# The Corpus of Historical Mapudungun

# The Corpus of Historical Mapudungun:
# Morpho-phonological parsing and the history of a Native American language

Benjamin Molineaux

(Accepted for the journal *Corpora* 18:2, due mid 2023)

## Abstract

The *Corpus of Historical Mapudungun* (CHM), which we present here, is a lemmatised, part-of-speech and grapho-phonologically parsed collection of texts in the ancestral language of the Mapuche people. This article gives an overview of the corpus materials (spanning 1606–1930), their processing and search capabilities. The TEI XML tags at the word and morpheme levels are shown to be suitable to account for the abundant agglutinative morphology of the language. The advantages of visualising sound-spelling equivalences across the various spelling systems in the corpus are also emphasised. Some uses and limitations of the corpus are surveyed too, with a particular emphasis on the contribution of typologically diverse languages to understanding language change and the importance of making heritage materials available to native speaker communities for revitalisation purposes.

## 1. Introduction

Minority, non-European languages – such as indigenous South American ones – are unevenly represented in the literature on language change. This not only narrows our view of the historical interaction of peoples and languages predating European expansion, but also limits our understanding of linguistic change as a whole. In the absence of the hundreds of years of philological study available for Old World languages, digital methods emerge as the best means available for systematically compiling and exploring the existing data for language change in the New World. This said, carefully tagged, historically-oriented corpora for Native American languages are still few and far between, despite the general growth in Digital Humanities scholarship throughout the region.

The *Corpus of Historical Mapudungun* (CHM – Molineaux & Karaiskos 2021), which we present here, provides a pioneering, digitally-based resource for exploring the history of a language of the Americas.[1] The corpus documents the first 324-years (1606-1930) of textual history for Mapudungun (ARN), the ancestral language of the Mapuche people of the Southern Cone of the Americas, also known as *Mapuche*, *Mapuchedungun*, *Mapuzungun* and *Araucanian* (the latter term now dispreferred). Through lemmatisation and tagging at the part-of-speech, morphological and phonic levels, the CHM allows users to draw links between individual related forms over time, effectively writing the morphological and phonological history of the language from the bottom up. Such research represents a qualitative leap in the study of Mapudungun, while at the same time laying the groundwork for historical corpus methods to be applied to minority languages more broadly.

With 415 years of documentation, Mapudungun is an excellent candidate for a corpus-based historical account. While some work has been done on change in the language, the field

[1] Important historical materials Mesoamerican languages are already available as digital resources, including for Classical Nahuatl (https://aclanthology.org/L16-1666/), Colonial Zapotec (https://ticha.haverford.edu/en/), and Mayan language (https://mayawoerterbuch.de/?lang=en). However, we believe the CHM is a first linguistically tagged corpus of historical Native American texts.

remains underexplored. The language, furthermore, has no agreed family membership,[2] and is often treated as an isolate. As a result, comparative work is bound to be unfruitful. There is, nevertheless, a good baseline for historical exploration, since present-day Mapudungun is fairly well described, with three linguistically well-informed grammars (Salas 1992, Zúñiga 2006, Smeets 2008), alongside numerous studies on word and sound structure.

There are also theoretically interesting reasons for working on Mapudungun historical linguistics. Features such as nominal incorporation, verb serialisation, reduplication, and abundant slot-based verbal suffixation raise interesting questions about the diachrony of units of sound and meaning which cannot easily be probed by better-studied, yet typologically distinct languages. Indeed, Mapudungun unambiguously fits all major criteria for polysynthesis, including a high morpheme-to-word ratio (in verbal and deverbal elements); obligatory, polypersonal head marking, and productive non-inflectional concatenation (see Zúñiga 2017, and Molineaux *in press b*, for an overview). Further to this, Mapudungun displays canonical features of agglutination: close mapping between individual morphemes and single semantic or syntactic features and clear-cut morphological boundaries (cf. e.g. Comrie 1989:43). These facts (exemplified in 1) are key both to our analysis of language change in the textual record and to the practicalities of corpus design.

