associated with the verb. These alternations are also unpredictable and could not be treated as phonology. In this case also we used filter tags to match verb entries to the appropriate set of subject markers.

The result is that there are currently 173 entries in the subject lexicon (excerpt in Code Block 2), which includes the morphology for all person categories matched to each combination of classifier and conjugation marker, as well as variant forms.

Besides indexing the relevant tags in each entry of each lexicon, tags must be matched at the level of the pattern (pattern example with tags shown in Code Block 3).

3.3 Distributed morphology

The implementation of the transducer needed to model the distribution of subject morphology across three slots of the verb structure (Table 1). This was done by making multi-column lexicons for verb morphology, as shown in Code Block 1. The verb lexicon currently includes four columns: one for disjunct prefixes associated with the given verb, one for the directive prefix associated with the given verb, one for the theme prefix associated with the given verb, and one for the stem. This treats the lexical entries for verbs as consisting of all four parts.

The different columns and associated morphology (e.g., Code Block 2) are referenced from a pattern that follows the structure of verbs in Hän. The pattern for transitive verbs currently in the transducer is shown in Code Block 3. This pattern does not yet implement the disjunctive prefix, or the gender/qualifier slot.

3.4 Spelling variation and tone spreading

There are a number of challenges for analysis related to orthography.

First of all, due to the small number of remaining speakers, as well as inconsistencies among our data sources, tokens of the same word often vary in spelling. We add variant forms of an entry to the lexicon in a way where only one variant (the one determined to be canonical) is included in the generator, but all variants are included in the analyser. This is done by simply including a control sequence (`Dir/LR`, a convention established by the Apertium community) in the comments of all but the canonical form in the `lexd` file, and including code in our compile script to strip all lines containing that control sequence when compiling the generator. Currently there are 56 instances of this control sequence in the transducer code.

Additionally, the tone system of Hän features interlexical tone spreading: if the last (or only) syllable of a word has a low tone, this low tone can spread to the following syllable of a subsequent word, if that syllable is not then followed by another low tone (Lehman, 2018). Notably, this spreading skips over schwas. In many instances however, this standard is not strictly adhered to in the orthography. Practically, this means that the first non-schwa vowel of a token may be written with an otherwise unexpected low tone (e.g., à for expected ä).

A related challenge is the differing encodings of various characters. For example, the 'ą̀' character may be encoded as the character 'a', followed by a combining ogonek, followed by a combining diaeresis, followed by a combining grave (which we treat as the canonical encoding).[4] However, it may also be rendered with any order of combining diacritics, or with a precomposed character (such as 'ä' or 'ą') with only the additional diacritics added as combining characters (again, in any order). Normally the transducer will only recognise characters in the particular encoding that material is entered with, and not visually similar characters with different encodings.

To overcome regular spelling alternations, differing encodings, and the possibility of an additional low tone, we implemented a layer of "spellrelax" rules (which allow for alternative spellings), implemented as a list of `foma`-style rules (each its own mini transducer). Each rule allows alternate character sequences for a given canonical character sequence, and the combined ruleset is compose-intersected with the base transducer to create the final analyser. An example of two spellrelax rules is provided in Code Block 4. Currently there are 28 implemented spellrelax rules for the Hän transducer.

3.5 Guesser

By leveraging the morphological patterns of the transducer and a regular expression, a transducer may be used as a guesser. A guesser is a transducer which analyzes forms of stems which are not part of the transducer's lexicon. The output when analyzing such a

[4]The canonical order in the transducer is based on Unicode NFKD (Normalisation Form Compatibility Decomposition); we do not perform the additional composition required of NKFC (NFKD, followed by Canonical Composition) in order to maintain compatibility with the Hän keyboard available from the Yukon Native Language Centre (`https://ynlc.ca/fonts-keyboards/`), and so that the low-tone diacritic may be directly manipulated in cases of tone sandhi.

```
LEXICON VerbStem-Iv(4)

[0cl,impf,0cm,sg]:  [0cl,impf,0cm,sg]:  [0cl,impf,0cm,sg]:n>  nähaa:haa[0cl,impf,0cm,sg]
[0cl,perf,n,sg]:    [0cl,perf,n,sg]:    [0cl,perf,n,sg]:n>    nähaa:zhaa[0cl,perf,n,sg]
[0cl,fut,0cm,sg]:   [0cl,fut,0cm,sg]:   [0cl,fut,0cm,sg]:n>   nähaa:haw[0cl,fut,0cm,sg]
[0cl,perf,n,pl]:    [0cl,perf,n,pl]:    [0cl,perf,n,pl]:n>    nähaa:jeww[0cl,perf,n,pl]
[0cl,fut,0cm,pl]:   [0cl,fut,0cm,pl]:   [0cl,fut,0cm,pl]:n>   nähaa:däẅ[0cl,fut,0cm,pl]
```

Code Block 1: An example of a verb entry for the verb *nähaa* 'go, come, arrive'. Filter tags are specified within `[]` and separated by commas, and are used to encode grammatical properties of the lines (e.g., `[0cl,impf,0cm,sg]` encodes 0-classifier, imperfect, 0-conjugation marker, singular). Columns (discussed in §3.3) are disjunct prefix (empty with this verb), directive (empty with this verb), thematic prefix (*n*-), and stem (varying by imperfective, perfective, future, as well as singular and plural). The plural imperfective stem is not in our data sources. Content outside filter tags is separated by colons: the left side contains elements of the analysis (e.g., in the last column containing the lemma, *nähaa*) and the right side contains elements of the form (e.g., the thematic prefix *n*- and the individual stems).

```
LEXICON subject(4)

[ɬ,impf,0cm,non3Ssub,sg]:         [ɬ,impf,0cm,non3Ssub,sg]:           [ɬ,impf,0cm,non3Ssub,sg]:ök>  [ɬ,impf,0cm,non3Ssub,sg]<s_1sg>:
[ɬ,impf,0cm,non3Ssub,pl]:         [ɬ,impf,0cm,non3Ssub,pl]:tr'{E}{¨}> [ɬ,impf,0cm,non3Ssub,pl]:oh>  [ɬ,impf,0cm,non3Ssub,pl]<s_1pl>:
[ɬ,impf,0cm,non3Ssub,pl]:h{E}{¨}> [ɬ,impf,0cm,non3Ssub,pl]:           [ɬ,impf,0cm,non3Ssub,pl]:oh>  [ɬ,impf,0cm,non3Ssub,pl]<s_3pl>:
```

Code Block 2: Some examples of entries in the subject lexicon. The first column provides content for plural marking in third person, the second column provides content for plural marking in first person, the third column provides remaining subject marking, and the fourth column provides the relevant morphological tags. Filter tags currently must be included in every column (a limitation of `lexd`), and in this case specify that this morphology patterns with ɬ-classifier verbs, imperfective aspect, a non-third-person-singular subject, and singular or plural subject (cf. Code Block 1).

```
(subject(1) object?(1) subject(2) object?(2) :VerbStem-Tv(2) aspect(1) VerbStem-Tv(3)
subject(3) [ :{NOV} ] VerbStem-Tv(4) [ <v><tv>: ] VerbStem-Tv(2): aspect(2) subject(4)
object?(3))[^[3Ssub,non3Ssub],^[impf,perf,incp,fut,opt],^[sg,pl],^[l,d,0cl,ɬ],^[0cm,dh,gh,n]]
```

Code Block 3: The current pattern for transitive verbs (no line breaks in the transducer entry). The elements before the verb stem (`VerbStem-Tv(4)`) reference the content of the various prefix slots. Material after the anonymous lexicon that consists of `<v><tv>` tags reference grammatical tags matching the prefixes, as well as the filter tags used to match elements of lexicons to one another.

```
.o. [ ?* [ ¸¨(->) ¨¸] ?* ]
.o. [ ?* [ i ¸(->) į ] ?* ]
```

Code Block 4: Two spellrelax rules currently used in the transducer. The `.o.` character is the compose operator, to compose each rule with the other rules. The first example allows either order of ogonek and diaeresis combining diacritics. The second rule allows a precomposed 'į' character for what is encoded in the transducer as an 'i' character followed by a combining ogonek diacritic.

form is the stem in place of the lemma, a full analysis, and information about the paradigm the form was successfully analyzed using.

Initial attempts at a guesser were implemented for some of Hän's verbal morphology by adding wildcard entries to the verb lexicon (excluded from normal compilation) with filters matching each of Hän's four verb classifiers with the zero conjugation marker. These four patterns were repeated twice, once with no thematic prefix, and once with an *n* thematic prefix, for a total of 8 entries. (Quite a few more would be needed for a complete set of entries.) An example is shown in Code Block 5.

An example of output from the guesser is shown in Code Block 6, using the example *shënähtthee* 'you all are barking at me' (the verb stem of which is not in the transducer). The returned set of analyses includes the correct one (`<GUESSER_ł_0cm_nthm>tthee<v><tv><impf><s_2pl><o_1sg>`), correctly revealing that the input token is a second-person plural subject form of an imperfective stem *tthee* of an ł-classifier verb with an *n* thematic prefix and a first-person singular object. However, other analyses are returned as well.

The guesser often returns a 3rd singular imperfective of a ∅-classifier verb. This is due to the fact that the 3rd singular imperfective subject prefix is null for ∅-classifier verbs. The entirety of the input form is then guessed as the root. Removing these extraneous analyses was done by implementing `twol` rules (Code Block 7) that restrict the possibilities for guessed roots. No verb roots in the language appear to begin with a vowel or 'h' or 'n' followed by a consonant. Additional work is needed to further restrict the options, perhaps by prioritising more complex ones using weights.

4 Evaluation

The transducer was evaluated for naïve coverage (§4.2) using available texts and elicitation sentence data (§4.1) and on its runtime and space requirements (§4.3).

4.1 Corpora

The coverage of the transducer was evaluated against several texts. The first set of texts come from two collections of short stories written by native speaker Ruth Ridley (Ridley, 1983, 2018), totaling ~3.3k tokens. The stories were manually transcribed (with some augmentation using OCR) to ensure accuracy and proper encoding.

The second set of text came from elicited sentences accompanying verb paradigms in de Reuse (2015b). Sentences were extracted using a script to filter out English, author comments, organizational codes, and Hän data that was not in sentence format. After filtering, the document contained ~11.5k words.[5]

4.2 Naïve coverage

Naïve coverage was measured as the raw percentage of tokens that were analyzed by the transducer, regardless of accuracy. Coverage numbers are shown in Table 4.

The higher coverage numbers on the stories corpus can be accounted for by several factors. First of all, the elicited sentences include a full range of verbs in the language, as opposed to handful of common and domain-specific verbs as in the stories. Additionally, the stories include common nouns, prepositions, and other uninflected parts of speech that are much less common in the sentences corpus (and which were easily included in the transducer lexicon).

Other reasons the sentences corpus has lower coverage include that (1) there was minimal punctuation in the corpus, especially since the sentences did not include sentence-final punctuation; (2) there were many words with differently encoded symbols (using private-use-area code points, presumably for a custom font) which we have not yet integrated into spellrelax; and (3) this corpus contains examples from multiple speakers and dialects, and much of the attested variation has not yet been incorporated into the transducer.

Overall, the verb paradigms were the principal source of data for implementing the transducer lexicon, so it is a good sign that it does analyze a large portion of the examples in the data. Coverage on this corpus can be increased by adding more verb stems to the lexicon (the existing morphology should be robust enough to support most cases), implementing more spellrelax rules to account for differences in encoding and orthography, and including more phonolog-

[5]There are ~4.5k sentences; i.e., they are on average very short.

```
[ɬ,0cm]:  [ɬ,0cm]<GUESSER_ɬ_0cm_nthm>:   [ɬ,0cm]:n>   /([a-z'\̃\̃])+/[ɬ,0cm]
[l,0cm]:  [l,0cm]<GUESSER_l_0cm_nthm>:   [l,0cm]:n>   /([a-z'\̃\̃])+/[l,0cm]
```

Code Block 5: Two guesser entries in the verb lexicon: one for l-classifier verbs and one for ɬ-classifier verbs. The columns match those in Code Block 1; an *n* thematic prefix is included in the third column. The regular expression in the fourth column occupies both sides of the separator, so the transducer includes the matching stem on both the analysis and the morphological side. For this reason, the guesser tag must be included in a different column than (and hence occurs before) the stem.

```
<GUESSER_0cl_0cm>nähtthee<v><tv><impf><s_3sg><o_1sg>
/<GUESSER_0cl_0cm_nthm>tthee<v><tv><impf><s_2pl><o_1sg>
/<GUESSER_d_0cm_nthm>tthee<v><tv><impf><s_2pl><o_1sg>
/<GUESSER_ɬ_0cm_nthm>tthee<v><tv><impf><s_2pl><o_1sg>
```

Code Block 6: Analyses returned by the guesser given the input *shënähtthee* 'you all are barking at me' (a verb form whose stem is not in the transducer). The correct analysis is the fourth one, highlighted in bold for presentation purposes.

```
"restrict guessed forms with vowel-initial stems"
Vowel:Vowel /<= %{NOV%}: _ ;

"no hC- or nC- initial stems guessed by guesser"
C1:C1 /<= %{NOV%}: _ Cons:Cons ;
         where C1 in ( h n ) ;
```

Code Block 7: `twol` rules that restrict guesser possibilities. `{NOV}` ("no vowel") is a control symbol included in the transitive verb pattern before the stem (Code Block 3). The `/<=` operator excludes from the compiled transducer any path matching the pattern.

corpus	tokens	ambiguity	coverage
stories	3 275	1.08	60.40%
elicitation data	11 479	1.10	21.87%

Table 4: Naïve coverage results by corpus. Corpus size is presented in number of tokens as determined by the analyser. Ambiguity (average number of analyses per form) is also included.

ical rules to better predict the morphophonology of long sequences of prefixes.

4.3 Size and speed

As of publication, the generator has 19 824 states and 23 105 arcs and a non-cyclical expansion of the generator[6] yields 4 286 analysis-form pairs, taking approximately 280ms to expand on a 3.5GHz Intel i9-9900X CPU, and running a simple coverage script on the 3.3k-token stories corpus takes approximately 125ms.[7] The compiled generator is 367kB the compiled analyser is 859kB and the compiled guesser is 6.7MB

Compilation of the entire transducer—including morphology, morphophonology, guesser, and spellrelax—using a single thread on a 10-core 3.5GHz Intel Core i9-9900X CPU takes approximately 30 seconds total and uses a maximum of 652MB of RAM. Use of additional threads brings compile time down to around 14 seconds.

While these are encouraging numbers given the complexity of the existing morphology, it is difficult to know how size and speed will scale as the lexicon is expanded and additional morphology is added.

5 Next Steps

The most pressing next steps are to continue to expand the transducer in all ways, including lexicon, morphology, and phonological alternations.

The primary motivation for creating this transducer is pedagogical. Specifically, we envision the transducer's use in tools that can be used by language learners, such as a verb-form generator, a paradigm generator, or a translator working at the sentence level rather than the word level (examples for other languages cited in §2.3). Such resources would be incredibly valuable to Hän language learners, many of whom do not have the opportunities for frequent contact with the few remaining speakers. Existing revitalization materials, being limited to slide shows and printed physical materials, do not cover many verbs or full conjugation paradigms. Hence any of these resources would be a significant addition to current revitalisation efforts, but would have to be built for use by non-technical audiences (e.g., avoiding linguistic terminology wherever possible). Community leaders have expressed excitement at the prospect of materials like these becoming available to the community.

As with all resources created for Hän, prototypes will be presented to the Hän community to allow their

[6] `hfst-expand -c0 haa.autogen.hfst`

[7] All data files and utilities are stored on a 2019-era Samsung 970 Pro NVMe SSD.

preferences to guide resource development, so that the resulting resources are only those that are deemed beneficial by the speakers and learners themselves.

Finally, we also plan to account for systematic spelling and vocabulary differences found between the the Eagle (Alaska) and Moosehide (Yukon) dialects of Hän, so that any pedagogical resources produced will be equally accessible to both communities.

6 Conclusion

To our knowledge, we are publishing the first morphological transducer for a Dene language written in `lexd`. Not only have we shown that it is possible to implement Dene morphology in `lexd`, but that it has many advantages over previous approaches to Dene morphology using `lexc` (see §3): the code is much cleaner (and hence the transducer is more easily maintained and expanded), and the resulting transducer is small and its compilation and runtime speeds are fast. Our hope is that an efficient transducer will allow us to create helpful and easy-to-use language resources to aid the revitalization of the Hän language.

Acknowledgments

Thanks first and foremost go to the speakers of Hän who have shared their language with us over the past 20 years, including many who are no longer with us. Speakers involved in the data used in this project are, alphabetically: Angie Joseph-Rear, Adeline (Juneby) Potts, Archie Roberts, Bertha Ulvi, Charlie Silas, Charlie Stevens, Danny David, Doris Roberts, Edith Josie, Edward Roberts, Ethel Beck, Eliza Malcolm, Geoffrey O'Grady, Harry David, Jr., Isaac Juneby, Joseph Susie Joseph, Louise Paul, Matthew Malcolm, Percy Henry, Richard Nukon, Richard Silas, Ruth Ridley, Sarah Malcolm, Stanley Roberts, Susie Paul, Timothy Malcolm, and Willie Juneby. We also thank linguists whose work with Hän has helped us at various stages, including Michael Krauss, Jordan Lachler, Blake Lehman, Gordon Marsh, John Ritter, and David Shinen, as well as community members Georgette McLeod and Eagle Chief Karma Ulvi, who have been heavily involved in the revitalization of Hän. We also recognize work done by research assistants Ryan Baldwin and T Sallie to expand the lexicon of the transducer, and thank the ComputEL-8 organizers and four anonymous reviewers for their feedback.

References

Antti Arppe, Christopher Cox, Mans Hulden, Jordan Lachler, Sjur N. Moshagen, Miikka Silfverberg, and Trond Trosterud. 2017. Computational modeling of verbs in Dene languages: The case of Tsuut'ina. *Working Papers in Athabascan Linguistics*.

Cody Scott Christopherson. 2023. A finite-state morphological analyzer for Q'eqchi' using Helsinki Finite-State Technology (HFST) and the Giellatekno infrastructure. Master's thesis, Brigham Young University.

Willem de Reuse. 2015a. A guide to the Hän verb paradigms. Unpublished ms.

Willem de Reuse. 2015b. Hän Athabascan verb paradigms. Unpublished ms. Fairbanks: Alaska Native Language Archives, University of Alaska.

Willem de Reuse and Kathy Joy Las. 2014. Han verb prefix paradigms: digitized and rechecked version of field notes from Jeff Leer (1974, 1977, 1980) and Leer and Ridley (1982). Unpublished ms.

Theodore B. Fernald, Nabil Kashyap, and Jeremy Fahringer. 2016. Navajo verb generator. Development version.

Mikel L. Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Atticus G. Harrigan, Katherine Schmirler, Antti Arppe, Lene Antonsen, Trond Trosterud, and Arok Wolvengrey. 2017. Learning from the computational modelling of Plains Cree verbs. *Morphology*, 27:565–598.

Joshua Holden, Christopher Cox, and Antti Arppe. 2022. An expanded finite-state transducer for Tsuut'ina verbs. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5143–5152, Marselha, França. European Language Resources Association.

Sardana Ivanova, Anisia Katinskaia, and Roman Yangarber. 2019. Tools for supporting language learning for Sakha. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 155–163, Turku, Finland. Linköping University Electronic Press.

Anisia Katinskaia, Javad Nouri, and Roman Yangarber. 2018. Revita: a Language-learning Platform at the Intersection of ITS and CALL. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Anna Kazantseva, Owennatekha Brian Maracle, Ronkwe'tiyóhstha Josiah Maracle, and Aidan Pine. 2018. Kawennón:nis: the wordmaker for Kanyen'kéha. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 53–64, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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*, 35:475–502.

András Kornai. 2013. Digital language death. *PLoS ONE*, 8.

Blake Lehman. 2018. Tone-prominence interaction in Hän. Master's thesis, University of California, Los Angeles.

Blake Lehman and Maura O'Leary. 2019. Unexpected Athabaskan pronouns. *UCLA Working Papers: Schuhschrift: Papers in Honor of Russell Schuh*, pages 122–137.

Krister Linden, Miikka Silfverberg, Erik Axelson, Sam Hardwick, and Tommi Pirinen. 2011. *HFST—Framework for Compiling and Applying Morphologies*, volume 100 of *Communications in Computer and Information Science*, pages 67–85. Springer.

Jonathan Manker. 2014. The syntax of sluicing in Hän. In *Proceedings of the 2012 Athabascan Languages Conference.*, Fairbanks, AK: Alaska Native Language Center.

Keren Rice. 2000. *Morpheme Order and Semantic Scope: Word Formation in the Athapaskan Verb*. Cambridge University Press, Cambridge.

Ruth Ridley. 1983. *Eagle Han Huch'inn Hòdök (Stories in Eagle Han Huch'inn)*. Alaska Native Language Center, University of Fairbanks, Fairbanks, Alaska.

Ruth Ridley. 2018. Hän children's stories. Unpublished manuscript.

Conor Snoek, Dorothy Thunder, Kaidi Lõo, Antti Arppe, Jordan Lachler, Sjur Moshagen, and Trond Trosterud. 2014. Modeling the noun morphology of Plains Cree. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 34–42, Baltimore, Maryland, USA. Association for Computational Linguistics.

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.

Jonathan Washington, Felipe Lopez, and Brook Lillehaugen. 2021. Towards a morphological transducer and orthography converter for Western Tlacolula Valley Zapotec. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 185–193, Online. Association for Computational Linguistics.

Multilingual MFA: Forced Alignment on Low-Resource Related Languages

Alessio Tosolini
Yale University
Department of Linguistics
alessio.tosolini@yale.edu

Claire Bowern
Yale University
Department of Linguistics
claire.bowern@yale.edu

Abstract

We compare the outcomes of multilingual and crosslingual training for related and unrelated Australian languages with similar phonological inventories. We use the Montreal Forced Aligner to train acoustic models from scratch and adapt a large English model, evaluating results against seen data, unseen data (seen language), and unseen data and language. Results indicate benefits of adapting the English baseline model for previously unseen languages.

1 Introduction

Forced Alignment (the matching of textual annotations with audio and/or video data, particularly at the level of phonological segments) is a very useful step in language analysis. Software such as ELAN (Wittenburg et al., 2006) allows straightforward (but mostly manual) transcription and alignment at the granularity of utterances. Alignment algorithms such as the Montreal Forced Aligner (McAuliffe et al., 2017) take utterances and align them at the level of words and segments, allowing a much greater array of analytical possibilities.

Forced Alignment requires an acoustic model and information about the mapping between the transcription system and the phonemes in the language (g2p). Acoustic models require training data, and the paucity of available materials for low-resource languages leads to lower model performance. Low-resource language materials are disproportionately created in naturalistic environments (outside quiet, controlled lab settings) and so in addition to having smaller amounts of data, the data that is there may be disproportionately difficult to work with.

Various methods exist for increasing performance, including a) using a very high resource language (mostly English) and adapting phoneme mappings to the high resource language; b) adapting a high-resource language model; c) using a closely related high-resource language model; d) using pretrained spoken term detection to identify particular words (San et al., 2021); or e) training a language-specific model despite small amounts of data and correcting manually. Chodroff et al. (2024) compared these techniques and found that for small amounts of data (under approximately 25 minutes for their Urum and Evenki datasets), large cross-language and language-specific acoustic models were effective, but where the amount of low-resource data is larger than about 25 minutes, a model trained on that data is as effective. Findings by San et al. 2024 show that crosslingual transfer from models, as one might expect, is more effective when the languages are phonologically similar.

For forced aligning Australian Indigenous corpus data, however, the question is somewhat different. In this case, we have a large number of phonologically similar (Round, 2023) but small corpora, which vary by number of contributors, circumstances and dates of recording, and language phonotactics. Since the languages are phonologically (and perhaps phonetically; Fletcher and Butcher 2014; Tabain et al. 2016) similar, pooling data should lead to more robust and accurate alignment models. Conversely, since the languages differ in phonotactics (Macklin-Cordes et al., 2021) and comprise different speakers, the increase in heterogeneity may limit improvements in model performance. Moreover, since even pooling data does not make the model "large" by "large corpus" standards, it may still be preferable to use or adapt a large model.

For small corpora, overfitting is seldom a problem; model performance on the data at hand is often the sole criterion. In this case, however, we care about performance increases on both held-out data and held-out languages, as we will continue to develop the corpus and hope the release models for others working with Australian language data.

In this paper, we describe results of model train-

Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 100–109

Language	Language Family	Reference	Collector	Minutes
Bardi	Nyulnyulan	A: Bowern_C05	Claire Bowern	108
Gija	Jarrakan	E: 0098MDP0190	Frances Kofod	157
Kunbarlang	Gunwinyguan	E: 0384SG0324	Isabel O'Keefe; Ruth Singer	16
Ngaanyatjarra	Pama-Nyungan	P: WDVA1	Inge Kral	53
Yan-nhangu	Pama-Nyungan	E: dk0046	Claire Bowern	290
Yidiny	Pama-Nyungan	A: A2616	R.M.W. Dixon	50

Table 1: Corpus information. A: AIATSIS; E: Elar; P: Paradisec

ing and evaluation for 5 MFA acoustic models. While previous work (DiCanio et al., 2013; Johnson et al., 2018; Babinski et al., 2019) has compared different alignment methods, here (as in Chodroff et al., 2024) we focus on comparing different acoustic language models within MFA.

We find general agreement between all but the model trained on the smallest amount of data. Adapting the English model for a crosslingual Australian dataset improves performance on held-out languages more than for held-out data from languages already in the dataset. Measurements of vowel space are equivalent for all except the smallest model when applied to seen languages, but there is more variation when applying models to a new language. This suggests that similarity among Australian language phonetics should be further investigated.

2 Methods and Data

2.1 Datasets

Datasets for this paper were downloaded from non-restricted collections in the ELAR[1] and Paradisec[2] digital language archives, along with materials previously received from AIATSIS.[3] These materials are a subset of the collections which were used in Babinski (2022). Corpus references are in Table 1. The total amount of training data for the current study is roughly 10 hours, with individual languages ranging from 15 minutes to nearly 5 hours of audio.

Some of the materials used here were initially used to compare forced alignment algorithms in Babinski et al. (2019), and the full cleaned dataset was used for Babinski (2022). The data pipeline involved word-level segmentation with the p2fa forced alignment suite (based on HTK) and subsequent manual correction in Praat (Boersma and Weenink, 2021). Manual correction included realigning substantially misaligned segments (for example, segment boundaries placed in the wrong word) and moving boundaries placed where no human annotator would place them. In this paper, those manually reviewed files are the comparator against which we evaluate the accuracy of the force aligned files.However, we acknowledge that it is misleading to claim that there exists a single correct boundary between two phones due to smooth transition between phones that results from overlap (e.g. Liberman et al. 1967) and that even expert annotators vary in regard to where they place phone boundaries. Moreover, some segments (such as word-initial glottal stops) might not have any detectable onset boundary. We compare models to human annotated data but acknowledge that such datasets are themselves subject to further scrutiny.

The languages that form the basis of this comparison do not have identical phoneme inventories. They differ as to whether they have phonemic vowel length (or not) and whether they have two series of stops or one. For languages with two stop series, the contrast is between voicing, length, or perhaps tense/laxness (or some combination of these features).

2.2 Preparing Input

Since the audio data collected for this experiment comes from a variety of sources, we preprocessed the data to standardize it and to ensure that (a) the data is processed as expected by the various MFA models we created and trained and (b) all datasets created have the same formatting. This processing included the removal of partially transcribed words, cleaning the transcription tier of analytical comments, and some transcript adaptation (such as the removal of hyphens). TextGrids processed for model evaluation underwent additional processing to match it with the expected output of the MFA model, such as removal of words shorter

[1] www.elararchive.org
[2] www.paradisec.org.au
[3] mura.aiatsis.gov.au

than 0.1s in duration. We did not alter transcripts.[4] Two datasets were created for model training. One of them is the Yidiny-Train corpus, comprised of 38 minutes of audio data. The other is the Big5 dataset, comprised of the entirety of the Bardi, Gija, Ngaanyatjarra, and YanNhangu corpus, and the Yidiny-Train corpus. The Big5 has a total duration of 646 minutes. Three datasets were created for model evaluation. The first is the same as the Yidiny-Train corpus, and the second is the comprised of the remaining 12 minutes of Yidiny data the models didn't train on. The last test corpus is the Kunbarlang corpus, which no model has trained on.

It must be noted that although similar, the phonemic inventories of these languages are different enough to play a significant role in the alignments generated. Notably, Kunbarlang's phones are not a proper subset of Yidiny's, with the tense stops /p t ʈ c k/, mid-vowels /e o/, and the retroflex nasal /ɳ/ all present in Kunbarlang but absent in Yidiny. All of Kunbarlang's phones are present in at least one of the models in the Big5 training set, with the exception of the mid front vowel /e/ which is not present in any of the Big5 languages. The impact of these inventory asymmetries is discussed in the section on results.

2.3 Acoustic Models

Five acoustic models were used for this experiment. The first two of these acoustic models were trained from scratch on the Yidiny-Train and Big5 corpora. The remaining three models use the English MFA 3.1.0 acoustic model (McAuliffe and Sonderegger, 2024), which is trained on over 3,600 hours of global English. In order to use the English models for non-English data, we used the methodology described in Dolatian (2024). The English base model was used in its off-the-shelf form as one of the models we evaluated. The other two English-based models were created by adapting the English model to the Yidiny-Train and Big5 corpora.

The dictionary was created from the corpora by creating a single wordlist of all language data to be included and replacing graphemes in the language orthographies with equivalents from the International Phonetic Alphabet. Since all languages used phonemic transcription systems this was straightforward.

[4]These transcripts do not mark pauses or hesitations in the original. We did not review transcripts for accuracy beyond what was completed for earlier publications.

2.4 Eval

We evaluated models against two criteria in three different testing settings. The first criterion is precision, defined as the distance in milliseconds between the human annotated onset boundary and the MFA aligned interval's onset boundary, where a positive value indicates the aligned onset boundary is placed after the human annotated onset boundary. To check for both accuracy and precision, we look at the mean and standard deviation for these values, which we call "diffs".

The second criterion is analysis comparison. Since the aligned output of a model is used for phonetic analysis, we compare vowel charts created from these model against those created by the human annotated files. To test for accuracy and precision, we plot ellipses centered around the mean formant values for the data using the `matplotlib` package in Python (Hunter, 2007). Formants were extracted using `Parselmouth` (Jadoul et al., 2018) and measurements averaged across each vowel.

The three testing settings correspond to the three human annotated datasets described in the previous subsection. They are: Yidiny-seen (comprised of Yidiny data the model has trained on), Yidiny-unseen (comprised of Yidiny data the model has not trained on), and Kunbarlang (comprised of Kunbarlang data). Note that all models except English-base have trained on some amount of Yidiny data, and that no models have trained on any Kunbarlang data.

3 Results

3.1 Precision

Figures in this section present the rules of mean differences in onset alignment of segments in milliseconds, where a positive value means that the forced-aligned onset boundary has a greater timestamp than the human annotated onset boundary (i.e. it is further on in the file).

Figure 1 shows 15 histograms, where each row is a different model and each column is a different testing setting. The histograms only plot values in the range of [-205, 205] milliseconds, with the percentage in the top left equal to the percent of total tokens that were excluded from the histograms due to being out of this range. The number below represents the test tokens per testing setting that were in the range.

From this, we see that all diffs are approximately normally distributed with a mean near 0. Models

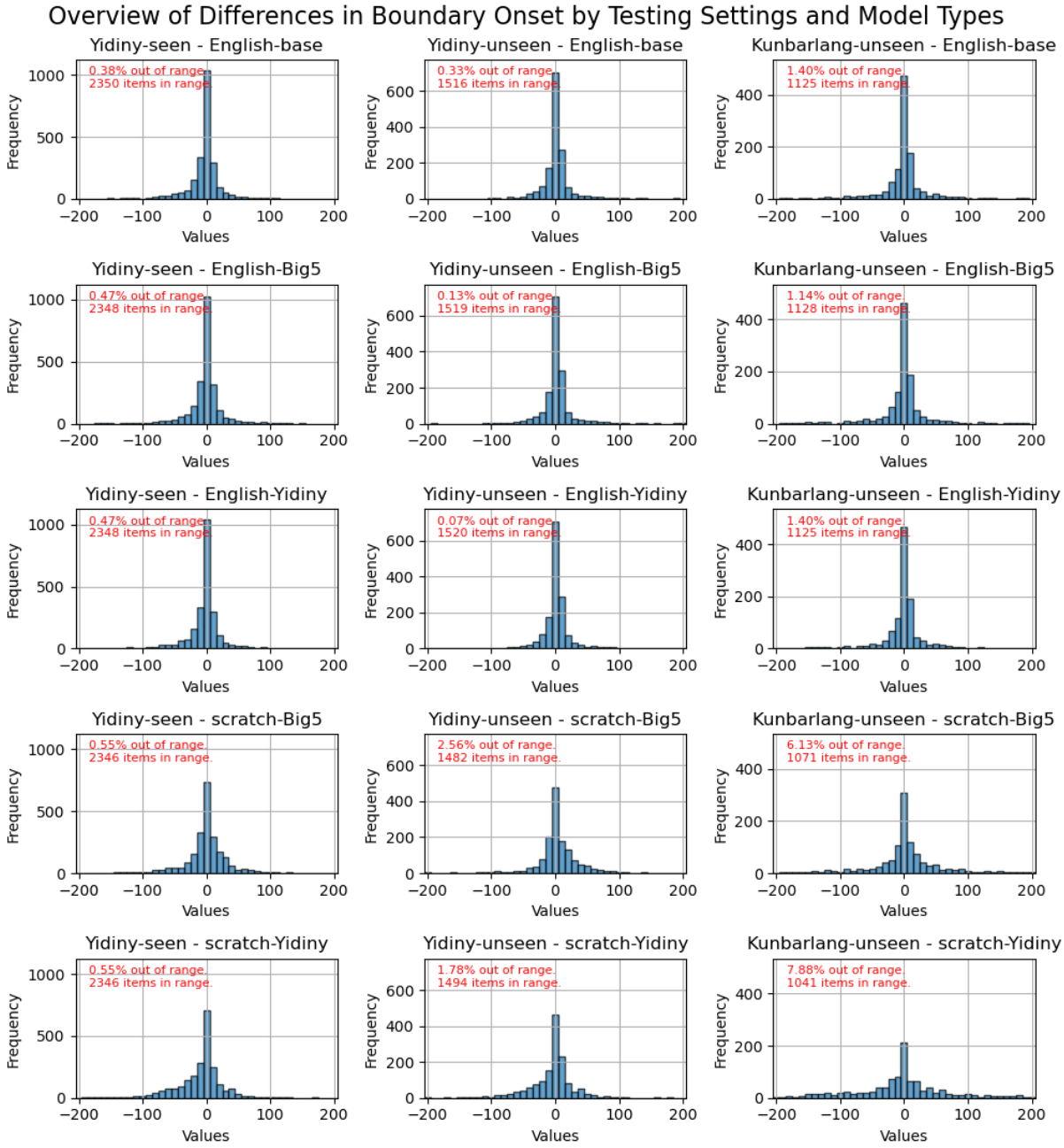


Figure 1: Onset boundary differences for all models across all testing settings.

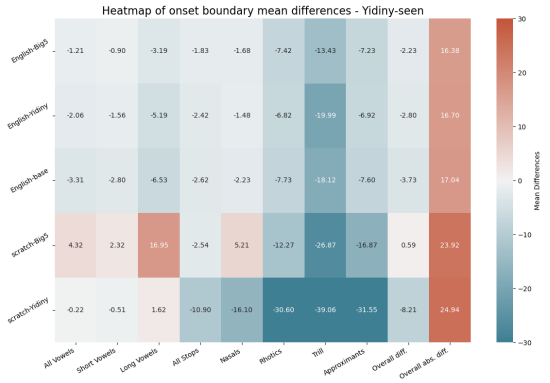


Figure 2: Yidiny seen data; mean precision

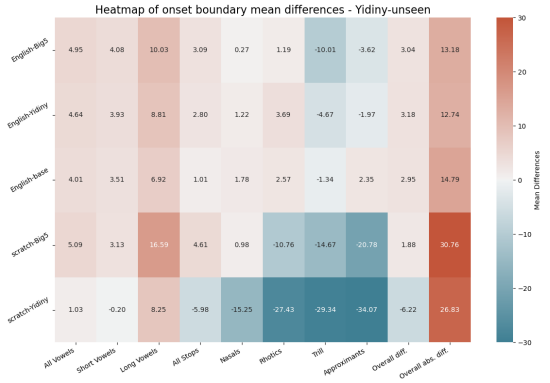


Figure 3: Yidiny unseen data; mean precision

trained from scratch have more spread, which is especially notable with the lower number of boundaries that differ from the human annotated boundaries by [-5, 5] ms. Although the histograms seem roughly symmetrical, there is tendency for English-based models tested on unseen data to place boundaries slightly ahead of the human annotated ones, as is seen by the higher number of tokens falling in the [5, 15] bin than the [-15, -5] bin for those models.

Figure 2 gives the results for seen language data. In this condition, the best performing model is the English model adapted to other Australian languages; however, adaptation only gives marginal improvements compared to the unadapted English model. Unsurprisingly, the best gains arise from segments which are not well represented in the English data (trilled rhotics, IPA /r/), while the gains over using a model trained only on Australian languages are those segments which are rare (long vowels) or difficult to identify boundaries for (approximants).

For held-out Yidiny data, results are similar (see Figure 3). Here are there larger gains from adapting, but the adapted model does worse than the unadapted one on trills and long vowels. This might imply that there are characteristics of individual audio files that are affecting the results (we made no attempt to control for constant background noise, for example). Interestingly, the mean absolute diffs across only the models trained from scratch is greater for testing on unseen data than seen data.

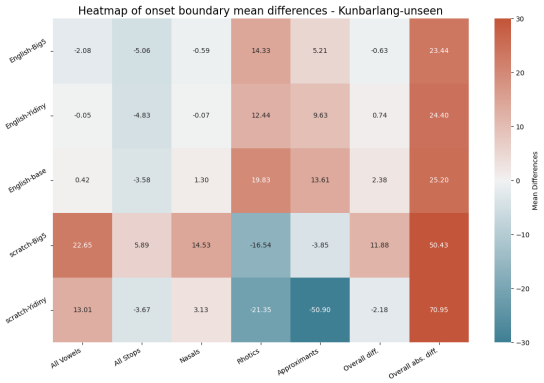


Figure 4: Kunbarlang (unseen), mean precision

Fig. 4 shows results for Kunbarlang, a language that no model trained on. Overall, all models perform worse on Kunbarlang than Yidiny data in either setting. For Kunbarlang rhotics and approximants, English-based models consistently predict the boundary is ahead of the human annotated boundaries while from-scratch models consistently predict the opposite. The Kunbarlang setting is also the setting where we see the greatest difference in the accuracy of the models trained from scratch on the Big5 dataset and the Yidiny dataset. This is not surprising, as there are many phones in Kunbarlang which are not present in Yidiny but are present in one of the Big5 languages.

Models trained from scratch on multilingual Australian data do very poorly on held out data, implying, perhaps, that there is not as much similarity between Australian languages as has been previously asserted, or that at least models are not able to take advantage of the similarities that do exist between languages.

Since a model with high accuracy and low precision would give an illusion of excellent performance, heatmaps for the standard deviation of onset boundary per natural class is provided below. Figure 5 shows the standard deviation of the diffs for models tested on seen Yidiny data.

In this condition, the most consistent model is again the English model trained on the data from

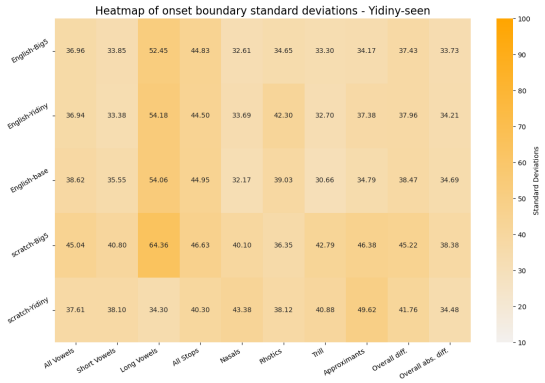


Figure 5: Yidiny (seen language) seen data, standard deviations of precision

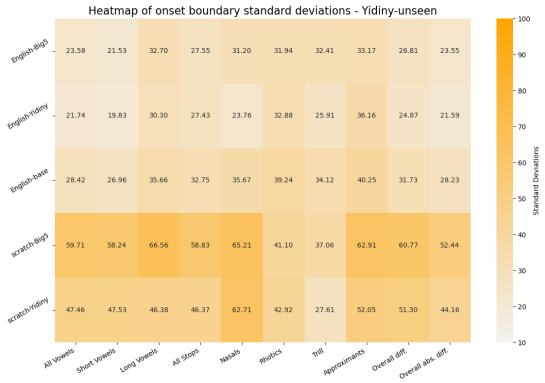


Figure 6: Yidiny (seen language) unseen data, standard deviations of precision

5 languages. Perhaps surprising however, is the comparable precision of all 5 models. The English based models also differ from the models trained from scratch in the natural class of phones that they align most precisely, with the English models' trill and approximant onset boundaries differences with the human annotated data having a lower standard deviation than the models trained from scratch.

As seen in Figure 7 testing on unseen language Kunbarlang, we find that all English models give more precise onset boundaries than their from-scratch counterparts fairly independently of the natural class of the phone. Adapting an English model also gives more precise measurements than the base English model, although for both English-adapted and from-scratch models, attempting to augment the training data with data from related languages lowers precision.

The biggest differences between the precision of the from-scratch and English-adapted models occurs for a language that was not in the training data (see Figure 7). In the unseen language setting, training on the Big5 dataset results in a notable

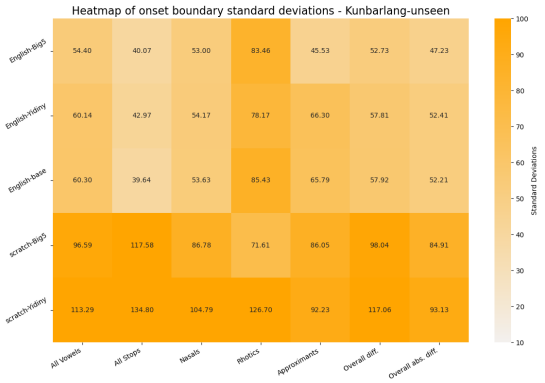


Figure 7: Kunbarlang (unseen language) unseen data, standard deviations of precision

improvement in precision in precision compared to training only on Yidiny data.

3.2 Analysis Comparison

As one might expect, given the overall similarity in precision of boundary identification discussed above, vowel dispersion plots show minimal differences between models. The exception is the model trained from scratch on a single Australian language which consistently has noticeably different vowel ellipses from the other from-scratch model and the English-based models.

For all plots, a character representing the standard IPA transcription for the vowel quality is placed at the mean F1, F2 of the vowel, and ellipses are drawn with the horizontal and vertical axes of the ellipse representing two times the standard deviation of F2 and two times the standard deviation of F1 respectively. The color of the character and the ellipse corresponds to the model used to generate alignment. Black solid lines represent the plots made with formant values extracted from the human annotated files.

Figure 8 shows vowel plots for the three short vowels in Yidiny. There is much similarity in the ellipses and mean values for each value, with the exception of the from-scratch model trained only on Yidiny data. All English-based models and the from-scratch model trained on more data produce analyses with similar ellipses as the human annotated files. The same trend of highly accurate means and ellipses can be seen with the short vowels in the Yidiny-unseen testing setting (see Figure 8). Again, the only model which seems to produce notably incorrect results is the Yidiny-only model.

