Educational Tools for Mapuzugun

Cristian Ahumada[1] **Claudio Gutierrez**[1] **Antonios Anastasopoulos**[2]
[1]Department of Computer Science, Universidad de Chile
[2]Computer Science Department, George Mason University
ahumada.860@gmail.com cgutierr@dcc.uchile.cl antonis@gmu.edu

Abstract

Mapuzugun is the language of the Mapuche people. Due to political and historical reasons, its number of speakers has decreased and the language has been excluded from the educational system in Chile and Argentina. For this reason, it is very important to support the revitalization of the Mapuzugun in all spaces and media of society. In this work we present a tool towards supporting educational activities of Mapuzugun, tailored to the characteristics of the language. The tool consists of three parts: design and development of an orthography detector and converter; a morphological analyzer; and an informal translator. We also present a case study with Mapuzugun students showing promising results.

Short abstract in Mapuzugun: Tüfachi küzaw pegelfi kiñe zugun küzawpeyüm kelluaetew pu mapuzugun chillkatufe kimal kizu tañi zugun.

1 Introduction

Recent years have seen unprecedented progress for Natural Language Processing (NLP) on almost every NLP subtask. Along with research progress, several tools have been developed and are currently aiding millions of users every day. However, most of this progress is limited on a handful of languages (Joshi et al., 2020). For example, learners of English can nowadays avail themselves to tools like Grammarly; English speakers can use Duolingo to start learning 38 languages, including Hawaiian, Navajo, as well as High Valyrian and Klingon.[1] The only option a Mapuzugun speaker would have in practice, though, would be to use language technologies in a language other than her own (likely Spanish).

Despite Duolingo's commendable inclusion of Hawaiian and Navajo for English speakers, and of Guaraní for Spanish speakers,[2] learning resources for Indigenous languages are hard to come by, let alone ones that incorporate language technologies in the educational setting in order to aid learners. In particular, it is undeniable that the development of NLP tools that reach the users lags further behind that NLP research itself (Blasi et al., 2021).

In this work, we develop a tool for educational use in an Indigenous language of south America, Mapuzugun. This tool was created by a speaker and instructor of the language and as such is tailored specifically to the instructional needs and linguistic characteristics of Mapuzugun.

Importantly, this work shows how linguistic research (grammars), minimal community resources (dictionaries), and NLP research (e.g. FST-based morphological analyzers) can be transformed into tools useful to Indigenous communities, in particular for efforts towards preservation and revitalization of endangered languages. Our tool is publicly available through an online interface (in Mapuzugun and Spanish) at crahumadao.pythonanywhere.com.[3]

2 The Mapuzugun Language

Mapuzugun (iso 639-3: arn) is an indigenous language of the Americas spoken natively in Chile and Argentina, with an estimated 100 to 200 thousand speakers in Chile and 27 to 60 thousand speakers in Argentina (Zúñiga, 2006, 41–3). It is an isolate language and is classified as threatened by Ethnologue, hence the critical importance of all documentary efforts. Although the morphology of nouns is relatively simple, Mapudungun verb morphology is highly agglutinative and complex. Some analyses provide as many as 36 verb suffix slots (Smeets, 1989). A typical complex verb form may consist of five or six morphemes. See example in Table 1.

[1] As of March 2022.

[2] Which are due to immense efforts by the Indigenous communities themselves.

[3] Username: epu and Password: meli

Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), pages 183 - 196
July 15, 2022 ©2022 Association for Computational Linguistics

Word	Kim mapuzuguyekümelleaiñ
Segmentation	Kim mapu-zugu-yekü-me-lle-a-iñ
English Transl.	We are indeed going to learn the Mapuche language.

Table 1: Segmentation of a Mapuzugun verb phrase.

Mapudungun has several interesting grammatical properties. It is a polysynthetic language in the sense of Baker (1996); see (Loncon Antileo, 2011) for explicit argumentation. As with other polysynthetic languages, Mapudungun has Noun Incorporation; however, it is unique insofar as the Noun appears to the right of the Verb, instead of to the left, as in most polysynthetic languages (Baker et al., 2005). One further distinction of Mapudungun is that, whereas other polysynthetic languages are characterized by a lack of infinitives, Mapudungun has infinitival verb forms; that is, while subordinate clauses in Mapudungun closely resemble possessed nominals and may occur with an analytic marker resembling possessor agreement, there is no agreement inflection on the verb itself. One further remarkable property of Mapudungun is its inverse voice system of agreement, whereby the highest agreement is with the argument highest in an animacy hierarchy regardless of thematic role (Arnold, 1996).

Beyond morphology and other interesting typological properties, an additional challenge in the computational processing of Mapuzugun is the lack of a single standardized orthography. In particular, the community uses three different alphabets, namely the “Unificado", “Ragileo", and “Azümchefe" alphabets.[4]

3 System Overview

The system is comprised of the following components, with the pipeline shown in Figure 1:

1. the orthography detector, which detects which of the three alphabets is used in the input;
2. the orthography transliterator, which can convert between orthographies if conversion is needed;
3. the morphological analyzer, which produces the possible segmentations of a word or phrase;
4. the mapping of the analyzed morphemes to user-friendly notation/phrases; and
5. the final presentation of the output.

The user can use these tools through an interface available both in Mapuzugun and in Spanish. A screenshot of the landing page of the interface is shown in Figure 2.

4 Orthography Detection and Transliteration

The differences between the three orthographies are showcased in Figure 3. where “Jampvzken” is written in Ragileo, “Llampüdken” in Unificado and “Llampüzken” in Azümchefe, all three referring to the same Mapudungun phonetics of the English word “butterfly”. This example shows the relationship between the ‘`J`’ in Ragileo with the ‘`Ll`’ in Azümchefe and Unificado.

We identified and constructed the conversion tables between these orthographies. In total, for the Unified-Ragileo relationship, there are 10 differences that are shown in the Table 4, in the following case Unified-Azümchefe there are 8 differences (Table 5) and for the Ragileo-Azümchefe relationship there are 8 differences, outlined in Table 6.