(1) *vamtipalduamqueymn* (Valdivia 1621) [3]
fem-tripa-l -duam-ke -e -y -m-ün (standardised spelling)
this-exit- CAUS-DESID -HAB-3.A-IND-2 -P
'they$_{S/P}$ usually want to make you$_{P}$ leave here'

An additional reason for focusing on Mapudungun is that this type of scholarship could have a real impact on maintenance and revitalisation. Although the language is endangered, it has a fighting chance. Mapudungun endangerment is not due to particularly small numbers of speakers – estimated at about 250,000, with varying degrees of proficiency (Zúñiga & Olate 2017) –, but rather to poor transmission in the face of Spanish bilingualism, a long history of marginalisation/invisibilisation, and both territorial and cultural loss (see Caniuqueo 2006, Gundermann 2014). Given ongoing lexical and morphological impoverishment (Chiodi & Loncon 1999), a well-structured repository of root and affixal morphemes in their historical usage will, we hope, provide key support for the revitalisation of the language's lexicon and native word-formation strategies (cf. Villena 2019).

## 2. Materials

Most early Native American writing comes through missionary efforts to evangelise the local population via sermons, catechetical and confessional texts in their languages, as well through grammars and vocabularies used to train missionaries (Zwartjes, 2000, Hanks 2010). This means that our earliest witnesses are rarely older than the 16th century. These texts bear the marks of second-language usage, representing in structure, topic and genre, the colonizers'

[2] For an overview of possible genetic affiliations for Mapudungun see Pache (2014).
[3] The first line here represents the original spelling, the second a standardised spelling based on the Catrileo Alphabet (see §3.2). Gloss abbreviations are as follows: A=agent; CAUS=causative, DESID=desiderative, FUT=future; IMP=imperative; IND=indicative; NEG=negative; NMLZ=nominaliser; P=plural; S=singular; 1,2,3=first, second, third person; '>' indicates that the preceding person is the agent and the following the patient.

worldview, objectives and biases. Much has been done in the tradition of missionary linguistics to try to tease these influences apart from the native linguistic repertoire, however this work is not usually integrated with the study of other non-missionary sources for said languages. The CHM – which currently contains nearly 150,000 words – attempts to provide tools in order to facilitate this integration, looking at the broadest range of sources available across time, space and genre.

Following Villena (2017) Mapudungun texts can be divided into three main periods, as shown in Figure 1, which also overlays the dates covered by the CHM.