There is more variation in the analyses of long

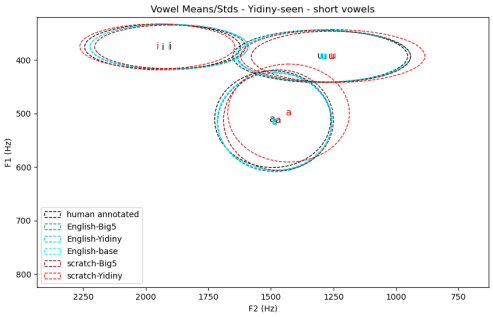


Figure 8: Comparison of vowel space measurements (F2:F1), short vowels, Yidiny seen data

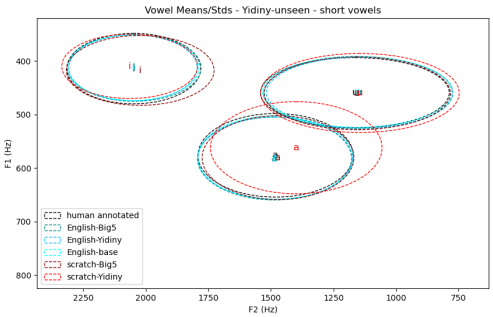


Figure 9: Comparison of vowel space measurements (F2:F1), short vowels, Yidiny unseen data

vowels than short vowels, as is seen in Fig. 10. The tendency for the model trained on five Australian languages from scratch to perform similarly to the English-based models is no longer observed, with the ellipses of long vowels being not only larger than the English-based models but also larger than the model trained from scratch on less data. Again, the predictions of English-based models are nearly identical to those derived from human annotated data.

The relationship between Yidiny-unseen short

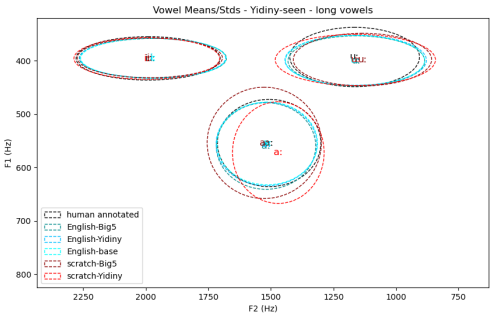


Figure 10: Comparison of vowel space measurements (F2:F1), long vowels, Yidiny seen data

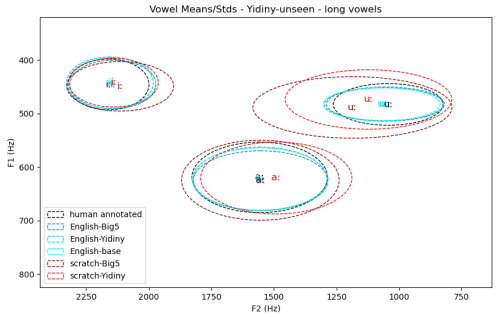


Figure 11: Comparison of vowel space measurements (F2:F1), long vowels, Yidiny unseen data

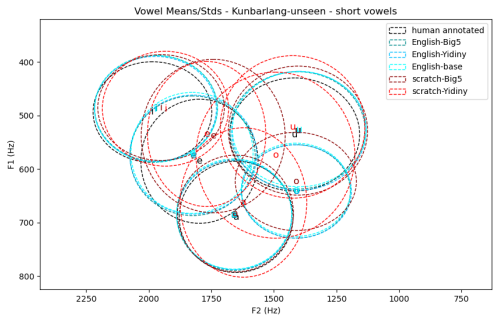


Figure 12: Comparison of vowel space measurements (F2:F1), short vowels, Kunbarlang (unseen language) vowel ellipses

and long vowels mirrors that between Yidiny-seen short and long vowels (see Figure 11). English-based models give similar vowel analyses to the human annotated data, but the models trained from scratch are noticeably inaccurate, with ellipses that are notably larger than the ellipses generated from human annotated data.

Kunbarlang has a five vowel system that does not contrast for vowel length. No model provides an analysis almost identical to the human annotated standard, but all English-based models demonstrate high accuracy with means approximating the human annotated boundaries well. The model trained from scratch on Yidiny is notably inaccurate for the mid and low vowels. The model trained from scratch on the Big5 dataset provides similarly inaccurate results for /e/, which is also not present in any Big5 language, but shows better results for /o/ which is present in Bardi. The models trained from scratch are imprecise in this setting, with ellipses that do not approximate the human annotated ellipse. The vowel ellipses of English-based models approximate the human annotated ellipses more closely than the from-scratch models.

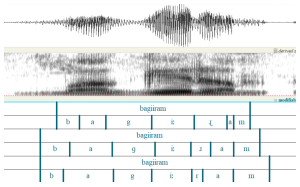


Figure 13: Typical set of textgrids of Yidiny seen data, with the scratch(big5) (top), English-adapted(big5) (middle), and human corrected (bottom) textgrids, illustrating errors in alignment.

3.3 Further Comments on Errors

In order to further investigate common types of errors, we manually compared and spot-checked alignments in Praat. Figure 13 shows an example textgrid of Yidiny seen data, with the scratch(big5) (top), English-adapted(big5) (middle), and human corrected (bottom) textgrids placed on top of one another for comparison. To investigate the major sources of misalignments, we tagged the first 100 items in the file that varied from human-annotated data by more than 100 ms. Errors in the test (Yidiny-unseen) data are similar in kind and relative frequency, but more copious. While human annotators may differ in the placement of boundaries in continuous data, the mismatches studied here are considered errors because for these cases, the alignment models place boundaries in areas where no human annotator would do so.

One third of the tagged errors arose from difficulties in identifying stop boundaries: onsets to stops in initial position, or onset or offset of closures in medial position. Yidiny stops (as has been reported for other Australian languages) can have debuccalized realizations, with extensive lenition and no clear closure or release burst (see, for example, Ennever et al. 2017). Almost another third arise from nasal boundaries in medial or final position. Most of the rest of the errors come from intervocalic rhotic, lateral, or glide identification vis à vis adjacent vowels. Such errors (except for those with initial and final segments) affect measurements of surrounding segments; 2/3 of the errors involved word-internal segments.

4 Discussion and Conclusion

Overall, we find that the most accurate models across all testing settings were the models with the global English model as a base. For seen data, English-based models slightly outperformed models in mean diff compared to models trained from scratch. However, when aligning unseen data from a seen language, English-based models produced mean diffs equal to about half of their from-scratch counterparts. This is consistent with the robustness of English models, a result of training on extremely large amounts of data. For the unseen language setting, English-based models have about half and a third of the mean absolute diff than the multilingual and monolingual models trained from scratch respectively. These findings suggest that English-based models are consistently more accurate than models trained from scratch in settings where there is little to no data for the target language.

In terms of adapting, we find that adapting the English-base model on the Big5 corpus provided marginal improvements for the Yidiny-seen and Kunbarlang settings compared to adapting on only the Yidiny-train corpus, but not for the Yidiny-unseen setting, suggesting that adapting on more data from related languages might "dilute" the effects of training on the language being tested on. All models struggled with rhotics, trills, and approximants, which is probably a result of the lack of good correspondences for the rhotics and trills present in Australian languages and a lack of a clear transition from the onset and offset for sounds belonging to these natural classes. However, across all settings the improvements of adapting an English-based model are marginal.

Of the three testing settings, we find that training a model from scratch on a multilingual dataset provided a notable 29% improvement when testing on a language that the models have never seen before. This fits with the intuition that a model trained on more languages has more flexible representations for what each phone may look like, and is thus better able to leverage that knowledge in a new setting. This effect is much more noticeable in situations where a phone in the testing language is not present in the monolingual dataset but is present in at least one language from the multilingual dataset. This is exemplified with the vowel plots on Kunbarlang data, where both models trained from scratch struggle with plotting /e/ to its absence in the training data but the Big5-trained model gives a much better analysis of /o/ due to its presence in Bardi. It should be noted that /o/ still appears infrequently in Bardi, with it being the least frequent vowel quality and the only one to lack a long counterpart.

The improvement from training on more multilingual data is minimal for the Yidiny-seen setting, and actually negative for the Yidiny-unseen setting,

suggesting that more data from related languages won't necessarily increase performance when testing on seen data and may actually hinder performance when testing on unseen data from a seen language. This can be explained using the same logic of "diluting" the data described in the previous paragraph.

The above results are mirrored when looking at the vowel analyses produced by the alignments output by the various models. For all testing settings, the plots generated from the alignments from the English-based models closely resembles the ones generated by the human generated alignments. Multilingual models trained from scratch performed comparably to the English-based models for short vowels, but produced visibly more imprecise measurements for long vowels, possibly due to long vowels having less tokens for these models to train on.

Ultimately, models trained from scratch on low-resource languages suffer from the small amount of data and the resulting lack of variety in training examples. Future research should explore whether there exist data augmentation methods that may alleviate data scarcity by providing a slightly acoustically modified version of the input audio, artificially increasing the amount of data a model sees during training. Additionally, overfitting is not an issue in most low-resource settings due to model performance on seen data being the most relevant metric for downstream tasks. Future research may thus explore the effects of hyperparameter tuning a model to encourage overfitting, sacrificing model generalizability for a performance boost on seen data.

The findings presented in this paper are useful in the context of language documentation and revitalization, because they highlight the effectiveness of using a pretrained global English model on field data. The availability of the global English pretrained models and ease of adapting them to other languages means that high quality forced alignment is accessible to any fieldworker. The similarity of the vowel plots for the Big5 models trained from scratch and the English-based models also show promise that medium-sized multilingual training datasets can provide a boost in low-resource setting.

References

Sarah Babinski. 2022. *Archival Phonetics & Prosodic Typology in Sixteen Australian Languages*. Ph.D. thesis, Yale University.

Sarah Babinski, Rikker Dockum, J. Hunter Craft, Anelisa Fergus, Dolly Goldenberg, and Claire Bowern. 2019. A Robin Hood approach to forced alignment: English-trained algorithms and their use on Australian languages. *Proceedings of the Linguistic Society of America*, 4:3–1.

Paul Boersma and David Weenink. 2021. Praat: doing phonetics by computer [Computer program]. Version 6.1.38, retrieved 2 January 2021 http://www.praat.org/.

Eleanor Chodroff, E. Ahn, and Hossep Dolatian. 2024. Comparing language-specific and cross-language acoustic models for low-resource phonetic forced alignment. *Language Documentation & Conservation*.

Christian DiCanio, Hosung Nam, Douglas H. Whalen, H. Timothy Bunnell, Jonathan D. Amith, and Rey Castillo García. 2013. Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment. *The Journal of the Acoustical Society of America*, 134(3):2235–2246. Publisher: AIP Publishing.

Hossep Dolatian. 2024. interlingual-mfa. https://github.com/jhdeov/interlingual-MFA.

Thomas Ennever, Felicity Meakins, and Erich R. Round. 2017. A replicable acoustic measure of lenition and the nature of variability in Gurindji stops. *Laboratory Phonology*, 8(1). Number: 1 Publisher: Open Library of Humanities.

Janet Fletcher and Andrew Butcher. 2014. 3. Sound patterns of Australian Languages. In Harold Koch and Rachel Nordlinger, editors, *The Languages and Linguistics of Australia*, pages 91–138. DE GRUYTER.

J. D. Hunter. 2007. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Yannick Jadoul, Bill Thompson, and Bart de Boer. 2018. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71:1–15.

Lisa M. Johnson, Marianna Di Paolo, and Adrian Bell. 2018. Forced alignment for understudied language varieties: Testing Prosodylab-Aligner with Tongan data. Publisher: University of Hawaii Press.

Alvin M Liberman, Franklin S Cooper, Donald P Shankweiler, and Michael Studdert-Kennedy. 1967. Perception of the speech code. *Psychological review*, 74(6):431.

Jayden L. Macklin-Cordes, Claire Bowern, and Erich R. Round. 2021. Phylogenetic signal in phonotactics. *Diachronica*, 38(2):210–258.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502.

Michael McAuliffe and Morgan Sonderegger. 2024. English mfa acoustic model v3.1.0. Technical report, https://mfa-models.readthedocs.io/acoustic/English/EnglishMFAacousticmodelv3_1_0.html.

Erich R. Round. 2023. Segment inventories. In Claire Bowern, editor, *The Oxford Guide to Australian Languages*, Oxford Guides to the World's Languages, page Chapter 10. Oxford University Press, Oxford, New York.

Nay San, Martijn Bartelds, Mitchell Browne, Lily Clifford, Fiona Gibson, John Mansfield, David Nash, Jane Simpson, Myfany Turpin, Maria Vollmer, Sasha Wilmoth, and Dan Jurafsky. 2021. Leveraging Pre-Trained Representations to Improve Access to Untranscribed Speech from Endangered Languages. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1094–1101.

Nay San, Georgios Paraskevopoulos, Aryaman Arora, Xiluo He, Prabhjot Kaur, Oliver Adams, and Dan Jurafsky. 2024. Predicting positive transfer for improved low-resource speech recognition using acoustic pseudo-tokens. Publisher: arXiv Version Number: 1.

Marija Tabain, Gavan Breen, Andrew Butcher, Anthony Jukes, and Richard Beare. 2016. Stress Effects on Stop Bursts in Five Languages. *Laboratory Phonology*, 7(1):16.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In *5th international conference on language resources and evaluation (LREC 2006)*, pages 1556–1559.

Creating an intelligent dictionary of Tsuut’ina one verb at a time

Christopher Cox
Tsuut'ina Gunaha Institute / Carleton University
christopher.cox@tsuutina.com

Bruce Starlight
Tsuut'ina Nation
spottedeagle1947@yahoo.com

Janelle Crane-Starlight
Tsuut'ina Nation
janelle.crane@tsuutina.com

Hanna Big Crow
Tsuut'ina Gunaha Institute
hanna.bigcrow@tsuutina.com

Antti Arppe
University of Alberta
arppe@ualberta.ca

Abstract

In this paper, we discuss the development of a long-term partnership between community and university-based language workers to create supportive language technologies for Tsuut'ina, a critically endangered Dene language spoken in southern Alberta, Canada. Initial development activities in this partnership sought to rapidly integrate existing language materials, with the aim of arriving at tools that would be effective and impactful for community use by virtue of their extensive lexical coverage. We describe how, as this partnership developed, this approach was gradually superseded by one that involved a more targeted, lexical-item-by-lexical-item review process that was directly informed by other community language priorities and connected to the work a local language authority. We describe how this shift in processes correlated with other changes in local language programs and priorities, noting how ongoing communication allowed this partnership to adapt to the evolving needs of local organizations.

1 Introduction

Tsuut’ina (ISO 639-3: `srs`, Glottocode: `sars1236`) is a Dene language spoken by members of the Tsuut'ina Nation, a signatory to Treaty 7 in present-day southern Alberta, Canada. Together with Plains Apache, Tsuut'ina is one of only two Dene languages spoken on the Great Plains, and is separated from other members of the Dene language family by surrounding Algonquian and Siouan-speaking Indigenous nations. As of October 2024, there are 18 first-language speakers of Tsuut'ina, all over the age of 75 and almost all residing at the Tsuut'ina Nation (Tsuut'ina Gunaha Institute, p.c.). While Tsuut'ina is thus critically endangered, strong connections between Tsuut’ina language and community identity and culture have fostered equally strong retention of Tsuut’ina language proficiency among present-day speakers. These same connections have also encouraged community-based language documentation, education, and revitalization initiatives, including those supported by collaborations with individuals and organizations outside of the Tsuut'ina Nation, as discussed in this paper.

From a linguistic perspective, Tsuut'ina closely resembles other Dene languages, with complex, prefixing, polysynthetic verbal morphology (cf. Cook, 1984; Rice, 2000; Rice, 2020). Tsuut'ina also relies heavily on tone to convey both lexical and grammatical distinctions, having one of the largest inventories of tone contrasts attested in the Dene language family (cf. Sapir, 1925; McDonough et al., 2013; Starlight & Cox, 2024). While previous research on the language conducted by both Tsuut'ina and non-Tsuut'ina linguists has resulted in notable collections of textual and grammatical documentation (e.g., Goddard, 1915; Onespot & Sapir, 1922; Cook, 1984; Starlight, Moore & Cox 2018; among others), linguistic research into aspects of Tsuut'ina grammar is ongoing, with many areas of grammatical organization still under active investigation. Both the presence of open questions concerning basic grammatical features of

Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 110–119

the language (e.g., how many tone and vowel length contrasts should be recognized; Starlight & Cox, 2023) and the degree to which the overall profile of the language differs from neighbouring Indigenous languages and from English present particular challenges, both for current Tsuut'ina language learners and teachers who are aiming to acquire and convey the language effectively and for efforts to develop approaches and resources that support 'front-line' language revitalization work.

2 History – How the partnership came about

The partnership described in this paper has deep roots in language education, documentation, and revitalization initiatives at Tsuut'ina Nation. Since the early 1970s, members of the Tsuut'ina Nation, recognizing a significant shift in the number of first-language speakers, began implementing programs aimed at supporting Tsuut'ina language retention and intergenerational language transmission. Over several decades, these efforts resulted in the establishment of K–12 school-based language education programs, Tsuut'ina literacy programs for L1 speakers, and the adoption of a standard Tsuut'ina orthography (cf. Cook, 1984: 1–2), alongside concurrent work to develop classroom resource materials, documentation with Tsuut'ina Elders, and an initial language curriculum (Calgary Roman Catholic Separate School Division, 1996).

While the direction of these initiatives was determined and led by the Tsuut'ina Nation, on several occasions, members of the Tsuut'ina Nation also sought out partnerships with individuals and organizations outside of the Nation. The second author of this paper, Dr. Bruce Starlight, a linguist and fluent Tsuut'ina speaker who had been involved in language revitalization and documentation initiatives since 1972, worked extensively to develop such relationships, collaborating with non-Tsuut'ina colleagues to support local language programs and projects. This included extensive work with Gary Donovan at the University of Calgary on the creation of pedagogical resources for Tsuut'ina and Sally Rice at the University of Alberta on Tsuut'ina language documentation and revitalization programs. Bruce's involvement in university-based programs also extended to linguistic field methods courses at the University of Alberta in 2007 and 2009, where the first author of this paper, Christopher Cox, a linguist with an interest in community-based language work, became involved in Tsuut'ina language programs as a student volunteer during his graduate studies.

Relationships such as these continued to develop in parallel with language programs at Tsuut'ina Nation, where community interest in Tsuut'ina language revitalization continued to grow. In 2008, the Tsuut'ina Nation established the Tsuut'ina Gunaha Institute, the body within Tsuut'ina Nation tasked with supporting the full revitalization of the Tsuut'ina language. Bruce served as the Institute's founding director until 2012, when he was invested as the first Tsuut'ina Language Commissioner, a position that oversaw the development of Tsuut'ina language documentation, contributed to the visibility of the language (e.g., through the translation of public signage into Tsuut'ina), and ensured the continued integrity of the language. The creation of both of these offices was accompanied by a substantial expansion in the resources and positions available for local language revitalization programs, providing opportunities for many younger Tsuut'ina Nation members to engage with local language work on a full-time basis. It was during this period that the third and fourth authors of this paper, Janelle Crane-Starlight and Hanna Big Crow, joined the Tsuut'ina Gunaha Institute, eventually coming to serve as the Executive Director of Language and Culture for Tsuut'ina Nation (Janelle) and the Director of the Tsuut'ina Institute (Hanna).

As language programs continued to expand at Tsuut'ina Nation over the past decade, both the Office of the Tsuut'ina Language Commissioner and the Tsuut'ina Gunaha Institute noted an increased demand for resources that supported Tsuut'ina language education, documentation, and revitalization activities in digital contexts, particularly as activities in all of these areas moved increasingly into the digital realm. This shift not only resulted in more emphasis being placed on developing new Tsuut'ina language resources in digital formats, suitable for use in community-based programs, but also increased access to information found in existing, non-digital language materials; support for continued teacher training for Tsuut'ina language educators; and tools that could assist in creating such resources quickly and reliably, such as spell-checkers, predictive text systems, and text-to-speech applications. Through the network of relationships that had been

developed with the University of Alberta previously, colleagues at Tsuut'ina Nation were introduced to the fifth author of this paper, Antti Arppe, a linguist who had been involved in recent years in supporting the development of language technologies and morphologically aware online dictionaries for other Indigenous languages in North America, drawing in part on computational infrastructures developed for Indigenous language technologies in northern Eurasia (Trosterud, 2006).

Intelligent online dictionaries combine a lexical database, with entries organized under citation forms and their (English) translations, with a computational model of the word-structure of a language (Johnson et al., 2013). Firstly, this "intelligence" allows the online dictionary to recognize all inflected word-forms for the entries (which the model covers, of course), to provide linguistic analyses for these word-forms and to link those to their citation forms. Secondly, one can use the computational model in reverse and generate full inflectional paradigms for each of the citation forms; for verbs, such inflectional paradigms are often called *conjugations* (following French via Latin); for nouns, the corresponding paradigms would be known as *declinations*. For languages with a rich (inflectional) morphology, as is the case for many Indigenous languages spoken in North America, and in particular the Dene languages, such "intelligence" is indispensable, as any lexeme can have tens if not hundreds or thousands of inflected word-forms, which would be impossible to harvest from corpora of any size, and impractical to store exhaustively as their own dictionary entries.

Over a period of two years, the authors of this paper began to meet informally and discuss a potential collaboration to expand such tools to support Tsuut'ina. This began modestly by arranging initial, in-person meetings between all partners at Tsuut'ina Nation, which focused on becoming better acquainted with one another, sharing information about current priorities for language programs at Tsuut'ina Nation (for Bruce, Janelle, and Hanna), and what a partnership to develop digital tools could realistically contribute (for Antti and Chris). For Tsuut'ina partners, developing an intelligent dictionary had the potential to respond to several priorities for supporting second-language learners and local language revitalization programs. First, an online dictionary was seen as potentially improving access to Tsuut'ina documentation materials from previous and ongoing/future language projects for second-language learners. Such materials serve a crucial purpose for Tsuut'ina language learners and teachers as a resource for language education programs, self-study, and other language revitalization resource development initiatives. Second, it was also recognized that tools that could model and present inflectional patterns could be particularly valuable for Tsuut'ina second-language learning and teaching, since verbs and verb paradigms are critical to using and understanding the language.

While these initial meetings suggested that a collaboration might indeed be desirable, having concrete discussions around tools and technologies that did not yet exist for Tsuut'ina and that had few familiar precedents among other Indigenous languages sometimes proved challenging. While it was possible to discuss what already been accomplished for other, neighbouring Indigenous languages, it was with the preparation of still mock-ups of what Tsuut'ina-specific digital tools could look like (e.g., screenshots of a browser window showing an example paradigm from the intelligent online dictionary for another language, with all paradigm entries replaced with Tsuut'ina word-forms and the layout adapted to fit Tsuut'ina tense and aspect categories) that the group found a way to effectively conceptualize and discuss what these tools could accomplish. For example, in Figure 1, the Tsuut'ina verb form *nàguts'idáátłíl* is recognized and analyzed as the Progressive Fourth Person form of the Intransitive Verb *nàgudiitłod*, meaning roughly "he/she/it jumps down", for which an entry exists in the lexical database, and to which this inflected word-form is linked. If the user then would click on the entry, this would yield an inflectional paradigm, giving all the person forms in the various aspects, of which an exemplary sample is provided here. The mockup in Figure 1 was created by taking an earlier version of an intelligent dictionary for another, unrelated Indigenous language spoken in Canada, and replacing the content with correct Tsuut'ina elements.

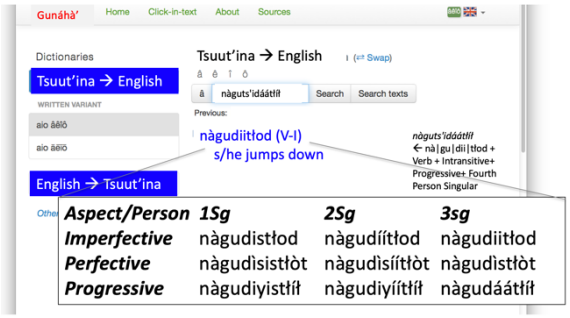


Figure 1: A mock-up of an intelligent dictionary entry for the Tsuut'ina lemma *nàgudiitłod* 'he/she/it jumps down'.

These initial meetings and co-design sessions quickly led to exchanging further ideas and information, with university-based partners drawing on the Giella infrastructure to prototype a preliminary computational model of Tsuut'ina verbal morphology and bootstrap working demos of a number of text-proofing tools and an online dictionary. Preparing and sharing presentations with Tsuut'ina Nation leadership of these tools in action—for example, with videos of spell-checking suggestions being offered for Tsuut'ina words while editing a document in LibreOffice, or of searching for morphologically complex Tsuut'ina words and being presented with their lemmas in a morphologically aware online Tsuut'ina dictionary—both teams found support for an application to the Social Sciences and Humanities Research Council of Canada (SSHRC) for a seven-year Partnership Grant, "21st Century Tools for Indigenous Languages", in which Tsuut'ina Nation would serve as one of two lead Indigenous partner nations. The awarding of this grant in 2019 allowed collaboration on these tools to move ahead on an expanded scale, with the promise of stable financial support for project activities until 2026.

3 Documentation, description, and deliberation

With support from the above grant in place, team members turned their attention to determining how best to expand the existing prototypes into applications that could meaningfully support access to Tsuut'ina language for local language programs. Since the Office of the Tsuut'ina Language Commissioner and its collaborators had recently been working on a number of substantial language resources for school-based programs that were available in digital format, it was recommended that these be prioritized for inclusion in an online dictionary and related tools. These resources included draft copies of a 'modernized' edition of the extensive list of elicited Tsuut'ina word-forms that linguist Edward Sapir and Tsuut'ina speaker John Whitney-Onespot developed together in 1922 (ca. 11,000 items; Whitney-Onespot & Sapir, 1922; Starlight et al., 2016), as well as two 100-page collections of Tsuut'ina verb paradigms (the Tsuut'ina Verb Phrase Dictionary, Books 1 and 2; cf. Starlight & Donovan, 2019). Through the contributions of Josh Holden during a postdoctoral fellowship at the University of Alberta, as well as Karoline Antonsen and Ruben Mögel, Master's students at the University of Alberta, work began to prepare to incorporate the information in these resources into a lexical database that would underlie all of these technical tools.

While this initial work on processing these materials got underway in the early days of the COVID-19 pandemic, regular video teleconference meetings were scheduled with all team members to discuss grammatical issues. For partners at Tsuut'ina Nation, these meetings often provided an opportunity to present and discuss issues around the interpretation of particular morphemes and constructions that had been encountered in recent language projects, with non-Tsuut'ina partners sharing comparisons with similar forms in other Dene languages and/or contributing to analysis together. For university-based partners, these sessions also provided opportunities to share regular updates about ongoing work on transferring information from the language resources into a database, as well as to seek advice on forms whose grammatical analysis or meaning seemed unclear or that needed to be confirmed by first-language Tsuut'ina speakers before being included in the computational morphological model. The mutual support and connection afforded by these meetings served important functions at the outset of the SSHRC Partnership Grant, contributing a sense of ongoing collaboration even when pandemic restrictions precluded any in-person gatherings.

As these regular meetings and efforts to incorporate existing language resources into a comprehensive lexical database moved forward, more of the characteristics of the latter materials' scope and coverage became apparent. By assigning each inflected Tsuut'ina verb word in these resources to a corresponding lemma, it became

possible to determine which lemmas contained information on all of the tenses/aspects/modes that are associated with regular Tsuut'ina verb phrases and which contained substantial gaps in documentation. It soon became apparent that, across the ca. 1,577 verbal lemmas attested in the Onespot-Sapir resource, the majority of lemmas showed at least one gap in a regular tense/aspect/mode form, with at least 700 having only a single tense/aspect/mode attested. A similar review of the verb phrase dictionary books revealed fewer missing paradigm forms, but brought attention to potential inconsistencies in tone marking in Tsuut'ina forms that eventually led the Office of the Tsuut'ina Language Commissioner to request that these preliminary books be set aside. While information from these analyzed resources would later prove valuable, regular project meetings underscored concerns over incomplete and potentially inaccurate information being circulated out of these provisional resources, as well as over the challenge of addressing such significant documentary gaps without systematic support from a much larger number of fluent Tsuut'ina speakers.

At the same time as these issues with the available language resources came to be discussed, other language initiatives at Tsuut'ina Nation continued to advance, including efforts to develop a new curriculum for use in core Tsuut'ina language programs at all age levels (i.e., Headstart, K–12, and adult education). This curriculum aimed to help reorient Tsuut'ina language learning and teaching from the noun-focused approaches that had generally been adopted in previous programs (e.g., beginning with teaching and learning lists of nouns at all age levels) to introducing and emphasizing verb phrases early on, recognizing how important verb-based patterns are in Dene languages like Tsuut'ina. With language learners and teachers being among the primary intended audiences for the tools being developed in this partnership, it was decided to set the previous language resources aside and attempt to align work on the online dictionary as closely as possible with the needs of curriculum users—that is, Tsuut'ina language educators, language learners, and curriculum team members. All partners recognized that the intelligent dictionary could be an essential resource to support this new curriculum, especially in its focus on verb paradigms. Encouragingly, this reorientation allowed members of the partnership team to draw on parts of the documentation analyzed in previous stages of this project to fill in portions of the vocabulary needed for the curriculum, thus saving time and effort. This work also brought attention to other gaps in existing documentation, this time in vocabulary related to both everyday activities and cultural practices that were either incompletely recorded in previous resources or entirely absent from prior documentation (e.g., specialized vocabulary related to hanging up meat on drying racks, pounding chokecherries, or other important cultural practices). Team members drew additional inspiration for curriculum vocabulary from several sources, including input from Tsuut'ina language teachers and advanced language learners and pedagogical resources developed for other Dene and non-Dene Indigenous languages. Connecting the development of the lexical contents of digital tools with the needs of Tsuut'ina language learners and teachers thus helped expose (and, in turn, contribute to addressing) significant gaps in the domain coverage of existing Tsuut'ina language materials in several high-priority areas.

Systematically addressing these gaps in lexical documentation—whether encountered in existing language resources or made apparent by requests from language teachers and members of curriculum development teams—and ensuring the accuracy of the information that would be represented in the tools developed in this partnership presented a standing challenge. This was addressed in part by the development of processes within Tsuut'ina Nation that sought to ensure that curriculum materials and other language resources reflected the understandings of fluent, first-language Tsuut'ina speakers. This involved the formation of the Tsuut'ina Language and Culture Committee, an advisory body consisting of six Tsuut'ina-speaking Elders that had within its mandate the review and approval of Tsuut'ina language resources prior to their use in the community. The establishment of a review process that supported nearly a third of the present-day first-language speakers of Tsuut'ina in gathering to offer constructive feedback on Tsuut'ina language matters proved important to addressing the above concerns over accuracy and coverage in language resources, with committee members often helping one another to recall less frequent Tsuut'ina terms and expressions that were previously in more active use. This new review

process required careful, item-by-item (or, in the case of sets of paradigmatically related forms, summarized paradigm by summarized paradigm; see Appendix A for an example) review to ensure that all members of the committee were in agreement that these language resources were acceptable for further use. This consensus-driven process of collectively reviewing lexical items one at a time, while requiring more time than previous attempts to incorporate existing language resources wholesale, helped not only to ensure that any inaccuracies or inadvertent gaps in Tsuut'ina forms or their English translations were systematically addressed, but also that the approved materials could be taken to reflect the collective understanding and priorities of the Tsuut'ina speech community, thereby fostering greater inclusion and a sense of collective ownership and investment in these collaboratively developed resources.[1]

4 Steps towards an intelligent online Tsuut'ina-English dictionary

As sets of verb phrases are identified for inclusion in the new Tsuut'ina language curriculum, members of the partnership team now have a comparatively straightforward workflow for ensuring that they are incorporated systematically in the online Tsuut'ina dictionary:

1. The second author and/or the Language and Culture Committee are consulted to recommend suitable Tsuut'ina equivalents, with the second author providing a brief overview of the regular aspectual forms.

2. These aspectual forms are added to a preliminary lexical database used to hold as-of-yet unapproved lexical items, then compiled as the lexical component of the current Tsuut'ina finite-state morphological model (`https://github.com/giellalt/lang-srs/`; Holden et al., 2022), producing a temporary finite-state transducer (FST) model (according to the Xerox-style specifications, cf. Beesley & Karttunen 2003; Hulden 2009).

3. The temporary FST is used to populate a Word document template, producing a condensed (1–2 page) overview of inflected forms for each of the regular tense/aspect/mood categories. A separate Python module developed by the partnership team also provides provisional English free translations for each Tsuut'ina word-form in this document, converting FST tag sequences and an English translation template sentence into contextually appropriate translations (e.g., `to_english("+V+I+Pfv+SbjSg1", "he/she/it will run") => "I ran"`, mapping the `+Pfv` perfective aspect tag to past inflection and the `+SbjSg1` first-person singular subject tag to "I" in English).

4. These automatically populated 'paradigm review sheets' are then reviewed and edited by the second author. Once that initial round of editing is complete, the first and second authors meet to review and record all of the Tsuut'ina word-forms together, producing high-quality WAV audio recordings of all entries in the paradigm review sheets that can later be incorporated as audio clips into the online dictionary.

5. On the basis of the paradigm review sheets emerging from this process, the preliminary lexical database is updated to reflect the corrected forms, then used again to produce an FST that is used to populate an overview of the recorded paradigms for the Language and Culture Committee to review. Any feedback from the committee members on this overview can then be incorporated and all lexical information moved into a permanent lexical database for approved material.

Importantly, the above review process is undertaken one verb lexeme at a time. That is, we ensure that the entire inflectional paradigm and all of its principal parts is fully validated for one

[1] The partnership has also provided financial support for the training of Tsuut'ina language instructors with Tsuut'ina-specific courses within the Community Linguist Certificate provided by the Canadian Indigenous Languages and Literary Institute (CILLDI) at the University of Alberta.

lexeme before we continue to the review of the next lexeme and its paradigm. Strict adherence to this ensures that no aspectual gaps are accidentally left in the paradigms (which would be facilitated by hopping from one lexeme to another in a less structured approach), nor do any mistranscriptions remain of individual word-forms in the paradigms. In this manner, we will be able to make available from the very onset an intelligent online dictionary for Tsuut'ina, with full features and functionality, e.g., the ability to (a) generate full and correct inflectional paradigms, (b) include full audio linked to all word-forms in these paradigms, as well as (c) recognize and linguistically analyzed each and every word-form in these paradigms—even if we can only implement this for a small set of verb lexemes, at least in the very beginning. This will allow for the informed examination of, and accurate feedback from various stakeholders for, a fully Tsuut'ina version of an intelligent online dictionary in terms of its linguistic content, rather than having to somehow explain (away) and go back to filling in missing sub-paradigms for some tense/aspect/mode, or correcting some incorrect word-forms in the paradigms. This resource is anticipated to gradually grow as more paradigms for verbal lexemes are individually created and reviewed. To date, this process has resulted in over 1,000 pages of completed paradigm review sheets for verb phrases requested for the Tsuut'ina language curriculum, with 45h18m of corresponding audio recordings of inflected word-forms. This work is still underway, with more material expected to be created again over the coming few months and committee review ongoing.

5 Lessons learned

In our experiences in this multi-year collaboration, we have noted positive outcomes from several practices that we have increasingly come to favour over time:

1. *Showing vs. telling:* In our initial conversations about this project, we found it valuable to be able to show what the tools we were discussing might look like, rather than simply talk about them in general terms. For new technologies with few precedents, or where the only precedents are currently available for unrelated languages, it can sometimes be difficult to picture what a particular tool or resource may look like in the target language and imagine how it might be useful. In this project, being able to share and discuss mock-ups of these tools, and later develop those into limited-but-working prototypes for preliminary evaluation, provided a valuable way forward for us in developing a common understanding of what we hoped to work towards.

2. *Change, communication, and responsiveness:* Since this partnership officially began in 2019, several project team members have transitioned into and out of key roles, new processes for the review of Tsuut'ina language materials have come into effect, and priorities have continued to shift for local language programs (here, in the direction of curriculum and teaching). Changes such as these have, at times, required significant deliberation to determine how best to proceed, including through extended pauses when members of the team needed to assess how these shifts might affect their planned contributions. Maintaining communication between project partners and deliberately expanding the circle of those involved—from an initially small group of community language leaders and university-based collaborators to a wider group of first-language speakers serving as reviewers, Tsuut'ina language teachers, and curriculum developers—has been crucial to seeing this project continue to develop, helping to ensure that it remains relevant in the context of local language work.

3. *Slow but steady wins the race:* In current work in developing collections of lexical material, computational morphological models, and related language technologies, it is not uncommon for breadth of lexical coverage and the rapid gathering of information to be presented as important features of useful, real-world resources and approaches (cf. Boerger

& Stutzman, 2018 on the motivations behind Rapid Word Collection methods). While this emphasis on broad-coverage and efficient lexicography is understandable, we would note here that initial attempts to draw on existing, relatively extensive language resources as a quick starting point were ultimately less successful than focusing on a much more restricted set of materials that were identified as the immediate needs of local language programs and processed in a slower, more deliberate manner. This approach has brought to light notable gaps between the vocabulary needed by current language education and revitalization programs and the outputs of previous generations of language documentation (cf. Mithun, 2007; Amery, 2009). We also find value in review processes that serve to build both consensus and community around such work, as is arguably the case here.

In this project, what began as a series of preliminary discussions among a small group of community and university-based language workers to see what could be possible to support Tsuut'ina language initiatives with new, digital tools has grown into a considerably broader partnership—one that now involves a much larger community of Tsuut'ina language teachers and curriculum developers, first-language speakers, and university-based researchers and students as key contributors. We look forward to seeing how this partnership continues to develop from here as these resources continue to grow—one verb phrase at a time.

Acknowledgments

We gratefully acknowledge the contributions made by many individuals who have been involved in this partnership, including Karoline Antonsen, Steven Crowchild, Josh Holden, and Ruben Mögel. This work has been funded by a Partnership Grant (895-2019-1012) from the Social Sciences and Humanities Research Council (SSHRC) of Canada.

References

Rob Amery. 2009. Phoenix or relic? Documentation of languages with revitalization in mind. *Language Documentation and Conservation* 3(2):138–148. http://hdl.handle.net/10125/4436.

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Publications, Stanford, CA.

Brenda H. Boerger & Verna Stutzman. 2018. Single-event Rapid Word Collection workshops: Efficient, effective, empowering. *Language Documentation & Conservation* 12:235–273. http://hdl.handle.net/10125/24766.

Calgary Roman Catholic Separate School Division. 1996. *Nanagusja: A Tsuut'ina (Sarcee) Language Development Program. Teacher's Guide*. Learning Resources Distributing Centre, Alberta Education. https://eric.ed.gov/?id=ED436084.

Eung-Do Cook. 1984. *A Sarcee grammar*. Vancouver, BC: University of British Columbia Press.

Pliny Earle Goddard. 1915. *Sarsi texts*. University of California Publications in American Archaeology and Ethnology 11, No. 3, pages 189–277. Berkeley, CA: University of California Press. http://digitalassets.lib.berkeley.edu/anthpubs/ucb/text/ucp011-004.pdf.

Joshua Holden, Christopher Cox & Antti Arppe. 2022. An expanded finite-state transducer for Tsuut'ina verbs. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, et al. (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5143–5152. Marseille, France: European Language Resources Association. https://aclanthology.org/2022.lrec-1.551.

Mans Hulden. 2009. Foma: a finite-state compiler and library. In *Proceedings of EACL*, pages 29—32, Athens: Association for Computational Linguistics.

Ryan Johnson, Lene Antonsen & 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*: 59–71.

Joyce McDonough, Jared O'Loughlin & Christopher Cox. 2013. An investigation of the three tone system in Tsuut'ina (Dene). In *Proceedings of Meetings on Acoustics* 19(1):060219. https://doi.org/10.1121/1.4800661.

Marianne Mithun. 2007. What is a language? Documentation for diverse and evolving audiences. *STUF – Sprachtypologie und*

Universalienforschung 60(1):42–55. https://doi.org/10.1524/stuf.2007.60.1.42.

John Onespot & Edward Sapir. 1922. *Sarsi (Tsuut'ina) notebooks #1–7*. Unpublished field notes, ms. American Council of Learned Societies Committee on Native American Languages. American Philosophical Society, Philadelphia, PA.

Keren Rice. 2000. *Morpheme order and semantic scope: Word formation in the Athapaskan verb* (Cambridge Studies in Linguistics 90). Cambridge: Cambridge University Press.

Keren Rice. 2020. *Morphology in Dene languages*. Oxford Research Encyclopedia of Linguistics. Oxford: Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.629.

Edward Sapir. 1925. Pitch accent in Sarcee, an Athabaskan language. *Journal de la Société des Américanistes de Paris* 17: 185–205.

Bruce Starlight, Gary Donovan & Christopher Cox. 2016. From archival sources to revitalization resources: Revisiting the Tsuut'ina notebooks of John Onespot and Edward Sapir. Paper given at the American Philosophical Society symposium 'Translating Across Time and Space: Endangered Languages, Cultural Revitalization, and the Work of History', Philadelphia, PA, October 13–15, 2016.

Bruce Starlight, Patrick Moore & Christopher Cox. 2018. *Documenting conversations in Tsuut'ina*. Endangered Languages Archive: http://hdl.handle.net/2196/00-0000-0000-0013-2FA3-9.

Bruce Starlight and Gary Donovan. 2019. *Tsuut'ina Verb Phrase Dictionary: Book One*. Calgary, AB: n.p.

Bruce Starlight & Christopher Cox. 2023. On vowel length contrasts in Tsuut'ina language work. Paper presented at the 2023 Annual Meeting of the Society for the Study of the Indigenous Languages of the Americas (SSILA), online, January 20–22, 2023.

Bruce Starlight & Christopher Cox (eds.). 2024. *Isúh Ánii: Dátł'ishí Ts'ìká áa Guunijà / As Grandmother Said: The Narratives of Bessie Meguinis*. Regina, SK: University of Regina Press.

Trond Trosterud. (2006). Grammatically based language technology for minority languages. In *Lesser-known languages of South Asia*, pages 293–316. De Gruyter Mouton, The Hague.

A Appendix A: Example verb paradigm summary

Table 1 presents a partial verb paradigm summary for the Tsuut'ina lemma *ànàyidiʔò* "he/she/it will lose it (solid obj.)", showing inflected Tsuut'ina word-forms associated with the Non-Past tense/aspect/mode category and their English free translations. In a complete paradigm summary for this lemma, similar tables would be included not only for the Non-Past, but also for the Past, Progressive, Repetitive, and Potential categories. The 13 verb forms shown in this table represent all possible subject person-number

Non-Past	**Ànàdisʔò.**	"I will lose it."
	Ànàdíʔò.	"You will lose it."
	Ànàyidiʔò.	"He/she/it will lose it."
	Ànàdaàʔò.	"We both will lose it (solid obj.)."
	Ànàdasʔò.	"You both will lose it (solid obj.)."
	Ànàgiyidiʔò.	"They both will lose it (solid obj.)."
	Ànàts'idiʔò.	"Someone will lose it (solid obj.)."
	Nominalized Verb Phrase	
	Ànàyidiʔò-hí	"the one who will lose it (solid obj.)"
	Ànàyidiʔò-hà	"the one that will lose it (solid obj.)"
	Distributive Plural	
	Ànàdàdaàʔò.	"Each and every one of us will lose it (solid obj.)."
	Ànàdàdasʔò.	"Each and every one of you will lose it (solid obj.)."
	Ànàdàgiyidiʔò.	"Each and every one of them will lose it (solid obj.)."
	Ànàdàts'idiʔò.	"Each and every one will lose it (solid obj.)."