Utilizing these conversion tables makes it straightforward to detect the orthography of any given input, by following a process of round-trip translation. For example, if we assume the input is in Ragileo, then if we convert to Azümchefe (or Unificado) and back to Ragileo and the final output is the same as the original input, then the input is declared to be Ragileo. If any of the intermediate translations fail it would have been exactly because our initial assumption of the input being in Ragileo was false. If no changes happen in the translation process, then all orthographies represent the input in a similar manner.

Orthography Converter Given the differences between the orthographies, special care must be taken in graphemes that have another grapheme as a substring. An example of this is the Unificado grapheme `Ng`, which also contains the grapheme `G`, which in turn is used in the same writing system for another phoneme. Or, cases in which three graphemes contain the same letter, such as the letter `"L"` in `L`, `Lh`, and `LL`. The only orthography that does not have this internal problem is Ragileo, because it uses unique letters for each Mapuzugun phoneme. This makes conversion from Ragileo to other orthographies straighforward, always taking care of the order of the transformations whose output can generate morpheme ambiguities during conversion.

[4] See Figure 5 in Appendix A.

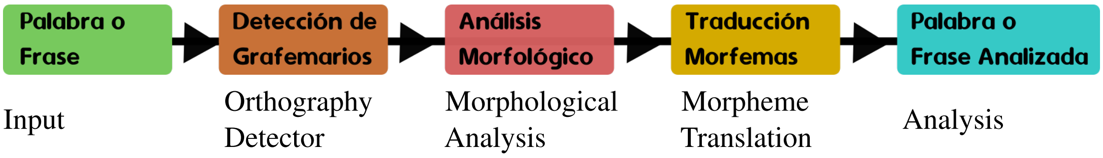
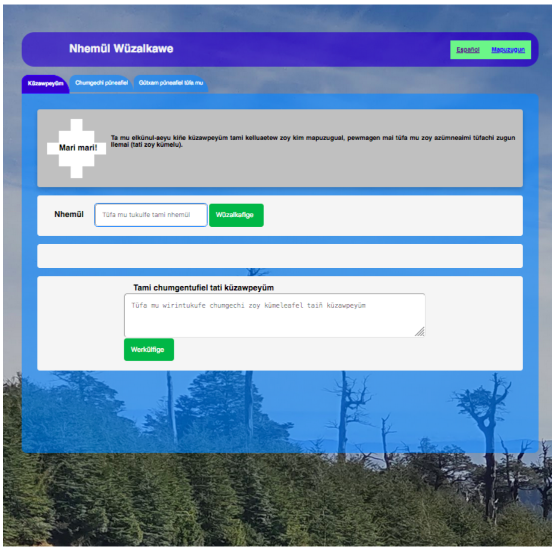


Figure 1: Pipeline of the full system.



Mari mari!

Ta mu elkünul-aeyu kiñe küzawpeyüm tami kelluaetew zoy kim mapuzugual, pewmagen mai tüfa mu zoy azümnealmi tüfachi zugun llemai (tati zoy kümelu).

Tüfa mu tukulfe tami nhemül

Tami chumgentufiel tati küzawpeyüm

Tüfa mu wirintukufe chumgechi zoy kümeleafel taiñ küzawpeyüm

Figure 2: Screenshot of the user interface.

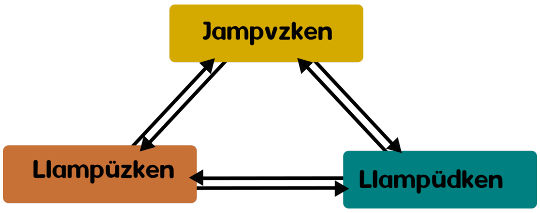


Figure 3: Conversions between orthographies for the Mapuzugun word for 'butterfly'. Top: Ragileo; bottom left: Azümchefe; bottom right: Unificado.

The order that must be taken into account because if a morpheme is contained by another, it must first be disambiguated and then continue with other changes. In the case of `Ng` and `g`, to go from Unificado to Ragileo or Azümchefe, as long as there is a `g` and there is no `N` preceding it, it can be changed to `Q`, therefore before making the transformations the `Ng` must be checked, saving the result `G`) in an auxiliary variable to be able to convert later all `Gs` of the Unificado to `Q`. Once this last step is done, the auxiliary variable is removed and the `G` resulting from the change is put back.

5 Morphological Analyzer

The morphological analyzer is responsible for producing the possible segmentations: separating words into a composition of morphemes.

5.1 Design

The analyzer is implemented through series of regular expressions, based on established grammars of Mapuzugun (Smeets, 1989; Cañumil, 2011; Chiguailaf, 1972). As another source, the compilation that was made in `azümchefe.cl` of the grammar of the language (Chiguailaf, 1972) was taken.

We worked with hand-crafted sets of regular expressions that contain the morphemes of the language. These sets separate, by function: in verb root, noun/adverb/adjective, suffixes, and endings. In addition, the position plays an important role, because each of the morphemes has a particular slot (Smeets, 1989).

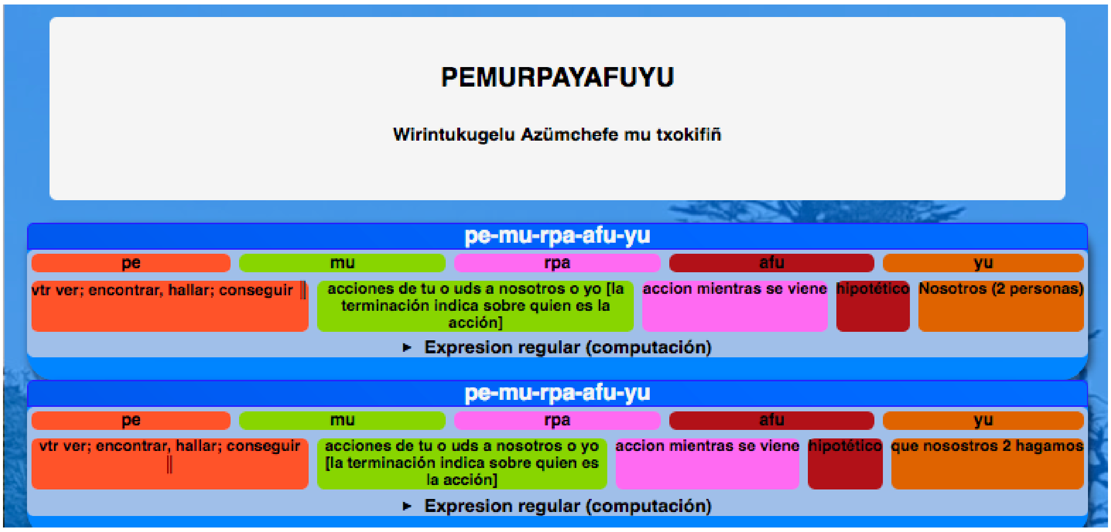
From these regular expressions, the chain of a word is traversed and possible derivations tree is generated. Only branches evaluated to be valid are passed on to the next "informal translator" step. The morphemes and their order must meet certain restrictions that have to do with the correct formulation of words in Mapuzugun, both in order, as mentioned before, but also in the compatibility of two morphemes being in the same word.