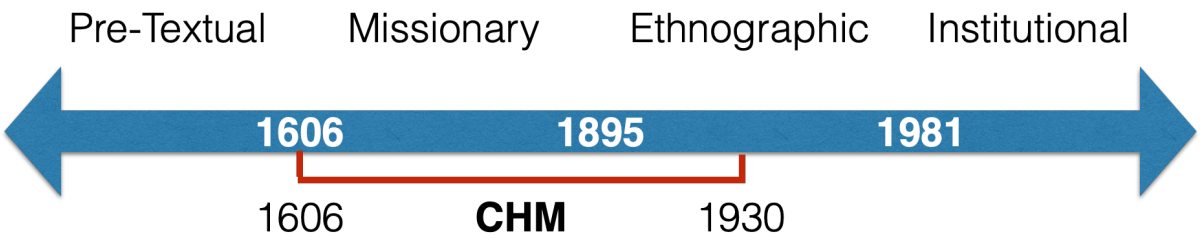



**Figure 1:** Periodization for Mapudungun texts, based on Villena (2017), with CHM dates overlain.

The textual record for Mapudungun begins with the 1606 publication of Spanish Jesuit Luis de Valdivia's *Arte y Gramática* ('art and grammar'– in Spanish). Two additional Jesuit grammars were published in the *Missionary* period: the Catalan Andrés Febrés' (1765 – in Spanish) and the Westphalian Bernard Havestadt's (1777 – in Latin). In all cases, the grammatical description of the language was followed by a vocabulary as well as Catholic doctrinal texts. Febrés and Havestadt also include secular texts, the first compiling formal speeches (*koyagtun*) and dialogues (*nütramkan*) – both fairly artificial in nature and clearly Christianised – and the second providing a lengthy translation of Fr. F. Pomey's *Indiculus Universalis* as well as a series of songs. Amongst the latter are a set of healer (*machi*) songs, the first genuinely Mapuche texts we know of. It is both these religious and secular texts that have been included in the CHM, alongside Valdivia's substantial *Sermón en Lengua de Chile* ('Sermon in the language of Chile') from 1621. Finally, three important, if brief, non-missionary texts are also found in the period: a 1640s Mapudungun-Latin vocabulary collected by Dutch explorers (cf. Schuller 1907), and the grammatical sketches by the Mancunian surgeon-turned-Jesuit, Thomas Falkner (1774 – in English), and the Argentinian Colonel Federico Barbará (1879 – in Spanish).

The publication of German-born linguist Rudolf Lenz's *Estudios Araucanos* (Araucanian Studies, 1895–1897) is a watershed in the language's textual history. Lenz not only made original transcriptions using the latest developments in phonetic science, but he elicited a body of dialectally diverse, culturally relevant narratives, descriptions and songs. This *ethnographic* approach to textual production was soon taken up by the Capuchin priest Félix de Augusta and his close collaborator and brother of the cloth, Sigfriedo de Fraunhäusel who produced a grammar (1903), compendium of traditional texts (1910) and dictionary (1916). Another important early 20th century partnership was that of Thomas de Guevara, the headteacher of the Secondary School of the city of Temuco and his extraordinary mentee, Manuel Manquilef, who went on to become the first Mapuche member of the Chilean congress.[4] With substantial

[4] On Manquilef and other early Mapuche intellectuals see the *Mapping Intercultural Conversations* website: http://interculturalconversations.com/#/.

input from Manquilef, Guevara published two ethnographic studies including Spanish-Mapudungun texts (1911, 1913). Supported by Lenz, Manquilef himself penned two studies using his mother tongue (1911, 1914). The most emblematic text within the ethnographic approach, however, is probably the lengthy autobiography of Chief Pascual Coña, transcribed by another Franciscan: Ernesto de Mösbach (1930).

Despite this late-19th-century change of focus, most of these studies viewed Mapuche language and culture as curiosities which required documentation before their inevitable doom under the influence of more civilised (read: *European*) peoples. Still, through the materials, we hear the voices of Mapuche themselves for the first time, as *nütramkafe* (narrators), *wewpife* (orators), *ülkantufe* (singers) and linguistic consultants. These voices, furthermore, come from different geographical areas (see Figure 2) and named individuals. Among these, Manquilef and Coña stand out, of course, but remarkable testimonies also come from Lenz's collaborators, Kallfün (Segundo Jara) and Domingo Kintupüray, as well as Augusta and Fraunhäusl's parishioners: Domingo Wenuñamko, Carmen Painemilla, Pascual Painemilla Ñamkucheu and José Francisco Kolün, amongst many more.