Table 1: Partial verb paradigm summary for the Tsuut'ina lemma *ànàyidiʔò* "he/she/it will lose it (solid object)."

combinations in Tsuut'ina (including forms with distributive plural marking and two distinct forms of deverbal nominalization) when appearing with a third-person singular direct object. This limited set of forms is sufficient to determine both the inflectional paradigm to which this lexeme belongs as well as its constituent morphemes. Moreover, by holding the person and number of any object marking constant across all subject forms, summary charts such as this are able to concisely represent verbs that mark one or more objects morphologically, which may otherwise have several thousand distinct inflected forms.

AILLA-OCR: A First Textual and Structural Post-OCR Dataset for 8 Indigenous Languages of Latin America

Milind Agarwal, Antonios Anastasopoulos
George Mason University
Correspondence: magarwa@gmu.edu

Abstract

It is by now common knowledge in the NLP community that low-resource languages need large-scale data creation efforts and novel contributions in the form of robust algorithms that work in data-scarce settings. Amongst these languages, however, many have a large amount of data, ripe for NLP applications, except that this data exists in image-based formats. This includes scanned copies of extremely valuable dictionaries, linguistic field notes, children's stories, plays, and other textual material. To extract the text data from these non machine-readable images, Optical Character Recognition (OCR) is the most popular technique, but it has proven to be challenging for low-resource languages because of their unique properties (uncommon diacritics, rare words etc.) and due to a general lack of preserved page-structure in the OCR output. So, to contribute to the reduction of these two big bottlenecks (lack of text data and layout quality), we release the first textual and structural OCR dataset for 8 indigenous languages of Latin America. We hope that our dataset will encourage researchers within the NLP and Computational Linguistics communities to work with these languages.[1]

1 Introduction

Latin America is home to a linguistically diverse set of hundreds of indigenous languages. Many of these are low-resource in terms of text and audio resources, and generally lack basic natural language applications such as spell checkers, part of speech (POS) taggers, etc. However, these languages have a large number of digital resources (not machine-readable) in the form of recordings, plays, stories, and dictionaries. One major repository of such materials is the Archive of the Indigenous Languages of Latin America (AILLA), whose raw materials and digitizations form the core of the dataset in our paper (Agarwal and Anastasopoulos, 2024).

[1]Relevant code and data are available here

Figure 1: The AILLA-OCR corpus covers 8 indigenous languages spoken across 6 countries in Latin America. Languages differ in terms of vitality, with only South Bolivian Quechua with over a million speakers and some official status, but most others exist as minority languages in the respective countries (Table 1).

Of particular interest to us are linguistic materials such as grammars, dictionaries, ethnographies, and field notes, that can serve as training data for NLP applications and Optical Character Recognition (OCR). The goal of releasing this digitized and corrected dataset is to preserve invaluable linguistic materials, promote research on downstream tasks such as language identification and machine translation, and encourage better OCR techniques that allow for more accurate extraction of data from such corpora at scale (Nguyen et al., 2021; Agarwal et al., 2023). Modern OCR systems specialize in extracting text from such documents, but this requires high-quality layout detection to make the extracted text usable for downstream NLP tasks (Bustamante et al., 2020; Neudecker et al., 2021). While progress has been made on correcting the OCR *text* outputs after extraction, no work has focused on automatically correcting the *layouts* themselves either before/after text post-correction due to lack of annotated data. We aim to address

Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 120–127

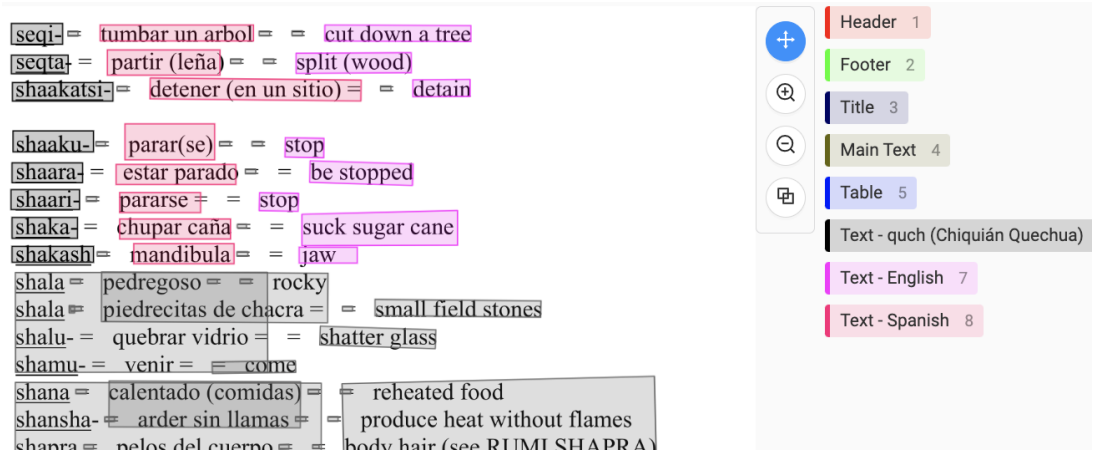


Figure 2: An in-progress annotation of a Chiquián Quechua language document (multilingual with Spanish and English) in our Annotation Workflow Portal. Here, the annotator is not only readjusting the detected bounding boxes (light grey), but is also correcting the textual errors in the new boxes, and labeling them (if language is known). Note that not all corrected bounding boxes need to be phrase or line-level. However, such organized post-corrected structure and text allows us to extract text more consistently.

this research gap by creating the first textual and structural OCR dataset for indigenous languages of Latin America. To summarize, our main contributions are:

1. OCR extractions from 8 Latin American indigenous languages from the AILLA collection.
2. Human-annotated text corrections for a sample of the digitized data, which can be used to model supervised post-correction of first-pass OCR output.
3. Structural post-corrections and associated metadata, including standard transformations like scaling, horizontal or vertical shifts, and creation of new gold-standard bounding boxes.

2 Language Profiles

South Bolivian Quechua (QUH) is a Quechuan language variety spoken primarily in Bolivia, but is also indigenous to some northern parts of Chile and Argentina. It is an agglutinative, polysynthetic language with a rich derivational morphology, is one of the most spoken indigenous languages in Bolivia with over 1.5 million speakers, and is constitutionally recognized. Ethnologue classifies South Bolivian Quechua's development as vigorous with standardized literature beginning to take shape. It is written in an extended Latin-based alphabet.

Mísqito/Mískito (MIQ) is a Misumalpan language spoken by more than 150K people (primarily Miskito) in Nicaragua and eastern Honduras. While orthographic conventions are not fully standardized, Miskito uses a subset of the Latin script for writing. Ethnologue's language vitalization hierarchy pegs Mísqito as threatened, since it is used for face-to-face communication within all generations, but it is losing young speakers to more dominant languages like Spanish and English.

Mam (MAM) belongs to the Eastern branch of the Mayan language family and is spoken by over 600K people mainly in Guatemala, where it is a recognized minority language. It is also called Qyo:l or Qyol Mam by its own speakers. Ethnologue classifies Mam's development as vigorous with standardized literature being steadily circulated. It is written in an extended Latin-based alphabet. Efforts to revitalize and preserve Mam have been ongoing, with initiatives such as bilingual education programs and the creation of written materials to strengthen literacy in both Mam and Spanish.

Chuj (CAC) is a Western Mayan (Q'anjob'alan branch) language spoken by about 60K people primarily in Guatemala. It uses the Latin alphabet and has two main dialects: San Mateo Ixtatán dialect and San Sebastián Coatán. It is heavily influenced by Spanish, the dominant and official language in Guatemala, and Chuj features heavy code-mixing and Spanish loan words. Ethnologue classifies Chuj's vitality as developing with standardized literature being developed due to language conservation and revitalization efforts taking place in San Mateo Ixtatán, through groups like the Academia de Lenguas Mayas de Guatemala.

Chimalapa/Oaxaca Zoque (ZOH) is an indigenous language primarily spoken in Oaxaca, Mexico

Language	693-3	Main Country	Speakers	Resource/Collection
South Bolivian Quechua	QUH	Bolivia	1.6M	Kalt (2016)
Mísqito	MIQ	Nicaragua, Honduras	150K	Bermúdez Mejía (2015)
Mam	MAM	Guatemala	600K	England (1972-1985)
Chuj	CAC	Guatemala	60K	Hopkins (1964)
Chimalapa Zoque	ZOH	Mexico	75K	Johnson (2000-2005)
Chiquián Quechua	QXA	Peru	<5K	Proulx (1968)
Sharanahua	MCD	Peru	<1K	Déléage (2002)
Tzeltal	TZH	Mexico	600K	Kaufman (1960-1993)

Table 1: A brief description of the 8 languages in our dataset, including their ISO 693-3 codes and other information about the primary country where it is spoken, and number of speakers. Along with this, we have also included references to the resources that are being released as part of the AILLA-OCR corpora's first release.

by about 75K speakers as per the 2020 report from the Mexican National Institute of Statistics and Geography. It is called Tzunitzame by its speakers and it belongs to the Zoquean language family. While it is written in the Latin script, there is no digital support for Chimalapa Zoque. As per Ethnologue, it's vitality is considered threatened as its face-to-face use among speakers is growing slowly.

Chiquián Quechua (QXA) belongs to the Central Quechuan language family and is spoken by less than 5K people primarily in Central Peru in the Bolognesi province. It does not have a standardized orthography and remains primarily oral. In AILLA records, it is transcribed in the Latin script like other American indigenous languages. Ethnologue classifies the language's vitality as *shifting* which means the language is no longer being consistently passed on to new generations, and speakers are instead shifting to Spanish.

Sharanahua (MCD) is an indigenous Panoan language spoken by less than 1000 people in Madre de Dios and Ucayali regions and the upper Purús river area in Peru. It is written in the Latin script and is spoken by all members of the small indigenous language community, who are also often bilingual in Spanish. Ethnologue classifies Sharanahua's vitality as developing with standardized literature being slowly developed due to low literacy rates and the small community size.

Tzeltal (TZH) is a Cholan–Tzeltalan Mayan language (also called Bats'il K'op Tseltal) spoken by about 600K people in Mexico. According to Ethnologue, it is a developing language, with increasing digital support, and a small amount of literature in its Latin-based orthography. Its usage is currently almost exclusively oral, and there is almost universal bilinguality in Spanish for younger speakers.

3 AILLA-OCR Corpora Creation

Language and Document Selection We selected 8 languages that have permissive licenses, use the Latin alphabet, whose special diacritics were available on the English keyboard, and which had typed documents (as opposed to handwritten ones) for this phase of the AILLA-OCR corpus. A uniform sample of pages, covering different layouts, is chosen for annotation per language.

Annotation Setup Annotators are trained to use the annotation platform using standardized guidelines (§A), and are allowed to label each corrected bounding box from several semantic categories (header, footer, title, main text, table, text - *lang_label* etc.), as shown in Figure 2.

Annotators When working with data in small indigenous languages for language documentation purposes, it can be extremely challenging to find native speaker data annotators. Previous work has shown that annotators without knowledge of the indigenous language can be reasonably adept at performing OCR corrections, provided they can read the script or are trained to read it (Rijhwani et al., 2023). So, for our 8 languages, we recruited 14 computer science graduate students as our annotators. The authors timed themselves annotating a small sample of pages and calculated an estimated commitment of 30 mins per 5 pages. Based on this, the payout rate was set at $20/10 pages (1 hour of work). Cumulatively, the annotation process itself costed ~$750 (~40hrs), not including time for recruitment, outreach, training, quality control etc.

In the current stage of the corpus, due to limited budget, we have *one* annotation per page, therefore inter-annotator agreement was not computed.

Manual Audit The lead author manually audited all annotators' annotations for all 8 languages. The author can easily identify Spanish, French, and English text in the documents. Moreover, since each multilingual document has document-level language identifiers, indigenous language text on a page was inferred and labeled by process of elimination and additionally confirmed by matching with the language's Universal Declaration of Human Rights text.

Annotated Corpus Table 2 shows the distribution of the annotated pages and other metadata. Overall, the annotators completed 340 pages. Previous work has used 10-30 pages (we share 50 for most languages) to train post-correction models and the first-pass OCR for unannotated pages can be used for pre-training (Rijhwani et al., 2020).

4 OCR Post-Correction

First-Pass OCR We use a high-quality commercial OCR system, Google Vision, that is known to work well on endangered-language documents (Fujii et al., 2017; Rijhwani et al., 2020). We define a document $\mathcal{C}$ as follows:

$$\mathcal{C} = \{p_i\}_{i=1}^{K} \quad (1)$$

where p_i denotes the i-th page of a K page document. Performing OCR on page p_i gives us a first-pass output, f_i in the form of n_i bounding boxes x and the texts within them a. Each x contains the set of coordinates for the bounding box, and the corresponding string a represents the text within the box.

$$f_i = [(x_1, a_1), (x_2, a_2), .., (x_{n_i}, a_{n_i})]$$

Structural Corrections Annotators are required to first structurally correct the first-pass OCR outputs. This would involve scaling, translating, merging, or splitting bounding boxes, while keeping the text within faithful to the each box's new coordinates. We frame the structure post-correction task as follows. For every OCR'd input page f_i, we output a corrected page

$$q_i = [(y_1, b_1), (y_2, b_2), .., (y_{m_i}, b_{m_i})]$$

where m_i denotes the number of new bounding boxes after post-correction (may be different from n_i). We consider human-corrected q_i as the ground-truth text and layout. Note that while this step mainly transforms the structure, it also involves transferring the first-pass text (x_i, x_{i+1}, etc) from the first-pass boxes that now make up the corrected box b_i, and therefore, the texts are labeled as y_i.

Text Corrections We frame the text post-correction task to follow the structural corrections made in the previous step. For every structure-corrected page q_i, we output a corrected page:

$$r_i = [(z_1, b_1), (z_2, b_2), .., (z_{m_i}, b_{m_i})],$$

where m_i indicates the gold bounding boxes, and z_i indicates the transformed and corrected text in box b_i as compared to the first-pass text in structure-corrected gold boxes, y_i. We use character and word-level error rates (CER and WER) to report the quality of the first-pass OCR and the post-corrected outputs from the annotators.

5 Correction Results

Text Corrections Based on the gold dataset created by our annotators, Table 2 shows an evaluation of the text quality of the first-pass OCR by Google Vision. We see that for almost all languages, the CER (character-level error rate) and WER (word-level error rate) are both reasonable (<10%, with the exception of MAM and MIQ). This range is to be expected for low-resource languages written in extensions of the Latin-script (even with diacritics or new characters) and those that don't have available language models for decoding in Google Vision (all selected languages). Since desired error rates for readability are usually less than 2%, the first-pass results are a great starting point and with efficient post-OCR correction modeling or alignment improvements, this error could be reduced further.

Structure Corrections We have included detailed statistics on structural annotations (Table 2) and the raw data contains detailed metadata. To the best of our knowledge, no previous work has explored modeling techniques for structure post-correction, and so we did not include a benchmark for this task. Classically, structure is learned and predicted as a first-step and more emphasis is laid on post-correcting the extracted text. We anticipate that with better alignment and structure, the CER/WER scores in Table 2 will decrease further and consistently across languages with post-correction.

				Structure				Text				
693-3	Multiling	P_{total}	P_{ann}	μ_1	μ_2	μ_3	$\mu_{\Delta b}$	$\mu_{\Delta l}$	μ_d	μ_i	CER	WER
ZOH	SPA,ENG	3744	50	1.02	4.85	4.93	-0.24	5.61	0.73	6.34	3.56	6.15
CAC	SPA,ENG	564	50	1.76	5.59	4.71	-1.20	-4.34	11.95	7.61	4.12	5.33
MAM	SPA,ENG	144	50	0.94	3.98	7.36	-7.74	7.55	17.34	24.89	10.56	19.66
MIQ	SPA,ENG	61	50	0.40	2.26	3.78	-7.16	8.04	16.20	24.24	10.47	12.34
MCD	FRA	209	50	1.45	4.08	4.72	-7.17	10.65	2.73	13.38	7.13	9.15
QUH	SPA,ENG	216	50	1.24	3.76	3.98	0.36	1.46	0.46	1.92	2.72	3.76
QXA	SPA,ENG	29	20	2.88	17.06	20.53	-41.00	7.06	60.82	67.88	6.64	9.60
TZH	SPA	38	20	1.69	6.77	4.62	-8.85	14.92	8.08	23.00	1.43	2.73
AVG				1.42	6.04	6.83	-9.13	6.37	14.79	21.16	5.82	8.59

Table 2: For each of the 8 indigenous languages, we report the number of pages that we have selected to be part of the first release of the AILLA-OCR corpora (P_{total}) and number of human-annotated pages (P_{ann}). Along with this, we report some metrics to gauge the quality of the first-pass OCR outputs and the corrections. For structural annotations, we report some metadata including transforming involving one, two, three coordinates of a first-pass bounding box (μ_1, μ_2, μ_3). Annotators reduced the aggregate number of boxes detected across languages, to simplify the detected layout to different extents ($\mu_{\delta b}$). For text-corrections, we report average change in length of page text ($\mu_{\delta l}$), character-level deletions (μ_d), and character-level insertions (μ_i), in addition to the achieved character and word-level error rates.

6 Related Work

OCR Resource Creation Text or image-based datasets and corpora are most commonly created by scraping or crawling the web; however, we would like to highlight a few OCR-created datasets, especially those that work with indigenous languages. Cordova and Nouvel (2021) addresses the lack of resources for Central Quechua, since resources exist mostly in the dominant Southern variety, using OCR technologies. Hunt et al. (2023) digitizes an Akuzipik (indigenous language spoken in Alaska and parts of Russia) dictionary parallel with Russian text, which is very valuable for downstream NLP tasks. Other relevant but non-OCR dataset creation efforts include Guarani-Spanish news articles' (Góngora et al., 2021), Nahuatl speech translation (Shi et al., 2021), and Mazatec and Mixtec translations (Tonja et al., 2023).

Post-Correction An ideal post-OCR text correction algorithm would model the error distribution of the OCR algorithm's output text and systematically correct it (Berg-Kirkpatrick et al., 2013; Schulz and Kuhn, 2017). This can be an extremely valuable tool when digitizing indigenous language documents because the OCR pipeline's decoder language model is often of low-quality due to the low-resource nature of indigenous and endangered languages. Across the digitization efforts that we've highlighted and amongst others, it is quite common to perform text-based automatic/human post-correction (Maxwell and Bills, 2017; Cordova and Nouvel, 2021; Rijhwani et al., 2021). However, as mentioned in § 5, for structure and layout detection, previous work has focused on layout detection as a first-step (Bustamante et al., 2020) and it has not been explored as a post-processing step. This is primarily because there is a lack of ground-truth structural data (which our dataset provides). Previously, two major studies (Blecher et al., 2023; Zhong et al., 2019) have used existing large-scale corpora like arXiv to extract large-scale ground truth (source-code); but, this approach is not scalable to resource-creation efforts involving low-resource languages.

7 Conclusion

We present the AILLA-OCR corpus covering 8 indigenous languages of Latin America spoken across 6 countries. Our dataset is the first textual and structural corrections dataset. All data has been audited carefully by the authors to maintain high-quality annotations and rich metadata for future researchers to build modeling approaches on top off our dataset. We train a popular post-correction model to benchmark the text-corrections that highlights the utility of our dataset and associated gaps in structure modeling approaches. We hope this dataset will serve as a starting point to researchers to build and test new modeling approaches for the unexplored task of structure post-correction. Future work can also explore what methods would work best for reducing the error rates (both text and structure). This could involve classic post-OCR neural correction methods or utilize current advances in multimodal large language models.

Limitations

The main contribution of this paper is a new resource for textual and structure OCR post-correction in 8 low-resource indigenous languages of Latin America. Since such a contribution is best suited to a short paper, we did not include more extensive benchmarking.

Ethics Statement

The raw data digitized and corrected as part of the AILLA-OCR corpus initiative is entirely hosted by AILLA. The data is freely available to the general public, with some files shareable through request. The data can be used without asking for permission, and without paying any fees, as long as the resource and collection is cited appropriately. We acknowledge the linguists, native and heritage speakers, and the AILLA team for creating such a valuable repository of raw data in indigenous languages of Latin America. Our dataset, by design, digitizes and augments the raw data, to allow researchers and language community members to utilize it for modeling, and for educational purposes. An ethical implication of this work is that it will allow for more sustainable and equitable work in language resource creation and natural language processing.

Acknowledgments

This work was generously supported by the National Endowment for the Humanities under award PR-276810-21 and the George Mason University's Doctoral Research Scholars Award 2024-25 The authors are also grateful to the anonymous reviewers for their valuable suggestions, feedback, and comments.

References

Milind Agarwal, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2023. LIMIT: Language identification, misidentification, and translation using hierarchical models in 350+ languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14496–14519, Singapore. Association for Computational Linguistics.

Milind Agarwal and Antonios Anastasopoulos. 2024. A concise survey of OCR for low-resource languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 88–102, Mexico City, Mexico. Association for Computational Linguistics.

Taylor Berg-Kirkpatrick, Greg Durrett, and Dan Klein. 2013. Unsupervised transcription of historical documents. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 207–217, Sofia, Bulgaria. Association for Computational Linguistics.

Tulio Bermúdez Mejía. 2015. Miskitu dance, food, and traditions: traditional miskitu food, dance, songs, festivities. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public. PID ailla:119700. Accessed February 15, 2024. Other Contributors include Waldan Peter, Wanda Luz (Speaker), Bermúdez Mejía, Tulio (Transcriber), Waldan Peter, Wanda Luz (Translator).

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents.

Gina Bustamante, Arturo Oncevay, and Roberto Zariquiey. 2020. No data to crawl? monolingual corpus creation from PDF files of truly low-resource languages in Peru. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2914–2923, Marseille, France. European Language Resources Association.

Johanna Cordova and Damien Nouvel. 2021. Toward creation of Ancash lexical resources from OCR. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 163–167, Online. Association for Computational Linguistics.

Pierre Déléage. 2002. Sharanahua language collection of pierre déléage. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public.Accessed February 15, 2024.

Nora England. 1972-1985. Mam language stories and grammars. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public. PID ailla:119520, ailla:119520,ailla:119520,ailla:119520,ailla:119520. Accessed February 15, 2024.

Yasuhisa Fujii, Karel Driesen, Jonathan Baccash, Ash Hurst, and Ashok C. Popat. 2017. Sequence-to-label script identification for multilingual OCR. In *14th IAPR International Conference on Document Analysis and Recognition, ICDAR 2017, Kyoto, Japan, November 9-15, 2017*, pages 161–168. IEEE.

Santiago Góngora, Nicolás Giossa, and Luis Chiruzzo. 2021. Experiments on a Guarani corpus of news and social media. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 153–158, Online. Association for Computational Linguistics.

Nicholas Hopkins. 1964. A dictionary of the chuj (mayan) language language community. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public. PID ailla:119647. Accessed February 15, 2024.

Benjamin Hunt, Lane Schwartz, Sylvia Schreiner, and Emily Chen. 2023. Community consultation and the development of an online akuzipik-English dictionary. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–143, Toronto, Canada. Association for Computational Linguistics.

Heidi Anna Johnson. 2000-2005. A grammar of san miguel chimalapa zoque. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public. PID: ailla:119500 Accessed February 15, 2024.

Susan Kalt. 2016. Entrevista con tomas castro v y santusa quispe de flores. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public. PID ailla:119707, ailla:119707 . Accessed February 15, 2024. Other Contributors include Waldan Peter, Wanda Luz (Speaker), Bermúdez Mejía, Tulio (Transcriber), Waldan Peter, Wanda Luz (Translator).

Terrence Kaufman. 1960-1993. Colección de idiomas mayenses de terrence kaufman. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public. PID ailla:119707, ailla:119707 . Accessed February 15, 2024.

Michael Maxwell and Aric Bills. 2017. Endangered data for endangered languages: Digitizing print dictionaries. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 85–91, Honolulu. Association for Computational Linguistics.

Clemens Neudecker, Konstantin Baierer, Mike Gerber, Christian Clausner, Apostolos Antonacopoulos, and Stefan Pletschacher. 2021. A survey of ocr evaluation tools and metrics. In *Proceedings of the 6th International Workshop on Historical Document Imaging and Processing*, HIP '21, page 13–18, New York, NY, USA. Association for Computing Machinery.

Thi Tuyet Hai Nguyen, Adam Jatowt, Mickael Coustaty, and Antoine Doucet. 2021. Survey of post-ocr processing approaches. *ACM Comput. Surv.*, 54(6).

Paul Proulx. 1968. Chiquian quechua vocabulary. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public.Accessed February 15, 2024.

Shruti Rijhwani, Antonios Anastasopoulos, and Graham Neubig. 2020. OCR Post Correction for Endangered Language Texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5931–5942, Online. Association for Computational Linguistics.

Shruti Rijhwani, Daisy Rosenblum, Antonios Anastasopoulos, and Graham Neubig. 2021. Lexically aware semi-supervised learning for OCR post-correction. *Transactions of the Association for Computational Linguistics*, 9:1285–1302.

Shruti Rijhwani, Daisy Rosenblum, Michayla King, Antonios Anastasopoulos, and Graham Neubig. 2023. User-centric evaluation of OCR systems for kwak'wala. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 19–29, Remote. Association for Computational Linguistics.

Sarah Schulz and Jonas Kuhn. 2017. Multi-modular domain-tailored OCR post-correction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2716–2726, Copenhagen, Denmark. Association for Computational Linguistics.

Jiatong Shi, Jonathan D. Amith, Xuankai Chang, Siddharth Dalmia, Brian Yan, and Shinji Watanabe. 2021. Highland Puebla Nahuatl speech translation corpus for endangered language documentation. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 53–63, Online. Association for Computational Linguistics.

Atnafu Lambebo Tonja, Christian Maldonado-sifuentes, David Alejandro Mendoza Castillo, Olga Kolesnikova, Noé Castro-Sánchez, Grigori Sidorov, and Alexander Gelbukh. 2023. Parallel corpus for indigenous language translation: Spanish-mazatec and Spanish-Mixtec. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 94–102, Toronto, Canada. Association for Computational Linguistics.

Xu Zhong, Jianbin Tang, and Antonio Jimeno-Yepes. 2019. Publaynet: Largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 1015–1022. IEEE.

A Annotation Setup

Annotation Guidelines We shared annotatator assignments over email and the guidelines were shared using a YouTube video due to the visual nature of the task. We share the email template below with anonymity-compromising links redacted temporarily.

Subject: Annotation Assignments - [[NAME]] (Annotator #[[ID]])

Hi [[NAME]],

Thank you for being a part of this annotation effort for AILLA (Archive of the Indigenous Languages of Latin America). We appreciate you taking out the time to help us digitize and document these valuable resources. From the information you shared with us on the Google Form, you have been assigned **[[N]] labeling tasks**. Once you've completed your annotation assignment, please let us know (by replying to this email) and I will send you a **$[[AMOUNT]]** Amazon gift card. If you like doing the annotations, you can also always request more assignments.

Assignments:
Your unique ID is still **Annotator [[ID]]**
(Example) Language assignments:

- **mam [Mam]**. 7 pages. Task IDs: 40743-40749
- **cac [Chuj]**. 8 pages. Task IDs: 40280-40287
- **zoh [Chimalapa Zoque].** 15 pages. Task IDs. 39457-39471

While you only need your ID and language codes (mam, cac, zoh) to find your assignments, I will encourage you to check your tab before annotating to make sure you're actually seeing the tasks I've assigned you. If you notice anything off, just let me know.

Setup Instructions:
To enable swift annotation, we will be utilizing a open-source data labeling platform, [[redacted]]. If you haven't already, we invite you to create a Community Edition account through the signup link given below. We request that you not share the link publicly. [[redacted]]

Get Started:
Once you have created your account, you can can use [[redacted] to login and begin your annotations! We've made a short 5-minute video to guide you through the interface, how the annotation process works, and our expectations. Please watch it here **[[redacted]]** before you start annotation. The video is English closed-captioned (CC).

If you have any followup questions (about a specific assignment, the process, account setup etc.), please feel free to contact us on this thread.

Connecting Automated Speech Recognition to Transcription Practices

Blaine Billings
University of Hawai'i at Mānoa
blainetb@hawaii.edu

Bradley McDonnell
University of Hawai'i at Mānoa
mcdonn@hawaii.edu

Johan Safri

Wawan Sahrozi

Abstract

One of the greatest issues facing documentary linguists is the transcription bottleneck. While the large quantity of audio and video data generated as part of a documentary project serves as a long-lasting record of the language, without corresponding text transcriptions, it remains largely inaccessible for revitalization efforts and linguistic analysis. Automated Speech Recognition (ASR) is frequently proposed as the solution to this problem. However, two issues often prevent documentary linguists from making use of ASR models: 1) the thought that the typical documentary project does not have sufficient data to develop an adequate ASR model and 2) that correcting the output of an ASR model would be more time-consuming for transcribers than simply creating a transcription from scratch. In this paper, we tackle both of these issues by developing an ASR model in the larger context of a documentation project for Nasal, a low-resource language of western Indonesia. Fine-tuning a larger pre-trained language model on 25 hours of transcribed Nasal speech, we produce a model that has a 44% word error rate. Despite this relatively high error rate, tests comparing speed of transcribing from scratch and correcting ASR-generated transcripts show that the ASR model can significantly speed up the transcription process.

1 Introduction

The use of Automated Speech Recognition (ASR) in language documentation and revitalization contexts has been met with almost universal enthusiasm due to its promise to loosen the transcription bottleneck (see e.g. Berez-Kroeker et al., 2023). The basic idea behind this approach is that while the limited data of low resource languages is often not able to produce highly accurate models comparable with those of high resource languages, it is more efficient to correct the output of an ASR model than to produce a transcription from scratch (see Foley et al., 2018; Bird, 2021). This approach crucially relies on an ASR model to generate transcriptions accurate enough that making corrections requires less time and effort than creating a transcription anew. More recently, there is a growing number of ASR studies that demonstrate how the accuracy of models with relatively little training data are able to be improved through the use of pre-trained models of high resource languages (Coto-Solano et al., 2022) or supplemental written corpora in the target language (Bartelds et al., 2023; San et al., 2023). Despite improvements to ASR models in language documentation and revitalization projects, it remains difficult to assess the usefulness of such models as there has been very little reported about how transcribers, who are often native speakers and members of the speech community, interact with the outputs of these ASR models.

In this paper, we address this issue through a case study of Nasal, an endangered, under-resourced Austronesian language of Sumatra, Indonesia. The study comprises two parts. The first discusses the development of an ASR model for Nasal through the fine-tuning of a pre-trained high-resource language model using Whisper (Whisper 2024). The second addresses the usefulness of such a model for Nasal transcribers by comparing the process of transcribing in ELAN from scratch against correcting transcriptions generated by the ASR model.

This paper is organized as follows. The remainder of this section provides an introduction to the Nasal speech community (§1.1) and the ongoing Nasal documentation project (§1.2). §2 describes the development of the ASR model for Nasal. §3 discusses the results from the model and the comparison between the two transcription methods. §4 provides some discussion on the viability of ASR models for documentation projects, and §5 gives some summary remarks.

Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 128–132

1.1 The Nasal speech community

Nasal [glottocode: nasa1239] is a Sumatran language spoken by approximately 3,000 people in southwest Sumatra, Indonesia (Billings and McDonnell, 2024). The Nasal speech community represents a fringe case of small-scale multilingualism (Pakendorf et al., 2021) where, in addition to Nasal, members of the community use two regionally significant varieties of Malay – Kaur [glottocode: kaur1269] and South Barisan Malay [glottocode: cent2053] – in daily life. Nasal was not known to linguists until 2007 (Anderbeck and Aprilani, 2013) and thus little documentation of the language existed until the authors, a team of outsider linguists and members of the Nasal speech community, initiated a documentation project in 2017 that continues to the present (McDonnell 2017; McDonnell et al. ongoing). At the onset of the project, several members of the Nasal community, including the third and fourth authors, were provided training in a simplified system of Discourse Transcription (Du Bois et al., 1993), methods for free translations into Indonesian (Schultze-Berndt, 2006), and ELAN (ELAN 2024) and Fieldworks Language Explorer (FLEx; Fieldworks 2024) software.

1.2 Documentation on Nasal

The documentation of Nasal consists of audiovisual recordings of everyday conversations, culturally important events, active elicitation sessions to elicit and discuss word meanings, and structured tasks to elicit aspects of Nasal phonology and grammar. In the vast majority of recordings, speakers were recorded on separate channels using lapel or headset microphones. Recording in this way better facilitates the ability to train, test, and use ASR models on conversational data by targeting individual speaker audio and reducing signal bleeding.

The largest portion of the documentation consists of everyday conversations, followed by active elicitation sessions. The majority of recordings that fall into the prior category have been transcribed using Discourse Transcription and translated into Indonesian using ELAN and later glossed in FLEx. However, the majority of the recordings that fall into the latter category have yet to be transcribed or translated. The aim of these active elicitation sessions is to discuss a large number of lexical items with their associated meanings and uses by facilitating conversations of various semantic domains. This documentation forms the basis of a Nasal dictionary project.

At the outset of this documentation project, project leaders (which includes the second author) hosted a series of meetings to discuss issues such as project outcomes and orthography development as well as training workshops in Discourse Transcription, ELAN, and FLEx at the Atma Jaya Catholic University of Indonesia. The third and fourth authors participated in the workshop. During the workshop, they began producing transcriptions, and with the help of project team members with linguistics training, they began transcribing recordings of conversations. Over subsequent years, Nasal transcribers have honed transcriptions and translations as well as their methods for transcribing. Currently, transcriptions and corresponding translations are produced by the third and fourth authors in ELAN with the following workflow:

1. **Segmentation**: Segment recording into Intonation Units in *Segmentation Mode* in ELAN
2. **Transcription**: Fill empty annotations with Nasal transcriptions
3. **Discourse & translation**: Input corresponding discourse transcriptions and Indonesian translations necessary for analysis
4. **Context**: Create additional annotations for various types of sporadic notes (speech context, code-switching, etc.)

Of these four steps, the one to be addressed by the ASR model is *transcription*. Depending on the granularity of transcription, it can take upwards of forty minutes to transcribe one minute of audio (Seifart et al., 2018), often requiring listening to each individual annotation up to five or ten times. Given its drastically greater time requirement over the other steps, we decided to work on implementing ASR for transcription first with plans to tackle the remaining three in the future.

2 Methods

2.1 Data preparation

The data used for training and testing the ASR model consist of transcribed audio from twenty-five recordings of various genres: everyday conversation (13), brief map game and role-play tasks originally recorded for prosody elicitation (10), and semantic domain active elicitation sessions recorded for dictionary development (2). Sessions lasted anywhere from fifteen minutes to three hours for approximately 25 total hours of recording time.

Each session included two to four speakers with 49 unique speakers in total (seven appeared in two recordings, one appeared in three recordings). While nearly all of the dictionary elicitation sessions feature both the third and fourth author as facilitators, one or two additional speakers, differing by recording, are also present, and thus training the model on a diverse speaker population accords with the intended use case. The transcriptions for these recordings were produced by the third and fourth authors over a period of four years. The text for these transcriptions constitute the corresponding text input for the language model. Audio segments corresponding to the timestamps for the annotations were extracted from each individual speaker audio file. This training data for the ASR model totaled more than 160,000 words in 66,500 annotations accounting for 17.5 hours of speech time (that is, excluding all silence, i.e. non-annotated segments, from each speaker audio file).

2.2 ASR training

The data was split into two sets, training data and testing data, with a simple 80/20 division of the annotations, respectively. The ASR model was built by fine-tuning the small model from Whisper (Whisper 2024), using the small model's pretrained tokenizer and feature extractor from Indonesian, a related language. Fine-tuning ran over 5,000 steps with evaluation according to word error rate (WER) taken at every 500-step checkpoint. The best of these checkpoints was used in generating the transcriptions for the transcription task.

2.3 Transcription task

In order to determine the viability of using ASR-generated transcriptions over transcribing recordings from scratch, we designed a short transcription task. In this task, the first author selected two excerpts, one from a conversation and the other from an active elicitation session and neither of which was used in training and testing the model. The 2 minute and 30 second excerpts were segmented in ELAN, leaving empty annotations. The third and fourth authors then each produced four transcriptions on the two excerpts. The third author transcribed the elicitation session first and the conversation second, while the fourth author transcribed the conversation first and the elicitation session second. Both started their first file by correcting the ASR-generated transcript and then transcribed from scratch, whereas with the second file, they first transcribed from scratch and then corrected the ASR-generated transcript. The task was designed in this way so as to balance any confounding influence from the order of session or transcription method. The third and fourth authors screen-recorded the process of transcribing each of the four files and later compared their experiences in each.

3 Results

3.1 ASR results

Over the 5,000 steps of fine-tuning the ASR model, the lowest error rate attained in the training checkpoints was 43.9%, a significant improvement over the 67.2% of the previous model trained on Nasal data (San et al., 2023). When tested against two segments of audio not included in the testing set (one from an everyday conversation, one from an active elicitation session), WER was higher at 60.1% (conversation) and 54.1% (active elicitation). Character error rate (CER) was similarly calculated for both to gain a better understanding of the kinds of errors made by the model. These came out to be 21.4% (conversation) and 20.4% (active elicitation), corresponding as expected with the WER above given the distribution of word length in Nasal. On further inspection of the ASR output, these rates were found to be inflated by errors with interjections (e.g. transcribing a single syllable *m* rather than a two syllable *mm*) and orthographic variations introduced by the training data (see *Limitations* at the end of this paper).

3.2 Transcription comparison results

After completing the transcription task, the third and fourth authors found that correcting the ASR-generated transcriptions was able to significantly speed up transcription time, with all four tests showing improvements. It is unsurprising that the gains were higher (23.49% compared to 11.30%, 32.29% compared to 21.92%) when the ASR-assisted method occurred second — as the authors had already been exposed to the media once — but the improvement is nonetheless apparent. Although many corrections needed to be made to the generated transcripts, changes were most often minor, single-letter or single-word edits and rarely required reannotating an entire IU. Furthermore, revising the automatic transcriptions was preferred over transcribing from scratch, since having a baseline of transcribed text meant the audio needed to

be listened to fewer times in order for the recorded speech to be accurately determined.

4 Discussion

Documentary projects typically result in the production of a large body of audio and video recordings of various genres, from narratives to conversation and elicitation. Whether to assist in the production of language materials or in linguistic analysis, creating transcriptions of these recordings is often a normal part of a documentary linguists' workflow. As has been demonstrated here for Nasal, once a small body of transcribed audio data has been produced, these transcriptions can be leveraged to fine-tune an ASR model to speed up the transcription of remaining or future documentary data. The authors, who themselves work directly with transcribing the Nasal data, have found that correcting the transcriptions generated by a model trained on such data significantly aids in the transcription process.

One of the primary motivations for developing this ASR model for Nasal is the ongoing compilation of the Nasal dictionary. A large corpus of active elicitation sessions, now totaling 70 hours of data, remains to be transcribed. Transcriptions of these sessions – many of which contain lexical items absent from the corpus of everyday conversation – would make the data more usable and more easily linked with dictionary outputs. Although recordings of active elicitation sessions contain a greater frequency of new lexical items, the ASR model developed here did not show significant differences in accuracy in transcribing the active elicitation recording and the everyday conversation recording, proving equally useful for the dictionary compilation process.

As discussed above (§1.2), the transcription workflow in the Nasal documentation project consists of four steps. Since transcription time is the most significant problem in this workflow, we decided to tackle it first. ASR-generated transcriptions for Nasal speech have already proven to significantly speed up the transcription process. Addressing the remaining three steps could further contribute to faster transcription of documentary data. For example, further AI models targeting prosody and intonation could be implemented to speed up the IU-based segmentation used in our transcriptions. Discourse transcription will likely need to remain manual, but machine translation has also shown promise in low resource contexts (see van Esch et al., 2019) and Whisper AI may even prove to be useful in this regard (see the description at Whisper 2024). Finally, the creation of additional contextual notes, while important for linguistic analysis, results in less than two percent of the total number of annotations and thus needs not be immediately addressed with computer-assisted methods.

5 Conclusion

We fine-tuned a pre-trained ASR model with 25 hours of data typical to a documentation project. Through comparing the processes of correcting transcripts generated from this model and transcribing from scratch, it was demonstrated that such a model proves effective for improving the transcription workflow and reducing the amount of time necessary for transcribing documentary data. We believe that such models are more accessible to documentary linguists than typically thought and can greatly assist in the transcription process.

Limitations and future prospects

In reviewing the ASR-generated transcripts, it was clear to the authors that a major contributor to the increased WER and CER was the lack of standardization in the Nasal orthography. For example, many words contain two-vowel sequences and can be written with or without a predictably inserted glide (e.g., *gauh*, *gawuh* 'just' are both valid written forms). In other cases, a shortened form of a word used in rapid speech is variably reflected in the transcripts either by the longer or the shorter form (e.g., either *jenu*, *nu* 'before, earlier' may be transcribed even if *nu* is uttered). For these and similar issues training ASR models in the documentary linguistics context, see Meelen et al. (2024).

In an effort to determine if better results could be easily attained, the authors addressed the first issue by standardizing spelling of vowel sequences throughout the transcriptions, included seven additional hours of recording, and used Whisper's medium baseline to generate a new model. Results from this model show an improved WER of 37.0% and CER of 14.4%.

Acknowledgments

We would like to thank people from the Nasal community who took part in this documentation project. We are also grateful to the National Research and

Innovation Agency (BRIN) for supporting this research in Indonesia as well as the Center for Culture and Language Studies at Atma Jaya Catholic University of Indonesia for sponsoring this research, especially the center's director Yanti. This material is based upon work supported by the National Science Foundation under Grant No. (1911641). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

Karl Anderbeck and Herdian Aprilani. 2013. *The improbable language: Survey report on the Nasal language of Bengkulu, Sumatra*. SIL International.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–729, Toronto, Canada. Association for Computational Linguistics.

Andrea L. Berez-Kroeker, Shirley Gabber, and Aliya Slayton. 2023. Recent Advances in Technologies for Resource Creation and Mobilization in Language Documentation. *Annual Review of Linguistics*, 9(1):–330342568.

Blaine Billings and Bradley McDonnell. 2024. Sumatran. *Oceanic Linguistics*, 63(1):112–174.

Steven Bird. 2021. Sparse Transcription. *Computational Linguistics*, 46(4):713–744.

Rolando Coto-Solano, Sally Akevai Nicholas, Samiha Datta, Victoria Quint, Piripi Wills, Emma Ngakuravaru Powell, Liam Koka'ua, Syed Tanveer, and Isaac Feldman. 2022. Development of automatic speech recognition for the documentation of Cook Islands Māori. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3872–3882, Marseille, France. European Language Resources Association.

John W. Du Bois, Stephan Schuetze-Coburn, Susanna Cumming, and Danae Paolino. 1993. Outline of discourse transcription. In Jane Anne Edwards and Martin D. Lampert, editors, *Talking Data: Transcription and Coding in Discourse Research*, pages 45–89. Lawrence Erlbaum Associates, Publishers, Hillsdale, NJ.

ELAN (version 6.8) [Computer software]. 2024. Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen.

FieldWorks (version 9.1.25) [Computer software]. 2024. SIL Global, Dallas, TX.

Ben Foley, Josh Arnold, Rolando Coto-Solano, Gautier Durantin, T. Mark Ellison, Daan Van Esch, Scott Heath, František Kratochvíl, Zara Maxwell-Smith, David Nash, Ola Olsson, Mark Richards, Nay San, Hywel Stoakes, Nick Thieberger, and Janet Wiles. 2018. Building Speech Recognition Systems for Language Documentation: The CoEDL Endangered Language Pipeline and Inference System (ELPIS). In *6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, pages 205–209. ISCA.

Bradley McDonnell. 2017. *Documentation of Nasal: An overlooked Malayo-Polynesian isolate of southwest Sumatra*. Endangered Languages Archive.

Bradley McDonnell, Blaine Billings, Jacob Hakim, Johan Safri, and Wawan Sahrozi. ongoing. *The languages of the Nasal speech community*. Collection BJM02 at catalog.paradisec.org.au [Open Access].