This module assumes input in the Ragileo orthography, therefore any word from another orthography must necessarily pass through the orthography converter. This decision has to do with Ragileo's advantage of 1-to-1 phoneme-to-grapheme mappings, making it easier to model morphemes.

5.2 Informal Analysis Translator

Once the segmentation is done, we implemented a module crucial for deploying the tool in educational revitalization settings: the "informal analysis translator". It assigns to each individual morphemes (or to combinations of them according to communicative role) a definition in plain Spanish. The rationale was to simplify the definition as much as possible leaving out technical linguistic features and jargon. For the case of substantives, verbs and adjectives, the definition was taken from the Mapuzugun-Spanish dictionary (Pérez, 2015).

As an example, we show the case of the word `txekayawkelai`. One of the possible segmentations



PEMURPAYAFUYU

Wirintukugelu Azümchefe mu txokifiñ

pe-mu-rpa-afu-yu

pe | mu | rpa | afu | yu

vtr ver; encontrar, hallar; conseguir ||

acciones de tu o uds a nosotros o yo [la terminación indica sobre quien es la acción]

accion mientras se viene

hipotético

Nosotros (2 personas)

▸ Expresion regular (computación)

pe-mu-rpa-afu-yu

pe | mu | rpa | afu | yu

vtr ver; encontrar, hallar; conseguir ||

acciones de tu o uds a nosotros o yo [la terminación indica sobre quien es la acción]

accion mientras se viene

hipotético

que nosotros 2 hagamos

▸ Expresion regular (computación)

Figure 4: Segmentation of the word "pemurpayafuyu" as presented by the tool.

is **txeka-yaw-ke-la-i**, with each component of the word being:

txekan- :	vi & vtr caminar, marchar, pasear || vtr medir con pasos *to walk, to take a walk*
-yaw- :	andar *to go*
-ke- :	habitualmente *usually*
-la- :	negación a modo "normal" indicativo *negation*
-i :	el / ella *he/she*

Given this,[5] the goal is that the learner deduces "el/ella no anda caminando habitualmente" *"he/she does not usually go for walks"*.

The challenges of this informal analyzer are many. Among them: how to give enough meaningful translations so that they can match the initial experience of learners, but as well, do not confuse them; how to deal with compositional morphemes (i.e. morphemes that have a different meaning when co-occurring than when occurring separately, for example transitions from second to first person); and how to include context to help the translation. We resolved these issues by relying on the expertise of an instructor of Mapuzugun.[6]

6 Usability Studies with Learners

The system (software) was tested on several groups of initial learners of Mapuzugun.

Study Design The first phase of the study design was to get access to study participants. As in the case of most endangered languages, it was difficult to identify test groups for various reasons. First, most current Mapuzugun courses are informal, given different types of social organizations with a great variety of methodologies, contents, levels. Second, students of Mapuzugun differ widely according to interests, degree of systematization and materials used. Third, there is a strong distrust by the interested community of learners in institutions, like academia, that historically have "used" aboriginal speakers as mere sources of information.

In a first preliminary round, more than 200 people (known to have been in courses or being students of Mapuzugun in the last 5 years) were contacted. From them, 30 people engaged to answer the questionnaire and from them, only 9 answers were obtained (3 of advanced knowledge of Mapuzugun).

With their feedback, the tool was refined. A second round was done by a public call in social networks related to Mapuzugun, and 32 people registered for the study, which were then classified in 5 groups:

- Group 0. Beginners (6 people);
- Group 1. Basic studies; able to greet but do not understand conversations (8 people);
- Group 2. Studies: able to understand conversations (7 people);
- Group 3. Studies; able to perform conversations (8 people); and
- Group 4. Speakers from early infancy (3 people).

The experiment consisted of giving a small set

[5] "vi" and "vtr" correspond respectively to intransitive and transitive verb.

[6] One of the authors is a speaker and instructor of Mapuzugun.

Word	Group 0	Group 1	Group 2	Group 3	Group 4	General
`elukelafimu`	2 / 2	2 / 1.6	2 / 1.86	2.75 / 2.86	2 / -	2.33 / 2.14
`pemurpayafuyu`	1 / 3	1 / 2	1.83 / 2.17	2.33 / 2.83	2 / -	1.94 / 2.44
`kujinerkeeiñmu`	0 / 1.33	0.67 / 1.4	1.75 / 2	2.43 / 2.86	3 / -	1.81 / 2.05
Phrase	1 / 3	2 / 2.25	2.57 / 2.71	3 / 3	2.67 / -	2.42 / 2.72

Table 2: Summary of the study with learners. showing the mean performance of each group for each task word. Scale goes from 0 (wrong translation) to 3 (perfect translation). The pairs A / B mean: without / with the tool.

	Group 0	Group 1	Group 2	Group 3	Group 4	General
Difficulty of use	2.67	2.71	1.86	1.86	3	2.33
Diff. of word transl.	3.17	2.86	3.0	2.43	3	2.87
Diff. of phrase transl.	3.33	3.71	2.71	3.14	1.67	3.0
Visual evaluation	3.83	3.29	4.14	3.86	2.33	3.63
General evaluation	3.17	4.0	4.71	4.29	3.33	4.0

Table 3: Summary of Usability Test. Scale goes from 1 (low) to 5 (top).

of Mapuzugun words (and one phrase[7]) to each participant. The task was to translate each word in Spanish, first without and then with the tool.