Marieke Meelen, Alexander O'neill, and Rolando Coto-Solano. 2024. End-to-end speech recognition for endangered languages of Nepal. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 83–93, St. Julians, Malta. Association for Computational Linguistics.

Brigitte Pakendorf, Nina Dobrushina, and Olesya Khanina. 2021. A typology of small-scale multilingualism. *International Journal of Bilingualism*, 25(4):835–859.

Nay San, Martijn Bartelds, Blaine Billings, Ella de Falco, Hendi Feriza, Johan Safri, Wawan Sahrozi, Ben Foley, Bradley McDonnell, and Dan Jurafsky. 2023. Leveraging supplementary text data to kickstart automatic speech recognition system development with limited transcriptions. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 1–6, Remote. Association for Computational Linguistics.

Eva Schultze-Berndt. 2006. Linguistic annotation. In Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, editors, *Trends in Linguistics. Studies and Monographs [TiLSM]*, pages 213–252. Mouton de Gruyter, Berlin, New York.

Frank Seifart, Nicholas Evans, Harald Hammarström, and Stephen C. Levinson. 2018. Language documentation twenty-five years on. *Language*, 94(4):e324–e345.

Daan van Esch, Ben Foley, and Nay San. 2019. Future directions in technological support for language documentation. In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 14–22, Honolulu. Association for Computational Linguistics.

Whisper (version 20240930) [Computer software]. 2024. OpenAI, San Francisco, CA.

Developing a Mixed-Methods Pipeline for Community-Oriented Digitization of Kwak'wala Legacy Texts

Milind Agarwal[1], **Daisy Rosenblum**[2], **Antonios Anastasopoulos**[1],
[1]George Mason University, [2]University of British Columbia,
Correspondence: magarwa@gmu.edu

Abstract

Kwak'wala is an Indigenous language spoken in British Columbia, with a rich legacy of published documentation spanning more than a century, and an active community of speakers, teachers, and learners engaged in language revitalization. Over 11 volumes of the earliest texts created during the collaboration between Franz Boas and George Hunt have been scanned but remain unreadable by machines. Complete digitization through optical character recognition has the potential to facilitate transliteration into modern orthographies and the creation of other language technologies. In this paper, we apply the latest OCR techniques to a series of Kwak'wala texts only accessible as images, and discuss the challenges and unique adaptations necessary to make such technologies work for these real-world texts. Building on previous methods, we propose using a mix of off-the-shelf OCR methods, language identification, and masking to effectively isolate Kwak'wala text, along with post-correction models, to produce a final high-quality transcription.[1]

1 Introduction

In this work, we focus on the Kwak'wala language (Wakashan, ISO 639.3 kwk), spoken on Northern Vancouver Island, nearby small islands, and the opposing mainland. Kwak'wala and several other Indigenous languages in this region have over a century of of legacy documentation created by early anthropologists, primarily in orthographies developed by Franz Boas to capture complex and typologically unusual phonetic and phonological inventories (Himmelmann, 1998; Grenoble and Whaley, 2005). Kwak'wala, for example, has 42 consonant phonemes represented with a selection of characters from the North American Phonetic Alphabet (cousin to the IPA), and over 13 possible vowel pronunciations represented with a heavy dose of diacritics and digraphs in all its scripts. During the first half of the 20th-century, scripts such as these were created and used by ethnologists, researchers, and collectors to transcribe the languages spoken in communities across North America. Between 1897 and 1965, an extensive series of texts in Indigenous languages was published by the United States Bureau of American Ethnology (BAE, now the Smithsonian). The collaboration between Franz Boas and George Hunt generated 11 volumes of published texts over 50 years, as well as extensive unpublished documentation. This script is difficult for anyone to read, amplifies phonetic complexity, and is primarily considered a legacy script, limiting access only to a few. However, many precious documents with detailed information of cultural value, were created in this script (see Figure 1), necessitating their accurate digitization and transliteration into modern Kwak'wala writing systems. Note that while Kwak'wala is classified as an endangered language with most first-language speakers over the age of 70, it has thriving language revitalization programs focused on creating new speakers among children and adults. Research progress for Kwak'wala and its three scripts (U'mista in the Northern communities, SD-72 in the Southern communities, and the legacy Boas-Hunt script) is urgent to better support revitalization and educational efforts led by community members. Currently, Kwak'wala, like many other 'low-resource' endangered and Indigenous languages, lags behind in the number of available computational tools (Agarwal and Anastasopoulos, 2024).

To remove this disadvantage and enable greater online community participation, in our project, we focus on digitization of valuable Kwak'wala texts, prioritized according to community needs, to enable building tools such as word processing, speech to text, predictive typing, etc. We create these resources by applying existing optical character recognition and language identification techniques,

[1]Relevant code and data resources are available here.

Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 133–138

and making necessary modifications to suit them to Kwak'wala. We use grapheme-to-phoneme technology (Pine et al., 2022) to transliterate texts into the U'mista orthography, one of two community-preferred modern Kwak'wala writing systems. A draft of the 1921 Boas-Hunt text produced through a previous collaboration was distributed to 50 community and academic experts for review, and the feedback we gathered through surveys and conversations informed our production of a second draft PDF for publication and distribution. This feedback assisted us in prioritizing highly-valued elements of the texts which had originally been overlooked or erased through the process, such as the text-referenced line numbers cited by Boas in his dictionary and grammar, creating an analog concordance and networking these Kwak'wala texts into the prototypical 'Boasian trilogy'. This research, conducted in consultation with community-based language programs and guided by community priorities, will greatly increase access to culturally significant documents, thus empowering the community to draw on these resources to propagate the language and culture to future generations (Lawson, 2004).

2 Data

We focus our effort on digitizing five books that include Kwak'wala text and, often, parallel translations in English. We chose these books due to their similar fontfaces, clean layout, typed content (as opposed to handwritten), and high-quality scans.

1. **Jesup Volume 5, Part 1 (Franz Boas and George Hunt, January 1902)**: This 280 page book is part of the Jesup North Pacific Expedition publication series and contains an anthology of Kwak'wala texts in Hunt-Boas orthography. The book primarily contains dictated Kwak'wala texts (with parallel running English translation), and an appendix with grammatical information, stems, vocabulary, and traditional songs sung by Kwak'wala communities (Boas and Hunt, 1902a).

2. **Jesup Volume 5, Part 2 (Franz Boas, December 1902)**: This 144 page book is mostly formatted similarly to Volume I, but this particular volume doesn't contain interlinear text, and instead has an abundance of monolingual single-column Kwak'wala prose (Boas and Hunt, 1902b).

3. **Jesup Volume 5, Part 3 (Franz Boas, 1902)**: This is the third and final part of Volume 5, and is formatted similarly to Part II. It also contains a substantial appendix with vocabulary and stems (Boas and Hunt, 1902c).

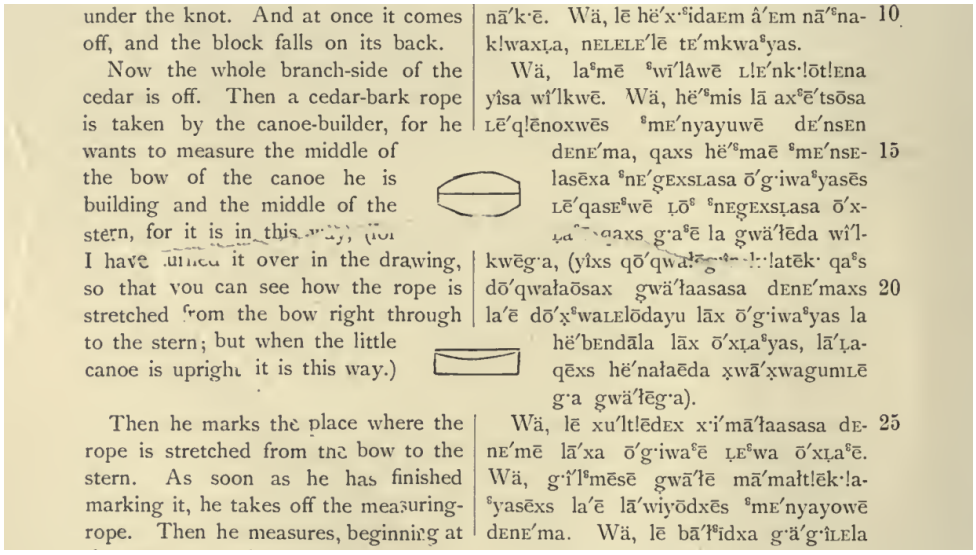
4. **Jesup Volume 10 (Franz Boas and George Hunt, 1906)**: This book contains valuable texts from the North Pacific Expedition in Kwak'wala and Haida (Masset dialect) languages. For the purposes of this project, we use only the first part of the first 282 pages of this book that contains the Kwak'wala texts (Boas and Hunt, 1906).

5. **The Kwakiutl Of Vancouver Island Volume II (Franz Boas, 1909)**: This book contains valuable texts in Kwak'wala on woodworking, weaving, hunting, fishing, clothing, measurements etc. Most of the descriptions are in English, with plenty of inline figures (that disrupt the layout extraction of the OCR), but there are also tens of pages of Kwak'wala dictated text (with parallel running English). The book alternates between a single and double column layout (Boas, 1909).

3 Related Work

Optical character recognition (OCR) is a multi-label classification problem, where a patch of pixels is shown to an OCR model, and its task it to classify it into one of n classes (usually the alphabet + punctuation). When extended to entire pages or documents containing textual material, this can allow us to digitize previously inaccessible materials. Since it is crucial for digitization of manuscripts, linguistic field notes etc., it is widely used in the humanities to render such texts accessible to researchers and to language community members (Reul et al., 2017; Rijhwani et al., 2021, 2020).

This technique has, over time, developed into a discipline, with many excellent surveys written covering the technical and applied aspects of OCR (Agarwal and Anastasopoulos, 2024; Nguyen et al., 2021; Neudecker et al., 2021; Memon et al., 2020; Hedderich et al., 2021). Today, many open-source (Tesseract and Ocular) and commercial systems (Google Vision and Microsoft OCR) exist for OCR and they can extract text from most images quite effectively, as long as they are in a language it has seen during training (Smith, 2007; Blecher et al., 2023; Berg-Kirkpatrick et al., 2013). Several research efforts before have tried to address the lack



under the knot. And at once it comes off, and the block falls on its back.

Now the whole branch-side of the cedar is off. Then a cedar-bark rope is taken by the canoe-builder, for he wants to measure the middle of the bow of the canoe he is building and the middle of the stern, for it is in this way, (for I have turned it over in the drawing, so that you can see how the rope is stretched from the bow right through to the stern; but when the little canoe is upright it is this way.)

Then he marks the place where the rope is stretched from the bow to the stern. As soon as he has finished marking it, he takes off the measuring-rope. Then he measures, beginning at

nā′k·ē. Wä, lē hë′x·ˢidaem â′em nā′ˢna- 10
k!waxɬa, nelele′lē te′mkwaˢyas.

Wä, laˢmē ˢwī′lâwē l!e′nk·!ōt!ena yîsa wī′lkwē. Wä, hë′ˢmis lā axˢē′tsōsa le′q!ēnoxwēs ˢme′nyayuwē de′nsen
denɛ′ma, qaxs hë′ˢmaē ˢme′nse- 15
lasēxa ˢne′gexsɬasa ō′g·iwaˢyasēs le′qaseˢwē ɬōˢ ˢnegexsɬasa ō′x-
ɬa′ˢ·qaxs g·aˢē la gwä′lēda wî′l-
kwēg·a, (yîxs qō′qwa·ˢ·l·!latēk· qaˢs dō′qwałaōsax gwä′łaasasa dɛne′maxs 20
la′ē dō′xˢwalelōdayu lāx ō′g·iwaˢyas la hë′bendāla lāx ō′xɬaˢyas, lā′ɬa-
qēxs hë′nałaēda x̣wā′x̣wagumɪlē g·a gwä′łēg·a).

Wä, lē xu′lt!ēdex x·i′mā′łaasasa de- 25
ne′mē lā′xa ō′g·iwaˢē ɬeˢwa ō′xɬaˢē. Wä, g·î′lˢmēsē gwā′łē mā′małtlēk·!aˢyasēxs la′ē lā′wiyōdxēs ˢme′nyayowē denɛ′ma. Wä, lē bā′lˢidxa g·ä′g·îlela

Figure 1: Example two-column text from the Kwakiutl of Vancouver Island (1909) collection. Notice the abundance of inline figures in this text that interfere with Google Vision's OCR pipeline.

of resources in certain indigenous languages using OCR to create machine-readable texts such as Central Quechua (Cordova and Nouvel, 2021) and Akuzipik (Hunt et al., 2023).

4 Methodology

4.1 First-Pass OCR

Google Vision is a well-maintained modern OCR tool that tends to work well on Latin/Roman orthographies and their extensions (Fujii et al., 2017; Rijhwani et al., 2020). Additionally, since our collections are composed of multilingual texts, it is vital to use a tool that can handle multilinguality within documents. It is a paid (per page) service at the rate of $1.25/1000 pages, but the first 1000 pages every month are free. Since our project and its digitization was conducted over several months, we did not incur any first-pass OCR charges. Open-source alternatives like Ocular or Tesseract may also be used for OCR, especially when data restrictions require local processing, instead of sending data through APIs to Google servers. However, note that they require manual training, computational expertise, preparation of training and evaluation data, and have a higher learning curve (Smith, 2007; Berg-Kirkpatrick et al., 2013) .

4.2 Language Identification

We use language identification (langID) to distinguish English from Kwak'wala as proxy for structure identification in our collections. LangID is also extremely important to enable masking of non-Kwak'wala text. To the best of our knowledge, no off-the-shelf language identification model supports Kwak'wala in the Hunt-Boas orthography. So, we use fastText as it allows easily training on custom data from scratch on CPU (Joulin et al., 2017; Agarwal et al., 2023). Our final model, trained on first-pass Kwak'wala and English texts (binary model, default fastText parameters, 1000 sentences per language), achieves a sentence-level accuracy of 99.84%. This model is applied on each page's bounding boxes, which allows us to reorganize text with improved layout.

4.3 Masking

The texts are diversely formatted, and contain additional information in illustrations, figures, line numbers and the like. Since the post-correction model (see Section 4.4) is trained to correct Kwak'wala text alone, we quickly realized that real-world digitization projects like ours require the development of a masking pipeline. Additionally, masking is preferable as post-OCR correction models are best trained for a single language, and English first-pass OCR quality is often extremely good without requiring post-OCR correction. Following the first-pass OCR, we apply a masking layer that temporarily hides/masks all English text (as labeled by the langID model), numbers, and certain punctuation like parentheses, that were impacting subsequent steps in the pipeline. This allows us to isolate, to the best ability of the language identification model and based off our overall structural cropping, the first-pass text in Kwak'wala that needs

post-correction. For each line, token-level indices of the masked tokens are stored in a separate file at this stage. This allows us to track exactly what tokens were masked so we can reintroduce them in the same spots after post-correction.

4.4 Post-Correction

Post-correction can allow us to automatically correct errors in very low-resource OCR settings, by training a correction model on a small sample of first-pass and reference text pairs (Kolak and Resnik, 2005; Dong and Smith, 2018). The post-correction model has a multi-source neural architecture, based on Rijhwani et al. (2021), which has been shown to reduce character error rates by 32–58%. We use the model from this paper directly for post-correction, with the weighted finite-state transducer setting for lexical induction turned off, as it was shown in Rijhwani et al. (2021) not to improve Kwak'wala post-correction in contrast to other low-resource languages. This is likely due to the polysynthetic structure of Kwak'wala words, leading to low lexical frequency of any one token. We train the model from scratch on the labeled Boas-Hunt dataset shared in the paper, with pre-training conducted on the unlabeled first-pass OCR outputs for the collection. We first replicated the character error rate results from the original paper to ensure reliability of the model. Then we applied our trained model to our test text. The unmasked Kwak'wala text from the previous stage is fed line-by-line to the post-correction model to obtain post-corrected Kwak'wala text.

4.5 Reconstruction

Next, we reinsert the masked tokens (English text, punctuation, line numbers in the margins, etc.) into the post-corrected sentence at the appropriate indices. This gives us the final reconstructed multilingual output, along with crucial indexical cross-referencing information such as page and line numbers. At this stage, the Kwak'wala text is also transliterated into the desired modern orthography (ex. U'mista or SD-72) using grapheme-to-phoneme conversion to allow for better readability and accessibility of the text.

4.6 Evaluation

We compare the reconstructed output to gold reference texts to evaluate the digitized texts' quality. We do this for two books at two levels:

	Jesup 5.1, 1902		Kwakiutl, 1909	
	CER	SER	CER	SER
First Pass	0.43	25	0.33	18
Corrected	0.18	2	0.15	3

Table 1: For both books, we find that using our pipeline greatly reduces not only textual errors (CER) but also greatly improves the layout and structure (SER)

- **Textual Errors**: To capture textual errors, such as misspellings, missing diacritics, tokenization etc., we use Character Error Rate (CER). This is a popular metric to understand character-level variations and error distributions in the output text, as compared to the gold-reference. For morphologically complex and polysynthetic languages like Kwak'wala, CER is a much better metric than word-level scores because a large amount of vocabulary would be unseen at test-time (Rijhwani et al., 2023).

- **Structural Errors**: We use the metric from Kanai et al. (1995) that measures insertion, deletion, and maximal move operations required across the output page to make it identical with the reference text. A weighted sum of these operations gives us the overall error, allowing us to quantify the structural quality of out outputs, and we normalize it to be between 0-100, with less being better.

Gold reference pages are created by inspecting the post-corrected output, comparing it with the source image, and manually correcting any errors. This is the most expensive part of the overall process and requires expertise in the language. So, for the moment, we evaluate on a few representative sample pages for two books. We showcase these results in Table 1, where we can observe a 50% decrease in character error rate and 87.5% reduction in structural error with our pipeline of language identification, masking, and automatic post-correction.

5 Conclusion

We apply the latest OCR techniques to a series of previously undigitized Kwak'wala texts, and demonstrate the challenges and unique adaptations necessary to make OCR work for real-world texts and collections. We propose using a mix of off-the-shelf OCR methods, language identification and

masking to effectively isolate Kwak'wala text, and post-correction models to produce a high-quality transcription. We plan to disseminate the digitized documents directly to the community members. Additionally, with consent of the community partners and data annotators, we plan to share the digitized and transliterated text (in three orthographies) with the data hosting institutions, such as the American Philosophy Society and Columbia University Rare Books and Special Collections, where a large collection of Boas-Hunt manuscripts have recently been digitized (Schlottmann, 2023). We hope to explore ways that this work in improving OCR for Kwak'wala and developing reliable digitization workflows for legacy texts can be transferable to other legacy orthographies, directly benefitting other language communities.

Limitations

Since our contribution type is best suited to a short paper, at the moment, we did not include more extensive benchmarking for language identification. As we continue to work with our language community collaborators, we will continue to add more gold reference texts for comparison and better evaluation of the transcriptions.

Ethics Statement

Though they derive from material in the public domain, the first-pass, gold reference texts, and corrected transcriptions of the selected Kwak'wala texts will only be released publicly with the consent of the language community members. An ethical implication of this work is that it will allow for more sustainable and equitable work in language resource creation and natural language processing, under the guidance of the language community members and their immediate and long-term needs for effective Kwak'wala revitalization.

Acknowledgments

This work was generously supported by the National Endowment for the Humanities under award PR-276810-21, George Mason University's Doctoral Research Scholars Award 2024-25 and the Stanford Initiative on Language Inclusion and Conservation in Old and New Media (SILICON) Practitioners 2024-25 Award. The authors are also grateful to the anonymous reviewers for their valuable suggestions, feedback, and comments.

References

Milind Agarwal, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2023. LIMIT: Language identification, misidentification, and translation using hierarchical models in 350+ languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14496–14519, Singapore. Association for Computational Linguistics.

Milind Agarwal and Antonios Anastasopoulos. 2024. A concise survey of OCR for low-resource languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 88–102, Mexico City, Mexico. Association for Computational Linguistics.

Taylor Berg-Kirkpatrick, Greg Durrett, and Dan Klein. 2013. Unsupervised transcription of historical documents. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 207–217. The Association for Computer Linguistics.

Taylor Berg-Kirkpatrick, Greg Durrett, and Dan Klein. 2013. Unsupervised transcription of historical documents. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 207–217, Sofia, Bulgaria. Association for Computational Linguistics.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents. *Preprint*, arXiv:2308.13418.

Franz Boas. 1909. *The Kwakiutl of Vancouver Island*. Leiden, New York: E.J. Brill; G.E. Stechert & Co.

Franz Boas and George Hunt. 1902a. *Volume 5, Part 1. Kwakiutl Texts - Memoirs of The American Museum of Natural History*. Leiden, New York: E.J. Brill; G.E. Stechert & Co.

Franz Boas and George Hunt. 1902b. *Volume 5, Part 2. Kwakiutl Texts - Memoirs of The American Museum of Natural History*. Leiden, New York: E.J. Brill; G.E. Stechert & Co.

Franz Boas and George Hunt. 1902c. *Volume 5, Part 3. Kwakiutl Texts - Memoirs of The American Museum of Natural History*. Leiden, New York: E.J. Brill; G.E. Stechert & Co.

Franz Boas and George Hunt. 1906. *Jesup North Pacific Expedition - Kwakiutl Texts, Second Series, Volume 10*. Leiden, New York: E.J. Brill; G.E. Stechert & Co.

Johanna Cordova and Damien Nouvel. 2021. Toward creation of Ancash lexical resources from OCR. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the*

Americas, pages 163–167, Online. Association for Computational Linguistics.

Rui Dong and David Smith. 2018. Multi-input attention for unsupervised OCR correction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2363–2372, Melbourne, Australia. Association for Computational Linguistics.

Yasuhisa Fujii, Karel Driesen, Jonathan Baccash, Ash Hurst, and Ashok C. Popat. 2017. Sequence-to-label script identification for multilingual OCR. In *14th IAPR International Conference on Document Analysis and Recognition, ICDAR 2017, Kyoto, Japan, November 9-15, 2017*, pages 161–168. IEEE.

Lenore A Grenoble and Lindsay J Whaley. 2005. *Saving languages: An introduction to language revitalization*. Cambridge University Press.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Nikolaus P Himmelmann. 1998. Documentary and descriptive linguistics.

Benjamin Hunt, Lane Schwartz, Sylvia Schreiner, and Emily Chen. 2023. Community consultation and the development of an online akuzipik-English dictionary. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–143, Toronto, Canada. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

J. Kanai, S.V. Rice, T.A. Nartker, and G. Nagy. 1995. Automated evaluation of ocr zoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(1):86–90.

Okan Kolak and Philip Resnik. 2005. OCR post-processing for low density languages. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 867–874, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Kimberley L. Lawson. 2004. *Precious fragments: First Nations materials in archives, libraries and museums*. Ph.D. thesis, University of British Columbia.

Jamshed Memon, Maira Sami, Rizwan Ahmed Khan, and Mueen Uddin. 2020. Handwritten optical character recognition (ocr): A comprehensive systematic literature review (slr). *IEEE Access*, 8:142642–142668.

Clemens Neudecker, Konstantin Baierer, Mike Gerber, Christian Clausner, Apostolos Antonacopoulos, and Stefan Pletschacher. 2021. A survey of ocr evaluation tools and metrics. In *Proceedings of the 6th International Workshop on Historical Document Imaging and Processing*, HIP '21, page 13–18, New York, NY, USA. Association for Computing Machinery.

Thi Tuyet Hai Nguyen, Adam Jatowt, Mickael Coustaty, and Antoine Doucet. 2021. Survey of post-ocr processing approaches. *ACM Comput. Surv.*, 54(6).

Aidan Pine, Patrick William Littell, Eric Joanis, David Huggins-Daines, Christopher Cox, Fineen Davis, Eddie Antonio Santos, Shankhalika Srikanth, Delasie Torkornoo, and Sabrina Yu. 2022. G_i2P_i rule-based, index-preserving grapheme-to-phoneme transformations. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 52–60, Dublin, Ireland. Association for Computational Linguistics.

Christian Reul, Uwe Springmann, and Frank Puppe. 2017. LAREX - A semi-automatic open-source tool for layout analysis and region extraction on early printed books. *CoRR*, abs/1701.07396.

Shruti Rijhwani, Antonios Anastasopoulos, and Graham Neubig. 2020. OCR Post Correction for Endangered Language Texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5931–5942, Online. Association for Computational Linguistics.

Shruti Rijhwani, Daisy Rosenblum, Antonios Anastasopoulos, and Graham Neubig. 2021. Lexically aware semi-supervised learning for OCR post-correction. *Transactions of the Association for Computational Linguistics*, 9:1285–1302.

Shruti Rijhwani, Daisy Rosenblum, Michayla King, Antonios Anastasopoulos, and Graham Neubig. 2023. User-centric evaluation of OCR systems for kwak'wala. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 19–29, Remote. Association for Computational Linguistics.

Kevin Schlottmann. 2023. Description and digitization of the george hunt kwak'wala ethnographic manuscripts. Accessed: 2025-01-10.

R. Smith. 2007. An overview of the tesseract OCR engine. In *9th International Conference on Document Analysis and Recognition (ICDAR 2007), 23-26 September, Curitiba, Paraná, Brazil*, pages 629–633. IEEE Computer Society.

AI for Interlinearization and POS Tagging: Teaching Linguists to Fish

Olga Kriukova[1] *, Katherine Schmirler[2] *, Sarah Moeller[3], Olga Lovick[1], Inge Genee[2], Alexandra Smith[2], Antti Arppe[4]

[1]University of Saskatchewan, [2]University of Lethbridge, [3]University of Florida, [4]University of Alberta
*These authors contributed equally
Correspondence: olga.kriukova@usask.ca

Abstract

This paper describes the process and learning outcomes of a three-day workshop on machine learning basics for documentary linguists. During this workshop, two groups of linguists working with two Indigenous languages of North America, Blackfoot and Dënë Sųłiné, became acquainted with machine learning principles, explored how machine learning can be used in data processing for under-resourced languages and then applied different machine learning methods for automatic morphological interlinearization and parts-of-speech tagging. As a result, participants discovered paths to greater collaboration between computer science and documentary linguistics and reflected on how linguists might be enabled to apply machine learning with less dependence on experts.

1 Introduction

During this time of increased AI-assisted language documentation, more and more studies emphasize the necessity and importance of collaborative efforts between documentary and computational linguists (Gessler, 2022; Flavelle and Lachler, 2023; Opitz et al., 2024). Additionally, Gessler and von der Wense (2024) point out the lack of interdisciplinary educational initiatives that could introduce specialists from both fields to the specific and general context of each other's work and, thus, bring mutual understanding and effective collaboration. In this paper, we describe our experiences hosting and participating in a "Machine-in-the-Loop" (MitL) workshop, held in Edmonton at the University of Alberta during November 14-16, 2023, which addresses this lack. The workshop curriculum Moeller and Arppe (2024) aims to introduce documentary linguists to machine learning (ML) and natural language processing (NLP) and to provide Python-savvy linguists with ML skills relevant to Indigenous language research and resource development. The workshop focused on foundational concepts underpinning machine learning and its application in NLP. In practical sessions, we worked in two teams focusing on two Indigenous languages of North America, Blackfoot and Dënë Sųłiné, each working toward a different project goal using a different machine learning model and NLP task. A Transformer deep learning model was trained to perform automatic interlinear morphological glossing of Blackfoot texts. A Conditional Random Fields (CRF) model was used to build a parts-of-speech (POS) tagger for Dënë Sųłiné.

The paper does not provide any ground-breaking solutions for computational linguistics but rather describes how already-established techniques can facilitate linguists' work with truly under-resourced languages. Notably, the workshop outcomes demonstrate that gaining awareness and a basic understanding of foundational ML concepts, combined with basic programming skills, enables linguists themselves to use NLP for the study and annotation of endangered languages.

This paper advocates for active collaboration between documentary and computational linguists in a way that enables documentary linguists to automate their own work efficiently, thereby reducing reliance on NLP experts to advance language technology for Indigenous communities. We feel that such a collaboration does not happen very often because linguists and computer scientists both assume it takes years of education before one can practically apply machine learning. This, combined with a below-average interdisciplinary dimension in NLP (Wahle et al., 2023), means many attempts at collaboration become inefficient interactions that seem more like data extraction to linguists (Flavelle and Lachler, 2023). This not only raises concerns about data security and sovereignty but also excludes the linguists' and language communities' perspectives from the NLP development. We believe that an approach where collaborators do not assume the technicalities are beyond linguists'

Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 139–149

grasp leads to the effective sharing of knowledge as well as results. For example, we found that, while NLP experts can automate their solutions, documentary linguists can immediately identify the problems in NLP model output, leading to increased problem-solving and benefits to both NLP and documentary goals.

Overall, by describing our workshop experience and our reflections on the interactions, we provide a positive example of collaboration between documentary and computational linguists, showing how much can be achieved in just three days by communicating needs, challenges, problems, and new terminology. We think that such collaborations can benefit both disciplines and support endangered language revitalization and documentation. We take inspiration from the proverb "Give a man a fish and you feed him for a day; teach a man to fish and you feed him for a lifetime" and use the metaphor of teaching a linguist to fish to illustrate the perspectives of both groups in section 2 followed by a description of the workshop in section 3 and the languages in section 4 followed by the outcomes from our three-day "teach a linguist to fish" approach to AI in sections 5 and 6.

2 Perspectives on Machine Learning for Documentary Linguistics

During the MitL workshop, we found ourselves falling into three main groups, each of which has different concerns regarding language data and ML, and thus takes a different approach to "fishing." The first group (co-authors Kriukova and Schmirler) consists of the documentary linguists with an interest in and familiarity with computational methods, who want to actively participate in creating, evaluating, and testing computational models (i.e., learning to fish, see section 2.1). We will refer to them as computationally-minded linguists or CMLinguist1 and 2. The second group (co-authors Genee, Lovick, Smith, to be referred to as DocLinguist1, 2, and 3) is comprised of documentary linguists with less interest in undertaking the computational data processing themselves, but who instead want to be familiar enough with the methods to communicate their needs and evaluate the outputs effectively (i.e., being on board, see section 2.2). The third group (co-authors Arppe and Moeller – to be referred to as CompLinguist1 and 2) contains the computational linguists who organized the workshop and who were primarily concerned with how participants might gain access to sufficiently powerful computing resources and make use of these resources (i.e., the tools used for fishing, see section 2.3). They chose the computational models and code used in the workshop to match the level of technical skills of the linguists in the second group, assuming they would have help after the workshop from those in the first group. Sections 2.1 to 2.3 are written by each group, respectively. Additionally, we want to emphasize that the teaching part of the metaphor is not intended in a strict sense, i.e., we did not expect to turn documentary linguists into computational linguists in a three-day workshop, but rather we aimed to bridge the knowledge gap sufficiently so that documentary linguists could initiate and foster effective collaboration with or without the direct guidance of NLP experts.

2.1 Teach a linguist to fish

Linguists who are primarily trained in language documentation and description, have basic programming skills and have an interest in computational methods are happy to be directly involved in the development of ML applications for endangered languages. We are also interested in working together with computational linguists. However, even when there is an interest, we find barriers to direct involvement or effective collaboration. When we look for help, we find guides for ML online that are either oversimplified or too focused on the mathematical foundations of the algorithms at hand (Vajjala, 2021). Meanwhile, we just want to know enough about ML methods to apply them to our data and to understand how the data and models (or at least the outputs) interact, allowing us to evaluate and improve the models ourselves. Moreover, most guides and books are focused on major languages and thus leave our questions about morphosyntactically different languages unanswered. They are insufficient for beginning work with endangered languages. For example, "Sentiment Analysis Using Python"[1] never mentions "English" but it becomes clear that the guide assumes the language already has a tokenization model, a list of stop words, and morphological model for lemmatization. Also, guides may assume we need the latest and most advanced language models. Sometimes, simple and time-tested methods are sufficient for work with limited data (see section 2.3) and their simplicity and reduced computing demands can

[1] https://www.analyticsvidhya.com/blog/2022/07/sentiment-analysis-using-python

significantly reduce our workload.

Another problem we encounter is that many ML tutorials use pre-built, standardized datasets. Little attention is paid to the description of how to create a dataset for a particular model from scratch (Vajjala, 2021). At the same time, questions such as what file format is needed, what pre-processing is required, or how the metadata file should be organized, are very important to us documentary linguists, who rarely possess "sterile" ready-made datasets.

As computationally-minded linguists who do language documentation work, we are also well-positioned to serve as translators between computational and documentary linguists, who have less or no interest in developing skills in the computational side of our work. This middle-ground understanding allows us to effectively communicate with both computational and documentary linguists about the modeling process and data annotation.

2.2 Get a linguist on board

Documentary linguists with less interest in undertaking computational data processing often take a somewhat ambiguous stance toward NLP. We are interested in utilizing customized NLP tools to manage our data and speed up our analytical work. When relying exclusively on manual annotation by trained individuals, this analytical work is time and labour-intensive, and as a result, can be very expensive as well. We also perceive serious interest in the communities we work with to benefit from the outputs of computational work, in particular in the areas of Automatic Speech Recognition, machine translation, talking dictionaries, and anything that will support the development of pedagogical materials. On the other hand, those communities are concerned about data sovereignty issues with respect to Indigenous language data (see for example, Rainie et al., 2019 or Junker, 2024). We also know from experience that computational linguists routinely underestimate just how "messy" language documentation is at all levels: from noisy multi-speaker audio recordings to code-switching and inconsistent or erroneous transcription and annotation. While we may not be best-suited to learning NLP techniques ourselves, our direct involvement and guiding role in NLP development for the languages we work with has clear benefits for the processes discussed below, particularly as it allows us to act as advocates for the communities who are most likely to suffer any negative impact.

2.3 A fishing rod vs. an industrial trawler

In the context of documentary linguistics, NLP researchers equipped with advanced AI techniques are like industrial fishing trawler operators, aiming to maximize scale, capabilities, and efficiency. We work in a field that often prioritizes publications of cutting-edge performance. However, we find that the needs of documentary linguists which could be served by NLP often require fundamental NLP tasks, such as POS tagging, or are best served by models which are not state-of-the-art. From our experience, linguists' most needed tasks often seem underwhelming. The computational problems involved in documentary work may be viewed in NLP as already solved even in low-resource settings, or the main workload may consist of basic data processing. Therefore, undertaking NLP work to benefit documentary linguistics and minority communities can leave one feeling that we are being asked to leave the trawler and sit on the shore with a bamboo rod.

Yet we found these seemingly mundane tasks are often out of reach for documentary linguists even if their training does include introductory programming skills. They may be unaware of the simplest NLP tools or common low-resource techniques. Designing the workshop we hypothesized that social scientists who discover the regular and irregular structures of language and can describe how they fit in a complex system of previously unstudied languages, all without being able to speak the language, are capable of grasping fundamental concepts of ML. We gambled that linguists could bridge the knowledge gap sufficiently in three days to empower them to design and direct their own collaborative computational projects, even if they could not code one line of Python. We feel the outcomes, whether in a POS tagger, F_1 scores, or the participant's intelligent use of new vocabulary, justified our assumptions. We emphasize that the next steps described in sections 5 and 6 were proposed, explained, and are being independently executed by the linguists themselves.

3 The Workshop

Just as hiring a fishing guide might be advantageous over buying one's own oceangoing trawler, collaboration with NLP experts can be highly beneficial for documentary linguists. However, the advantages of relying on NLP expertise for computationally intensive tasks must be weighed against long-term

dependence on domain experts who have different long-range goals. Also, if the short-term need of the documentary linguist or language community is critical enough to outweigh the downsides of "being given a fish," quick-fix NLP solutions are appropriate. The ideal situation, however, is that linguists themselves would be able to perform the required NLP work. The workshop description below illustrates how this long-term ideal situation can be created in practical terms.

3.1 Summary of the Machine-in-the-Loop Workshop

This workshop aimed to introduce linguists from non-technical backgrounds to the use of NLP for language-related tasks. In the mornings, lectures and interactive activities introduced the participants to the general principles of ML algorithms with clarification of specific relevant topics or terminology such as unsupervised vs. supervised learning and classical machine learning vs deep learning. Special attention was paid to those ML methods that work well in low-resource settings and to precision, recall, and F1 scores for evaluation. In the afternoons, two teams of linguists were able to apply what they learned by training a model on their own data and improving it during the workshop. Discussions and questions were encouraged, as well as sharing progress and roadblocks between the two teams.

3.2 ML for language documentation

While a fuller account of the curriculum of our workshop can be found in Moeller and Arppe (2024), here we briefly summarize our understanding and use of ML in the workshop. We define ML as a type of AI, wherein a computer makes use of an algorithm and statistical model to do something "intelligent". Data are mapped as points in space, and ML creates a statistical model based on that data, which can then be used for prediction. Predictions are made by learning patterns from data. This pattern recognition somewhat mimics how humans learn, and thus the computer can help improve on a manual task.

ML is already used for some linguistic or language-related tasks, such as clustering for dialectology or n-grams for predictive text, but can be also useful for documentary linguistics, particularly in the data annotation bottleneck. Taking audio or transcribed data from its raw form to a fully interlinearized corpus is a time-consuming process. Since linguists already create some transcribed and annotated data as part of their basic analysis, ML offers the opportunity to use those annotations as training data for predictive models to speed annotation for the remaining data.

In the workshop, we take a machine-in-the-loop approach (active learning) that allows human linguistic expertise to annotate new data selected based on the marginal probabilities of a CRF, for example. This approach assists simultaneously in completing the annotation process and more quickly improving the model's output. The workflow involved gathering and preparing our data for training (ideally ahead of time for preprocessing), choosing a model, and then training, testing, and evaluating the model. The linguists evaluated the model, decided what changes to the data were needed, and updated the training dataset.

4 Languages and Data

4.1 Blackfoot

Niitsi'powahsin or Siksikai'powahsin, usually called Blackfoot (ISO: bla) in English, is an Algonquian language spoken in Alberta and Montana by perhaps less than 5,000 people out of a total population of around 40,000 (Genee and Junker, 2018, 301–302). The data used in the workshop is a collection of stories containing ∼1,000 words, drawn from several sources (Russell and Genee, 2014; Ermineskin and Howe, 2005; Genee, 2009; Frantz, 2017; Glenbow Museum, n.d.; Many Feathers et al., 2013), interlinearized as in (1).[2]

(1) Ninna iikaahsitapiiwa.

n-inn-wa
1-father-AN.SG
iik-yaahs-itapii-wa
very-kind-be_person.VAI-3SG

'My father was a very kind person.' (Russell and Genee, 2014, 12)

Blackfoot is a polysynthetic language with many possible morphemes per word. While generally concatenative, these morphemes also display considerable allomorphy and surface variation due to morphophonological processes, which, as we will discuss in our outcomes, can cause issues with data

[2]Most sources provided the analyses for the interlinear glossing. For Russell and Genee (2014), the analyses were provided by DocLinguist1 and for Glenbow Museum (n.d.), the analyses were provided by Heather Bliss, November 2010.

preprocessing for machine learning and no doubt offers an extra challenge for the model itself by introducing considerable variation and ambiguity.

The team working with Blackfoot data consisted of three members. DocLinguist1 is working on Blackfoot language documentation and revitalization and has more than a decade of experience with this language. CMLinguist2 is a postdoctoral scholar who specializes in Algonquian linguistics with a focus on morphosyntactic and phonological modeling and currently works with DocLinguist1. DocLinguist3 is an MA student of DocLinguist1 and a member of the Piikani Nation, who also works in the field of documentation and revitalization of Blackfoot, with a focus on corpus creation and textual annotation.

4.2 Dënë Sųłné

Dënë Sųłné (ISO: chp; hence: DS) is a Dene (Athabaskan) language spoken in Manitoba, Saskatchewan, Alberta, and the Northwest Territories (Cook, 2004). The 2021 census indicates that there are around 10,000 speakers of DS (Statistics Canada, 2022), making it one of Canada's most vital Indigenous languages. More than half of these speakers reside in Northern Saskatchewan. The data used in the workshop is a sub-corpus of the audiovisual corpus compiled during the *Talking Dene* project [3], which collected 70 hours of naturalistic DS representing 100 speakers ranging in age from 13 to 83 years of age. Most of this corpus has been transcribed and translated (at the utterance- and word-level) by speakers fluent in DS and English as shown in (2). The dataset is not made available here due to community preferences.

(2) grade two *dé́ dlą́t'ı së*teacher *nı sí́ bënasní=lë hotıé́ dódí*

grade	*two*	*dé́*	*dlą́t'ı*
grade	two	when	who
së-teacher		*nı*	*sí́*
1SG.PSR-teacher		PST1	EMPH

bë-n-a-s-ní=lë
3SG.P-LX-LX-IPFV:1SG.S:VV-remember=NEG
hotıé́ dódí
very NEGEX

'In grade two, I don't remember who my teacher was at all.' ITN-ETM-2022-11-28-AB

[3] A University of Saskatchewan research project (PI - Olga Lovick) in partnership with the Clearwater River Dene School and the University of Zurich.

From a typological perspective, DS can be described as highly synthetic and fusional, particularly in the verbal domain. The language is overwhelmingly prefixing, with lexical/derivational and inflectional morphemes interspersed within the verb word. It is head-marking and has SOV word order although the fact that arguments are marked on the verb means that full noun phrases are used more sparingly than in languages such as English.

Example (2) illustrates one of the more challenging aspects of our DS corpus: the extensive use of English. Almost all speakers of DS nowadays are bilingual, and switches ranging from one word to multiple sentences, as well as English stems with DS affixes, are extremely common.

The DS team consisted of three individuals. DocLinguist2 is a specialist in Dene/Athabaskan linguistics with over two decades of experience in the description and documentation of this language family. CMLinguist2 is DocLinguist2's and CompLingust1's Ph.D. student. Her research area is in harnessing computational tools for educational and documentary purposes in low-resource language settings, in particular for DS. In addition, we had one graduate student observing the workshop and the work of the DS team, though their own research concerns neither DS nor Blackfoot.

5 Modeling outcomes

5.1 Blackfoot

5.1.1 Process

Our Blackfoot team used the Transformer deep learning model (Vaswani et al., 2017) for the automatic interlinearization of Blackfoot text, specifically morphological segmentation and morpheme glossing. Approximately 10% of the manually annotated data was set aside for testing, with the remaining 90% used for training.

5.1.2 Outcome

For the first training iteration, our Blackfoot team found that a small number of closed-class morphemes achieved promising precision and recall scores. These included the demonstrative stem *ann-* 'that' (0.75 precision and recall, n = 4 in the test dataset) and the demonstrative suffix *-hka* 'invisible' (0.67 precision and recall, n = 6). However, we quickly learned that the glossing in our training data was less consistent than we had thought. Our team thus needed to identify and correct inconsistencies in the glossing. The same morphemes

may have been given slightly different English glosses (e.g., *aakii* 'woman' or 'lady'; *sook-* 'suddenly' or 'unexpected'), allomorphs were separated (e.g., *-nnaan*, *-innaan*, *-(i)nnaan* '1PL', *n-*, *ni-*, *nit-*, *ni(t)-* 'first person'), or different abbreviations were used (e.g., NONSP, NSPEC for 'nonspecific'). Especially with such small datasets, consistency in morphemes and glosses can drastically increase the number of different training examples of each feature.

In the second iteration, many of the morphemes our team corrected showed improvements, such as *aakii* 'woman' (0.67 for both precision and recall, n = 3), demonstrative stems *am-* (0.75 for precision and 1.00 for recall, n = 3), *amo-* (1.00 for precision and 0.67 for recall, n = 3), and *ann-* (precision increased from 0.75 to 1.00, recall unchanged at 0.75, n = 4). The suffix *-hka* also improved (from 0.67 for precision and recall to 0.71 and 0.83 respectively, n = 6), as did the first person prefix *nit-*, now combined with its allomorphs (0.90 for precision and 0.82 for recall, n = 11).

More glossing issues were found and adjusted before the third iteration, but fixing inconsistencies no longer seemed to affect the training, so our team expanded the training data by adding five new sentences, which is a 3.2% increase in training tokens. For our fourth and final iteration of the workshop, some improvements were seen, such as for *ann-* (0.80 precision, 1.00 recall, n = 4) and *nit-* (1.00 precision, 0.50 recall, n = 12). Other affixes also showed promise, such as the third person prefix *ot-* (0.67 precision, 0.50 recall, n = 4), the animate singular suffix *-wa* (0.67 precision, 0.15 recall, n = 13), and the singular suffix for inanimate nouns *-yi* (0.44 for both precision and recall, n = 9).

Overall, our team found that consistent and frequent inflectional morphemes and frequent stems without considerable allomorphy were recognized well. Some demonstrative stems, noun and verb stems, person morphology, and the particle *ki* demonstrated decent precision and recall scores, some were correctly recognized correctly from the beginning and others after glossing was made more consistent. However, much of the inflectional morphology and most stems were still unrecognized or very poorly identified by the model, and much remains to be done before a useful morphological model is available.

5.1.3 Next Steps

In the future, the Blackfoot team plans to tackle two main issues. First, we intend to develop two sets of strict glossing standards across the existing analyzed texts, with a clearly defined correspondence between them. One set will be for linguistic analysis, with each morpheme represented separately and homophonic morphemes are marked differently for nouns and verbs. The other set will be geared toward machine learning, where frequently occurring strings of morphemes can be chunked together and morphemes are marked the same regardless of the word class they attach to. For example, a morpheme string like *-aanaana* can be broken down into *-a* 'direct', *-innaan* '1PL', and *-wa* '3SG', but for the sake of statistical modeling, it may be worthwhile to consider this frequent sequence of morphemes as one unit glossed '1PL>3SG'. For some homophonic person and number suffixes, such as *-wa*, the linguistic analysis will give different glosses depending on whether it attaches to a verb (3SG) or a noun/pronoun (AN.SG). However, for machine learning, one label (e.g. '3SG') may be more effective, especially for a relatively small dataset.

Additionally, our team has an option for generating Blackfoot words with an FST-based morphological model (Kadlec, 2023). With this option, we can generate potentially (hundreds of) thousands of Blackfoot words for inclusion in the training data, increasing the data exponentially. As the token counts in the previous subsection indicate, an increase in data is much needed. In doing this, we intend to explore to what extent synthetically generated paradigms improve this process, e.g., when do we see diminishing returns with increased data.

5.2 Dënë Sųłné

5.2.1 Process

Our DS team undertook the task of parts-of-speech (POS) tagging for DS using the Conditional Random Fields (CRF) model (Lafferty et al., 2001), chosen because of its ability to learn from very small datasets and to demonstrate the role of features of the data for training. The training data consisted of two files comprising 582 DS utterances and 2961 DS words. DocLinguist2 created a controlled parts-of-speech vocabulary in ELAN (ELAN (Version 6.7) [Computer software], 2023) informed by her grammatical research on Dene/Athabaskan languages (Lovick, 2020) and

tailored to DS. ELAN files were manually annotated for POS by an undergraduate student, hand-corrected by DocLinguist2, and exported as FlexText to facilitate further data extraction by CompLinguist2.

The list of POS tags used by our team comprised 18 items. In a sample of 2,817 words taken from dialogue and monologue, nouns and verbs were the most frequent with 521 and 520 tokens, respectively. Particles, adverbs, postpositions and conjunctions were the next most frequent categories with more than 250 tokens each.

5.2.2 Outcome

For our team's first iteration, we achieved an accuracy score of 0.71. Similarly to the Blackfoot team, inconsistent training data annotation was a major source of our model's poor performance in the beginning. This inconsistency was partly due to grammatical differences between DS and English (the language spoken by the undergraduate annotator). Property concept words, for example, are typically adjectives in English but verbs in DS. Other inconsistencies resulted from the fact that some lexical items are polyfunctional and therefore often annotated for the wrong function in context; e.g. *dë́* can function as a postposition 'when, at the time of' (cf. (2) above) or as a clause conjunction 'if' (see also Cook, 2004, 375–380). To simplify the modeling task, we decided to reduce the number of tags, which led to a slight accuracy improvement to 0.73 over several iterations.

For our fourth iteration, we modified the CRF model features. Initially, we used default word feature extraction parameters designed to capture English POS-specific prefixes (e.g., *re-*, *un-*, *mis-*) and suffixes (e.g., *-ed*, *-s/-es*, *-er*. To address the radically different verb morphology of DS, where a verb stem may be preceded by multiple prefixes, we experimented with different numbers of word-initial and -final characters. The settings that gave us the best results captured up to six word-initial and word-final characters. This change in the word-to-features Python function improved the overall accuracy and verb and noun recall value (from 0.84 to 0.87 and from 0.73 to 0.80 respectively).

After the fourth iteration, we found that further feature engineering led to an improvement in the recall for certain parts of speech, at the cost of recall for others. For instance, the recall of verbs improves from 0.87 to 0.91 when we include up to 5 final characters of a word. However, these settings lower the recall of nouns to 0.73, that of postpositions from 0.73 to 0.65, and that of conjunctions from 0.75 to 0.50. This experimentation taught us that we can adjust the model feature parameters to refine the results in specific areas.

Careful examination of the predicted POS for our team's best iteration revealed that a major source of errors was the presence of English and mixed-language lexical items (such as *sëteacher* 'my teacher' in (2)) present within the DS discourse. Tailoring our feature parameters in the CRF to capture DS morphological features caused the model to perform poorly when faced with English or mixed-language words. Consequently, the overall POS tagging performance for DS words is, in fact, higher than the numbers above suggest.

5.2.3 Next Steps

Given the persistence of code-switching and code-mixing in the DS corpus, it appears that the easiest way to improve the accuracy of POS tagging is to add an intermediate step of language identification. The language recognizer could employ a CRF or another non-neural classifier model such as a Support Vector Machine (SVM) to classify each word as DS, English, or Mixed.

POS tagging will then proceed differently depending on the language of each lexical item. English words will be tagged by a pre-trained tool such as spaCy (Honnibal and Montani, 2017). DS and Mixed items will be tagged by our CRF-based tagger. This will also allow our team to evaluate the 'real' accuracy of this tagger. Additionally, in order to facilitate further linguistic data analysis and to improve word search in ELAN, we need to develop a workflow to import the predicted POS tags back into the ELAN with the *lxml* Python package.

6 Learning outcomes

In this section, we move onto the outcomes that we deem of even more interest than the modeling outcomes—the knowledge, understanding, and skills we gained over the course of the workshop and the methods we learned. We call back to section 2 and reflect on these outcomes by each subgroup at the workshop.

6.1 For computationally-minded linguists

It is very significant to us computationally-minded linguists that we not only made a functioning model but also learned how to adapt it to different needs. The CRF model we developed is not perfect and

probably not optimal, but now we have the knowledge sufficient to maintain, modify, and improve it. Moreover, we have a better understanding of how to use our expertise in the languages at hand for feature engineering. As a result, after the workshop, the DS team trained several CRF-based models for different annotation needs.

Working with a Transformer model for interlinearization gave us a better idea of how the annotation may need to differ between computational and documentary linguistics. Though the computational FST modeling of Blackfoot had already demonstrated this to some degree, the chunking and standardization of morphemes and tags became even more apparent when training a Transformer model on a small data set, and will inform both documentation and computational modeling of Blackfoot in the future.

Finally, we realized that all we needed to launch our independent work with ML-based tools was guidance appropriate to our skill level and field of application and a gentle push in the right direction to use our data for our goals.

6.2 For documentary linguists

From the perspective of documentary linguists without programming skills, an important advantage of the workshop approach is the establishment of trust relationships. By forming small teams including both documentary and computational linguists, we were able to ensure that the data did not leave the servers approved by our community partners and University Research Ethics Boards, which protects the data from unauthorized use. We could also see and control what happened with the data because we were in the same space.[4] We think this should be expanded in future workshops and collaboration to include an even more important relationship: that between language communities and the academic community. Including representation from the language communities will foster transparency and create confidence in the process.

A second advantage of the workshop approach lies in the ability to jointly and immediately look at model output and identify problems. The computational linguist may look at numerical indicators of model results, but only someone intimately familiar with the language under analysis can determine that a particular set of errors is perhaps due to inconsistent glossing within the training data. What's more, we can immediately correct some of the errors or suggest improvements to the model based on our understanding of language's fundamental principles. This effect is maximized by goal-setting and preparation in advance of the workshop (i.e., preparation of training data by linguists).

The ultimate strength of the workshop format is that it allows all participants to bring their unique expertise to the table. Rather than force a computational linguist to clean a dataset, or a language documentation specialist to use the command line, we all perform those tasks that we are best suited to. Documentary linguists do not necessarily want to learn to fish ourselves—we want to see that the boat is going in the right direction and to help you know what fish are worth fishing for.

We may lack the time or inclination to learn how to apply NLP ourselves. However, a basic introduction to NLP concepts enables us to communicate our needs effectively and evaluate results when collaborating with NLP experts. Continuing the metaphor, we now have the general knowledge so we can navigate the "fish market" and choose the best species—one that delivers high-quality nutrients and is ethically sourced, i.e. procured in a fashion that maintains the viability of the resource (language), rather than dynamiting the fishing grounds for spectacular but one-time hauls.

6.3 For computational linguists

First, we discovered a spectrum of skills among documentary linguists that supported their quick grasp of ML principles and ability to work with NLP models. For example, a prominent skill among descriptive linguists is complex pattern recognition, which is also a cornerstone of ML. The field of linguistics has traditionally placed less emphasis on statistical patterns of language usage and instead focuses on generalizing from specific patterns in order to describe a language's structure and from there building abstract theoretical models. Nonetheless, we find linguists readily embrace statistical methods and bring their expertise in pattern recognition and data analysis to bear once they see the value of a machine-in-the-loop approach for their goals.

Second, the pressure of academic publishing, or commercial interests for those in industry, may lead to the prioritization of novelty. At the same time what is novel for NLP may not be valued by another field. This disconnect in perceived common

[4] We are aware that the learning and outcomes would not require us to be physically in the same room, but personal interaction certainly helps in creating trust.

goals among academics may lead to miscommunication between computational linguists and documentary linguists who do not care about the novelty of models as long as they are relatively simple, work reliably, and reduce the annotation workload necessary to discover novel linguistic phenomena. We prioritized documentary linguists' immediate needs; even though models like CRF are not state-of-the-art, they were ideal for connecting principles of linguistics to ML concepts and better-equipped documentary linguists to continue using what they learned after the workshop.

The third lesson for computational linguists was not new to us, but bears repeating. An ethically operating NLP project using minority language data should entail a willingness to engage in long-term collaboration. Crucially, long-term collaboration allows one to assess the benefit not to only NLP research or documentary efforts but also to the language communities whose data we are using. How to elicit language data while giving value to the community has been discussed in linguistics literature for the past 50 years and more (D'Arcy and Bender, 2023). Collaboration with experienced documentary linguists is one way to discover how "fishing" for data might become a fair market.

6.4 For all participants

Finally, our key observation at the workshop was that genuinely listening to the divergent concerns of the computational and documentary linguists and adjusting one's approaches to accommodate each others' goals and felt needs was able to overcome prejudices based on prior less-than-optimal interactions. The paramount concern for the linguists was that the language data—collected together with the language communities—would not just disappear somewhere, to reappear as part of someone's research with no connection to or benefit for the language communities in question or in an application the communities would be expected to pay for. For the computational linguists, the primary concern was not to gain access to the data as such, but rather whether the documentary linguists would have sufficient resources to run the ML algorithms (e.g. access to GPUs), wherever the documentary linguists wished to keep the data, without needing the NLP experts' support and time to rerun and adjust the code. In this end, one positive experience where concerns were voiced and understood changed what both groups feel is possible for AI in documentary linguistics.

7 Discussion & Conclusion

The workshop proved to be successful at equipping linguists to do their own AI "fishing" for several reasons. First, the interaction allowed the linguists to close the existing gaps in their knowledge of ML and its application in endangered languages. Second, the workshop format allowed the linguists to put this knowledge into practice right away. They worked on solving real research problems both teams faced—the need for more morphological interlinearization and quick POS tagging. After the three-day workshop, participants had a trained model in their hands. Third, both teams had constant support from computational linguists, who helped to fix errors in data and gave valuable suggestions on model or workflow optimization.

The composition of the two research teams played a large role in the success of the model development. Each team had at least one linguist trained in Algonquian or Dene linguistics, and at least one with basic programming skills. This allowed both teams to 1) quickly identify and correct mistakes in the training data and the model; 2) devise and implement solutions tailored to each language; and 3) expand and incorporate new training data by correcting the models' predictions. Having NLP experts in the room reduced the time needed for troubleshooting and fostered confidence.

Although this workshop's main goal was to educate linguists, it was also an exciting and educative experience for the NLP experts. So many NLP tasks that documentary linguists face are as simple as "shooting fish in a barrel." Hence, in three days the NLP experts saw maximum positive impact. As a result, the impact of our AI workshop has gone beyond a three-day event, leading to further collaborations and grant applications.

By describing the results of our workshop, we want to emphasize that progress in NLP does not always depend on inventing new methods; rather, it often lies in the meaningful application of established methods to different languages. After all, each language brings new and often unique challenges to old tools. We hope that this workshop's outcomes will set a positive trend of impactful collaborations between documentary and computational linguists and lead to better communication between these two fields. The models might seem complicated and intimidating at first, but all participants discovered that linguists do not need a degree in computer science to use AI.

Acknowledgments

We are grateful to the Blackfoot and Dënë Sųłiné communities for the opportunity to work with their languages. In particular, we want to thank Niitsitapii communities in Alberta, Canada and the Clearwater River Dene Nation in Saskatchewan, Canada. This workshop was supported by SSHRC Partnership Grant 895-2019-1012 "21st Century Tools for Indigenous Languages". Blackfoot language documentation work was supported by SSHRC IG 435-2021-0562 and SSHRC PDF 756-2022-0428. Dënë Sųłiné language documentation work was supported by SSHRC IG 435-2020-1197.

References

Eung-Do Cook. 2004. *A Grammar of Dëne Sųłiné (Chipewyan). Algonquian and Iroquoian Linguistics – Special Athabaskan Number, Memoir 17*. Algonquian and Iroquoian Linguistics, Winnipeg.

Alexandra D'Arcy and Emily M. Bender. 2023. Ethics in Linguistics. *Annual Review of Linguistics*, 9(Volume 9, 2023):49–69. Publisher: Annual Reviews.

ELAN (Version 6.7) [Computer software]. 2023. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. [link].

Rachel Ermineskin and Darin Howe. 2005. On Blackfoot syllabics and the law of finals. Paper presented at the 37th Algonquian Conference.

Darren Flavelle and Jordan Lachler. 2023. Strengthening relationships between indigenous communities, documentary linguists, and computational linguists in the era of NLP-assisted language revitalization. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 25–34, Dubrovnik, Croatia. Association for Computational Linguistics.

Donald G. Frantz. 2017. *Blackfoot grammar*. University of Toronto press.

Inge Genee. 2009. What's in a morpheme? Obviation morphology in Blackfoot. *Linguistics*, 47(4):913–944.

Inge Genee and Marie-Odile Junker. 2018. The Blackfoot Language Resources and Digital Dictionary project: Creating integrated web resources for language documentation and revitalization. *Language Documentation & Conservation*, 12:274–314. Publisher: University of Hawaii Press.

Luke Gessler. 2022. Closing the NLP gap: Documentary linguistics and NLP need a shared software infrastructure. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 119–126, Dublin, Ireland. Association for Computational Linguistics.

Luke Gessler and Katharina von der Wense. 2024. NLP for language documentation: Two reasons for the gap between theory and practice. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 1–6, Mexico City, Mexico. Association for Computational Linguistics.

Glenbow Museum. n.d. Kaitapiitsinikssiistsi / Traditional Stories.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Marie-Odile Junker. 2024. Data-mining and extraction: the gold rush of AI on indigenous languages. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 52–57, St. Julians, Malta. Association for Computational Linguistics.

Dominik Kadlec. 2023. A computational model of blackfoot noun and verb morphology. Master's thesis, Lethbridge, AB, Canada.

John Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289. Cited in Synthesis and paper for CRF.

Olga Lovick. 2020. *A Grammar of Upper Tanana, Volume 1: Phonology, Lexical Classes, Morphology*. University of Nebraska Press, Lincoln.

Sandra Áístainskiaakii Many Feathers, Brent Issapóíkoan Prairie Chicken, Wes Áínnootaa Crazy Bull, and David Osgarby. 2013. Aakíípisskani / the women's buffalo jump. In *Papers of the 48th International Conference on Salish and Neighbouring Languages*, volume UBCWPL35, pages 1–21. University of British Columbia.

Sarah Moeller and Antti Arppe. 2024. Machine-in-the-loop with documentary and descriptive linguists. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 27–32, St. Julians, Malta. Association for Computational Linguistics.

Juri Opitz, Shira Wein, and Nathan Schneider. 2024. Natural language processing relies on linguistics.

Stephanie Carrol Rainie, Tahu Kukutai, Maggie Walter, Oscar Luis Figueroa-Rodriguez, Jennifer Walker, and Per Axelsson. 2019. Issues in open data - Indigenous data sovereignty. In T. Davies, S. Walker, M. Rubinstein, and F. Perini, editors, *The State of Open Data: Histories and Horizons*. African Minds and International Development Research Centre, Cape Town and Ottawa.

Lena Heavy Shields Russell and Inge Genee. 2014. *Ákaitsinikssiistsi: Blackfoot Stories of Old (First Nations Language Readers Blackfoot)*. University of Regina Press.

Statistics Canada. 2022. Mother tongue by geography, 2021 [data visualization tool].

Sowmya Vajjala. 2021. Teaching NLP outside linguistics and computer science classrooms: Some challenges and some opportunities. In *Proceedings of the Fifth Workshop on Teaching NLP*, pages 149–159, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jan Philip Wahle, Terry Ruas, Mohamed Abdalla, Bela Gipp, and Saif Mohammad. 2023. We are Who We Cite: Bridges of Influence Between Natural Language Processing and Other Academic Fields. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12896–12913, Singapore. Association for Computational Linguistics.

Universal Dependencies for the Amahuaca language

Candy Angulo **Pilar Valenzuela** **Roberto Zariquiey**

University at Buffalo Chapman University Pontificia Universidad del Perú

candyang@buffalo.edu valenzuela@chapman.edu rzariquiey@pucp.edu.pe

Abstract

This paper presents the creation of a Universal Dependency (UD) treebank for Amahuaca (Peru), marking the first UD treebank within the Headwaters subbranch of the Panoan family. While the UD guidelines provided a general framework for our annotations, language-specific decisions were necessary due to the rich morphology of the Amahuaca language. The paper also describes specific constructions to initiate a discussion on several general UD annotation guidelines, particularly those concerning clitics and morpheme-level dependencies.

1 Introduction

This paper describes the methodology employed in the creation of the UD treebank for the language. On the one hand, this treebank aims to enhance the future development of an NLP toolkit for this language as well as contribute to its revitalization. On the other hand, this work aims also to contribute to the discussion on how to integrate polysynthetic languages into the lexically oriented framework of Universal Dependencies (UD). Following Park et al. (2021), we argue that adopting a morpheme-level framework is indispensable due to the morphosyntax of Amahuaca. Specifically, it is crucial to accurately capture the intricate morphological relationships and dependencies within the language, particularly considering the unique characteristics of clitic behavior and their interaction with other morphemes. By focusing on morpheme-level annotations, we aim to provide a clearer understanding of the syntactic structure and the grammatical functions of various elements. This approach facilitates a deeper exploration of the language's complexity, ultimately contributing to more effective natural language processing applications and linguistic analysis.

The structure of the paper is as follows: Section 2 provides a brief overview of some notable features of the Amahuaca language. Section 3 explains the reasons behind our choice of morpheme-level analysis and presents the dependency relations found. Section 4 details the data collection process as well as the composition of the corpus. The following sections present the POS tags and the dependency relations. Section 7 focuses on the comparison between the morpheme-level annotation scheme and the word-level annotation scheme.

2 The Amahuaca language

The Amahuaca people are primarily concentrated in some provinces of the Ucayali region, in Peru. In the Atalaya province, they reside in the basins of the Yurúa River (Yurúa district), Inuya and Mapuya Rivers (Raymondi district), and Sepahua River (Sepahua district). In the Purús province, they occupy a community in the Purús River basin, within the district of the same name. Some settlements in the Upper Inuya and Mapuya regions host Amahuaca populations in "initial contact situations." For more information on Amahuaca society and culture, see Dole (1998) and Hewlett (2014). As mentioned before, this language is endangered, with approximately 400 speakers, most of whom are over 40 years old, and children are no longer learning it.

Amahuaca is a language, characterized by rich morphology. While there are works that describe this language (see Sparing-Chávez 2012, Clem 2019), we base the analysis on Valenzuela et al. (in prep.), which focuses more on the behavior of clitics in the language. Similar to other Panoan languages (for more information about Shipibo-Konibo and Kakataibo languages, see Valenzuela 2003, Zariquiey 2018), this language is characterized by being postpositional and predominantly agglutinative. A notable feature of the language is the absence of deverbal derivation and the use of auxiliaries to convert a noun into a verb; consequently, some nouns may carry verbal inflection markers. We will discuss this point in more detail later.

The language primarily follows a basic constituent order of SOV, but this order is flexible. Constructions like (1) can be found, where the subject *michito chaho* ‘black cat’

Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 150–154

precedes the object ‘Paco’, and the verb is at the end carrying the inflectional clitic.

1. Mishito chahonmun Paco ratuuxonu.

mishito chaho=n=mun Paco ratuu=xo=nu

cat black=A=FOC Paco scare=PFV.3=DECL

‘The black cat scared Paco.’

However, sentences with final subjects are found, as shown in (2). The subject *vaku maxko* ‘little baby’ appears at the end, and the verb *oyo* ‘suck’ precedes it. What is interesting about this free word order behavior is that the inflectional morphology is not always attached to the verbal root. Additionally, when S or A is not in the unmarked position, it loses its case marking and takes the form of the copy pronoun. This language is characterized by the presence of doubling pronouns in constructions with transitive verbs.

2. Jaton jaha chochomun oyoni vaku maxkokinu.

jaton jaha=n chocho=mun oyo=niko vaku maxko=ki=nu

3SG.POSS mother=GEN breast=FOC suck=ENDEAR baby=IPFV.2/3=DECL

‘The babies are sucking their mothers' breasts.’

Comparing (1) and (2), it can be observed that the clitic *=ki*, which encodes an aspectual meaning, in the first sentence is attached to the verbal root *ratuu* ‘to scare’, but in the second sentence, it is attached to the noun phrase *vaku maxko* ‘baby’.

3 Morpheme-level annotation scheme

Universal Dependencies (UD) traditionally employs a word-level annotation scheme (Nivre et al. 2017, 2020), which works well for many languages with relatively straightforward morphological structures. Shipibo-Konibo (2018) and Kakataibo, other Panoan languages, have UD treebanks. Consequently, we have based our guidelines for Amahuaca on these resources. However, Amahuaca's rich morphological system and the significant role of clitics require a different approach. After reviewing studies on handling phenomena in polysynthetic languages, such as noun incorporation (Tyers & Mishchenkova, 2020), as well as more general works like Park et al. (2021) and Çöltekin (2016), we decided to follow the direction of morpheme-level annotations proposed in the second paper, as will be explained later.

Table 1. Amahuaca bound morphemes behavior.

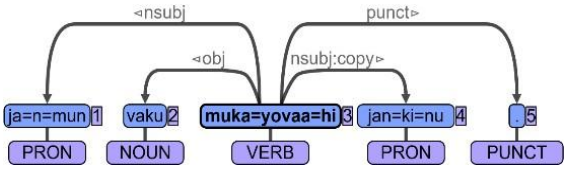
Morpheme Type	Within a Constituent	At the edge of phrases	Fixed Position	Selective for Host	Without Host
Case markers	NO	YES	YES	YES	NO
=mun	possible	usually	usually	NO	NO
Perfective aspect	YES	NO	YES	NO	NO
Degree of remoteness	usually	Possible	NO	NO	NO
Person markers	YES	NO	YES	NO	NO
Declarative markers	NO	YES	usually	NO	NO
=kiha	possible	usually	usually	NO	possible
Switch-reference markers	NO	YES	usually	usually	possible

Unlike Shipibo-Konibo, Amahuaca morphemes sometimes do not require an open-class word as a host for their pronunciation. Additionally, clitics can attach to various parts of speech and carry important grammatical information such as tense, aspect, mood, and case. These clitics often do not function as standalone words but as bound morphemes that modify the meaning and function of their host words. Table 1 summarizes the behavior of such bound morphemes.

Firstly, case markers are selective for a host and are attached to them. However, the topic marker *=mun* can appear without a host, but it must follow another clitic; if it appears alone or in the first position, it is not allowed. This restriction applies to aspect, tense, and mood markers. But, switch-reference markers, the hearsay marker *=kiha*, as well as degree of remoteness markers, can appear without an open-class word as a host. “Within a constituent” refers to being inside a phrase, which could be a noun or verb phrase. “At the edge of phrases” means at the end of a syntactic constituent. “Fixed position” indicates if the clitic always occupies the same position in relation to the host. For example, case markers always come immediately after the nucleus of the constituent they modify (whether it is just a noun or a noun phrase). “Selective for a host” indicates if it can serve as the base morpheme where a clitic can be attached. Finally, "without host" means that it cannot appear without a host. From our

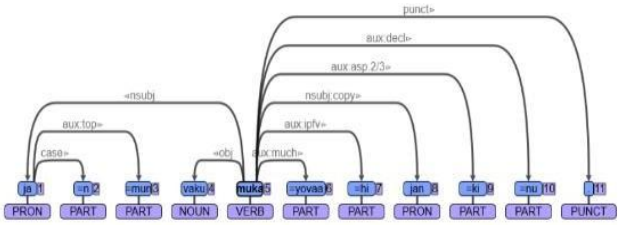
perspective, a word-level annotation would fail to capture the dependency relationships between these clitics and their hosts accurately. Compare the representations of the sentence *Janmun vaku mukayovaahi janhkinu*. 'He is laughing a lot at his baby'. (3) corresponds to the lexicalist representation, namely Word-level. As observed, only 4 dependency relations are shown. While it captures the grammatical relations of subject and object as a transitive sentence, it overlooks the fact that inflection does not occur entirely on the verb – only the aspectual marker *=hi-* is shown, but it fails to indicate its complement *=ki*, which attaches to the doubling subject *jan.*

3. Word-level representation



(4) shows the dependency relation at a morpheme-level, a total of 11. This analysis adequately captures the fact that the aspectual marker *=ki* attaches to the doubling subject. Even though *=ki* corresponds to a grammatical person, it works together with the aspectual marker *=hi*, because the latter clitic requires to be with a person marker within the same clause. In other words, if there is no *=ki*, the sentence would be agrammatical.

4. Morpheme-level representation



In this section, we presented examples that demonstrate the necessity of morpheme-level annotation. We show how clitics interact with other morphemes and how their roles are more clearly defined in a morpheme-level framework. This approach not only provides a more accurate representation of Amahuaca syntax but also helps in understanding the language's morphological richness. In Section 6, we will explore in greater depth the clitics and their corresponding dependency relations that we have assigned to them.

4 Corpus

The annotated corpus for Amahuaca consists of sentences that were translated from Spanish into Amahuaca. This work is part of a broader initiative to compare treebanks of various Peruvian Amazonian languages, including Amahuaca. Each language was assigned a set of 60 sentences, resulting in a total of 420 sentences to translate. Three native speakers of Amahuaca participated by translating all 420 sentences into their language, after which each translation was reviewed with them. Of these, 202 sentences have been manually annotated for Amahuaca, while the remaining sentences are still awaiting verification. Our corpus contains two sets of the same 202 sentences but annotated from different perspectives. The first one, corresponding to Word-level, has 1028 words, while the second one, corresponding to Morph-level, has 1928. For annotations following the word-level notation paradigm, the process was not done manually. Instead, a Python script was used to automatically attach all "words" in the original text that start with "=" to the preceding word. Finally, manually, it was necessary to double-check the number assigned for each dependency relations.

5 POS Tags

The difference between word-level and morpheme-level POS tags is illustrated in Table 2. We should note that the primary distinction between the two schemes lies in the PART category. This is expected since clitics, which are often overlooked in word-level annotations, have been explicitly labeled as PART in the morpheme- level scheme.

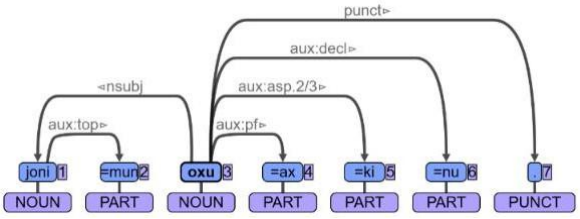
Table 2. POS Frequency

POS	Word-level	Morph-level
NOUN	268	268
ADJ	41	41
PART	-	889
PROPN	31	31
VERB	201	201
PUNCT	210	210
PRON	169	169
DET	70	70
ADV	29	33
NUM	5	5

While the language has clear nominal and verbal bases, it is important to note that there is no

deverbal derivation, so nouns may carry "verbal" morphology, as seen in (5), where *oxu* 'moon' has no morphological derivation, but it means 'turn into the moon'. In these cases, we maintain the grammatical category of the base, as it is a property of the language.

5. The man turned into the moon.



6 Dependency Relations

Our annotation scheme utilizes 56 types of dependency relations. Generally, we have adhered to the guidelines provided by UD, except for cases involving clitics. In the morpheme-level scheme, there are a total of 1,927 dependency relations, while for the word- level scheme, there are 1,031.

Table 3. Clitic and its dependency relation label Frequency

Clitic	Dependency relation Label	Frequency
=mun	aux:top	183
=nu	aux:decl	178
=n	Case	93
=ki	aux:2/3	70
=xo	aux:pfv.3	67
=hi	aux:ipfv	36
=ku	aux:pfv.1/2	22
=x	Case	17
=ka	aux:1	15
=ra	aux:int	14

While Universal Dependencies (UD) aims to provide "a universal inventory of categories and guidelines to facilitate consistent annotation of similar constructions across languages" (Nivre et al., 2017), it also accommodates language- specific subtype relation labels when necessary. Following Vásquez et al. (2018), we have chosen to treat clitics as distinct syntactic entities. Consequently, connections between words and clitics are regarded as syntactic and annotated using the appropriate dependency structure. In fact, Amahuaca grammatical elements, specifically clitics, exhibit such a free distribution that they resemble words. We employ the label "aux" for non-nominal clitics, as illustrated in Table 3. Except for *=n* and *=x*, which are clitics for cases, the other more frequent clitics are non-nominal: topic (*aux:top*), declarative (*aux:decl*), verbal persons (*aux:2/3*, *aux:1*), perfective (*aux:pfv.3*), imperfective (*aux:ipfv.2/3*), and interrogative (*aux:int*).

We considered introducing new subtype relation labels corresponding to verbal inflection, mood, and focus clitics. However, to ensure that the label reflects the syntactic meaning of the dependency relation, we decided to use "aux" followed by the gloss corresponding to the clitic. For example, if it is *=nu*, marking declarative mood, the corresponding label would be "aux:decl". Additionally, we found it necessary to include the "nsubj:copy" relation due to the doubling pronouns mentioned earlier in preceding sections (See (4)).

7 Conclusions

This paper presented the results obtained from the manually annotated corpus following both a morpheme-level and a word-level annotation schema for the Amahuaca language. As explained in detail in Section 3, annotating according to a morpheme-level schema is more convenient for Amahuaca, a language with rich morphology characterized by complex morphosyntactic relations among morphemes and interactions with clitics. For instance, in the sentence *Janmun jan ruratixon machitoxon nixohnu*, meaning "He made machetes and axes," where =ni, the temporal clitic, functions as the root of the sentence, this interaction would not be adequately captured in a word-level analysis.

The evaluation of accuracy between these two schemas using UDPipe remains pending, allowing for a comparison of whether there is a significant difference between them. While the morpheme-level annotation may require more linguistic resources, such as a morphological analyzer and morphological segmentation, it provides a deeper insight into the language and has the potential to improve automatic parsing. Ultimately, it is expected that a morpheme-level syntactic dependency annotation may be a more effective way to represent polysynthetic languages within the framework of Universal Dependencies.

References

Alonso Vasquez, Renzo Ego Aguirre, Candy Angulo, John Miller, Claudia Villanueva, Željko Agić, Roberto Zariquiey, and Arturo Oncevay. 2018. Toward Universal Dependencies for Shipibo-

Konibo. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 151–161, Brussels, Belgium. Association for Computational Linguistics.

Cristopher Hewlett. 2014. *History, kinship and comunidad: learning to live together amongst Amahuaca people on the Inuya River in the Peruvian Amazon* (Doctoral dissertation, University of St Andrews).

Çağrı **Ç**öltekin. 2016. (When) do we need inflectional groups? In *Proceedings of The 1st International Conference on Turkic Computational Linguistics*, page (to appear).

Emily Clem. 2019. Amahuaca ergative as agreement with multiple heads. *Natural Language & Linguistic Theory,* 37, 785-823.

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, Barcelona, Spain (Online). Association for Computational Linguistics.

Gertrude Dole. 1998. Los amahuaca. *Guía etnográfica de la Alta Amazonía*, 3, 125-273.

Hyunji Hayley Park, Lane Schwartz, and Francis Tyers. 2021. Expanding Universal Dependencies for Polysynthetic Languages: A Case of St. Lawrence Island Yupik. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 131–142, Online. Association for Computational Linguistics.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. 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. 2020. 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.

Margarethe Sparing-Chávez. 2012. Aspects of Amahuaca grammar: An endangered language of the Amazon basin. *Dallas: SIL International.*

Pilar Valenzuela. 2003. *Transitivity in shipibo- konibo grammar*. University of Oregon.

Pilar Valenzuela, Roberto Zariquiey and Candy Angulo. In preparation. *A grammar sketch of Amahuaca (Pano, Peru)*

Roberto Zariquiey, Claudia Alvarado, Ximena Echevarría, Luisa Gomez, Rosa Gonzales, Mariana Illescas, Sabina Oporto, Frederic Blum, Arturo Oncevay, and Javier Vera. 2022. Building an Endangered Language Resource in the Classroom: Universal Dependencies for Kakataibo. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3840–3851, Marseille, France. European Language Resources Association.

Data augmentation for low-resource bilingual ASR from Tira linguistic elicitation using Whisper

Mark Simmons
University of California San Diego
mjsimmons@ucsd.edu

Abstract

This paper explores finetuning Whisper for transcribing audio from linguistic elicitation of Tira, a Heiban language of Sudan. Audio originates from linguistic fieldwork and is bilingual in English and Tira. We finetune Whisper large-v3 using hand-labeled Tira audio and evaluate the resulting model on bilingual audio. We show that Whisper exhibits catastrophic forgetting of English after only a small amount of training, but that including automatically annotated English spans of audio in the training data dramatically reduces catastrophic forgetting of English while largely preserving ASR performance on monolingual Tira audio. This work is relevant to the study of automatic speech recognition for under-resourced languages and for contexts of bilingualism in a high and low-resourced language.

1 Introduction

Automatic speech recognition (ASR) tools convert speech into text, enabling rapid transcription or captioning of audio. Recent ASR models have reached or exceeded human performance at transcription on high-resource languages such as English (Radford et al., 2022), however performance lags in under-resourced languages and in contexts of code-switching (where multiple languages are used in a single conversation). While research on expanding Whisper's performance on low-resource languages exists (e.g. Lord and Newman, 2024; Liu et al., 2024; Williams et al., 2023; Qian et al., 2024), less work has been done on improving performance in code-switched scenarios. Code-switching is an under-addressed topic in ASR and in NLP in general, and research there often focuses on a few high-resource language pairs, such as Spanish-English, Mandarin-English or Hindi-English (Winata et al., 2023). Peng et al. (2023) evalute Whisper on Mandarin-English code-switched audio and Kulkarni et al. (2023), on Mandarin-English, Arabic-English and Hindi-English, for example.

The majority of languages in the world can be classified as 'ultra low-resource' in terms of the amount of NLP research and tools available for them (Liu et al., 2022). While ASR research for such languages exists (e.g. Prud'hommeaux et al., 2021; Adams et al., 2018; Amith et al., 2021; Mitra et al., 2016), the only work we are aware of that addresses ASR with an ultra low-resource language paired with a high resource language is San et al. (2022), which uses a corpus of single-speaker audio in English and Muruwari, though they only use ASR for English in their corpus. Thus, we are not aware of any work that directly addresses code-switched ASR involving at least one ultra low-resource language.

In this paper, we evaluate Whisper on bilingual audio in English and Tira, an ultra low-resource language of the Heiban family spoken in the Nuba mountains region of Sudan, before and after finetuning on monolingual audio in Tira. Audio comes from linguistic elicitation on Tira conducted by the authors and other colleagues in the Tira language project in collaboration with native Tira speaker Himidan Hassen. Linguistic elicitation refers to the process of studying the grammar of a language by "asking questions" from native speakers (Mosel, 2008). This often involves use of a metalanguage, a language spoken in common between the linguist and language speaker, in this case English, to ask for translations of words, paradigms, or sentences into the target language, or to elicit morphological paradigms for a given word in the target language. Audio from the Tira language project, then, contains speech both in Tira and English. While elicitation is different than classical code-switching, where interlocutors use multiple languages to communicate (often within the same utterance), the challenges faced in ASR for bilingual elicitation are largely the same as those faced in ASR for code-switched audio, thus, we use the term "bilingual

Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 155–161

audio" to refer to either.

The contributions of this paper are as follows. We describe our process for using fieldwork data from linguistic elicitation of Tira, an out-of-domain language for Whisper, to create an ASR dataset. We then finetune Whisper on this dataset, and evaluate on bilingual audio in Tira and English. We also compare this with fine-tuning Whisper on Tira and English simultaneously by using existing hand-labeled annotations for Tira and automatically generated labels for English.

2 Dataset

We first created a Tira audio corpus using existing fieldwork recordings. Tira is a tonal language, meaning that pitch can distinguish words and morphemes. Tone has historically been difficult for ASR, as it is realized suprasegmentally, that is, simultaneous with the production of phonological segments such as consonants and vowels (Adams et al., 2018; Mortensen et al., 2016).

Audio labels for Tira come from pre-existing annotations recorded in ELAN (Sloetjes and Wittenburg, 2008), a software for annotation of multimedia recordings. The annotations relevant to this work are narrow IPA transcription and free translation into English. IPA transcriptions along with timestamps were extracted using the Python `pympi-ling` package[1]. A total of 28k annotated utterances were found from across 202 elicitation session recordings, totalling to 16 hours of audio. As these annotations were made to be used as reference for the purposes of linguistic documentation, certain noise is present in the labels relative to ASR training data. For example, 2123 records that did not have tone marked were excluded from the dataset. Sometimes Himidan's metacommentary in English is included alongside a Tira utterance in a single annotated label. We used the `pyenchant`[2] library to look for any sentences containing English words in the transcription and discarded these sentences from the monolingual Tira dataset. Sometimes, Himidan hums or whistles a Tira sentence for purposes of hearing the tones. Many records were explicitly labeled as such either in the transcription or translation tier, i.e. [kəvəlɛðéŋí únɛrɛ] "Whistling: I pulled him here yesterday.", but some whistled or hummed speech is included with no overt indication. To account for this, we used PyAnnote voice activity detection[3] (VAD) (Bredin and Laurent, 2021; Bredin, 2017) to determine the percentage of total duration for each record that was detected as speech. We found that the majority of records contained $\geq 60\%$ speech, so we excluded all records beneath this threshold, 825 records in total. Manual inspection showed that records beneath the 60% threshold were often completely silent, contained humming, whistling, excessive noise, or static.

Another metric we use for assessing audio quality is cosine similarity of text and audio embeddings using CLAP-IPA (Zhu et al., 2024). CLAP-IPA consists of an audio encoder, which takes audio input in any language and returns an acoustic embedding $\mathbf{s}$, and a phoneme encoder, which takes a sequence of IPA characters as input and returns a phoneme embedding $\mathbf{p}$ such that the speech embedding for a given word should have a small cosine distance to the phone embedding for its respective IPA sequence. CLAP-IPA was intended for keyword spotting (the task of identifying a given word, or in this case phoneme sequence, in a stream of speech) and forced alignment (the task of mapping each unit in a given word or phoneme sequence to its timestamps in the audio). However, we adopt it here as a metric for summarizing transcription noise with the assumption that audio which is clearer and is free of noise, cross-talk or other artefacts will have a high cosine similarity to its respective transcription. We calculated the cosine similarity of the embedding for each audio record with the phoneme embedding for its respective transcription. We found that most records in the dataset were above or equal to the threshold $\mathrm{sim}(\mathbf{s}, \mathbf{p}) = 0.6$, so we excluded any record whose cosine similarity fell beneath this value, 2156 records in total. Manual inspection of excluded records indicated that they generally contained significant noise or echo, or included commentary in English run along with the Tira utterance where only the Tira utterance had been transcribed in the label.

Annotations were made in a narrow phonetic transcription rather than in an established orthography, which can introduce variation as transcribers are required to make subjective decisions of how to represent phonetic variation (cf. Michaud et al. 2018). We compensated for this by normalizing the set of IPA symbols used in the dataset. For

[1] https://pypi.org/project/pympi-ling/
[2] https://pypi.org/project/pyenchant/
[3] https://huggingface.co/pyannote/voice-activity-detection

example, the phoneme /ɟ/ might be transcribed [ɟ, j, dʒ, dʑ, dʐ]. Each of these symbols were replaced with [ɟ], and similarly for other phonemes. We also used NFKD normalization from the Python `unicodedata`[4] package.

Data splits should be chosen so as to minimize overlap between partitions. For fieldwork audio datasets, splits may be decided on speaker identity or grouped by narrative. For the Tira dataset, only one speaker is present, and different recordings may have significant overlap in their content. For example, across several elicitation sessions focusing on syntactic structure utterances may begin with [ùrnɔ̀ kəlɛ̀ŋ̟ì̟t̪ɔ́ àprí…] 'grandfather knows that the boy…'. To maximize the difference across data partitions, we calculated the phone embedding for each transcription using CLAP-IPA and sampled records so as to maximize the cosine distance of embeddings between the train, validation and test splits. Statistics for the size of this and other datasets are given in Table 1. We refer to this dataset as the "hand-labeled monolingual" or "hand-labeled" dataset.

To evaluate the model's generalization to bilingual audio, we hand annotated labels from two elicitation recordings containing both Tira and English speech. We picked one recording that supplied Tira labels used for training (the "in-domain" bilingual set) and one recording that was not used in training (the "out-of-domain" bilingual set). Note that English audio for both recordings will be unseen for the model. Each label was taken from up to 30 seconds of speech, always segmented to end at the end of a speech turn.