We additionally collected information on usability of the software tool: difficulty of use, difficulty to translate words, difficulty to translate phrases, evaluation of visual interface and finally, a general evaluation. Last, we requested open-ended general qualitative feedback.

Translation Results Table 2 summarizes quality of the produced translations, with and without the tool, for each user group.[8] For two words, `pemurpayafuyu` and `kujinerkeeiñmu`, using the tool improves the translation capabilities for all user groups. The word `elukelafimu` is a word that is typically accessible in basic levels of Mapuzugun, and hence, the segmentation plus the translation could have confused users (they realized that the word was more complex than they thought). Another encouraging sign is that the translation of the phrase also improved for the first three groups when using the tool. Last, we found that experienced learners (group 4), preferred not to use the tool because they felt secure in their knowledge.

Usability Results Table 3 summarizes the scores received by the users (in a Likert scale). User groups 2 and 3 seem to be the ones showing less difficulty to use the tool, and also those that can take more advantage of it. Beginners got stuck with instructions (many were in Mapuzugun; they will also be provided in Spanish in future iterations) and ability to compose particles. We suspect that experienced speakers (group 4) probably did not invest effort because they did not need the tool.

All groups except experienced speakers rated the phrase as more difficult to translate than single words. The visual aspects of the interface and the tool in general mostly received very positive scores.

As a summary, our small study shows that, at its current stage of development, our tool is appropriate *and* useful for intermediate learners.

Qualitative Feedback We summarize here the qualitative feedback we received from user groups.

In general, all groups were particularly positive about the tool's presentation of the segmentation of the words. All groups were also very positive towards our informal translator that provides the explanations of each word segment.

In general, comments in the beginners' group (group 0) mentioned the difficulty to produce the translations, even though each part of the segmentation could be understood, a note that highlights the utility and importance of our proposed "informal translation". What was liked the most was the possibility of "see" in a graphical form the composition of words. This group also struggled with certain labeling words like VTR, VI, that are not widely known.

Users in group 1 positively mentioned the possibility to see the different segmentation options. Some people signaled that there should be exam-

[7] The words are shown in Table 2. The phrase was: `Pichikalu iñche , amukefun chillkatuwe ruka mew , fewla chillkatuwekelan.`

[8] The translations were rated for accuracy by an instructor.

ples of the usage.[9]

Group 2 was the one that gave most comments. Some mentioned that a scenario when a morpheme occurs duplicated with different communicative functions was confusing. They also indicated that they would have wanted the ability to actually see the the correct translation, not just the segmentation and its explanation; unfortunately, the current state of MT for Mapuzugun does not allow this, but it provides a concrete avenue for future work.

They also liked the segmentation and its explanation, and suggest that give the possibility to practice conjugation. On the other hand, words without context can be used in different forms and this could confuse beginners.

Last, there were comments about the choice of colors of the interface, as well as a suggestion for turning the tool into a mobile app.

Group 3 suggested that beginners could get confused by the amount of options that are shown for certain words. Some of them mentioned that the program helped them to understand certain particles. They also mentioned the need of context for the words. Regarding negative issues, some persons mentioned the need to have a translation besides morphemes, although one person liked the idea that you must make efforts to compose instead of receiving the translation immediately. Group 4 did not made relevant comments.

It is worth noting, last, that many of the comments reflected the excitement that such a tool was even available for Mapuzugun.

7 Related Work

Computational Work on Mapuzugun Today there are various initiatives of computational linguistics on Mapuzugun. There is an orthographic normalizer and a morphological analyzer (Chandía, 2012), but its accuracy is low, since it is rule-based. Another aspect that could be improved is that, currently, there is no possibility of choosing the output alphabet, restricting it to only one form of writing. This is still inconvenient today, as there is still no agreement on orthography standardization. This implementation is based on a set of rules through regular expressions, with a finite state transducer, which have been released on the author's website.

The purpose of another project, called AVENUE, in which the Universidad de la Frontera, the Intercultural Bilingual Education Program and the Language Technologies Institute of Carnegie Mellon University (CMU) collaborated, was to generate simple and low-cost translations, in addition to helping to preserve the Indo-American languages. This project first developed an alphabet that was used to transcribe (but not fully revise) a 170-hour audio corpus along with Spanish translations (Duan et al., 2020), and last deployed prototype translation systems and base spell checkers that are available for OpenOffice.

In the educational field, there is software to learn Mapuzugun called MAPU from a project at the Pontificia Universidad Católica de Valparaíso that also includes voice recognition to control the application, which works correctly, but is not robust to pronunciation (Troncoso, 2012). This work also refers to another Mapuzugun-to-Spanish voice-text translation prototype, based on recordings, and to a chatbot from the Pandora project.

Last, we refer the reader to Appendix B for an additional discussion of further computational work on other south American Indigenous languages.

8 Conclusion and Future Work

We have presented a system comprised of set of NLP tools appropriate for educational purposes in Mapuzugun, an Indigenous south American language, and we have demonstrated its usefulness through a small user study. Our study also provided a guide for future improvements. As more data will hopefully become available in Mapuzugun, we plan to incorporate more recent statistical machine learning components, both for the orthography converted and the morphological analyzer. We will also hopefully be able to deploy full-fledged machine translation systems to provide free-form translations of words or phrases to learners. Many users would benefit by the incorporation of a text-to-speech component (as long as it is of high quality), that would also allow the teaching of Mapuzugun pronunciation.

Going further, the tool could be complemented with a system that permits annotation of words and/or phrases in order to collect data for future tasks, as more users adopt it – especially if language instructors use our tool in their courses. We are also hoping to create an offline version of the tool to make it accessible in areas with low connectivity. We will also attempt to incorporate any available corpora of Mapuzugun such as the those

[9] Examples are provided as part of the documentation, but they probably did not find them.

of Levin et al. (2000) and Duan et al. (2020) to use as educational examples.

We release our code[10] in the hopes that more Indigenous communities are able to use it to develop similar systems for their languages.

Ethical Considerations

Working with endangered/Indigenous languages and language data, there is always substantial risk of unwittingly perpetuation of colonial harms (Bird, 2020). This is obviously an extremely complex issue, but according to Bird (2020) and other working in the space of NLP for endangered/Indigenous languages, perhaps the most critical aspect in working with Indigenous language data is that researchers actively develop meaningful relationships with members of these respective language communities.