We also created bilingual labels through data augmentation. For bilingual label creation, we used PyAnnote VAD to detect regions of speech from the longform elicitation recordings that were not included in the hand-labeled dataset. Since not all Tira utterances from the elicitation recordings were hand-labeled, several of these detected utterances contain speech in Tira. To distinguish between Tira and English audio, we trained a logistic regression model on hand-labeled Tira and English spans from the elicitation corpus to perform language identification (LID), using embeddings from the SpeechBrain ECAPA TDNN for language identification (Ravanelli et al., 2021), similar to the protocol outlined in (San et al., 2022). We trained LID on a dataset of 3637 Tira and 3637 English utterances, and it achieved 90% classification accuracy on a test dataset of 1818 Tira and 1818 English utterances. Tira utterances were all taken from Himidan, whereas English utterances were sampled from both Himidan and other speakers. Once utterances were segmented and labeled for language identity, English utterances were transcribed using Whisper large-v2 (which we found to perform better than Whisper large-v3 on English) and Tira utterances were transcribed using the fine-tune of Whisper large-v3 on monolingual Tira audio, as described in the following section. We then used these annotations to make a bilingual ASR dataset. For each hand-transcribed label from the monolingual dataset, we concatenated adjacent transcribed speech regions in the same elicitation recording to create a new label of up to 30 seconds. We excluded utterance transcriptions with excessive repetition (e.g. "Yeah. Yeah. Yeah. Yeah…"), a known failure mode of Whisper. For all training labels from the in-domain elicitation recording used for bilingual evaluation, we only included the hand-labeled Tira utterances to ensure both the model trained on the monolingual dataset and the bilingual dataset have seen the same set of data from the in-domain elicitation recording during training. We refer to this dataset as the "augmented bilingual" or "augmented" dataset.

Textual analysis of the labels revealed 14,017 words (5.3% of the whole dataset) were not identified as English (using `pyenchant` as above) or Tira (defined as any word containing only Tira IPA characters). Manual inspection of such words suggests that several are Tira words that were missed by the VAD+SLI pipeline and thus were transcribed by Whisper large-v2 rather than the checkpoint trained on Tira, e.g. "kukungapitito" instead of [kúkù ŋgápɪ̀t̪ɪ̀t̪ɔ́] 'Kuku hunted (in someone's place)', or "ngiyol" instead of [ŋìjɔ́l] 'eat'.

3 Experiment

We finetuned Whisper large-v3 using a learning rate of $3e-4$ with 500 warmup steps. We used the AdamW optimizer (Loshchilov and Hutter, 2019) with betas of 0.9 and 0.99, and trained with a batch size of 4 with 2 gradient accumulation steps for an effective batch size of 8. All models were trained with an Nvidia GeForce RTX 4090 with 24 gigabytes of VRAM. Due to GPU VRAM limitations, we were not able to finetune all of the weights

[4] https://docs.python.org/3/library/unicodedata.html

Dataset	Split	N records	Length (total)	Avg record len	%Tira	%Unk
Monolingual	train	16,384	9h29m	2.08s	100	
Monolingual	dev	2,048	1h8m	1.99s	100	
Bilingual	train	16,384	51h48m	11.38s	21.0	5.3
Bilingual in-domain	test	88	39m	26.25s	8.60	
Bilingual out-domain	test	65	29m	26.31s	2.86	

Table 1: Size of datasets used for training, validation (dev) and testing.

of Whisper large-v3, and had to rely on parameter efficient finetuning (Han et al., 2024; Houlsby et al., 2019). We used LoRA (Hu et al., 2021) applied to the query and value weights of the attention modules for parameter-efficient finetuning, following the example given in PEFT (Mangrulkar et al., 2022)[5], similar to Liu and Qu (2024). Models are evaluated in terms of word error rate (WER) and character error rate (CER). As Tira is out-of-domain for Whisper, labels were prefixed with a language ID for Yoruba (another tonal African language) for purposes of knowledge transfer (Qian et al., 2024), though we leave more thorough investigation of language ID choice for later research. We compared finetuning Whisper large-v3 using a LoRA, Whisper medium using a LoRA, and a full finetune of Whisper medium, by training each model size on the Tira dataset for 4 epochs. We found the best results came from Whisper large-v3 with LoRA, so we use this configuration for our experiment. We finetune one model for 10 epochs on each dataset respectively (hand-labeled monolingual and augmented bilingual).

4 Results

We evaluate Whisper large-v3 out of the box and compare it to a finetune using LoRA at each epoch of training using both monolingual hand-labeled data and augmented bilingual data, evaluating on both monolingual Tira data and bilingual Tira-English data. WER and CER on each evaluation set across training are given in Figure 1, where "epoch 0" corresponds to Whisper large-v3 with no finetuning.

In general, the model trained on augmented bilingual data outperforms the monolingual model when evaluated on bilingual data. When evaluated on monolingual data, both models perform similarly, with the monolingual model slightly outperforming the bilingual model.

For monolingual data, we see a precipitous drop in CER (0.86 to 0.15 for the monolingual model, 0.20 for the bilingual model) and WER (1.70 to 0.59 for the monolingual model, 0.72 for the bilingual model) in epoch 1, with much smaller improvements each subsequent epoch. For bilingual data, we see conflicting results with the model trained on monolingual Tira data. On the out-of-domain bilingual dataset, the monolingual model underperforms Whisper large-v3 at all epochs of training. For the in-domain bilingual dataset, there is a slight reduction in WER and CER by epoch 2, likely owing to the model's ability to transcribe Tira it has recognized in training, followed by a decline in performance in all subsequent epochs. Unlike the monolingual model, the augmented bilingual model's performance improves on both monolingual and bilingual datasets with training, achieving the best WER and CER at epoch 4 for the in-domain dataset and 10 for the out-of-domain dataset.

Dataset	Model	WER	CER	Epoch
Tira monoling	Tira only	**0.48**	**0.11**	8
	Augmented	0.53	0.13	10
In-domain biling	Tira only	0.83	0.57	2
	Augmented	**0.55**	**0.34**	4
Out-domain biling	Tira only	0.57	0.83	0
	Augmented	**0.49**	**0.34**	10

Table 2: Best WER and CER on validation sets

In Figure 2, we break down CER and WER by language. This plot confirms the trend suggested in Figure 1, namely that both the monolingual and augmented bilingual models perform similarly on Tira, but the augmented bilingual model significantly outperforms the monolingual model on English, likely owing to the inclusion of an English transcription task in training, even on synthetic labels.

Manual inspection gives further evidence that the worsening performance following epoch 2 for the monolingual model is due to catastrophic forgetting of English. For example, the span "on the computer" uttered by Himidan is transcribed correctly

[5] https://github.com/huggingface/peft/tree/main/examples/int8_training

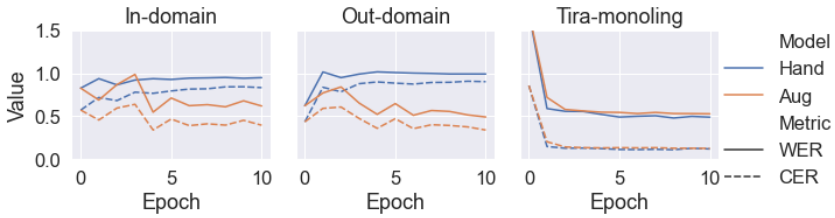


Figure 1: CER and WER on Tira validation sets. $\text{Epoch} = 0$ is equivalent to Whisper large-v3 with no finetuning.

in English by Whisper large-v3 and the model finetuned on monolingual hand labels after the first epoch, but in the second epoch is already transcribed as "aða kʊɔɛmpídɔ". This happens even to linguist's speech in English, particularly in proximity of Himidan giving a Tira production e.g. the span "[Himidan] ŋɔ́ɽɔ́n [linguist] Yeah I saw in the Stephen dictionary it was written as ŋicəlo" from the out-of-domain dataset was rendered "ŋɔ́ɽɔ́n, jà ìs ìn sn̪ìjən̪ dìkʃənɛ̀ ìwɛ̀s rɛ̀n̪ìŋ ìcəlò" after only one epoch of training. Manual inspection of the output of the augmented bilingual model shows that it is more common for Tira spans to be transcribed in a non-IPA pseudo-English orthography, even if the same span is correctly transcribed in the same proximity, e.g. "I'll pull them... okay **La lovela. lál ló válɛ̀ða nd̪ɔ̀bà**". This is likely due to the presence of similar spans in the augmented dataset, owing to the imperfect nature of the VAD>SLI pipeline, and could likely be ameliorated by improving the quality of the augmented dataset.

5 Conclusion

We describe the steps to create an ASR dataset from linguistic elicitation of an ultra low-resource language, Tira, including various strategies for data cleaning. We use the dataset to train Whisper large-v3, and evaluate on bilingual audio in Tira and English. We compare training on hand-labeled monolingual Tira audio with training on an augmented dataset where English (and additional Tira) audio is included with machine-generated labels. We show that the model exhibited catastrophic forgetting of English and overfitting after only one epoch of training, but this can be minimized by adding synthetic labels in English.

6 Limitations and future directions

Our training dataset comes from one speaker alone and is limited in its subject domain. As such the models produced in this work are overfit to his speech and to the domain of linguistic elicitation,

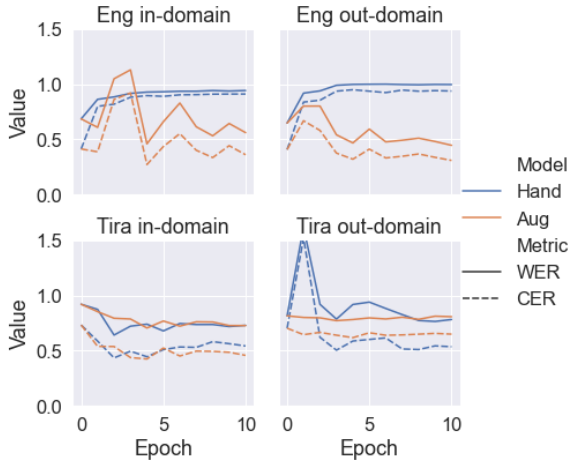


Figure 2: Language-specific WER and CER for bilingual datasets.

and would not generalize well to conversational speech in Tira or to other Tira speakers. However, our goal is a model suited to transcribing audio specifically from a context of linguistic fieldwork or pedagogy. We hope that our method can be extended to aid documentation and revitalization efforts on other low resource languages.

Future directions include comparing machine learning techniques for preventing catastrophic forgetting to training with artificial bilingual labels to see which causes the least degradation of English ASR performance, improving the quality of the augmented dataset, and reproducing these experiments with other datasets of bilingual audio from fieldwork corpora.

7 Ethical considerations

Data gathered on Tira were recorded with the consent of the speaker and the permission of UC San Diego's IRB (Protocol 805624). Annotations were produced by the authors and other academic colleagues. Data in English come from Himidan as well as the authors and other linguists present during elicitation sessions. No other data were used.

References

Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, and Alexis Michaud. 2018. Evaluation Phonemic Transcription of Low-Resource Tonal Languages for Language Documentation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jonathan D. Amith, Jiatong Shi, and Rey Castillo García. 2021. End-to-End Automatic Speech Recognition: Its Impact on the Workflowin Documenting Yoloxóchitl Mixtec. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 64–80, Online. Association for Computational Linguistics.

Hervé Bredin. 2017. Pyannote.metrics: A Toolkit for Reproducible Evaluation, Diagnostic, and Error Analysis of Speaker Diarization Systems. In *Proc. Interspeech 2017*, pages 3587–3591.

Hervé Bredin and Antoine Laurent. 2021. End-to-end speaker segmentation for overlap-aware resegmentation.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey. *Preprint*, arXiv:2403.14608.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. *Preprint*, arXiv:1902.00751.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *Preprint*, arXiv:2106.09685.

Atharva Kulkarni, Ajinkya Kulkarni, Miguel Couceiro, and Hanan Aldarmaki. 2023. Adapting the adapters for code-switching in multilingual ASR.

Yunpeng Liu and Dan Qu. 2024. Parameter-efficient fine-tuning of Whisper for low-resource speech recognition. In *2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*, pages 1522–1525.

Yunpeng Liu, Xukui Yang, and Dan Qu. 2024. Exploration of Whisper fine-tuning strategies for low-resource ASR. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024(1):29.

Zoey Liu, Crystal Richardson, Richard Hatcher, and Emily Prud'hommeaux. 2022. Not always about you: Prioritizing community needs when developing endangered language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3933–3944.

Laurel Lord and Mark Newman. 2024. Automatic Speech Recognition Variance: Consecutive Runs of Low-Resource Languages in Whisper. *International Journal of Machine Learning*, 14(2).

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. *Preprint*, arXiv:1711.05101.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Alexis Michaud, Oliver Adams, Trevor Cohn, Graham Neubig, and Séverine Guillaume. 2018. Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone toolkit. *Language Documentation & Conservation*, 12:393–429.

Vikramjit Mitra, Andreas Kathol, Jonathan D. Amith, and Rey Castillo García. 2016. Automatic Speech Transcription for Low-Resource Languages — The Case of Yoloxóchitl Mixtec (Mexico). In *Interspeech 2016*, Interspeech_2016. ISCA.

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. PanPhon: A Resource for Mapping IPA Segments to Articulatory Feature Vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484, Osaka, Japan. The COLING 2016 Organizing Committee.

Ulrike Mosel. 2008. Chapter 3 Fieldwork and community language. In *Essentials of Language Documentation*, pages 67–86. De Gruyter Mouton.

Puyuan Peng, Brian Yan, Shinji Watanabe, and David Harwath. 2023. Prompting the Hidden Talent of Web-Scale Speech Models for Zero-Shot Task Generalization. In *INTERSPEECH 2023*, pages 396–400. ISCA.

Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Karin Michelson. 2021. Automatic Speech Recognition for Supporting Endangered Language Documentation. *Language Documentation & Conservation*, 15:491–513.

Mengjie Qian, Siyuan Tang, Rao Ma, Kate M. Knill, and Mark J. F. Gales. 2024. Learn and Don't Forget: Adding a New Language to ASR Foundation Models. *arXiv preprint*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv preprint*.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba,

Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. SpeechBrain: A General-Purpose Speech Toolkit.

Nay San, Martijn Bartelds, Tolulope Ogunremi, Alison Mount, Ruben Thompson, Michael Higgins, Roy Barker, Jane Helen Simpson, and Dan Jurafsky. 2022. Automated speech tools for helping communities process restricted-access corpora for language revival efforts. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*. Association for Computational Linguistics.

Han Sloetjes and Peter Wittenburg. 2008. Annotation by Category: ELAN and ISO DCR. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Aiden Williams, Andrea Demarco, and Claudia Borg. 2023. The Applicability of Wav2Vec2 and Whisper for Low-Resource Maltese ASR. In *2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023)*, pages 39–43. ISCA.

Genta Winata, Alham Fikri Aji, Zheng Xin Yong, and Thamar Solorio. 2023. The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978, Toronto, Canada. Association for Computational Linguistics.

Jian Zhu, Changbing Yang, Farhan Samir, and Jahurul Islam. 2024. The taste of IPA: Towards open-vocabulary keyword spotting and forced alignment in any language. *Preprint*, arXiv:2311.08323.

Integrating diverse corpora for training an endangered language machine translation system

Hunter Scheppat[1], Joshua K. Hartshorne[2], Dylan Leddy[1],
Éric Le Ferrand[1], Emily Prud'hommeaux[1]
[1]Boston College, [2]MGH Institute of Health Professions
{scheppat,leferran,prudhome}@bc.edu,joshua.hartshorne@hey.com

Abstract

Machine translation (MT) can be a useful technology for language documentation and for promoting language use in endangered language communities. Few endangered languages, however, have an existing parallel corpus large enough to train a reasonable MT model. In this paper, we re-purpose a wide range of diverse data sources containing Amis, English, and Mandarin text to serve as parallel corpora for training MT systems for Amis, one of the Indigenous languages of Taiwan. To supplement the small amount of Amis-English data, we produce synthetic Amis-English data by using a high quality MT system to generate English translations for the Mandarin side of the Amis-Mandarin corpus. Using two popular neural MT systems, OpenNMT and NLLB, we train models to translate between English and Amis, and Mandarin and Amis. We find that including synthetic data is helpful only when translating to English. In addition, we observe that neither MT architecture is consistently superior to other and that performance seems to vary according to the direction of translation and the amount of data used. These results indicate that MT is possible for an under-resourced language even without a formally prepared parallel corpus, but multiple training methods should be explored to produce optimal results.

1 Introduction

The potential of language technology to support endangered language documentation and revitalization efforts is well established though not always effectively realized (van Esch et al., 2019). Machine translation (MT) in particular has been cited as a useful tool (Zhang et al., 2020; Bird and Chiang, 2012). First, translation from an indigenous language into a more widely spoken language is a common, if not required, part of generating linguistic documentation. This also ensures that understanding of the language will continue even if the language ceases to be used regularly (Bird and Chiang, 2012). Second, MT is often proposed as a way to make languages more accessible to language learners in Indigenous communities where younger generations were not raised speaking the language(Pinhanez et al., 2024). Finally, MT is appealing to NLP researchers because generating a new dataset only requires the expertise of a speaker to produce a translation; translation does not require complex software to control audio playback or alignment of audio with transcription, as speech transcription might, or extensive annotator training as part-of-speech tagging or parsing would.

Unfortunately, building a reasonable MT system with the quantity of parallel data typically available for an endangered language is remarkably challenging. There are few existing parallel corpora, and since nearly half of the world's languages lack an established writing system or written tradition (Eberhard et al., 2024), there are generally very few texts in the target language that can be translated in order to create a parallel corpus.

In this paper, we describe a broad effort to compile two parallel corpora – one with English and one with Mandarin – for Amis, one of the 16 recognized indigenous languages of Taiwan. We use nine different public sources[1] for our parallel data, which range from digital dictionaries to pedagogical materials to websites with user-contributed translations to YouTube videos curated and translated by Taiwan's Indigenous Language Research and Development Foundation. Since very little of this data includes English translations, we also generate synthetic Amis-English parallel data by using Mandarin as a pivot language, using high-quality MT to produce English translations of the Mandarin side of the Amis-Mandarin parallel data.

Using two different popular neural MT archi-

[1]Please see the Ethical Considerations section for details about our data use agreements.

Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 162–169

tectures – the end-to-end OpenNMT framework (Klein et al., 2020) and the No Language Left Behind architecture for fine-tuning from a large pretrained multilingual model (Costa-jussà et al., 2024) – we build models to translate between Amis and Mandarin, and Amis and English. Even with our small and diverse "found" datasets, we are able to achieve reasonable BLEU and chrf++ scores. Supplementing the Amis-English parallel corpus with pivot data yields improvements when translating to English but not to Amis. Interestingly, we find that neither architecture consistently outperforms the other. The results suggest that compiling parallel data from diverse sources can create a corpus sufficient for training reasonable MT models. The interactions between architecture, corpus size, and translation direction, however, require additional study.

2 Background

2.1 Amis language

Though spoken in Taiwan, Amis (ISO 639-3 language code ami) is unrelated to Mandarin or other Sinitic languages. Rather, it is a member of the Formosan branch of the Austronesian language family, one of the largest families in the world both by number, with around 1,200 extant languages, and by geographic spread, ranging from Malagasy in the West to Rapanui in the East, and from New Zealand in the South to Taiwan in the North. Amis, the most widely spoken of the 16 Formosan languages with just over 100,000 speakers, has five officially-recognized dialects and is classified by Ethnologue as *threatened* (Eberhard et al., 2024).

Amis, like other Formosan languages, has a number of typologically unusual features (Li et al., 2024). It has primarily VSO word order. It has a limited phonetic inventory, with a only three vowels. It makes use of reduplication as part of the grammar, and its lexical roots are not easily categorized by part of speech. Most famously, Amis and other Formosan languages have a complex grammatical voice system. In short, Amis bears little resemblance to the languages used to train most multilingual models, including the multilingual NLLB model.

2.2 Endangered language MT

As noted previously, MT is recognized as a potentially useful tool for language documentation (Bird and Chiang, 2012; van Esch et al., 2019). A number of MT systems for endangered and indigenous languages have been developed for research or demonstration purposes, including (among many others) Cherokee-English (Zhang et al., 2020), Kotiria (Kann et al., 2022), Quechua (Ortega et al., 2020), Highland Puebla Nahuatl, and Ainu (Miyagawa, 2023).

There is a small amount of prior work on MT for Amis specifically. Zheng et al. (2022) created a small parallel Amis-Mandarin corpus using a subset of the ILRDF data that we use and an associated dictionary (see Section 3.1). Using an mBART-based transformer model, they trained models to translate between Amis and Mandarin. In a follow-up paper, Zheng et al. (2024) continued this work with multiple Formosan languages exploring the impact of including additional data, particularly dictionaries and lexica, as well as synthetic data. While their results are not directly comparable to our Amis-Mandarin results given the very different training corpora, their results without data augmentation are comparable to those we present here. We note, however, that we consider translation both to Mandarin and to English. In addition, unlike our work, this prior work does not compare the performance of different MT architectures.

3 Data

3.1 Data sources

The following data sources were used to create parallel corpora for our experiments translating between Amis and English and between Amis and Mandarin.

1. **ILRDF Videos**: Videos and manually-produced captions created by the Indigenous Language Research and Development Foundation (ILRDF) of Taiwan. The content primarily includes short-form, casual conversations with Amis speakers, with translations provided in Mandarin. The videos typically range from 1 to 5 minutes in length.

2. **Presidential Apology**: An official apology issued by the president of Taiwan to the Indigenous people of Taiwan. Although brief, the document contains high-quality text with long sentences, available in Amis, English, and Mandarin.[2]

[2]https://www.president.gov.tw/NEWS/20603

3. **Bible**: A short, user-generated and unverified section of the New Testament in Amis with English translations.

4. **ePark** (Aboriginal Language Research and Development Foundation, 2023b): A large electronic education website supported by ILRDF. All texts are available in Amis and Mandarin; many are also available in English and one or more Amis dialects.

5. **Glosbe**: An online community-developed dictionary similar to Wikipedia, which includes user-contributed Mandarin translations and definitions.[3]

6. **ILRDF Dictionary** (Aboriginal Language Research and Development Foundation, 2023a): An electronic dictionary published by ILRDF which contains extensive example sentences in Amis and translations into Mandarin.

7. **Zheng Corpus**: Zheng et al. (2022) made available a dataset based on the above ILRDF Dictionary resource that contains some new text, with translations into Mandarin.

8. **NTU Corpus** (Su et al., 2008): 16 narratives and 2 conversations in Amis, with free translations in English and Mandarin.

9. **Fey Dictionary** (Fey, 1986): Amis dictionary compiled by Virginia Fey, with example sentences translated into Mandarin and English.

3.2 Data acquisition and alignment

The Glosbe, Bible, and ILRDF Video texts were downloaded from the Web, and the Presidential Apologies were extracted from PDFs. The ILRDF Dictionary texts were obtained through the ILRDF API. The NTU Corpus and ePark were provided to the authors by the owners.

The ILRDF video data posed several challenges. First, many instances of code-switching occurred, where Amis speakers switched to Mandarin mid-sentence, resulting in Chinese characters appearing within the Amis text. Some sentences contained only Mandarin, as speakers switched languages for extended periods. Additionally, the quality of the translation pairs was sometimes inconsistent. For instance, many pairs included the Mandarin word for "unknown", indicating that the translator was unsure of the Amis speaker's meaning. Furthermore, the translations often included descriptions of non-verbal sounds, such as "leaves rustling" which did not appear in the original Amis text, which we attempted to automatically filter out.

The quality of Glosbe data was also challenging. The text often included unusual formatting, and there was significant overlap between Glosbe and other texts, such as the Bible. The Glosbe website was scraped by searching for common words in the Amis language, as direct access to all sentences for one language was unavailable. After the search, duplicates were removed, and the remaining sentences were formatted into parallel text.

The text of the Presidential Apology was manually aligned separately for each translation pair. As a result, the sentence counts differ in the two parallel corpora.

The NTU Corpus was prepared by linguists specializing in Amis and fully bilingual in English and Mandarin, yielding reliable and high quality translations. Similarly, because the ePark corpus consists of officially-produced and verified language educational material, it is of very high quality. We note that most of the ePark texts, rather than being produced initially in Amis, were instead translated *to* Amis from Mandarin or English.

The Bible data used here – a user-generated subset of Bible verses – was available online pre-aligned. We note that the English side of this corpus features the archaic language found in the King James version of the Bible, which may render this corpus less useful for translation to contemporary English.

The Fey dictionary included translations of sentences, short sentence fragments, and individual words. Occasionally, a single Amis sentence had multiple valid English translations. In these cases, we included each English translation as a distinct pair with the original Amis sentence.

3.3 Data preprocessing

Data processing focused primarily on ensuring reliable alignments and translations and correct formatting. To effectively address the issue of noisy translation pairs – those that either do not accurately reflect the source content – we implemented a fertility heuristic. This heuristic was designed to filter out sentence pairs exhibiting significant discrepancies in length between the source and target texts. The assumption underlying this approach is

[3] https://glosbe.com/

Corpus	Language	Sentence Pairs (w/ pivot)	Tokens (w/ pivot)	Types (w/ pivot)
ILRDF Videos	eng	(38,780)	(268,006)	(10,149)
	ami	38,780	227,074	21,469
Presidential Apology	eng	92	1,559	532
	ami	92	1,573	422
Bible	eng	512	11,676	1,482
	ami	512	11,469	1,679
ePark	eng	21,699 (27,904)	87,693 (143,319)	5,074 (8,138)
	ami	27,904	127,328	13,298
Glosbe	eng	(1,305)	(18,732)	(2,759)
	ami	1,305	18,340	2,752
ILRDF Dictionary	eng	(17,763)	(99,639)	(7,100)
	ami	17,763	80,012	8,756
NTU Corpus	eng	922	8,881	1,252
	ami	922	8,881	1,595
Fey Dictionary	eng	2,180	11,827	2,248
	ami	2,180	9,621	2,273

Table 1: Corpus sentence pair, token, and type counts for the English-Amis corpora. For English, counts without pivot data appear outside parentheses, while counts including pivot data appear inside parentheses. Token counts are all word-based. Note that the pivot data includes only a subset of the full Mandarin-Amis dataset.

that a large difference in length could indicate a potential misalignment or a translation that deviates considerably from the source material. We also implemented hard-coded detection mechanisms to identify noisy or incorrect translation pairs. For instance, in many cases in the ILRDF Video data where the translator failed to understand the original speech, the translation was labeled as "indistinct" or "no Chinese record". Such sentence pairs were removed from the corpus. Further processing was centered on preparing data for machine translation. We utilized the Moses library to perform spellchecking and to harmonize punctuation.

Due to the overlap between sources and the inclusion of multiple dialects of Amis in some of the sources, the corpus contained both duplicate translations and many-to-one translation mappings, where a single word or phrase was associated with multiple possible translations in the other language. Duplicate translation pairs were retained but exclusively allocated to the training data. This approach ensures that no sentence pairs appears in both the training and testing sets.

We note that we did not attempt to distinguish among the five dialects represented in the corpora. Given the very limited amount of data for training, we treated all Amis data as one language. We plan to address the complexities of dialectal variation in our future work.

3.4 Pivot data creation

While all of the Amis words, phrases, and sentences in the datasets had Mandarin translations, only a small percentage had English translations. To augment the much smaller Amis-English corpus, we used Mandarin as a pivot language to create new Amis-English pairs by translating the Mandarin side of the Amis-Mandarin pairs into English. For this task, we used the DeepL API[4], which offers a free tier of 1,000,000 characters per month. The Mandarin text from each corpus was submitted to the DeepL API in batches, specifying English as the target language and Mandarin as the source language. This process increased the size of the English-Amis corpus from 25,405 pairs to 89,458 pairs. While this was an efficient way to synthesize new training data, we did observe that the translations did not always faithfully render the general style or tone of the original Mandarin text.

[4]https://www.deepl.com/en/pro-api

Corpus	Language	Sentence Pairs (w/ pivot)	Tokens (w/ pivot)	Types (w/ pivot)
ILRDF Videos	ami	41,459	241,295	24,354
	cmn	41,459	395,066	3,303
Presidential Apology	ami	33	1,929	530
	cmn	33	3,536	580
ePark	ami	48,071	319,138	19,397
	cmn	48,071	543,948	3,039
Glosbe Amis	ami	5,860	91,201	4,242
	cmn	5,860	160,315	1,748
ILRDF Dictionary	ami	5,482	37,140	8,054
	cmn	5,482	61,383	2,462
Zheng Corpus	ami	15,022	48,764	11,994
	cmn	15,022	93,734	2,822
NTU Corpus	ami	742	7,718	1,282
	cmn	742	11,650	904
Fey Dictionary	ami	2,478	10,619	2,436
	cmn	2,478	17,706	1,924

Table 2: Corpus sentence pair, token, and type counts for the Mandarin-Amis corpora. Token counts for Mandarin are based on subword-unit token counts from the NLLB tokenizer, while token counts for Amis are word-based.

Architecture	NLLB		OpenNMT	
Eval Metric	**BLEU**	**chrf++**	**BLEU**	**chrf++**
Amis -> English without pivot data	11.35	25.50	14.68	28.14
Amis -> English with pivot data	14.55	29.59	20.44	32.74
English -> Amis without pivot data	10.78	37.13	N/A	N/A
English -> Amis with pivot data	10.38	37.10	8.34	31.87
Mandarin -> Amis	15.15	36.25	17.10	36.70
Amis -> Mandarin	23.83	26.64	28.10	31.70

Table 3: MT output evaluation across architectures (NLLB vs. OpenNMT) and training corpora (Amis-English with and without pivot data, Amis-Mandarin). N/A indicates that the model was apparently unable to learn, yielding output consisting entirely of <unk> tokens.

3.5 Data partitioning

The Amis-English corpus contained 25,405 without the pivot data and 89,458 sentence pairs including the pivot data. The Amis-Mandarin corpus contained 119,147 sentence pairs. We partitioned the datasets as follows. First all duplicate pairs were removed from each corpus. From the remaining sentences pairs, approximately 5% of the pairs from each corpus were selected to form the test set for that corpus. Duplicate sentences were added back to the corpus, and the remaining sentences pairs of each corpus formed the training data.

4 Method

While there are various approaches to MT in extreme low-resource settings, we focus on two popular approaches: an older but reliable end-to-end sequence-based MT architecture, OpenNTM (Klein et al., 2020), and fine-tuning with the multilingual No Language Left Behind (NLLB) architecture (Costa-jussà et al., 2024). For the OpenNMT training, we followed the most recent version of the OpenNMT tutorial.[5]. For NLLB, our starting

[5]https://github.com/ymoslem/OpenNMT-Tutorial

point was a notebook[6] originally used to fine-tune the NLLB-200 distilled 600M model to translate between the Turkic language Tyvan and Russian.

When using this NLLB framework, we initialized the tokenizer for Amis with a base configuration set for the only available related language, Tagalog, noted internally as `tgl_Latn`. We additionally added a new language token specific to Amis, identified within our system as `amis_Latn`. This required modifying the tokenizer's vocabulary to include Amis and adjusting the model's embeddings to accommodate this addition. The embedding for the new Amis token was initialized using the embeddings of the Tagalog language, leveraging the linguistic similarities to enhance the model's performance without extensive retraining from scratch. This process also involved repositioning certain tokenizer elements, such as the mask token, to maintain the tokenizer's integrity and functionality after the introduction of new language components.

Mandarin in Taiwan is written using traditional Chinese orthography. The NLLB tokenizer documentation indicates that the `zho_Hant` tag can accommodate both simplified Chinese and traditional Chinese orthography, but we found that only 60% of the traditional characters in our data were accounted for by the tokenizer. We addressed this problem by training a custom SentencePiece (Kudo and Richardson, 2018) tokenizer on our Mandarin data and then inserting the resulting missing tokens into the token set for the NLLB `zho_Hant` tokenizer.

As NLLB models are inherently bidirectional, we built three models: Amis-English with no pivot data; Amis-English with pivot data; and Amis-Mandarin. Within OpenNMT, whose basic design is for unidirectional training, we trained six models: Amis->English and English->Amis without pivot data; Amis->English and English->Amis with pivot data; Amis->Mandarin; and Mandarin->Amis. We evaluate the output of these models on our test data using two metrics: BLEU, the long-standing MT evaluation metric based on word n-gram precision, and chrf++ ("CHaRacter-level F-score"), which uses character-based, rather than word-based, n-grams.

[6]https://cointegrated.medium.com/how-to-fine-tune-a-nllb-200-model-for-translating-a-new-language-a37fc706b865

5 Results

Table 3 presents the BLEU and chrf++ scores for each model trained, in each direction, for each of the two MT architectures. We first consider the impact of including synthetic pivot data. We see that when translating from Amis to English, the inclusion of pivot data increases performance as measured by both metrics – at times rather dramatically – under both MT architectures. Strangely, including pivot data when translating to Amis does not yield improvements. With OpenNMT, the very small amount of unpivoted Amis-English data was insufficient to train a model. The addition of pivot data facilitates the production of actual output, but BLEU scores are weak.

We now turn to a comparison of NLLB with OpenNMT. When translating from Amis to English, OpenNMT outperforms NLLB by several points in terms of both BLEU and chrf++. When translating to Amis from English, NLLB outperforms OpenNMT, which fails to yield output at all for the condition with no pivot data. For the Amis-Mandarin models, OpenNMT again holds a small advantage over NLLB, with higher BLEU and chrf++ scores in both directions. Interestingly, translation to Amis using NLLB, whether from English or Mandarin, yields very high chrf++ scores. We conclude that both architectures show promise for translation in these low-resource settings, with each architecture outperforming the other under certain conditions.

Table 4 shows a few example outputs for OpenNMT and NLLB trained on the larger Amis-English dataset that included pivot data. Recall that OpenNMT yields higher BLEU and chrf++ scores than NLLB when translating to English. In all cases, the translations are reasonable. We see that OpenNMT appears more likely to produce verbatim matches for the reference or text with more shared unigrams. We also observe that the NLLB output sometimes includes words found in the Amis input, something that almost never happens in the OpenNMT output. Overall, while the BLEU and chrf++ scores reported for this translation direction are higher for OpenNMT, the NLLB translations often capture the gist of the reference without necessarily generating a verbatim match, which will negatively impact BLEU scores. The final example is typical of output produced from input sentences drawn from the Bible. These consistently yielded Biblical language unrelated to the

Amis	Reference	OpenNMT	NLLB
Matataruturud kita tu tayal	Let us pass on the work from one to the other	one by one of our work is continued	let us work together
Maulah kaku aci kaka ^aku a rumadiw	Both my brother and I like to sing	both my brother and i like to sing	i also like to sing with my brother and sister
Aciyah adihayay tu ku tayal isu	Ouch you have got a lot on your plate	wow you guys are a lot of work	wow i have got a lot of work
Adihay ku heci nu kilang nira	His fruits were plenty	he has many fruit	the fruits of its trees are numerous
Ahecid ku nanum nu liyal	The sea is salty	the sea is salty	the sea water is salty
Hay fangcal ku keru ^nu maku	Yes i dance well	yes i dance well	yes my dances are great
Ira ku rengus I umaumahan	There is grass in the field	is there any grass here	rengus is in the field
Matngiltu namu ku nanu sapi'met a limuut nu itiyaayhu a tamdaw tuya aka pipatay tu tamdaw o mipatayay a tamdaw i u mamasawkit sanay	You have heard that it was said to the people long ago, 'Do not murder, and anyone who murders will be subject to judgment.'	amen i say to you whatsoever you shall bind upon earth shall form and shall cast on enemy	and when jesus was come into the house of the ruler and saw the minstrels and the multitude making a rout

Table 4: Example Amis-English sentence pairs, along with the predicted output of OpenNMT and NLLB when trained on data including the pivot data.

actual content of the reference sentence for both architectures.

6 Conclusions

While accurate machine translation offers utility for supporting language documentation, training a robust model requires a large parallel corpus that would be difficult to acquire for most endangered and indigenous languages. In this paper we mined a wide variety of diverse existing corpora containing parallel data in order to produce MT-ready corpora for Amis-English and Amis-Mandarin translation. We were able to achieve promising BLEU and chrf++ scores under some conditions, but many questions remain about the utility of the pivot data for translation into Amis and about the performance contrasts between NLLB and OpenNMT.

In our future work, we plan to carry out data ablation studies to determine the individual contributions of each of the component corpora. In particular we suspect that the Bible data may not be appropriate given the style and content of the other corpora. Given the ethically and practically dubious value of Bible data in low-resource MT and other NLP tasks (Liu et al., 2021; Domingues et al., 2024), we may see improvements while recognizing the potentially harmful effects of using culturally irrelevant texts. We also would like to incorporate dictionary entries in a more effective way following (Zheng et al., 2024), who showed success across a large number of languages in the Formosan family.

Ethical considerations

When working with an indigenous language, it is necessary to ensure that the community to whom the language belongs is a willing and active participant in the research. While the data used in our project is freely available for download, we have taken extra steps to gain the explicit permission of the Indigenous Language Research and Development Foundation (ILRDF) and the managers of the ePark indigenous educational organization to use their Amis data. All of our models will be shared with these organizations and the Amis community.

Acknowledgments

The authors thank Yuyang Liu, Li-May Sung, and the Indigenous Languages Research and Development Foundation, especially Akiw, Lowking Nowbucyang, and Yuyang Liu, for generously providing data. This material is based upon work supported

by the National Science Foundation under Grant #2319296. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Aboriginal Language Research and Development Foundation. 2023a. Online dictionary of aboriginal languages. https://e-dictionary.ilrdf.org.tw.

Aboriginal Language Research and Development Foundation. 2023b. yuanzhumin yuyan leyuan (epark). https://web.klokah.tw/.

Steven Bird and David Chiang. 2012. Machine translation for language preservation. In *Proceedings of COLING 2012: Posters*, pages 125–134.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630:841–846.

Pedro Henrique Domingues, Claudio Santos Pinhanez, Paulo Cavalin, and Julio Nogima. 2024. Quantifying the ethical dilemma of using culturally toxic training data in ai tools for indigenous languages. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages@ LREC-COLING 2024*, pages 283–293.

David M Eberhard, Gary F Simons, and Charles D Fennig. 2024. *Ethnologue: languages of the world, 27th Edition*, volume 22. SIL International.

Virginia Fey. 1986. Amis dictionary. *Taipei: The Bible Society*.

Katharina Kann, Abteen Ebrahimi, Kristine Stenzel, and Alexis Palmer. 2022. Machine translation between high-resource languages in a language documentation setting. In *Proceedings of 1st Workshop on NLP applications to field linguistics*, page 26.

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.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *Preprint*, arXiv:1808.06226.

Paul Jen-kuei Li, Elizabeth Zeitoun, and Rik De Busser, editors. 2024. *Handbook of Formosan Languages: The Indigenous Languages of Taiwan*. Brill's Handbooks in Linguistics. Brill.

Ling Liu, Zach Ryan, and Mans Hulden. 2021. The usefulness of bibles in low-resource machine translation. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1, pages 44–50.

So Miyagawa. 2023. Machine translation for highly low-resource language: A case study of ainu, a critically endangered indigenous language in northern japan. In *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages*, pages 120–124.

John E Ortega, Richard Castro Mamani, and Kyunghyun Cho. 2020. Neural machine translation with a polysynthetic low resource language. *Machine Translation*, 34(4):325–346.

Claudio Pinhanez, Paulo Cavalin, Luciana Storto, Thomas Fimbow, Alexander Cobbinah, Julio Nogima, Marisa Vasconcelos, Pedro Domingues, Priscila de Souza Mizukami, Nicole Grell, et al. 2024. Harnessing the power of artificial intelligence to vitalize endangered indigenous languages: Technologies and experiences. *arXiv preprint arXiv:2407.12620*.

Lily I-wen Su, Li-May Sung, Shuping Huang, Fuhui Hsieh, and Zhemin Lin. 2008. Ntu corpus of formosan languages: A state-of-the-art report. *Corpus Linguistics and Linguistic Theory*, 4(2):291–294.

Daan van Esch, Ben Foley, and Nay San. 2019. Future directions in technological support for language documentation. In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 14–22.

Shiyue Zhang, Benjamin Frey, and Mohit Bansal. 2020. Chren: Cherokee-english machine translation for endangered language revitalization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 577–595.

Francis Zheng, Edison Marrese-Taylor, and Yutaka Matsuo. 2022. A parallel corpus and dictionary for amis-mandarin translation. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 79–84.

Francis Zheng, Edison Marrese-Taylor, and Yutaka Matsuo. 2024. Improving low-resource machine translation for formosan languages using bilingual lexical resources. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11248–11259.

Comparing efficacy of IPA vs Pinyin romanisation transcriptions for complex tonal languages: A case study in Baima

Katia Chirkova
CNRS-CRLAO
katia.chirkova@gmail.com

Rolando Coto-Solano
Dartmouth College
rolando.a.coto.solano@dartmouth.edu

Rachael Griffiths
EPHE-PSL
rachael.griffiths@ephe.psl.eu

Marieke Meelen
University of Cambridge
mm986@cam.ac.uk

Abstract

How is automated tone transcription affected by the choice of transcription orthography? In this paper we present a range of experiments that indicate that, even when the tonal representations are kept the same, the way vowels and consonants are transcribed can affect tonal character outputs. Our results also indicate that using a Language Model (LM) for decoding can mitigate problems with tonal outputs, but tones remain the most difficult part of the transcription. In doing this we also present the first Automatic Speech Recognition (ASR) models for the Baima language, spoken in Sichuan and Gansu, China. We hope to use these models to contribute to ongoing documentation efforts.

1 Introduction

Researchers who start documenting endangered languages without writing systems often face the challenge of a race against the clock to collect and transcribe as much data as possible before the language disappears. With extremely limited access to native speakers who are not only essential when gathering, but also when transcribing and interpreting data, linguists and community members interested in preserving the language have to make crucial choices on how to spend limited time with informants. Is it worth the tremendous amount of time and effort to preserve every detail using the International Phonetic Alphabet (IPA) to facilitate further research in the sound system of the language? Or should they choose a local and/or romanised script to speed up transcription and to increase the possibilities of language revitalisation?

In this paper we present several ASR experiments to gain further insight into these important practical questions, focusing on the Baima language, spoken in Sichuan and Gansu, China. With three native tones, tone sandhi and tonal borrowings as well as complex consonantal onsets and epiglottalisation, this language forms the perfect case to test the trade-off of different transcription systems. In addition to tests with different base models, LMs and transcription systems, we will also do an in-depth error analysis of each of the tones to gain insight into which are more challenging for specific models. The results will therefore not only further work on ASR for tonal languages but also help researchers and speaker communities working on language documentation and revitalisation to choose how to best spend limited time and resources in order to get the best possible results.

1.1 Baima Language

Baima (/pêkê/, Chinese 白马语 *báimǎyǔ*, ISO-639 code *bqh*) is a Tibeto-Burman (Tibetic) language spoken at the border of Sichuan and Gansu provinces in China. The language has approximately 10,000 speakers, who reside in the counties of Pingwu, Songpan (Tib. *Zung chu*), and Jiuzhaigou (Tib. *Gzi rtsa sde dgu*) in Sichuan, and in the counties of Wenxian and Zhouqu (Tib. *'Brug chu*) in Gansu.

The area of distribution of the Baima language lies at the historical Sino-Tibetan border, in a multiethnic and multilingual region. In all counties of its present distribution, immediate linguistic neighbours of Baima include varieties of Mandarin (mostly Southwestern Mandarin) and Tibetic languages. To our knowledge, there are no longer any monolingual speakers of Baima, as all age groups are bilingual in the local varieties of Mandarin. Mandarin (both the local varieties and the closely related Standard Mandarin, the official language of the People's Republic of China) also dominates the education system and work in public domains. Baima is not used in writing or education and its use is mostly restricted to family and community events. For those reasons, it is severely endangered.[1]

[1] https://www.ethnologue.com/language/bqh/

Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 170–181

Baima is little-studied. To date, most linguistic fieldwork on this language has concentrated on the Baima variety as spoken in Baima Township of Pingwu County, which is also the focus of the present study (Huang and Zhang, 1995; Chirkova, 2017; Sun et al., 2007). A small set of audio-visual non-annotated recordings of Pingwu Baima is available on the Pangloss archive of the Centre national de la recherche scientifique (CNRS).[2] Speakers of Baima are keen to preserve their language and cultural traditions and would greatly benefit from the development of tools that can facilitate this effort.