In our case, our work is lead by an instructor of Mapuzugun and member of the Chilean Mapuche community, who knows first-hand the oppression the Mapuche people have sufferred and the harms they have undergone by being forced to operate in Spanish. This work is also partially funded by a program dedicated to addressing the long-standing colonial harms in Chile, by specifically helping Indigenous students through their studies.

We do not anticipate any serious harms by the development of our system, and we believe that the positive reception by the Mapuche volunteers who participated in our case study will be mirrored by its reception by the wider Mapuche community. It is also important, though, to acknowledge its limitations and make it clear that our tool is meant to be a companion tool for learning and can by no means substitute instructors of the language.

No indigenous language data were collected or are released through this project. We re-used existing, publicly available tools and corpora. The tool is provided for free: it is currently behind a simple username and password setting to ensure that its traffic is not overwhelming, so that the tool remains available to the Mapuzugun instructors and learners that need it the most (and who already have access to it).

Acknowledgements

We are thankful to the Mapuzuguletuaiñ organization for their support and initial feedback, and to the volunteer Mapuzugun speakers who took part in our study. The first author was funded by the Programa de Pueblos Indígenas, FCFM, Universidad de Chile. Antonios Anastasopoulos is generously supported by the NSF through project BCS-2109578.

References

Carlo Alva and Arturo Oncevay. 2017. Spell-checking based on syllabification and character-level graphs for a peruvian agglutinative language. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 109–116.

Marco Espinoza Alvarado. 2019. El 'nativohablantismo'en la investigación de las lenguas indígenas: el caso del mapudungun en chile. *Trabalhos em Linguística Aplicada*, 58:795–825.

Gabriel E Alvarado Pavez. 2020. Glotopolítica de la desigualdad: Ideologías del mapudungun y el español en chile (2009–2019).

Belén Villena Araya, María Teresa Cabré Castellví, and Sabela Fernández-Silva. 2019. Noun formation in mapudungun: Productivity, genuineness and language planning. *Revista Signos*, 52(100):615.

Jennifer Arnold. 1996. The inverse system in Mapudungun and other languages. *RLA: Revista de Lingüística teórica y aplicada*, 34:9–47.

Mark Aronoff and Janie Rees-Miller. 2020. *The handbook of linguistics*. John Wiley & Sons.

Antti Arppe, Jordan Lachler, Trond Trosterud, Lene Antonsen, and Sjur N Moshagen. 2016. Basic language resource kits for endangered languages: A case study of plains cree. In *Proceedings of the 2nd Workshop on Collaboration and Computing for Under-Resourced Languages Workshop (CCURL 2016), Portorož, Slovenia*, pages 1–8.

Alicia Alexandra Assini. 2013. Natural language processing and the mohawk language: creating a finite state morphological parser of mohawk formal nouns.

Mark C. Baker. 1996. *The Polysynthesis Parameter*. Oxford University Press, Oxford.

Mark C. Baker, Roberto Aranovich, and Lucía A. Golluscio. 2005. Two types of syntactic noun incorporation: Noun incorporation in Mapudungun and its typological implications. *Language*, 81(1):138–176.

Ximena Bertin. 2016. Encuesta cep: 67% de la población mapuche no habla ni entiende el mapuzugun.

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.

[10] https://github.com/crahumadao/kaxvkaam

Damián Blasi, Antonios Anastasopoulos, and Graham Neubig. 2021. Systematic inequalities in language technology performance across the world's languages. arXiv:2110.06733.

María Catrileo. 2017. *Diccionario Lingüístico Etnográfico de la Lengua Mapuche: Mapudungun-Español-English*, volume 7. Ediciones Universidad Austral de Chile.

Tulio Fernando Cañumil. 2011. *Estudio del idioma mapuche, Mapucezugun ñi gvnezuam.* Florencio Varela: Editorial Xalkan.

Andrés Chandía. 2012. *Dungupeyem _alfa 2 _v0. 1: un prototipo de analizador morfológico para el mapudungun a través de transductores de estados finitos*. Ph.D. thesis, Tesis de Máster sin publicar, Universitat Pompeu Fabra, Barcelona.

Maria Rayen Catrileo Chiguailaf. 1972. *A tagmemic sketch of Mapuche grammar*. The University of Texas at El Paso.

Francesco Chiodi and Elisa Loncón. 1999. Crear nuevas palabras. *Editorial Pillan, Universidad de la Frontera, Corporación Nacional de Desarrollo Indígena, Temuco*.

Pascual Coña. 2019. *Kuyfi mapuche chumgechi ñi azmogekeel egün.* Editorial Genlol. In Mapuzugun.

Robert A Croese. 2014. Tiempo verbal en mapudungun. *Lenguas y Literaturas Indoamericanas*, (1).

Mingjun Duan, Carlos Fasola, Sai Krishna Rallabandi, Rodolfo Vega, Antonios Anastasopoulos, Lori Levin, and Alan W Black. 2020. A resource for computational experiments on mapudungun. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2872–2877, Marseille, France. European Language Resources Association.

Alexandra Espichán-Linares and Arturo Oncevay-Marcos. 2017. Language identification with scarce data: A case study from peru. In *Annual International Symposium on Information Management and Big Data*, pages 90–105. Springer.

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.

Katharina Kann, Jesus Manuel Mager Hois, Ivan 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.

Cristián Lagos. 2012. El mapudungun en santiago de chile: vitalidad y representaciones sociales en los mapuches urbanos. *RLA. Revista de lingüística teórica y aplicada*, 50(1):161–184.

Lorraine Levin, Rodolfo M Vega, Jaime G Carbonell, Ralf D Brown, Alon Lavie, Eliseo Cañulef, and Carolina Huenchullan. 2000. Data collection and language technologies for mapudungun.

Ariadna Font Llitjós. 2005. Developing a quechua-spanish machine translation system.

Elisa Loncon. 2010. Derechos educativos y lingüísticos de los pueblos indígenas de chile. *ISEES: Inclusión Social y Equidad en la Educación Superior*, (7):79–94.

Elisa Loncon Antileo. 2011. *Morfología y Aspectos del Mapudungun*. Universidad Autónoma Metropolitana, México, D.F.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69.

Manuel Manquilef. 1911. Comentarios del pueblo araucano (la faz social). In *Anales de la Universidad de Chile*, pages ág–393.

Manuel Manquilef. 1914. Comentarios del pueblo araucano ii, la jimnasia nacional (juegos, ejercicios i bailes). In *Anales de la Universidad de Chile*, 72, pages ág–239.

José Millalén, Pablo Marimán, Rodrigo Levil, and Sergio Caniuqueo. 2006. Escucha winka. *Santiago de Chile: LOM Ediciones*.

Christopher Moseley. 2012. *The UNESCO atlas of the world's languages in danger: Context and process*. World Oral Literature Project.

Viktor Naqill Gomez. 2016. Lengua y emancipación nacional.

Aldo Olate Vinet, Paula Alonqueo Boudon, and Jaqueline Caniguan Caniguan. 2013. Interactividad lingüística castellano/mapudungun de una comunidad rural bilingüe. *Alpha (Osorno)*, (37):265–284.

John Ortega and Krishnan Pillaipakkamnatt. 2018. Using morphemes from agglutinative languages like quechua and finnish to aid in low-resource translation. In *Proceedings of the AMTA 2018 workshop on technologies for MT of low resource languages (LoResMT 2018)*, pages 1–11.

César Pérez. 2015. *Diccionario Mapuzugun-Castellano*.

Ricardo Rosasa, María Isabel Larab, Victoria Espinozaa, María Paz Ramíreza, Felipe Porflitta, and Catalina Benaventea. Mapudungun mew: Software para la enseñanza del mapudungun en la escuela1.

Alex Rudnick. 2011. Towards cross-language word sense disambiguation for quechua. In *Proceedings of the Second Student Research Workshop associated with RANLP 2011*, pages 133–138.

Scott Sadowsky, Héctor Painequeo, Gastón Salamanca, and Heriberto Avelino. 2013. Mapudungun. *Journal of the International Phonetic Association*, 43(1):87–96.

Felipe Hasler Sandoval, Aldo Olate Vinet, and Guillermo Soto Vergara. 2020. Origen y desarrollo del sistema evidencial del mapudungun. *CÍCULO de Linguística Aplicada a la Comunicación*, 81:9.

Catharina Johanna Smeets. 1989. *A Mapuche grammar*. Ph.D. thesis, Rijksuniversiteit Leiden.

Marco Antonio Alvarado Troncoso. 2012. *Sistema para el Aprendizaje del Mapudungun. Incluyendo características de reconocimiento de voz y bot conversacional.* Ph.D. thesis, Pontificia Universidad Católica de Valparaíso.

uatv.cl. 2020. Organizaciones mapuche convocan a novena marcha por el mapuzugun.

Fernando Wittig. 2009. Desplazamiento y vigencia del mapudungún en chile: un análisis desde el discurso reflexivo de los hablantes urbanos. *RLA. Revista de lingüistíca teoríca y aplicada*, 47(2):135–155.

Fernando Zúñiga. 2006. *Mapudungun: El Habla Mapuche*. Centro de Estudios Públicos, Santiago, Chile.

A Notes on Mapuzugun

In this section, to understand the context and the need for this work, it will be explained how Mapugugun went from being a language of a million speakers in the 16th century to becoming, according to UNESCO, an endangered language today.

History of Mapuzugun The Mapuche people have their origin in the territory known as Wallmapu (which can be considered as the Mapuche Country (Millalén et al., 2006)). This territory ranges from Coquimbo to Chiloé, also including areas on the "other side" of the mountain range, such as Neuquén, in a vast area demarcated by the Río Negro. Throughout this territory different denominations for this people can be found, shareing many cultural aspects (Millalén et al., 2006). This is the area in which the scope of the language known as Mapuzugun can be framed – also in accordance with what the Spaniards defined at the end of the 16th century.

Mapuzugun, at its height, at the arrival of the Spaniards, was spoken by around a million people (Millalén et al., 2006). One of the first published books on this language is entitled "Art and Grammar of the General Language that runs throughout the Kingdom of Chile, with a Vocabulary and Confessionary" published in 1606 by Luis de Valdivia (Alvarado Pavez, 2020). In addition, toponyms with a clear Mapuzugun origin are still preserved, such as Huente Lauquen in the north, Puchuncaví, Curacaví, Pudahuel, Vitacura, with examples even in Puel Mapu (or what we know today as Argentina), and Chiloé in the south.

During the interaction of Mapuche with Spaniards during the Colony, the place of the Mapuzugun in all spheres of society can be appreciated, from the family, to international political relations, as were the Koyagtun (or Parliaments) mainly with the Spanish Empire, the Chilean and Argentine States, but also with the French, Dutch, and English. In all of these, the figure of the '*lenguaraz*' stood out, who acted as a translator to try to faithfully reproduce the ideas that were held in Mapuzugun to foreign representatives.

It was during the construction of the Chilean and Argentine national states in the 19th century -which initially did not include Mapuche territory- with the so-called "Campaign of the Desert" and "Pacification of Araucanía", when these states politically subjected the Mapuche people. Along with this, as part of the construction of the identities of Chile and Argentina, space was taken away from Mapuzugun and the Mapuche culture through schools and the church, which, through evangelization and punishment, denied indigenous identity along with their language.

Then at the beginning of the 20th century, after that territorial dispossession, there was a strong Mapuche migration to the most important cities of Chile, in search of better living conditions. This translated into cultural loss, often due to racism and discrimination. However, some efforts were made by the Mapuche themselves to maintain the culture and language, as shown by publications such as those by Coña (2019) and Manquilef (1911, 1914), which were written in both Mapuzugun and Spanish.

During the dictatorship and since the 90's, the Mapuche people began to have a greater political position. With this, the Mapuche language was recovered hand in hand with a recovery of identity in various areas, in addition to maintaining territories in which Mapuzugun is spoken as the first language. Today, according to the 2017 census, the majority of the Mapuche population would be in Santiago, but most do not speak or understand their language.