Baima is remarkable for its phonological complexity, and for a number of features that are typologically uncommon. These include non-modal phonation type contrasts in both consonants and vowels and a tonal system characterised by syllable-level contrasts, with redundant use of pitch, voice quality, and vowel length. The Baima consonant inventory consists of 57 phonemes, including 11 epiglottalised prenasalised, nasal, and approximant phonemes. The vowel inventory consists of 11 monophthongs, three native diphthongs, and one diphthong that only occurs in loanwords from Mandarin (/ua̲/). The three contrastive tonal categories are high falling (53), mid (44), and low (213). The high falling tone is correlated with a high falling pitch contour, tense vowel quality, and short vowel duration. The mid and low tone categories are correlated with long vowel duration. The mid tone has a mid level pitch contour and a modal voice quality. The low tonal category has a low falling-rising pitch contour and a breathy-like or lax voice quality. Detailed phonological analyses can be found in Chirkova et al. (2023) and Chirkova (2025), and examples of the tones are given in Table 1.

Baima Tones		
Category	**Example**	**Meaning**
1. Contrastive tones in the native lexicon		
High falling	no^{53}	inside
Mid level	no^{44}	sky, heaven
Low rising	no^{213}	exist, have
2. Tone sandhi		
Compound change	$no^{31}mba^{53}$	possessions
3. Tones in Chinese loans		
High level	t^ha^{55}	he/she/it
Mid rising	$ts̪^hɑ^{35}$	examine

Table 1: Tones in Baima in IPA

1.2 ASR for No-Resource Tonal Languages

As there are no NLP efforts, corpora, dictionaries or other resources available for Baima, we have to resort to techniques to address the well-known transcription bottleneck for extremely low- (or no-)resource languages. Recent work by (Stoian et al., 2020; Prud'hommeaux et al., 2021; Coto-Solano et al., 2022) and others show getting good ASR results in those challenging situations is possible by relying on pre-trained models of acoustic data from other, high-resource languages. In addition, some techniques involve transfer learning or modification of the acoustic signal (Mitra et al., 2012; Meelen et al., 2024), data augmentation with written sources such as dictionaries and word lists (Hjortnaes et al., 2020; Arkhangelskiy, 2021).

Languages like Baima with complex phoneme inventories and tones are generally more challenging for any ASR model, especially when data and resources are scarce or non-existent. For ASR systems that evaluate the Character Error Rate (CER), it is therefore important to think carefully about the transcription method, as CER has been shown to strongly correlate with orthographic complexity (Taguchi and Chiang, 2024). Representations where the tone is marked explicitly but kept separate from the vowel (i.e. explicit tone recognition, as discussed by Lee et al. (2002)) are not often used for larger languages, but they are common in low-resource ASR systems, such as those for Yongning Na from China and Eastern Chatino from Mexico (Adams et al., 2018). Coto-Solano (2021) shows that manipulating the transcription input can improve results in a language like Bribri, where not marking the most common tone and separating the tonal markings from the vowel can lead to major improvements in performance. Bribri has only four tones, however, transcribed with a limited number of additional segments, and only when necessary. Baima, on the other hand has three native tones, tone sandhi as well as additional tones on loanwords for Chinese. Following sinological tradition, those are all represented with Chao tone numbers (Chao, 1930), which means they consist of at least 2-3 additional characters on every tonal syllable. The current use of complex tone notation in Baima is in line with the research tradition that characterises Baima as a tonal language defined by pitch, favouring Chao tone numbers over IPA diacritics for tone representation (see Section 2.2). The fact

[2] https://pangloss.cnrs.fr/corpus/Baima?lang=en&mode=pro

that Baima tones are produced with both a particular f0 specification and a voice quality specification has only been recently discovered (Chirkova et al., 2023; Chirkova, 2025).

In this paper we therefore focus on transcription systems and how they might impact the automatic transcription of complex tones, testing different base models as well as the usefulness of adding an LM in an extremely low/no-resource context.

2 Methodology

2.1 Data collection

The data for Baima used in the present study was collected during two fieldtrips in November-December 2003 and October-November 2004 in several villages in the Baima Township of Pingwu County. We collected ca. 20 hours of traditional narratives, interviews, and descriptions of traditional activities. 191 minutes (4 hours 5 minutes) are fully transcribed. All of the transcribed materials are from recordings of traditional narratives from three native speakers (all male, between 50+ and 70+ years old at the time of recording).

To enhance efficiency of fine-tuning the base models and to avoid potential confounds in the results due to differences in segment length, we excluded segments longer than 15 seconds, reducing the dataset to 186 minutes. The total corpus contains 27,417 words (2715 unique words).

2.2 Transcription methods

The original transcriptions of recordings in the Baima language were done in IPA capturing all phonetic details of the language, including nasalisation, epiglottalisation and tones. While nasalisation is not phonemic, there is variation between different speakers. Epiglottalisation and tones are phonemic, however, and the latter are indicated with Chao tone numbers in our transcriptions.

The Pinyin-style transcription was created with the primary objective of being comprehensible to native speakers of Baima. It is rooted in the Hanyu Pinyin system, the official romanization system in China (Committee, 1956). The choice to establish a romanization system for the Baima language on the basis of the Hanyu Pinyin system was influenced by two crucial factors: (i) its widespread familiarity, which is a result of its extensive usage in elementary school education and public life, and (ii) its ease of adaptation for electronic applications, particularly mobile phones.

Over the past few decades, the Hanyu Pinyin system has served as the foundation for romanization systems of numerous minority languages in China, including large languages such as Nuosu (see (Ma et al., 2008)). It has also been instrumental in our own work on the Duoxu and Ersu languages (Chirkova and Han (2016); Chirkova and Wang (2017); Wang et al. (2019)). It is worth noting that the issue of tone notation in the Hanyu Pinyin system is intricate. The official system employs diacritics to represent the four tones of Standard Chinese. Nevertheless, these diacritics are often disregarded in various contexts, such as when spelling Chinese names. Alternatively, tones can be indicated by placing a tone number (1 to 4) at the end of each individual syllable.

In essence, transcribing tone remains a challenging aspect for speakers of tonal languages, such as Mandarin Chinese speakers and those whose languages we developed romanization systems for in the past. Therefore, it is crucial to engage in careful consultation with potential users of the system to address the issue of tone representation. We chose Chao tone numbers for several reasons. First, the complexity of the tonal system of Baima has only recently begun to be unravelled. While recent research has provided a better understanding of contrastive tones on monosyllabic words, tone sandhi in polysyllabic words remains largely understudied. Consequently, Chao tone numbers offer the most accurate and reliable method for noting tone variation before a comprehensive understanding of the tonal system is achieved. Secondly, the tradition of using Chao tone numbers in IPA transcription is deeply rooted in the field. The vast majority of publications on Baima, including the only reference grammar with the most comprehensive vocabulary list to date (Sun et al., 2007), rely on this system. Therefore, Chao tone numbers provide convenience for cross-reference and comparison between our work and previous descriptions of that language.

To facilitate testing of different transcription systems, we wrote one-way conversion scripts to create Pinyin and Simple romanisation equivalents of the detailed IPA transcriptions with tones.[3]. These conversions can only be done from IPA, as certain details are simplified in both alternative transcription systems. Table 2 shows examples for each

[3]Code and models can be found on https://github.com/rolandocoto/baima-asr

Transcription	With tones	No tones
IPA	ɲə^{53}	-
Pinyin	nyii^{53}	nyii
Simple	nyə^{53}	nyə

Table 2: Possible transcriptions for [ɲə^{53}] 'man'

transcription type. All transcriptions have the same representation for tone: two numerical characters for contours (e.g. ^{53} for the falling tone) and three characters for the dipping tone (i.e. ^{213} dipping).

2.3 ASR Training

Our next step was to create ASR models for the Baima language. We carried out monolingual fine-tuning using the Baima data, and we chose three base models for this[4]: Wav2Vec2 XLSR-53 Large (Baevski et al., 2020), henceforth Wav2Vec2; MMS 1b-all (Pratap et al., 2024), henceforth referred to as MMS; and Whisper Medium (Radford et al., 2023). For Wav2Vec2 and MMS we tried versions of the models with and without an LM for decoding. We used KenLM (Heafield, 2011) to produce the LMs, which were trained using the text in the training and validation partitions of the data.

In order to train the models, we took the total 186 minutes of data and created 20 randomly ordered versions of it. We split these 20 versions into train/dev/test sets, with ratios of 80%, 10% and 10%. We used these partitions to train the models, and the checkpoint before overfitting was saved. These were used to evaluate the test sets, and from there calculate the median CER and Word Error Rate (WER) for each test set. In section 3 we report the average values of the median error for each randomly assigned test set.[5] The total sample only has three speakers, so the speakers in the train/valid sets are also present in the test set.

2.4 Calculation of Tonal Errors

In addition to reporting the standard CER and WER, we also calculated the metrics of tonal character error rate (tonal CER) and tonal word error rate (WER). Table 3 shows an example of this process. Let's assume we have the phrase [ɲə^{53} te^{53}] 'that man' as a human-transcribed phrase in the test set. It is transcribed [ɲə^{53} te^{53}] in IPA with Chao tone numbers. Let's then assume that one of the ASR systems produces the wrong automatic transcription [ɲə^{53} te^{44}]. Here, the falling (53) tone of the first word is correct, but the tone of the second word is incorrectly tagged as a mid level (44) tone. We then strip both phrases of their consonants and vowels, leaving only the tones. This would result in `53 53` for the human transcription, and `53 44` for the incorrect automatic transcription. The next step is to convert the tones into single units, to avoid counting the start and end points of the falling contour tone (e.g. 5,3) as separate errors. When we do this, the transcriptions could take the form `F F` for the human transcription, and `F H` for the erroneous automated transcription. It is at this point that we can calculate the distance between the human and automated transcriptions, using the standard CER and WER algorithms. The *tonal CER* is the percentage of characters in this transcription that are wrong. The *tonal WER* is the percentage of words that have a tonal error in them. These tonal WER and CER will be reported for the transcriptions that do have tone (i.e. IPA, Pinyin, Simple).

	Source	Hypothesis
1. Get hypothesis	[ɲə^{53} te^{53}]	[ɲə^{53} te^{44}]
2. Get only tones	53 53	53 44
3. One unit per tone	F F	F H
4. Calculate error	tCER=50, tWER=50	

Table 3: Example of the calculation of tonal CER and WER for the human-transcribed phrase [ɲə^{53} te^{53}] 'that man' and a potential automatic (and partially wrong) transcription of the phrase.

3 Results

3.1 ASR Training

First, we performed a simple comparison of the base models to determine which had the best performance **without** the use of an LM. We restricted this test to the IPA transcription. When we compare the three base models (Wav2Vec2, MMS and Whisper), **Wav2Vec2** had the lowest character error (CER=18.3±1.1), compared to Whisper (CER=19.3±1.8) and MMS (CER=25.1±0.7), but Whisper has the lowest word error rate (WER=33.6±2.0), com-

[4]Detailed hyperparameters can be found in Appendix B.

[5]This paper reports the averages for 20 sets of Wav2Vec2 regular models (IPA, Pinyin, Simple), 20 sets of the Wav2Vec2 no-tone models (Pinyin, Simple), 20 MMS IPA models, and 5 Whisper IPA models. Each Wav2Vec2 model took approx. 93 mins to train and test; each MMS model took approx. 105 mins, and the Whisper models approx. 8.5 hrs. The results reported here needed a total of 155 hrs of an Nvidia A100 Tensor 80GB PCIe GPU and 4 CPU cores in an HPC environment (W2V2), as well as 78 hrs of an Nvidia L4 GPU with 8 CPUs in a cloud-based environment (MMS+Whisper).

pared to Wav2Vec2 (WER=47.5±2.7) and MMS (WER=69.5±2.8). Figures 1 and 2 show the summary of the results for the Wav2Vec2 models, figures 3 and 4 show a comparison between Wav2Vec2 and MMS, two base models which have the same architecture, but which differ on the number of languages used during the training phase. Based on these figures, we will answer questions about the interaction of the transcription style, the LM, and the presence of tones in the transcription.

Figure 1: Character and tonal character error for models trained with Wav2Vec2.

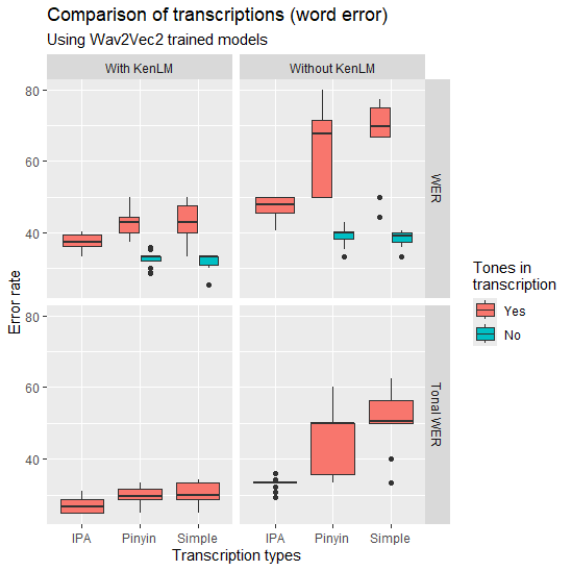


Figure 2: Word and tonal word error for models trained with Wav2Vec2.

Our next question was: **Does using an LM make a difference to the transcription?** The answer is **yes**; using a KenLM-style LM decreases the error rate. In this section we used paired Wilcoxon signed rank tests or paired t-tests to test significance, depending on whether the distributions met the assumption of normality or not, as determined by a Shapiro-Wilk test. The use of an LM reduces CER by 2.6±1.9 points (V=5050, p<0.0001) and WER by 13.9±8.8 points (V=5050, p<0.0001). The use of an LM also reduces tonal error: The tonal CER goes down by 8.8±5.2 points (V=1830, p<0.0001), and the tonal WER goes down by 14.1±8.1 points (V=1830, p<0.0001).

The third question is: **Does the amount of languages in the base model make a difference?** The answer is **yes**, but adding more languages does not seem to lead to an improvement in performance. Since Wav2Vec2 and MMS are based on the same architecture, but trained on a different number of languages (53 for Wav2Vec2 and 1162 for MMS), we decided to test this question. The answer was the opposite of what could be expected. The smaller model, Wav2Vec2, performed better. Its CER was lower by 5.5±1.6 points (V=820, p<0.0001), and its WER was lower by 15.3±7.4 points (V=820, p<0.0001). Wav2Vec2 also had a lower tonal error: the tonal CER was lower by 9.5±6.2 points (V=820, p<0.0001), and the tonal WER was lower by 14.8±8.0 points (V=820, p<0.0001). It seems that the additional languages in MMS did not aid in the transcription of Baima. Therefore, from this point on we will restrict the following tests to the Wav2Vec2 models.

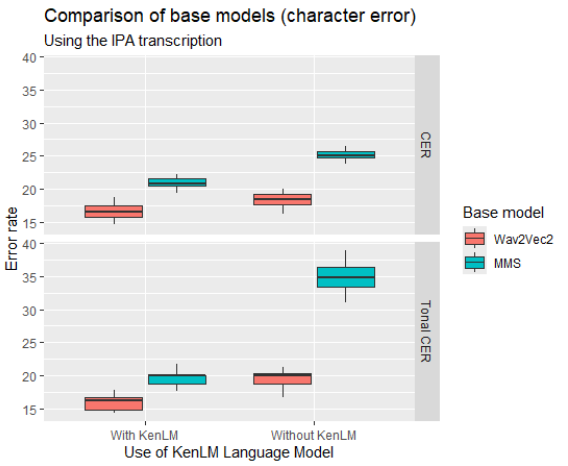


Figure 3: Comparison of Wav2Vec2 and MMS models for character and tonal character error.

The fourth question is: **Does adding tones**

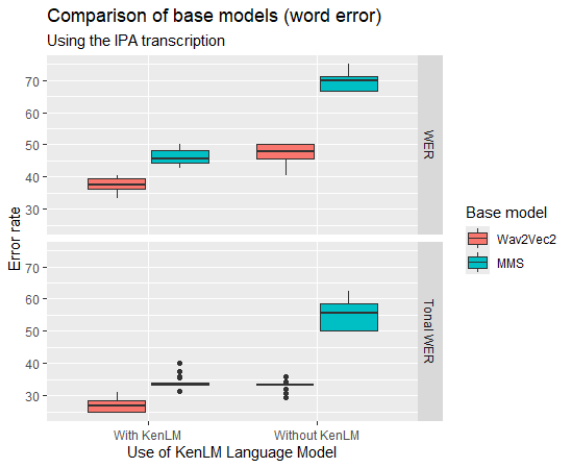


Figure 4: Comparison of Wav2Vec2 and MMS models for word and tonal word error.

to the transcription make transcribing Baima harder? Yes; the tones do take a toll on the transcription. When we studied the difference between the "tone" and "NoTone" versions of Pinyin and Simple romanisations, using tone increases the error rate. It increases CER by 6.9±3.6 points (t(79)=17, p<0.0001) and WER by 18.8±11.4 points (V=3240, p<0.0001).

Finally, **does the transcription style make a difference in the error rates? Yes**, the Pinyin transcription style leads to more errors overall, as well as more tonal errors, even if the tones themselves are the same in all transcription styles. Table 5 shows the (total) CER and WER, as well as the tonal CER and WER. (The consonants and vowels will be discussed in section 3.3). To test this we used the Kruskall-Wallis rank sum test to compare the error means between the three types of transcriptions (IPA, Pinyin, Simple). When the KenLM model is NOT used, there is a significant difference between transcriptions for the four metrics used. For example, IPA has CER=18, Pinyin CER=23 and Simple CER=24 ($\chi^2(2)$=19, p<0.0001). The difference is more pronounced for the word error; IPA has WER=48, Pinyin WER=65 and Simple WER=67 ($\chi^2(2)$=30, p<0.0001). This difference is attenuated by the use of the KenLM, but IPA still performs significantly better. As for the character error, IPA has CER=17, Pinyin CER=19 and Simple CER=20 ($\chi^2(2)$=19, p<0.0001). As for the word error, IPA has WER=37 and both Pinyin and Simple have WER=43 ($\chi^2(2)$=22, p<0.0001).

When we study the tonal errors without an LM, IPA is again the transcription with the least error. For tonal CER, IPA has tonal CER=20, compared to Pinyin tonal CER=28 and Simple tonal CER=31 ($\chi^2(2)$=28, p<0.0001). For tonal WER, IPA has tonal WER=33, compared to Pinyin tonal WER=46 and Simple tonal WER=50 ($\chi^2(2)$=32, p<0.0001). These differences are, again, attenuated by the use of an LM. In the case of character error, IPA has tonal CER=16, compared to Pinyin tonal CER=18 and Simple tonal CER=19 ($\chi^2(2)$=23, p<0.0001). For tonal WER, IPA has tonal WER=27, compared to Pinyin and Simple tonal WER=30 ($\chi^2(2)$=19, p<0.0001).

3.2 Tonal Errors

An additional question for this paper is: **Are there any tones that perform worse than others?** Table 4 shows the percentage of error for specific tones. It shows the average (across 20 models) of the percentage of all the occurrences of a certain tone (e.g. 53) that were predicted erroneously (e.g. 11% for 53, for Wav2Vec2+KenLM using IPA).

In order to understand the patterns in the table, we used an ANOVA test with the percentage of error as the dependent variable, and four independent variables and their interactions: (i) tone {53, 44, 213, 31, 35 and 55}, (ii) transcription style {IPA, Pinyin, Simple}, (iii) use or not of a KenLM LM, and (iv) base model (Wav2Vec2 vs MMS). There is a significant three-way interaction between tone, transcription and base model (F(5,912)=11.9, p<0.0001). In general the Baima and Sandhi tones have less error than the borrowed tones. The use of an LM decreases the error. On average, tones have an error of 47±32% when using KenLM, compared to 55±30% without it. As for the type of model, MMS transcriptions have more errors in general, but this depends on the tone: Wav2Vec2 and MMS have almost identical error rates for tone 53 (both of them 16%), but they have very different error rates for tone 44 (57% versus 40%).

Perhaps the most relevant for the present is the interaction between tone and transcription. There are tones that have much lower error rates than others. As can be seen in table 4, the dipping tone 213 has a lower error rate in the IPA transcription (24% for Wav2Vec2 with KenLM) than in the Pinyin and Simple transcriptions (29% and 30% respectively). This pattern is different from the falling tone 53, which has almost identical error rates across all transcriptions (approx. 12% when using KenLM). Tone 44 also shows large differences between transcriptions, whereas the tones in borrowed words,

			Baima tones				Sandhi tone	Borrowed tones	
	Model	**Transcription**	**213**	**53**	**44**	**31**	**35**	**55**
With LM	MMS	IPA	34	11	46	29	97	99
	Wav2Vec2	IPA	24	11	30	21	75	78
		Pinyin	29	12	36	26	89	96
		Simple	30	13	38	28	91	97
Without LM	MMS	IPA	59	21	68	36	100	100
	Wav2Vec2	IPA	28	14	34	23	80	83
		Pinyin	34	21	50	35	94	98
		Simple	40	26	55	36	96	100

Table 4: Percentage of errors for each tone by transcription, base model, and use of a KenLM LM. The underlined number is the largest error amongst the different transcription models per tone.

35 and 55, are almost identical (i.e. equally poor) for all transcription styles. These patterns will be further discussed in section 4.

3.3 Tonal versus consonant and vowel errors

We ask one final question which is important for anyone working in the documentation of a tonal language: **Are tones more difficult to transcribe than other parts of the phonology, like the consonants and the vowels? They are**, but mainly when an LM is NOT used. Table 5 shows the CER and WER when only the tones, consonants and vowels were considered. Using a technique similar to that described in section 2.4, we made versions of the transcriptions that had only the consonants and the vowels. For example, [ɲə^{53} te^{53}] 'that man' would be *ɲ t* in the IPA consonant transcription, and *ə e* in the IPA vowel transcription.

We used two ANOVA models, one for the character errors, and one for the word errors. Each of these had the percentage of error as the dependent variable, and three independent variables: transcription (IPA, Pinyin, Simple; all of them with tones), type of phone (Tone, Consonant, Vowel) and use or not of a KenLM LM. The CER model had a significant three-way interaction ($F(4,342)=3.6$, $p<0.01$), and the WER model had significant two-way interactions.

In the case of the LM, the CER shows a pattern where the use of a KenLM LM reduces the error, but it reduces it more for tones than for the other segments. This is also true for the WER ($F(2,342)=5.2$, $p<0.01$), where tones improved an average of 9 points, but consonants and vowels only improve by an average of 3 points.

In the case of the transcriptions, the use of KenLM led to a bigger improvement in the Pinyin and Simple transcriptions. This pattern is also true for the WER; where the Pinyin KenLM transcriptions improve by an average of 14 points and the Simple improve by an average of 15.7 points, compared to 5.3 for ($F(2,342)=32$, $p<0.0001$).

The main difference between CER and WER is in the way they interact with the transcriptions. The tones always have a larger CER when an LM is not used, and they always have amongst the highest CERs even if an LM is used. However, in the case of the WER, the tones are always the worst performers when an LM is absent, but the consonants and vowels behave slightly worse than the tones when an LM is present ($F(4,342)=3.0$, $p<0.05$).

4 Qualitative Error Analysis

In this section we shift our focus to the specific errors that the models make when transcribing, and how those might affect linguistic work.

4.1 Specific errors

Table 6 provides specific examples of transcription output. Further examples, including for the contrast between transcription systems are available in the Appendix. Examples (1) and (2) show the difference between the base models (without LM) and Wav2Vec2 with and without LM. It is clear that without an LM both base models, but especially MMS, struggle to get the right word boundaries for words that are acoustically merged together, like the copula [re^{213}] and the following question marker [a]. The target transcriptions actually give the original (lexical) tones of the two morphemes, whereas the models provide the actual pronunciation: a fused syllable with the overlaid interrogative intonation, which is closer to actual acoustic signal. Both models also appear to make errors at the end of the segment in (1). The acous-

		CER				WER			
	Transcription	**Total**	**Tone**	**Cons**	**Vowel**	**Total**	**Tone**	**Cons**	**Vowel**
With LM	IPA	17	16	13	16	37	27	25	28
	Pinyin	19	18	18	17	43	30	32	32
	Simple	20	19	17	19	43	30	34	31
Without LM	IPA	18	20	15	18	48	33	31	32
	Pinyin	23	28	22	21	65	46	45	45
	Simple	24	31	21	23	67	50	48	44

Table 5: Error for each type of character (tones, Cons=consonants, vowels) for Wav2Vec2 models. The underlined number is the largest error amongst the three types of characters.

	Different base models with IPA transcription (Without LM)		
1. SPX-bqh-018-193	"[He] asked (literally: said): "Is it the herdsman's horse or this young wanderer's horse.""		
Target transcription	ta^{53} $ndzʊ^{213}$ $s^he^{31}pu^{53}$ ta^{53} re^{213} a $tɕ^ho^{31}mba^{53}$ $go^{31}dʑy^{53}$ ta^{53} re^{213} te^{53} $dzɛ^{213}$ ʃə	CER	WER
MMS prediction	ta^{53} $ndzo^{53}$ $se^{31}pu^{53}$ ta^{53} ra^{3} $tɕ^ho^{31}mba^{53}$ $ŋgo^{31}ʑy^{53}$ ta^{53} re^{2} ə	27	67
Wav2Vec2 prediction	ta^{53} $ndzʊ^{213}$ $s^he^{31}pu^{53}$ ta^{53} re^{213} $tɕ^ho^{31}mba^{53}$ $ŋgo^{31}dʑy^{53}$ ta^{53} $re^{213}z^{2}$	14	42
	Wav2Vec2 IPA transcription Without vs With LM		
2. SPX-bqh-020-053	"When the two of them were hunting, [they accidentally] fired an arrow into a tree, and that tree turned into a young man [= a tree brother appeared], then they... "		
Target transcription	$ɲʂi^{53}$ $ŋgɛ^{53}$ $ndɛ^{53}$ sə $\tilde{o}^{213}$ $ʃ^he^{213}$ $kɛ^{53}$ nda^{53} $dʑʉ^{53}$ $ɕɛ^{44}$ $ʃ^he^{213}$ $ɲa^{31}ɲu^{53}$ ly^{213} ue^{44} $ɲi$ to^{44} $ʧo^{31}rə^{53}$	CER	WER
Without LM prediction	$ɲʂi^{53}$ $ŋgɛ^{53}$ $ndɛ^{53}$ ʃə $\tilde{o}^{213}$ $ʃ^he^{213}$ $kɛ^{53}$ nda^{53} $dʑʉ^{53}$ sə $ʃ^he^{213}$ $ɲa^{31}ɲu^{53}$ ly^{213} ue^{44} $ɲi$ to^{44} $ʧo^{31}rʊ^{53}$	6	18
With LM prediction	$ɲʂi^{53}$ $ŋgɛ^{53}$ $ndɛ^{53}$ ʃə $\tilde{o}^{213}$ $ʃ^he^{213}$ $kɛ^{53}$ nda^{53} $dʑʉ^{53}$ $ɕɛ$ $ʃ^he^{213}$ $ɲa^{31}ɲu^{53}$ ly^{213} ue^{44} $ɲi$ to^{44} $ʧo^{31}rə^{53}$	4	12
	Perfect CER and WER (even in detailed IPA with tone)		
3. SPX-bqh-011-121	"[You] need to go to my place, so [the emperor] said."	CER	WER
Target & Prediction	$k^hʉ^{53}$ tsa^{44} $ndʑi^{53}$ go^{53} re^{213} $ndzu^{53}$ $dzɛ^{213}$ ʃə	0	0
	Bad CER/WER most challenging IPA and 'easiest' Simple NoTone transcriptions		
4. SPX-bqh-002-277	"The big sister looked around, looked up, looked sideways, [then she] returned home, shook her head and said, there's nothing there."		
IPA			
Target transcription	pu^{44} $ʧ^he^{213}$ $ŋgo^{31}kɛ^{31}$ $tʂa^{53}$ $tyʉ^{44}$ mbo $tɕe^{53}$ $tyʉ^{44}$ $ndzɐ^{44}$ $tyʉ^{44}$ $ɕi^{53}$ tse^{53} a^{31} $ã^{53}$ $tʃo^{53}$ mu^{31}=no^{213}	CER	WER
With LM prediction	pu^{44} $ʧ^he^{213}$ te^{53} $ŋgo^{213}$ $ke^{31}ʧɑ^{53}$ te^{53} kumbo $tɕɛ^{213}$ te^{53} $ndzɐ^{53}$ $ɕi^{53}$ $ɑ^{213}ɑ^{213}$ $tʃo^{53}$ mu^{31} no^{213} $ndzu^{53}$ $dzɛ^{213}$ ʃə	56	87
Simple NoTone			
Target transcription	pu tsyhe nggookëtra tyue mboo tsyë tyue ndrqe tyue syi tse aã tsyoo mu noo	CER	WER
With LM prediction	pu tsyhë nyi ngoo ketsya te khu mboo tsyë te ndu aa tsyoo mu noo ndrqu dzë syə	58	93
Without LM prediction	pu tsyhë nyə ngoo ketsya te khumboo tsyë te nduë i aa tsyoo mu noo ndrqu dzë syə	58	93

Table 6: ASR results from various experiments for Baima - Part I: Base and Language Models

tic signal is actually deprecated here, showing the real benefit of adding an LM that can add words in often-seen contexts even if they are barely audible in the recordings. Finally, the MMS base model in particular seems to struggle with clusters at the start of syllable like [ʃ], [dz] and [dʑ].

Examples (3) and (4) illustrate that the models have outliers too, yielding both very good examples (in 3) or seemingly very bad examples (in 4), judging by the error rates. While the recording for (3) is rather short and clear, the articulation of the speaker uttering (4) is much less clear. The suggestion of the model to transcribe a particularly unclear part of the segment as [te^{53} kumbo] is actually probably more plausible than what the original transcriber first proposed. Furthermore, the final part of the utterance is completely 'swallowed' in the recording, but the model still proposes a very good transcription for those final words. Overall, zooming in on specific errors shows that even when results look very bad when simply calculating the CER and WER, in reality the models may actually be more useful than originally thought.

4.2 Tonal error analysis

Generally, the models for transcription types without the tones perform better. This could be due to the fact that it is genuinely 'easier' to ignore suprasegmental features like tones, and because the Chao tone numbers simply add further characters

to the target inventory, especially when they are counted as separate characters. When it comes to tonal errors we can make one clear observation from these qualitative data: some errors are due to the fact that transcriptions only note etymological tones and disregard sentence-level stress, that is, distinctive pitch contours that serve to mark words 'in focus' position and overlay the etymological tone of the word in focus.

As for specific tones, out of the six different options the 53 tone is the easiest to recognise, probably due to its high frequency, whereas the high tone 55, which only occurs on a handful of Chinese borrowings proves the most challenging.[6]

4.3 General errors

In general, based on the examples above, the main reasons for discrepancies between transcriptions and predictions are easily explained. For example, weakening in unstressed position can lead to the models predicting schwas, which is no doubt a frequent occurrence in any base model. Mainly, however, we note that all models suffer significantly from bad quality of the recording: background noise, unclear articulation, etc. lead to an increase in both CER and WER. However, when these increases are there because of incomplete or inexact original transcriptions, we also see that the models (especially those enriched with a Baima-specific KenLM LM) can actually yield transcriptions that are even better than the original.

5 Conclusion

In this paper we tested tonal accuracy and the effect of transcription type, base model as well as the option of adding a KenLM LM to the ASR pipeline of the Baima language, which has phonological features, including six tones, and is extremely limited in resources.

First, we found that more languages in a similar architecture for the base model (i.e. MMS vs Wav2Vec2) does not lead to better outcomes when transcribing smaller languages, perhaps because the extra languages are not phonologically similar to Baima. Wav2Vec2 has 5 tonal languages (Mandarin Chinese, Hakka, Cantonese, Lao and Zulu). MMS has these, plus many others, including small Indigenous languages with a wealth of tones. However, maybe the specific typology of the tonal system in Baima (where tones are consistently produced with both a particular f0 specification and a voice quality specification) poses a problem for the model. We furthermore showed that complex tones remain the most difficult part of the phonology to transcribe, despite the complexity of Baima vowel phonology. However, adding an LM to the decoding process can help to mitigate this problem.

Overall, non-tonal romanised transcriptions trained with a Wav2Vec2 base model and enhanced with a KenLM LM show the best results, but even detailed IPA models with Chao-numbered tones perform reasonably well, considering the very small amount of input data (186 mins). While it remains essential to reliably document and describe the sound system of the language using detailed IPA, it may at times be preferable to use a simplified romanisation system to speed up transcription of larger speech samples. Pinyin results are generally worse than both detailed IPA and Simple romanisation, but it would be naturally easier to learn for speakers as they are familiar with this type of transcription system thus facilitating language preservation. While conversion from Simple romanisation or Baima Pinyin to IPA is impossible as too many details are lost, it is possible to convert into Pinyin and Simple romanised script from the better-performing IPA model, making the latter potentially the most useful, not just for phoneticians, but also the local community.

To conclude, the way the language is transcribed can affect tonal outputs, even when the tonal markings themselves remain the same throughout different transcriptions. This underlines the difficulties in using deep-learning based technology, where the various orthographies produce opaque but significant differences in how the system outputs tone.

Ethics Statement

Ethics approval was obtained prior to data collection from the Research Ethics Office of CNRS.

Acknowledgements

We would like to thank the Baima speakers. We also gratefully acknowledge funding from the ANR (PhoTon, ANR-23-CE54-0003) and the ERC (PaganTibet Advanced Grant 101097364).

[6] Overall frequencies can vary slightly due to the different splits in training/validation/test data, but to give an impression, in the test set #3 the frequencies are (in descending order): Tone 53 - n=1626 (53%), Tone 31 - n=551 (18%), Tone 213 - n=495 (16%), Tone 44 - n=363 (12%), Tone 35 - n=22 (0.7%) and Tone 55 - n=15 (0.5%).

References

Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, and Alexis Michaud. 2018. Evaluation phonemic transcription of low-resource tonal languages for language documentation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Timofey Arkhangelskiy. 2021. Low-resource asr with an augmented language model. In *Proceedings of the Seventh International Workshop on Computational Linguistics of Uralic Languages*, pages 40–46.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

Yuen Ren Chao. 1930. ə sɪstəm əv "toun-lɛtəz". *Le Maître Phonetique*, 30(1):24–27.

Katia Chirkova. 2017. 14 evidentials in pingwu baima. *Evidential systems of Tibetan languages*, 302:445.

Katia Chirkova. 2025. Pitch, vowel duration, and phonation in baima and neighboring languages. *Language and Linguistics*, 26.2.

Katia Chirkova and Zhengkang Han. 2016. *Shiyong Duoxuyu Yufa* 实用多续语语法 *[Practical Grammar of Duoxu]*. Beijing: Minzu Chubanshe.

Katia Chirkova, Tanja Kocjančič Antolík, and Angélique Amelot. 2023. Baima. *Journal of the International Phonetic Association*, 53(2):547–576.

Katia Chirkova and Dehe Wang. 2017. Binwei yuyan diancang yu ersuyu pinyin fang'an 濒危语言典藏与尔苏语拼音方案 [endangered languages documentation and ersu romanization system]. *Xinan Renmin Daxue Xuebao* 西南民族大学学报 *[Journal of Southwest University for Nationalities]*, pages 69–75.

Chinese Script Reform Committee. 1956. Hanyu pinyin fang'an 汉语拼音方案 [scheme for the chinese phonetic alphabet].

Rolando Coto-Solano. 2021. Explicit tone transcription improves ASR performance in extremely low-resource languages: A Case Study in Bribri. Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas (pp. 173–184). Association for Computational Linguistics.

Rolando Coto-Solano, Sally Akevai Nicholas, Samiha Datta, Victoria Quint, Piripi Wills, Emma Ngakuravaru Powell, and Isaac Feldman. 2022. Development of Automatic Speech Recognition for the Documentation of Cook Islands Māori. Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 3872-3882). https://aclanthology.org/2022.lrec-1.412.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197.

Nils Hjortnaes, Timofey Arkhangelskiy, Niko Partanen, Michael Rießler, and Francis M Tyers. 2020. Improving the language model for low-resource asr with online text corpora. In *Proceedings of the 1st joint SLTU and CCURL workshop (SLTU-CCURL 2020)*. European Language Resources Association (ELRA).

Bufan Huang and Minghui Zhang. 1995. Baimahua zhishu wenti yanjiu [a study of the genetic affiliation of baima]. *Tibetology in China*, 1995:79–118.

Jesin James, Deepa P Gopinath, et al. 2024. Advocating character error rate for multilingual asr evaluation. *arXiv preprint arXiv:2410.07400*.

Tan Lee, Wai Lau, Y. W. Wong, and P. C. Ching. 2002. Using tone information in cantonese continuous speech recognition. *ACM Transactions on Asian Language Information Processing (TALIP)*, 1(1):83–102.

Linying Ma, Dennis Elton Walters, and Susan Gary Walters. 2008. *Nuosu Yi–Chinese–English glossary*. Beijing: Minzu Chubanshe.

Marieke Meelen, Alexander O'Neill, and Rolando Coto-Solano. 2024. End-to-end speech recognition for endangered languages of nepal. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 83–93.

Vikramjit Mitra, Horacio Franco, Martin Graciarena, and Arindam Mandal. 2012. Normalized amplitude modulation features for large vocabulary noise-robust speech recognition. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4117–4120. IEEE.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Karin Michelson. 2021. Automatic Speech Recognition for Supporting Endangered Language Documentation. *Language Documentation & Conservation*, 15:491–513.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Mihaela C Stoian, Sameer Bansal, and Sharon Goldwater. 2020. Analyzing asr pretraining for low-resource speech-to-text translation. In *ICASSP 2020-2020*

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 7909–7913. IEEE.

Hongkai Sun, Katia Chirkova, and Guangkun Liu. 2007. *Baimayu Yanjiu* 《白马语研究》. Beijing: Nationalities Press 民族出版社.

Chihiro Taguchi and David Chiang. 2024. Language complexity and speech recognition accuracy: Orthographic complexity hurts, phonological complexity doesn't. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15493–15503, Bangkok, Thailand. Association for Computational Linguistics.

Dehe Wang, Ke Wang, Xuan Wang, Katia Chirkova, and Tao Gu. 2019. *Ersu-Chinese Dictionary* 尔苏语词汇通释. Hefei 合肥: Anhui Publishing House 安徽出版社.

A Appendix: More Sample Outputs

This appendix provides additional transcription examples to enable full comparison between different transcription systems. It is clear that all five options struggle with the same Baima words and the same phonemes, namely the first vowel in [ʧʰu^{31}jo^{213}] and the onset of [wo^{44}]. The vowel [u] is in an unstressed position here, which may explain why all models predict a schwa (or similar). Similarly, all of the converted models (i.e. all apart from the original IPA transcription) appear to struggle with onset glides [j-] vs [w-] or zero. The WER in all models apart from the Simple NoTone version is mainly higher because of the failure to recognise [ʧʰu^{31}jo^{213}] as one word. Overall, WER is very similar for all transcription forms, which provides additional support for the importance of reporting both CER and WER, especially when it comes to ASR for extremely low-resource and highly-endangered languages (James et al., 2024).

B Appendix: Hyperparameters

The following are the hyperparameters for the Wav2Vec2 training, using the `wav2vec2-large-xlsr-53` base model:

1. `attention_dropout = 0.1`
2. `hidden_dropout = 0.1`
3. `feat_proj_dropout = 0.0`
4. `mask_time_prob = 0.05`
5. `layerdrop = 0.1`
6. `gradient_checkpointing = true`
7. `ctc_loss_reduction = mean`
8. `per_device_train_batch = 8`
9. `gradient_accumulation_steps = 2`
10. `evaluation_strategy = steps`
11. `num_train_epochs = 29` (4000 steps)
12. `fp16 = true`
13. `save_steps = 400`
14. `eval_steps = 100`
15. `learning_rate = 3e-4`
16. `warmup_steps = 500`
17. `kenlm_ngrams = 4`

The following are the hyperparameters for the MMS training, using the `mms-1b-all` model:

1. `attention_dropout = 0.0`
2. `hidden_dropout = 0.0`
3. `feat_proj_dropout = 0.0`
4. `ctc_loss_reduction = mean`
5. `per_device_train_batch = 2`
6. `evaluation_strategy = steps`
7. `num_train_epochs = 4` (4872 steps)
8. `gradient_checkpointing = true`
9. `fp16 = true`
10. `save_steps = 400`
11. `eval_steps = 100`
12. `learning_rate = 1e-3`
13. `warmup_steps = 100`
14. `kenlm_ngrams = 4`

The following are the hyperparameters for the Whisper training, using the `whisper-medium Multilingual` model:

1. `per_device_train_batch_size = 2`
2. `per_device_eval_batch_size = 1`
3. `gradient_accumulation_steps = 1`
4. `learning_rate = 1e-5`
5. `warmup_steps = 500`
6. `max_steps = 4001`
7. `gradient_checkpointing = true`
8. `evaluation_strategy = steps`
9. `predict_with_generate = true`
10. `generation_max_length = 225`
11. `fp16 = true`
12. `metric_for_best_model = wer`
13. `greater_is_better = false`

	Five different transcription systems (withoutLM, W2v2)		
1. SPX-bqh-018-453	"I have a buffalo hide soaked in water [if you can tan the hide...]."		
IPA			
Target transcription	$k^{h}ʉ^{53}$ la^{53} $ʧ^{h}u^{31}$ jo^{213} $ʃu^{31}$ mba^{53} wo^{44} $ʒɑ^{53}$ $ʐu^{53}$	CER	WER
Prediction	$k^{h}ʉ^{53}$ la^{53} $ʧ^{h}ə^{53}$ jo^{213} $ʃu^{31}$ mba^{53} wo^{213} $rʒɑ^{53}$ $ʐu^{53}$	15	57
Pinyin			
Target transcription	gue^{53} la^{53} chu^{31} yoo^{213} syu^{31} $nbba^{53}$ woo^{44} $ssha^{53}$ xxu^{53}		
Prediction	gue^{53} lu^{31} ei^{53} $chii^{53}$ oo^{213} syu^{31} $nbba^{53}$ oo^{213} $zzei^{213}$ xxu^{53}	36	57
Pinyin NoTone			
Target transcription	gue la chuyoo syunbba woo ssha xxu		
Prediction	gue la chii yoo syunbba oo ra xxu	21	57
Simple			
Target transcription	$khue^{53}$ la^{53} $tsyhu^{31}$ yoo^{213} syu^{31} mba^{53} woo^{44} zya^{53} zyu^{53}		
Prediction	$khue^{53}$ la^{53} $tsyhə^{31}$ yoo^{413} shu^{31} mba^{53} oo^{213} zyu^{53}	24	57
Simple NoTone			
Target transcription	khue la tsyhuyoo syumba woo zya zyu		
Prediction	khue la tsyhəyoo syumba oo dzya zyu	9	43

Table 7: ASR results from various experiments for Baima - Part II: Transcription systems

Kuene: A Web Platform for Facilitating Hawaiian Word Neologism

Sunny Walker **Winston Wu** **Bruce Torres Fischer** **Larry Kimura**
University of Hawai‘i at Hilo
{swalker,wswu,bruce42,larrykim}@hawaii.edu

Abstract

This paper presents Kuene, a web-based collaborative dictionary editing platform designed to facilitate the creation and publication of Hawaiian neologisms by the Hawaiian Lexicon Committee. Through Kuene, the Committee can create, edit, and refine new dictionary entries with a multi-round approval process, ensuring accuracy and consistency. The platform's technical features enable flexible access control, fine-grained approval states, and support for multimedia content and AI-assisted orthography modernization. Just in the past several months, Kuene has enabled the publication of over 400 new Hawaiian words. By streamlining the dictionary editing process, Kuene aims to alleviate the scarcity of modern Hawaiian words and facilitate the revitalization efforts of the Hawaiian language.