Today, there are various organizations that offer courses or tools that contribute to the revitalization of Mapuzugun. These instances have a milestone in a march that is organized during February, within the framework of the commemoration of the "International Mother Language Day" (uatv.cl, 2020), having, as a movement, important demands such as the officialization of Mapuzugun (Naqill Gomez, 2016).

Various sources estimate the number of Mapuzugun speakers to be between 100,000 and 300,000 (Bertin, 2016).[11] They constitute about 5 to 10% of the Mapuche population (1,745,147), who make up 9.9% of the total population of Chile (17,574,003).

According to UNESCO, a language is in danger when it is no longer used, when it is used in fewer areas and when it is no longer transmitted. From this it is stated that "about 90% of all languages "could be replaced by the dominant ones by the end of the 21st century". All this, added to insufficient documentation, generates that there are extinct or endangered languages. , which are

[11] https://news.un.org/es/story/2019/04/1454571

unrecoverable (Aronoff and Rees-Miller, 2020).

There are six degrees to define the state of danger of these languages. Within this classification is the Mapuzugun, referred to as Huilliche and, both in Chile and Argentina, as Mapuche, as can be seen in the Unesco Atlas (Moseley, 2012). They are in grade 1 ("In a critical situation", Huilliche), 2 ("Seriously in danger or threatened", Mapuche, Argentina) and 3 ("Clearly in danger or threatened", Mapuche, Chile).

But not only through UNESCO, research has been carried out on the state of the Mapuzugun. There are also various studies from the area of sociolinguistics to understand certain current language processes and their incorporation into public policies (Naqill Gomez, 2016; Loncon, 2010; Catrileo, 2017; Wittig, 2009; Lagos, 2012; Olate Vinet et al., 2013) [46].

Typological Notes Linguistically, Mapuzugun is defined as an agglutinative and polysynthetic language, which means that its expressions have a main root to which defined and distinguishable suffixes are added to form phrases. For example the word Kim mapuzuguyekümelleaiñ, which is explained in Table 2.1. Examples such as English, Chinese or Spanish are not in this category, and therefore the processing techniques used in those languages differ from the techniques that could be used for Mapuzugun.

Before colonization, Mapuzugun is described as a purely oral language. Today, until recently it was not formally taught or used by public and educational institutions in Chile. This has meant that it does not have a standardization of its writing or spelling. Today there are different ways of writing it and also, territorial orthographic variations, because in different regions there are phonetic differences for certain sounds and that translates, in general, into different writings. Today there are three main graphemaries to write Mapuzugun: Ragileo, created by Anselmo Raguileo in 1985; Unified, created by María Catrileo in 1989; and Azümchefe, created by Necul Painemal for CONADI (National Corporation for Indigenous Development) in 2008.

Among these graphemaries certain visible differences can be noted in Table 2.2. In the case of Ragileo, this grapheme uses only one grapheme per phoneme, and on certain occasions, the sounds associated with these graphemes do not correspond to those of Spanish, so it is a little more difficult to learn than the others. The Unified has a script more similar to Castilian. Although, although most of the graphemes are the same, there are phonemes that can be considered similar, but are not the same between Castilian and Mapuzugun. Finally, the Azümchefe is a kind of intermediate point, but it also presents difficulties and differences between graphemes and phonemes in Spanish. It is used by public institutions such as CONADI.

This lack of standardization of Mapuzugun brings complications to people who are studying the language and who only master a grapheme. This also affects the processing of Mapuzugun, since there would be inconsistencies when taking data from different sources or even from the same source, especially in topics such as automatic translation or semantic analysis, where the same word could have various forms and affect learning. some model. This probably affects the current lack of basic tools in this language.

In this direction, there are currently various works related purely to Mapuzugun linguistics: descriptions (Zúñiga, 2006; Smeets, 1989; Chiguailaf, 1972), but also specific academic articles on technical aspects of the language (Chiodi and Loncón, 1999; Olate Vinet et al., 2013; Sadowsky et al., 2013; Croese, 2014; Araya et al., 2019; Alvarado, 2019; Sandoval et al., 2020) and dictionaries (Catrileo, 2017).

A.1 Computational Work on Mapuzugun

Today there are various initiatives of computational linguistics on Mapuzugun. There is an orthographic normalizer and a morphological analyzer (Chandía, 2012), but it still has some errors derived from the fact that it directly applies a series of rules without analyzing the input it receives. These are aspects that could be improved. Another aspect that could be improved is that, currently, there is no possibility of choosing the output grapheme, restricting it to only one form of writing. This is inconvenient today that there is still no agreement on the standardization of writing. This implementation was made from a set of rules through regular expressions, with a finite state transducer, which have been released on the author's website. This author is also working on a prototype morphological analyzer and spell checker, based on Xerox finite state tools. There are also corpus exploitation interfaces annotated with these same tools, created in an interuniversity master's degree in Barcelona, (coordinated by the University of Barcelona) and an automatic Mapuzugun translator is being targeted. As he is in the process of doctoral work, the results of these tools have not yet been published, but they can be reviewed in his thesis proposal.

On the other hand, there was a project called AVENUE, in which the Universidad de la Frontera, the Intercultural Bilingual Education Program and the Institute of Language Technologies of Carnegie Mellon University (CMU) collaborated. The purpose of this project was to generate simple and low-cost translators, in addition to helping to preserve the Indo-American languages. This project resulted in four products: 1. In the first place, a graphemebook for the purposes of processing and computer development of the Project. 2. A 170-hour corpus that has been transcribed, but not fully revised. 3. A translation prototype consisting of a trained example-based translator. In addition, one based on transfer rules was worked on in parallel (both with Spanish as a pair). This prototype also has a morphological analyzer. After the Avenue project, CMU also worked on the automatic improvement of translations. 4. A spell checker that is said to contain an estimated 6,000,000 words, for OpenOffice. And that consists of two dictionaries, one for roots and the other for suffixes, which within OpenOffice's MySpell, correct a text in the typical way that the user is used to. This continues to have the limitation of the grapheme, in addition to not having the security that when writing a word in another grapheme it will convert it to the one used by the system.

In the educational field, there is software to learn Mapuzugun called MAPU from a job at the Pontificia Universidad Católica de Valparaíso that also includes voice recognition to control the application, which works correctly, but is not robust to pronunciation (Troncoso, 2012). In this work, it also refers to another Mapuzugun-to-Spanish voice-text translation prototype, based on recordings, and to a chatbot from the Pandora project.

In addition, the CEDETI of the Pontificia Universidad Católica, which is dedicated to working on technologies for integration, has language learning software called Mapudungun mew (Rosasa et al.).

A.2 The three orthographies currently used

See Figure 5 for a comparison.

Unificado	Ragileo
CH	C
D	Z
G	Q
L	B
Ll	J
N	H
Ng	G
Tr	X
T	-
Ü	V

Table 4: Differences and conversion between the Unificado and Ragileo orthographies.

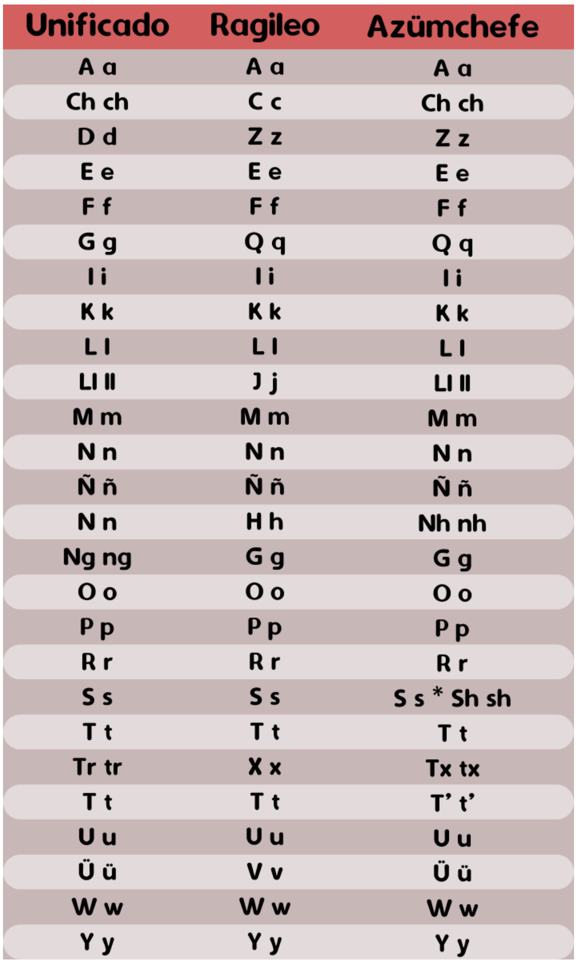
Unificado	Azümchefe
D	Z
G	Q
L	Lh
N	Nh
Ng	G
Tr	Tx
T	T’
S	Sh

Table 5: Differences and conversion between the Unificado and Azümchefe orthographies.

B Computational Work on South American Indigenous Languages

Mager et al. (2018) review the challenges for indigenous languages in America in terms of language technologies and NLP, which is also a review of the experiences that have been had for different languages throughout the continent. Beyond Mapuzugun, it also addresses languages such as Quechua, Nahuatl, Wixarika, Shipibo Konibo, Guaraní, among others. The challenges have to do mainly with the insufficient or not well developed corpora, translations, and morphological analyzers. In addition, experiences are named in the different common tasks for NLP.

Llitjós (2005) presents the most complete process of what would be the result of the AVENUE project, whose product was a Quechua - Spanish translator. This could not be completed for the Mapuzugun case, but there is a methodology with which it could continue. One can also see the use of Bayesian classifiers and K nearest neighbors (k-nearest neighbors, KNN) for disambiguation in



Unificado	Ragileo	Azümchefe
A a	A a	A a
Ch ch	C c	Ch ch
D d	Z z	Z z
E e	E e	E e
F f	F f	F f
G g	Q q	Q q
I i	I i	I i
K k	K k	K k
L l	L l	L l
Ll ll	J j	Ll ll
M m	M m	M m
N n	N n	N n
Ñ ñ	Ñ ñ	Ñ ñ
N n	H h	Nh nh
Ng ng	G g	G g
O o	O o	O o
P p	P p	P p
R r	R r	R r
S s	S s	S s * Sh sh
T t	T t	T t
Tr tr	X x	Tx tx
T t	T t	T’ t’
U u	U u	U u
Ü ü	V v	Ü ü
W w	W w	W w
Y y	Y y	Y y

Figure 5: Comparison of the three alphabets used by the Mapuche.

Ragileo	Azümchefe
C	Ch
B	Lh
J	Ll
H	Nh
S	Sh
X	Tx
-	T’
V	Ü

Table 6: Differences and conversion between the Ragileo and Azümchefe orthographies.

Quechua translation (Rudnick, 2011).

Also in Quechua, the improvement of morpheme recognition from its comparison with Finnish, due to the fact that they have similar structures, especially in the agglutination part (Ortega and Pillaipakkamnatt, 2018).

The closeness in typology also happens with other languages that are in Peru and the rest of the continent, such as Mexicanero, Nahuatl, Wixarika and Yorem Nokki (Kann et al., 2018). Or the Mohawk and Plains Cree (Arppe et al., 2016), from further north.

At the University of Limerick a thesis was developed on a morphological analyzer for the Mohawk case. This is done through finite states and their training from the language data (Assini, 2013).

Espichán-Linares and Oncevay-Marcos (2017) present a study of low-resource Peruvian languages. This is done from the construction of a vector space model for languages, from bigrams and trigrams, and a matrix from "term frequency - inverse document frequency" or (TF-IDF, for its acronym, in English). It is classified by sentences and a performance of over 96% is achieved in classification with support vector machine. Although these are good results, there is no way to know if it is exactly the orthography used or if it is just the closest.

Alva and Oncevay (2017) propose a corrector based on syllabification and characters for an agglutinating Peruvian language. This is done with graphs of syllables and characters from models extracted from the corpus. This method proposes three closest corrections for a misspelled word with distance metrics per character, also saving the previous corrections. This method is complete and takes into account the syllables and characters, which would be important in the case of orthographies which have subtle differences, as if they were spelling errors. Although the error can be improved (76%), it could be a solution for the normalizer, if it is extended to multiple languages (or in this case orthographies).