1 Introduction

Hawaiian is a critically endangered language in the Austronesian language family, spoken in the state of Hawaii, USA. Through most of the 1900s, Hawaiian was banned in schools, leading to a sharp decline in usage and a generation with nearly no native speakers. Only in the past 40 years have there been active efforts to revitalize the language through educational initiatives such as immersion schools, leading to a resurgence of usage. One of the many hindrances to the active use of Hawaiian in daily life today is the lack of words for many modern concepts. To remedy this issue, the Hawaiian Lexicon Committee, Kōmike Hua‘ōlelo, was formed in 1987 for the purpose of creating new words in the language. The Committee is composed of native Hawaiian speakers who meet regularly to discuss and create new words. As a result of their meetings, the Committee has published Māmaka Kaiao (Kōmike Hua‘ōlelo, 2003), a dictionary of modern Hawaiian words, which has been updated several times since. This dictionary, along with others (Pukui et al., 1976; Andrews, 1865; Pukui and Elbert, 1986), have been instrumental for students and learners of Hawaiian. However, due to several factors including the COVID pandemic, the Committee has not met in several years, and progress on updating Māmaka Kaiao with new words has stalled until very recently.

In this paper, we present Kuene, an online collaborative dictionary editing and publishing platform that facilitates the process of creating and publishing neologisms by the Hawaiian Lexicon Committee. Using Kuene, the Committee can propose new words and definitions. Then, other Committee members can review proposed entries, making edits as needed. Several rounds of approvals by different members can be completed through Kuene to ensure the accuracy of the new words, their translations, parts of speech, example usages, and other information associated with the new entry. After a final editorial review, a word can be seamlessly published using a one-click export to a public Hawaiian dictionary website, Wehewehe Wikiwiki[1], hosted at the University of Hawai‘i.

Kuene sports several technical features that facilitate the neologism process. User accounts with different permissions can limit access to users with different roles, e.g. one member responsible for creating the dictionary entry, or an editor responsible for proofreading for typos. An entry's headword and definition can have different approval states, allowing for finer distribution of effort when approving a new entry, particularly for headwords that have previously approved definitions. Users may also post internal comments for in-context asynchronous discussion about entries. Kuene takes advantage of the web-based medium to support embedding of media such as photos, audio, video, and taxonomic tagging to further add context to dictionary entries, enhancing comprehension for new

[1] https://hilo.hawaii.edu/wehe/

Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 182–187

and multimodal learners of Hawaiian. Furthermore, Kuene supports efficient checking for duplicate entries and existing related entries and integration with AI tools for NLP-assisted modernization of Hawaiian's 19th century printing press orthography.

Just in the past few months, the Kuene platform has been used to publish over 400 new Hawaiian words, and it is also being used to develop a dictionary for legal Hawaiian terms and a monolingual (Hawaiian-Hawaiian) dictionary. With Kuene, we envision a considerably shorter lead time from the proposal to the publication of new words by the Hawaiian Lexicon Committee, which will greatly alleviate the lack of missing words in Hawaiian as well as support the language revitalization efforts of this critically endangered language.

2 Related Work

To our knowledge, there are no existing software designed specifically to aid the revitalization of an endangered language through the creation and publication of new words. However, two related areas are language documentation and conlang creation.

Regarding language documentation, software such as FieldWorks Language Explorer[2] and WELT (Ulinski et al., 2014) have been developed for field linguists to elicit and document words in a language. These tools have a number of features that are useful in language documentation, but they are not designed for the creation of new words. Furthermore, these complex tools are designed for trained linguists, while several target users of Kuene, i.e. members of the Hawaiian Lexicon Committee, are elders with little technology experience. In addition, due to the potential lack of internet in documenting endangered languages, language documentation tools are often installed onto a computer, as opposed to Kuene's web-based interface. Some tools such as Linguistic Field Data Management and Analysis System (LiFE) (Singh et al., 2022) and Glam (Gessler, 2022) integrate NLP into the language documentation process, but these systems are also designed primarily for use by field linguists rather than native speakers of the language.

Conlangs (constructed languages) undergo a similar process as Hawaiian, where one goal of a conlang's creator is to expand the language's vocabulary. However, conlang vocabulary is often limited to the creator's needs, and the development of a conlang's vocabulary is often manageable with a simple excel spreadsheet. Some specialized software exists for keeping track of a conlang's vocabulary and grammar, such as PloyGlot[3]; language documentation tools mentioned above can also serve this purpose. These kinds of software include features not needed by the Hawaiian Lexicon Committee, such as encoding phonological rules. Furthermore, creating conlangs is often a one-person affair, and as such, conlang software often do not support collaborative editing.

In the NLP literature, there is a wealth of research in detecting neologisms (e.g. Cartier, 2017; Breen et al., 2018; McCrae, 2019; Ryskina et al., 2020) and computationally constructing neologisms (e.g. Özbal and Strapparava, 2012; Das and Ghosh, 2017; Wu and Yarowsky, 2018; Mizrahi et al., 2020). However, these tasks are not the current focus of Kuene, which aims to facilitate the *human* process of creating neologisms.

The decisions about how to coin new Hawaiian words are also out of the scope of this paper. Briefly, priority is given to a wide range of words commonly encountered by current Hawaiian language speakers, and from the curriculum content of Hawaiian language medium education programs. The guidelines created by the Committee to coin Hawaiian neologisms are published in the Māmaka Kaiao new Hawaiian words dictionary. A Hawaiian neologism requires two separate meeting approvals by the Committee before it is forwarded to the Kuene Committee to conduct its work.

3 System Description

We describe the components of the Kuene platform and how it facilitates the work of the Hawaiian Lexicon Committee. A screenshot of the entry editing page for the word *pena waha* (lipstick) is shown in Figure 1. The interface can be localized in Hawaiian or English and easily supports extension to other languages.

Entry. The main unit of data in Kuene is the entry. A user can enter a new entry in either the Hawaiian to English direction or the English to Hawaiian direction, which we refer to as the Headword entry and Definition entry, respectively. Both directions will ultimately need to be entered, but Kuene supports automatically creating the opposite direction entry and linking to existing entries. The main

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

[3] https://draquet.github.io/PolyGlot/

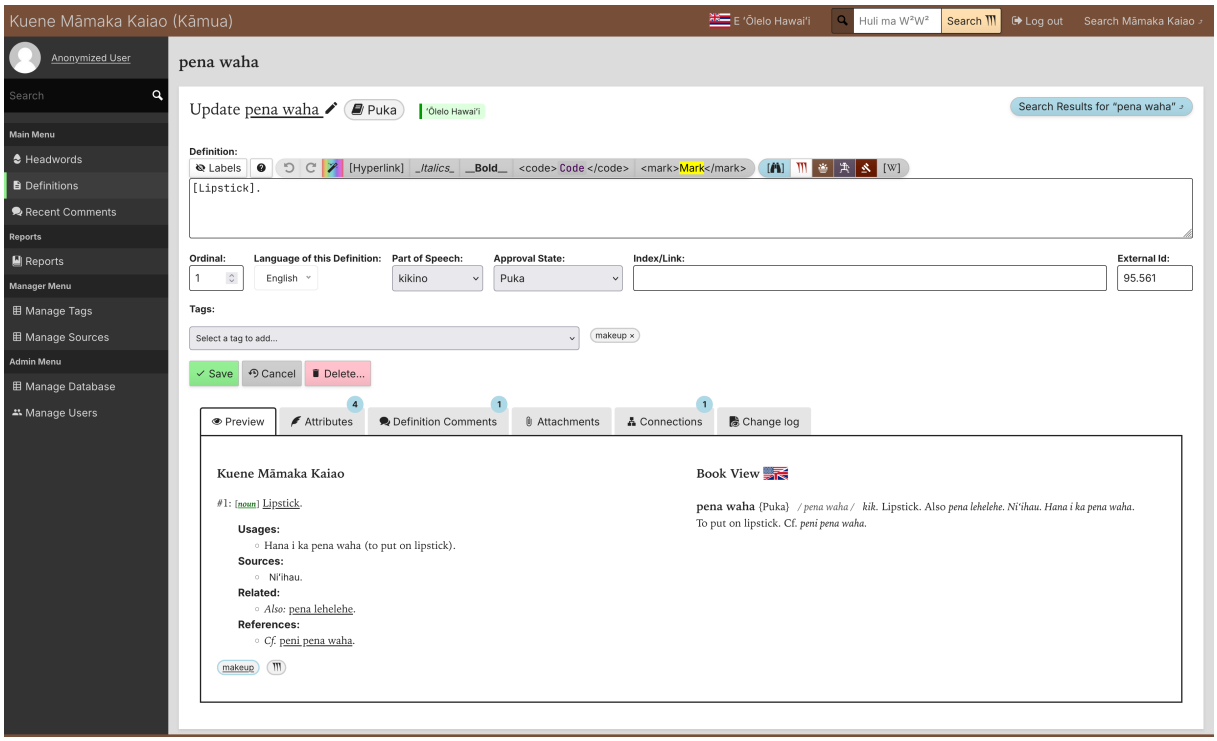


Figure 1: A screenshot of Kuene on the editing page for the word *pena waha* (lipstick), one of many words which was added to the Hawaiian dictionary in the past year. In Appendix A, Figure 3 shows the same interface for a user with lower access privileges.

components of a Headword entry are: the Hawaiian word, segmented syllables (to aid in pronunciation), approval state, and potential links to other headwords. The Definition entry is more complex, containing the English definitions, the part of speech, and several other attributes described below. Kuene currently supports 14 Hawaiian parts of speech defined in Kamanā and Wilson (2012). Kuene is also inherently multilingual, supporting headwords and definitions in English and Hawaiian, as well as other languages including Latin, French, Māori, Samoan, and Tahitian.

Approval. A dictionary entry must undergo numerous checks for quality by different people before being published in the dictionary. To support finer division of labor, the headword entry and definition entry can be separately approved. This is particularly useful for cases where definitions (English words) have already been approved (i.e. there is a demonstrated need for a Hawaiian neologism), but the Hawaiian word is still being considered by the committee. Kuene supports a variable number of approval states as designated by the Committee's needs, ranging from *introduced* to *needs supplemental attributes* to *published*.

Attributes. Entry attributes provide details about an entry that aid in comprehension of the word, including example sentences, sources where this word was used or found, related words, references, and tags to indicate the topical categories of an entry. Kuene has built-in functionality to format each attribute with appropriate HTML styling when published to the dictionary website or output to a printed version, and also includes a helpful feature to automatically add missing diacritics with AI integration (described below). In Appendix A, Figure 2 presents an interface (localized to Hawaiian) for editing the available tags.

Comments. Because users of Kuene live in various parts of the Hawaiian Islands, it is important that users can communicate about an entry without having to travel to the same place. The Comments section allows for in-context asynchronous discussions about an entry, facilitating the Committee's work.

Attachments. The Attachments section allows users to upload images, audio files, and video clips that can be displayed on the dictionary website or to support discussion about an entry.

Changelog. For accountability of online management, Kuene supports an in-context change log so users can view a timeline of entry modifications, preserving historical data.

Reports. Kuene can generate reports that summarize the current state of progress. Currently, Kuene supports generating reports for Approval States, Attachments, Attributes, Recent Changes, Duplicate Headwords, and Tags. This allows users, for example, to quickly list all entries that are at a specified approval state, with links to edit those entries.

Publication. Once an entry has obtained full approval, automated checks can be performed for existing duplicate or related entries, which can be manually fixed by editing the entry through Kuene. After all necessary edits have been made, Kuene can export the entry to an existing online Hawaiian dictionary[4] for use by the general public. If additional edits need to be made, the user can perform the edits through Kuene and then republish the entry.

3.1 Orthography Modernization

When Protestant missionaries first arrived in Hawaii around 1820, they introduced an orthography using Latin letters to the previously unwritten Hawaiian language. The use of the ʻokina (ʻ) to represent glottal stops, and the kahakō (macrons) for long vowels, was not standardized until about 100 years later. Today, text with the ʻokina and kahakō diacritics is particularly helpful for new Hawaiian language learners who are not able to easily discern between words that are spelled the same without diacritics by context alone.

Kuene also supports integration with AI tools for NLP-assisted modernization of the 19th century missionary orthography. When viewing selected example sentences from old sources written in the old orthography, users can have AI systems perform a first pass at modernizing the orthography of the entry. On the backend, this is implemented by prompting a locally hosted Llama 3.2 model to add ʻokina and kahakō to the provided sentence. Preliminary experiments show that this method is competitive with sequence-to-sequence Transformer translation models. After orthography modernization by the NLP model, the user can make necessary corrections before saving their edits to the dictionary entry. This process saves the user the effort of manually modernizing the sentence from scratch, and users can also indicate that their corrections will be saved for future retraining of the orthography modernization model.

[4] https://anonymized

4 Conclusion

We presented Kuene, an online collaborative dictionary platform that facilitates the work of the Hawaiian Lexicon Committee in coining Hawaiian neologisms. Kuene supports all the steps of coining new Hawaiian words, from its creation to its publication in the Māmaka Kaiao dictionary, which is accessible online through Wehewehe Wikiwiki. Kuene has already seen major successes, with over 400 words published through Kuene in the last few months. The design of Kuene is very modular and extensible, making it relatively easy to produce different dictionaries for different purposes or even different languages. Kuene is also currently being used to develop a dictionary for historical legal Hawaiian terms and a monolingual (Hawaiian-Hawaiian) dictionary.

Because Kuene supports multiple dictionary sources, a long-term goal is to develop a unified corpus of entries drawn from various Hawaiian dictionaries available today. We also plan to expand Kuene to incorporate more NLP methods and techniques, including improvements to the orthography modernization, and tools that can automatically generate neologism, e.g. Özbal and Strapparava (2012), which can potentially lessen the cognitive load of the Committee members. Thanks to the efforts of the Hawaiian Lexicon Committee, the new words added to the Hawaiian lexicon through Kuene are contributing significantly to the revitalization of Hawaiian and promoting wider use of the Hawaiian language in daily life.

Limitations

Our paper presents a web platform to assist the Hawaiian Lexicon Committee in creating neologisms for Hawaiian, a critically endangered and historically marginalized language. This type of system may not be applicable to all under-resourced languages in need of new vocabulary, especially if there is no committee or governing body responsible for introducing new words to the language's lexicon. We also recognize that there may be differing opinions about the best way to facilitate the neologism process. For example, Hornsby and Quentel (2013) describes conflicts of authenticity of neol-

ogisms from different sources in Breton, a Celtic language spoken in Brittany in modern-day France. In this paper, we have taken a pragmatic approach, focusing on creating a user-friendly and functional platform that meets the needs of the Hawaiian Lexicon Committee.

Acknowledgments

This work is partially supported by the National Science Foundation (Award No. 2422413). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. The authors would also like to thank the anonymous reviewers for their helpful feedback.

References

Lorrin Andrews. 1865. *A dictionary of the Hawaiian language, to which is appended an English-Hawaiian vocabulary and a chronological table of remarkable events*. Dalcassian Publishing Company.

James Breen, Timothy Baldwin, and Francis Bond. 2018. The company they keep: Extracting Japanese neologisms using language patterns. In *Proceedings of the 9th Global Wordnet Conference*, pages 163–171, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.

Emmanuel Cartier. 2017. Neoveille, a web platform for neologism tracking. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 95–98, Valencia, Spain. Association for Computational Linguistics.

Kollol Das and Shaona Ghosh. 2017. Neuramanteau: A neural network ensemble model for lexical blends. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 576–583, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Luke Gessler. 2022. Closing the NLP gap: Documentary linguistics and NLP need a shared software infrastructure. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 119–126, Dublin, Ireland. Association for Computational Linguistics.

Michael Hornsby and Gilles Quentel. 2013. Contested varieties and competing authenticities: neologisms in revitalized breton. *International Journal of the Sociology of Language*, 2013(223):71–86.

Kauanoe Kamanā and William H Wilson. 2012. *Nā Kai ʻEwalu*. Hale Kuamoʻo.

Kōmike Huaʻōlelo. 2003. *Māmaka Kaiao: A Modern Hawaiian Vocabulary*. University of Hawaiʻi Press.

John Philip McCrae. 2019. Identification of adjective-noun neologisms using pretrained language models. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 135–141, Florence, Italy. Association for Computational Linguistics.

Moran Mizrahi, Stav Yardeni Seelig, and Dafna Shahaf. 2020. Coming to Terms: Automatic Formation of Neologisms in Hebrew. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4918–4929, Online. Association for Computational Linguistics.

Gözde Özbal and Carlo Strapparava. 2012. A computational approach to the automation of creative naming. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 703–711, Jeju Island, Korea. Association for Computational Linguistics.

Mary Kawena Pukui and Samuel H Elbert. 1986. *Hawaiian dictionary: Hawaiian-English English-Hawaiian revised and enlarged edition*. University of Hawaii Press.

Mary Kawena Pukui, Samuel H Elbert, and Esther T Mookini. 1976. *Place Names of Hawaii: Revised and expanded edition*. University of Hawaii Press.

Maria Ryskina, Ella Rabinovich, Taylor Berg-Kirkpatrick, David Mortensen, and Yulia Tsvetkov. 2020. Where new words are born: Distributional semantic analysis of neologisms and their semantic neighborhoods. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 367–376, New York, New York. Association for Computational Linguistics.

Siddharth Singh, Ritesh Kumar, Shyam Ratan, and Sonal Sinha. 2022. Towards a unified tool for the management of data and technologies in field linguistics and computational linguistics - LiFE. In *Proceedings of the Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia within the 13th Language Resources and Evaluation Conference*, pages 90–94, Marseille, France. European Language Resources Association.

Morgan Ulinski, Anusha Balakrishnan, Daniel Bauer, Bob Coyne, Julia Hirschberg, and Owen Rambow. 2014. Documenting endangered languages with the WordsEye linguistics tool. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 6–14, Baltimore, Maryland, USA. Association for Computational Linguistics.

Winston Wu and David Yarowsky. 2018. Massively translingual compound analysis and translation discovery. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

A Additional Screenshots

Figure 2 and Figure 3 present additional screenshots showing the powerful functionality of Kuene.

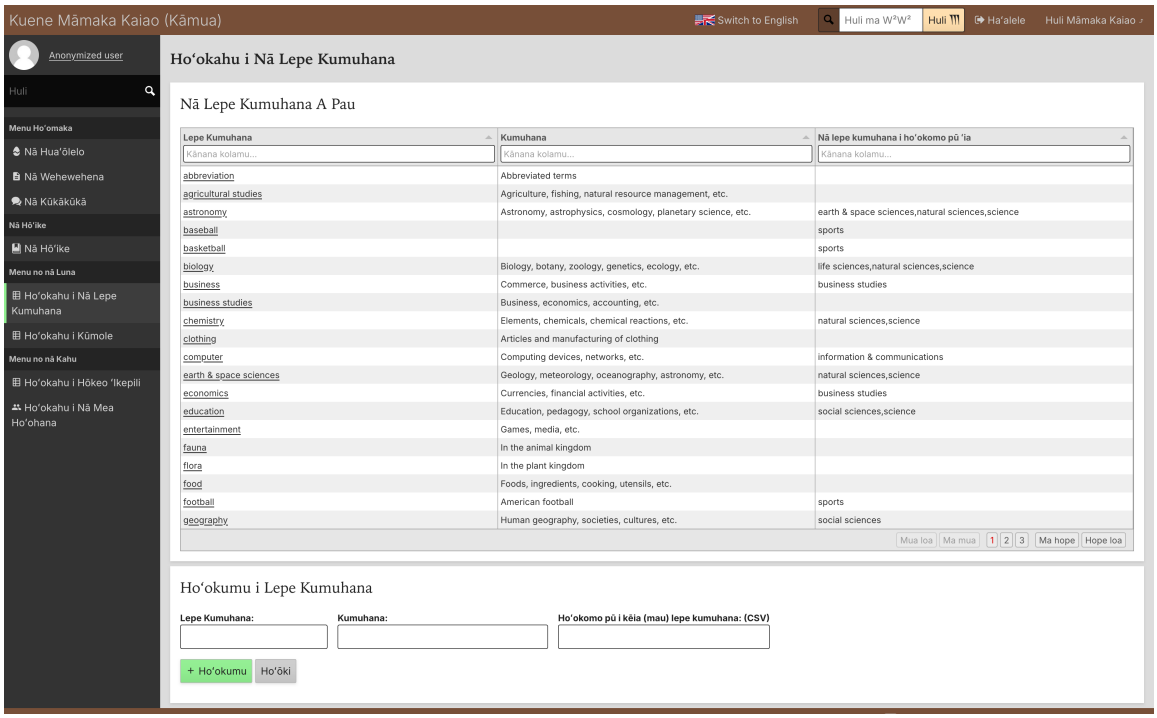


Figure 2: Kuene supports localized UI and assigning tags for categorizing entries.

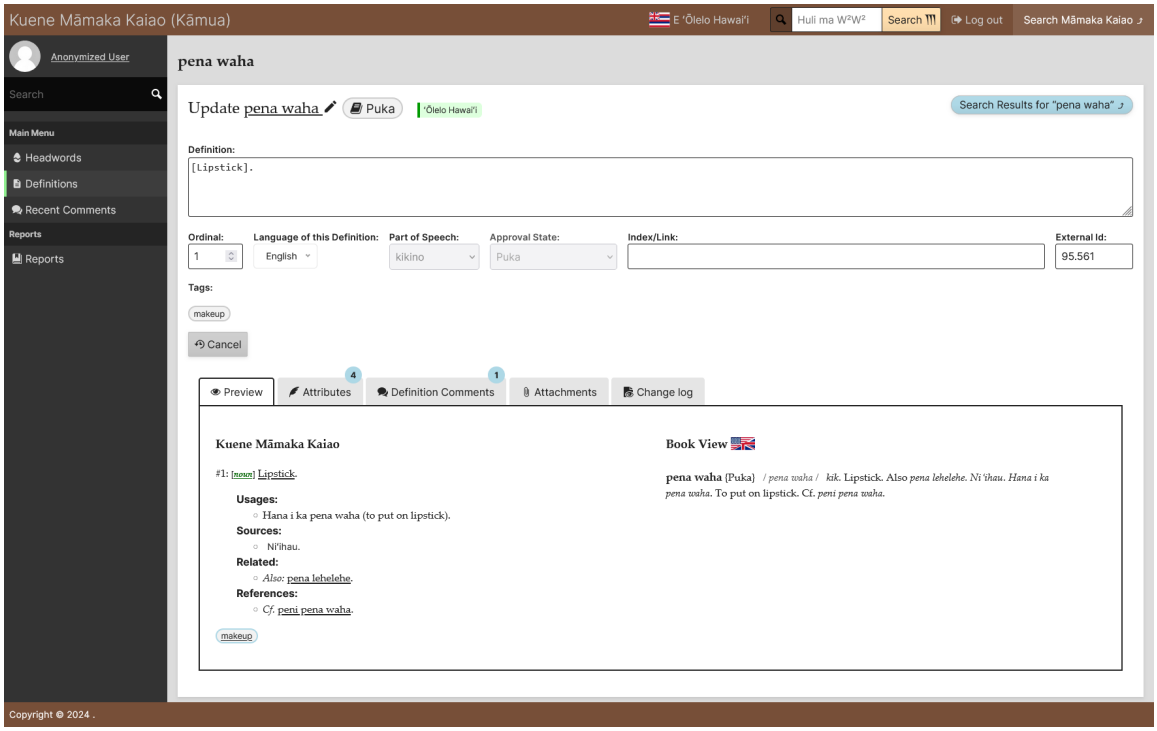


Figure 3: Editing the word *pena waha* from the perspective of a user with lower access privileges and an English interface.

Evaluation of Morphological Segmentation Methods for Hupa

Nathaniel Parkes
Department of Linguistics
University of Florida
n.parkes@ufl.edu

Zoey Liu
Department of Linguistics
University of Florida
liu.ying@ufl.edu

Abstract

Building downstream NLP applications with tokenization systems built on morphological segmentation has been shown to be fruitful for certain morphologically-rich languages. Yet, indigenous and endangered languages, which tend to be highly polysynthetic and therefore potential beneficiaries of this approach, pose additional difficulties in their limited access to annotated data for morphological segmentation tasks. In this study, we develop morphological segmentation models for Hupa, a Dene/Athabaskan language critically endangered to North America. With a total of 595 word types, we seek to identify an optimal morphological segmentation model and illustrate how those tested perform under different levels of training data limitation. We propose a simple method that casts morphological segmentation as a sequence binary classification task. While this approach does not outperform the established practice of multi-class classification, it outperforms neural alternatives. This work is conducted under the intention to act as a starting point for future technological developments with Hupa looking to leverage its morphological qualities, which we hope can serve as a reflection for work with other indigenous languages being studied under similar constraints.

1 Introduction

The Hupa people of the Hoopa Valley Reservation in Humboldt county California are a federally recognized indigenous group within the United States with over 3,000 documented descendants (Encyclopaedia Britannica, 2024). Despite resistance to policies or attempts at cultural erasure imposed by the American Government, the Hupa tribe has shown signs of gradual increase in American influence, noted in reports dating back to the mid-20th century (Bushnell, 1968). Today, many aspects of their culture and tradition are upheld, but modern descendants are exhibiting a declining trend in language retention with English taking over as the primary language (Spence, 2021). Efforts are being made to revitalize this piece of their culture, but relevant language data is limited and the Hupa language, of the Dene/Athabaskan language family, is currently recognized under endangered status (Campbell and Grondona, 2008).

With the support of community members and linguists with advanced knowledge on the language, recent work has started to leverage computational techniques to facilitate documentation of Hupa and creation of pedagogical materials for language teaching. However, said research has only focused on automatic speech recognition (Venkateswaran and Liu, 2024). In this paper, we intend to contribute to such efforts, focusing specifically on morphological segmentation for Hupa. The goal of morphological segmentation is to automatically segment a word into its individual component morphemes (e.g., *lemons* $\rightarrow$ *lemon* + *s*).

Like many other native American languages, Hupa has a highly complex, yet productive, polysynthetic morphology (Goddard, 1902-1907). As a result, the process of segmenting words into their morphological components in Hupa is likewise a difficult process when completed manually by seasoned linguists. Building computational models to segment words into sub-words, or morphemes, can be advantageous for such morpheme-rich systems. Furthermore, this can have major implications in the automation of language documentation processes (see also Zevallos and Bel (2023)).

With that in mind, this study makes two contributions. First, we evaluate the performance of four different model alternatives for morphological segmentation for Hupa; we purposefully create experimental settings with varying degrees of data limitations in order to probe the robustness of these models when faced with severely resource-constrained contexts. Second, we propose a simple

Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 188–193

augmentation to the sequence-tagging approach to morphological segmentation and show how it levels up to established neural techniques.

2 Related Work

The task of morphological segmentation has enjoyed popularity over the years for a number of reasons. First, morphological supervision has practical use in downstream NLP tasks such as dependency parsing (Seeker and Çetinoğlu, 2015) and language modeling (Blevins and Zettlemoyer, 2019). Morphological information has also been shown to be helpful for machine translation (Clifton and Sarkar, 2011; Mager et al., 2022) and automatic speech recognition (Afify et al., 2006), two tasks that are among some of the most useful for indigenous endangered speech communities (Zhang et al., 2021; Prud'hommeaux et al., 2021). In addition, morphological structures can be included in learning materials such as online dictionaries (Garrett, 2011).

Prior work has addressed morphological segmentation for low-resource morphologically complex languages, including cases such as Seneca (Liu et al., 2021) as well as Mexican indigenous languages (Kann et al., 2018). These studies largely focused on surface segmentation[1], where the concatenation of all the individual morphemes is the same as the initial surface word form (e.g., *lemons* → *lemon* + *s*). In this paper, we also concentrate on surface segmentation using orthographic representations of words in Hupa.

3 Experiments

3.1 Data and preprocessing

The data for this study consists of 595 word types (no duplicates), which were extracted from a set of nine unpublished Hupa texts drawn from archival manuscripts with handwritten transcriptions by Curtin (1888-1889), Goddard (1902-1907), Kroeber (1900-1906), and Woodward (1953), plus recorded and transcribed stories told by contemporary Hupa speaker Mrs. Verdena Parker and handwritten sources, both validated in consultation with Mrs. Parker. All transcriptions were rendered in the practical Hupa orthography originally developed in the 1980s by Victor Golla and the Hoopa Valley Tribe's language committee, which is featured in resources like the Hupa Online Dictionary and Texts Website[2] and the learner-oriented print dictionary on which it is based (Golla, 1996). The practical orthography uses conventions familiar to people who are already literate in English, and is accessible for a standard English keyboard, such as the use of the digraph *ch* for an alveopalatal affricate, *u* for a centralized schwa-like vowel in closed syllables, colon *:* for vowel length, and apostrophe *'* for glottalization of certain classes of consonants and glottal stops elsewhere. These orthographic representations were manually parsed into component morphemes. The complete dataset held an average of 3.10 morphemes per word, as well as an average of 4 characters per morpheme. Experiments were run using solely this practical orthographic transcription.

3.2 Dataset construction

To probe the impact of and the interaction between training data size and morphological segmentation methods, we create augmented datasets with varying training set sizes. We illustrate the dataset construction process with the following example.

Recall that the original dataset in orthographic representation for Hupa contains 595 unique items. We carry out the following procedures: (1) We first split this dataset evenly (roughly) into five folds; each time we select one fold as *the test set* and the concatenation of the other four folds as *the training data pool*. There are 595 / 5 = 119 items in each test set, thereby 476 items in each of the training data pools. (2) Based on the training data pool size, we decided on a range of training set sizes with mostly 100-item increment between each size: {100, 200, 300, 400, 476/training data pool size}. (3) With each training size, we randomly sample without replacement a training set of that size from a training data pool, 2 times, corresponding to two training sets of that size. (4) We repeated step (3) for each pair of training data pool + test set created from (1).

3.3 Model architectures

We study four model alternatives from two broad model classes: conditional random field (CRF) (Lafferty et al., 2001) and neural sequence-to-sequence (seq2seq) models.

CRF casts morphological segmentation as a sequence tagging task. Given a character w_t within a word w, where t indicates the index position of

[1] See Cotterell et al. (2016) for details on canonical segmentation.

[2] https://pages.uoregon.edu/jusp/dictionaries/hupa-lexicon.php

the character in the word, along with a curated feature set x_t that consists of n-grams of local (sub) strings, CRF gradually predicts the corresponding label y_t of the character using its feature set.

We curated the feature set for every character in each word as follows. We first appended each word with a start (<w>) and an end (</w>) symbol. The feature set for the character consists of the substring(s) occurring to the left and to the right side of the character up to a maximum length, δ. Consider the following Hupa word, $xotuq$, which consists of two morphemes: xo and tuq (them/people + between; together the word means "between them (people)"). If we were to set the value of δ to be 4 (which we did for model training), for the fourth character in the word, u, the sequence of substrings appearing to the left and to the right side of this character will be, respectively, {t, ot, xot, <w>xot} and {u, uq, uq</w>}. We concatenated these two sequences to be the full feature set of the fourth character u.

We implemented and compared two methods for character tagging here: multi-class classification, which is an approach applied before (Mager et al., 2022), and binary classification, inspired by Pranjić et al. (2024). With multi-class classification, for a character w_t at position t in word w, we assigned it one of six labels: `START` (for <w>); `END` (for </w>); `S` (for any single-character morpheme); and `B` (beginning); `M` (middle); or `E` (end) for characters in a multi-character morpheme. Based on the morpheme structure of the word $xotuq$, the segmentation labels are as follows:

<w>	x	o	t	u	q	</w>
START	B	E	B	M	E	END

In binary classification, said character w_t at position t in word w, if not set to `START` (for <w>) or `END` (for </w>), is assigned one of two labels: `B` (for any character bounded, or followed, by a morpheme boundary); and `U` (for characters unbounded, or not followed, by a morpheme boundary). Again, based on the morpheme structure of the word $xotuq$, labels are as follows:

<w>	x	o	t	u	q	</w>
START	U	B	U	U	U	END

We consider this form of classification as a simpler alternative to multi-class classification. If successful, breaking down the task of sequence tagging to a simple option of 0 or 1, bounded or unbounded, provides a more efficient data representation design that can possibly facilitate the model's training when faced with fewer resources.

We built first-order CRFs (Lafferty et al., 2001; Ruokolainen et al., 2013) for morphological segmentation. All models were implemented with the Python library `crfsuite`. This decision was motivated by two factors. First, prior work has demonstrated CRF to be superior to neural sequence-to-sequence models as well as different variants of unsupervised models such as Morfessor (Creutz and Lagus, 2002), when it comes to low-resource morphological segmentation for a variety of typologically diverse languages (Liu and Dorr, 2024; Liu and Prud'hommeaux, 2022; Cotterell et al., 2015). Second, CRF models, particularly those of lower orders (first-/second-order), are much faster and efficient to implement.

Our second model class is the neural-network models, specifically seq2seq. The models are expected to, given a word, produce an output of the equivalent word segmented by internal morpheme boundaries, indicated by the '!' delimiter below:

INPUT	x	o	t	u	q
OUTPUT	x	o	!	t	u	q

We made use of three seq2seq frameworks with the Python library `fairseq` (Ott et al., 2019), each under their default parameters: TRANSFORMER model (embedding size of 512, 6 encoder-decoder layers, 8 self-attention heads, and 2048 hidden units in the feed-forward layers); TRANSFORMER_TINY model, a less computationally demanding alternative contrary to the aforementioned (embedding dimension and feed-forward layer dimension both being 64; and a LSTM-based framework (embeddings of 512 dimensions and one hidden layer with 512 hidden units in both the encoder and the decoder).[3]

4 Results

We use $F1$ score as an evaluation metric for model performance. Table 1 shows the results of the CRF models for multi-class and binary classifications trained with differently sized training sets. Table 2 shows the results of the remaining three seq2seq models. Notably, the CRF models are most successful. Specifically, the multi-class classification CRFs outperform all other approaches/model architectures. While the binary classifier lags behind the multi-class alternative, it still performs notably better than any of the seq2seq models.

[3] https://fairseq.readthedocs.io/en/latest/models.html

Training Sample Size	*multiclass*	*binary*
100	70.07	62.08
200	76.18	68.96
300	78.40	70.13
400	80.27	71.16
Total	84.30	75.32

Table 1: Performance averages of the CRF-model architectures per Training Sample Size: multi-class classification and binary-classification; *Total* refers to the setting when all data from the training data pool is used for model training.

Sample Size	$Trans.$	$Tiny$	$LSTM$
100	7.41	15.46	9.63
200	13.96	28.64	15.49
300	20.56	39.58	25.37
400	29.54	46.15	34.60
Total	46.78	64.12	59.98

Table 2: Performance averages of the seq2seq-model architectures per Training Sample Size: TRANSFORMER ($Trans.$), TRANSFORMER_TINY ($Tiny$), and LSTM; *Total* refers to the setting when all data from the training data pool is used for model training.

Regarding the tendencies of the results between training sizes, we find that the CRF models showcase a gradual increase in performance capability as training set size increases. Despite CRF's sequence tagging strategy performing, comparatively, the most optimal in these low-resource environments, this trend demonstrates there is still a dependency on data set size to consider, with the dependency being stronger when training sizes are smaller (e.g., the largest $F1$ score increase occurs when training samples go from 100 to 200 word types).

The seq2seq models follow a similar trend, yet with much lower $F1$ score averages (Table 2). This is possibly due to the fact that neural-network models have much more complex training parameterization, which in turn can result in a reliance on much more extensive data resources (Wei and Ma, 2019). This conjecture is further supported by the results here that TRANSFORMER_TINY outperforms TRANSFORMER, with the former having a simpler architecture. The spread of $F1$ scores is also unique, with seq2seq models showing greater performance increase between larger training sets in comparison to what is observed with CRFs.

Learning that CRF models achieve the best performance in our experiments, we now ask: where do CRF models fall short? To address this question, we take a close look at the errors made by CRFs. Most remarkably, the CRF models struggle with words of 2 or more morpheme boundaries, especially those consisting of short, 1-3 character, morphemes. Around 66% of the time, the label-ers for both multi-class and binary classifications underestimate the number of morphemes in a word or simply predict words to be one single morpheme. Specifically, approximately 33% of all mistakes can be attributed to the later, in which CRFs fail to recognize the presence of any morpheme boundaries at all.

Another possible consideration of where the CRF models fall short is the lack of overlap between the training and the test sets. Almost none of the morphemes in the test sets can be found in the corresponding training data. With a lack of parallelism between model training and evaluation, this leaves ambiguity in certain morphological structural situations that segmentation models might fail to recognize. Yet, this challenge could be mended by data augmentation methods in the future.

5 Discussion & Future Directions

We attempt to provide evidence of the efficacy of various morphological segmentation models for Hupa and their level of robustness in response to different training set sizes. Our investigation identifies that CRF model performances shift in response to resource availability, yet they largely outperform neural alternatives in significantly low-resource settings. More notably, we also record a relatively successful CRF model using binary classification, again, outperforming all neural-network models. Despite not surpassing the multi-class classifier, the model averages are still relatively high and demonstrate a simple implementation which can be taken further in future work for Hupa and potentially other languages alike.

As mentioned prior, one caveat of model performance here is the recognition of words composed largely of short morphemes. To combat this issue, future work could consider experimenting with data in phonological representations in comparison to orthographic data. Phonological data formats may provide insight into phonetic environments for morpheme boundaries, providing suprasegmental details such as stress, tone, etc. Orthographic data may also falter as different sounds, varying in quality or length, are represented by the same symbol. Future experiments testing phonological datasets could help resolve ambiguity where morphological distinctions are created by phoneme variations that are not visible orthographically.

Another future direction for this study is to apply data augmentation methods to alleviate resource constraints. With a dataset of only 595 unique tokens, data augmentation could be implemented to strengthen validity of findings pertaining to the interaction of model performance and required data resources. In addition, while seq2seq models fell behind in this study, neural-networks may perform promisingly when trained under a larger artificially augmented dataset. We leave this for future work.

Finally, the findings reported in this paper and future avenues discussed are made with the purpose of continuously contributing to community-based efforts in language documentation. For the Hupa speech community, our plan for this line of work is to keep improving the performance of the morphological segmentation models, which will eventually be applied to automatically parse collected and digitized Hupa texts for use by the community. Additionally, we hope that our work will be helpful for other indigenous communities and academics engaged in similar efforts. To that end, we make all our code publicly available.[4]

Acknowledgments

We would like to express our gratitude to the Hupa indigenous community for their continuous support. We are grateful to Mrs. Verdena Parker and Dr. Justin Spence for their valuable efforts on Hupa language documentation over the years, and for giving us permission to work with the annotated data. Without their endorsement and expertise, this work would not have been possible. Lastly, we thank the reviewers for their thoughtful suggestions.

References

Mohamed Afify, Ruhi Sarikaya, Hong-Kwang Jeff Kuo, Laurent Besacier, and Yuqing Gao. 2006. On the use of morphological analysis for dialectal Arabic speech recognition. In *Ninth international conference on spoken language processing*.

Terra Blevins and Luke Zettlemoyer. 2019. Better character language modeling through morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1606–1613, Florence, Italy. Association for Computational Linguistics.

John H. Bushnell. 1968. From American Indian to Indian American: The Changing Identity of the Hupa. *American Anthropologist*, 70(6):1108–1116.

Lyle Campbell and Verónica Grondona. 2008. Ethnologue: Languages of the world. *Language*, 84(3):636–641.

Ann Clifton and Anoop Sarkar. 2011. Combining morpheme-based machine translation with post-processing morpheme prediction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 32–42, Portland, Oregon, USA. Association for Computational Linguistics.

Ryan Cotterell, Thomas Müller, Alexander Fraser, and Hinrich Schütze. 2015. Labeled morphological segmentation with semi-Markov models. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 164–174, Beijing, China. Association for Computational Linguistics.

Ryan Cotterell, Tim Vieira, and Hinrich Schütze. 2016. A Joint Model of Orthography and Morphological Segmentation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 664–669, San Diego, California. Association for Computational Linguistics.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning - Volume 6*, MPL '02, page 21–30, USA. Association for Computational Linguistics.

Jeremiah Curtin. 1888-1889. *Hupa vocabulary December 1888-January 1889*. National Anthropological Archives: NAA MS 2063.

Encyclopaedia Britannica. 2024. Hupa. https://www.britannica.com/topic/Hupa. Last updated: August 28, 2024.

Andrew Garrett. 2011. An online dictionary with texts and pedagogical tools: The Yurok language project at Berkeley. *International Journal of Lexicography*, 24(4):405–419.

Pliny Earle Goddard. 1902-1907. *Chilula field notes (Redwood Creek)*. American Philosophical Society Na20g.1.

Victor Golla. 1996. *Hupa Language Dictionary*. Hoopa, CA: Hoopa Valley Tribal Council.

Katharina Kann, Jesus Manuel Mager Hois, Ivan Vladimir Meza-Ruiz, and Hinrich Schütze. 2018. Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 47–57, New Orleans, Louisiana. Association for Computational Linguistics.

Alfred Kroeber. 1900-1906. *Untitled Hupa text*. Transcription in Kroeber's hand included in Goddard (1903-1906), notebook #4.

[4] https://github.com/ufcompling/hupa_morphseg

John Lafferty, Andrew McCallum, and Fernando C.N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.

Zoey Liu and Bonnie Dorr. 2024. The effect of data partitioning strategy on model generalizability: A case study of morphological segmentation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2851–2864, Mexico City, Mexico. Association for Computational Linguistics.

Zoey Liu, Robert Jimerson, and Emily Prud'hommeaux. 2021. Morphological segmentation for Seneca. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 90–101, Online. Association for Computational Linguistics.

Zoey Liu and Emily Prud'hommeaux. 2022. Data-driven Model Generalizability in Crosslinguistic Low-resource Morphological Segmentation. *Transactions of the Association for Computational Linguistics*, 10:393–413.

Manuel Mager, Arturo Oncevay, Elisabeth Mager, Katharina Kann, and Thang Vu. 2022. BPE vs. morphological segmentation: A case study on machine translation of four polysynthetic languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 961–971, Dublin, Ireland. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Marko Pranjić, Marko Robnik-Šikonja, and Senja Pollak. 2024. LLMSegm: Surface-level morphological segmentation using large language model. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10665–10674, Torino, Italia. ELRA and ICCL.

Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Karin Michelson. 2021. Automatic speech recognition for supporting endangered language documentation. *Language Documentation and Conservation*, 15:491–513.

Teemu Ruokolainen, Oskar Kohonen, Sami Virpioja, and Mikko Kurimo. 2013. Supervised morphological segmentation in a low-resource learning setting using conditional random fields. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 29–37, Sofia, Bulgaria. Association for Computational Linguistics.

Wolfgang Seeker and Özlem Çetinoğlu. 2015. A graph-based lattice dependency parser for joint morphological segmentation and syntactic analysis. *Transactions of the Association for Computational Linguistics*, 3:359–373.

Justin Spence. 2021. A Corpus Too Small: Uses of Text Data in a Hupa-English Bilingual Dictionary. *International Journal of Lexicography*, 34(4):413–436.

Nitin Venkateswaran and Zoey Liu. 2024. Looking within the self: Investigating the impact of data augmentation with self-training on automatic speech recognition for Hupa. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 58–66, St. Julians, Malta. Association for Computational Linguistics.

Colin Wei and Tengyu Ma. 2019. Data-dependent sample complexity of deep neural networks via Lipschitz augmentation. *CoRR*, abs/1905.03684.

Mary F. Woodward. 1953. *Survey of California and Other Indian Languages*. University of California Berkeley, Woodward.002.

Rodolfo Zevallos and Nuria Bel. 2023. Hints on the data for language modeling of synthetic languages with transformers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12508–12522, Toronto, Canada. Association for Computational Linguistics.

Shiyue Zhang, Benjamin Frey, and Mohit Bansal. 2021. ChrEnTranslate: Cherokee-English machine translation demo with quality estimation and corrective feedback. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 272–279, Online. Association for Computational Linguistics.

Author Index