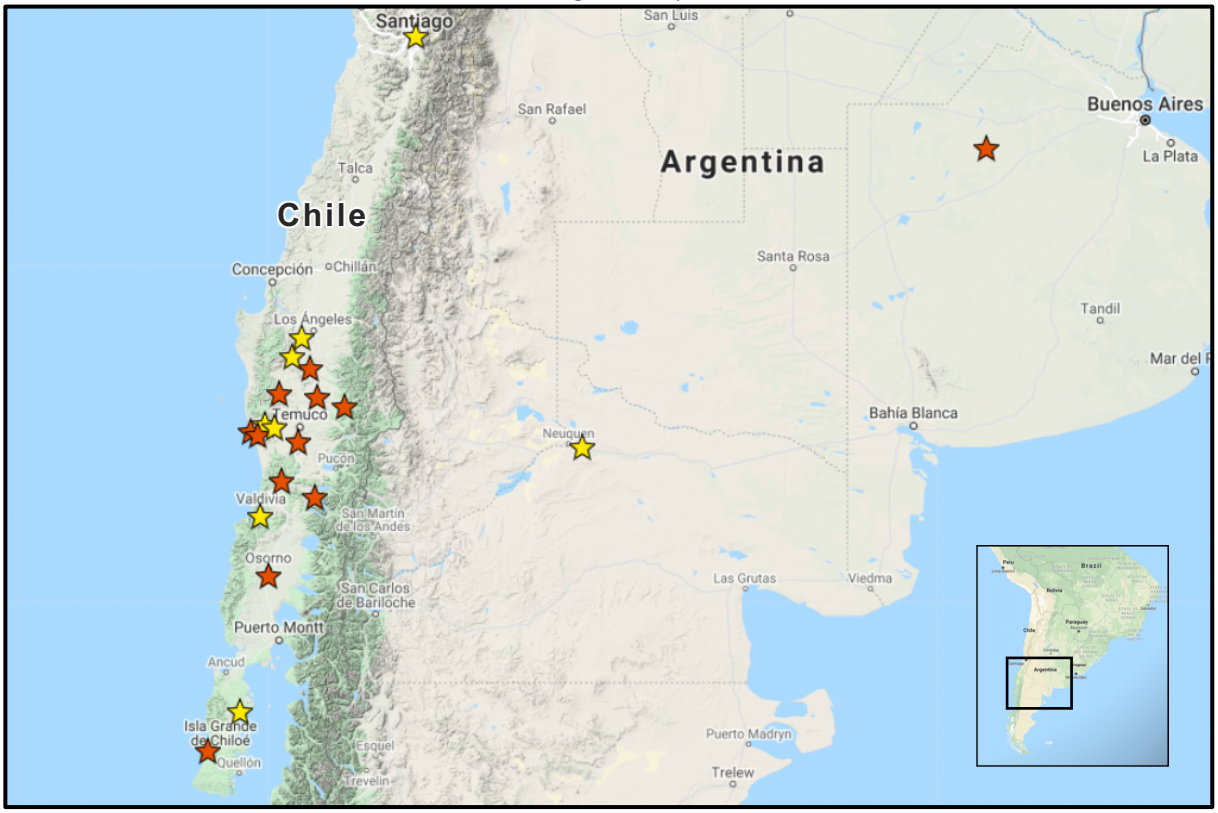



**Figure 2:** Locations associated to CHM texts. Yellow and orange markers represent missionary and ethnographic-period texts, respectively.

## 3. Corpus Building

### 3.1 From paper to digital text:

Despite reports of their existence (Medina 1894, Schuller 1907), no substantial older Mapudungun texts survive in manuscript form. As a result, all the materials currently in the CHM are based on printed editions. In most cases, image-based PDFs were available (see primary-source references) from whence machine-readable text could be produced. This was accomplished via OCR using Google Cloud Vision on a beta version of the *Digital Humanities*

*Dashboard* (Tarpley 2018). The OCR outputs were then meticulously hand-checked, a process that was particularly labour-intensive for the 17th and 18th century materials, which include non-standard page formatting and a number of diacritic devices to convey differences with European sound systems. Having concluded such checks, the materials were added to XML files, using TEI standards (2021). This format not only allows tagging at different levels of linguistic structure further down the corpus-building pipeline, but provided space for metadata entry.[5] Thus, where available, each file was tagged with information about its title, year of gathering or/and publication, compiler, speaker/orator/consultant, location (provenance of the speaker/location of mission) and genre. Writing in languages other than Mapudungun — such as notes or parallel translations — were also transcribed and made available in HTML versions of the texts.[6]

### 3.2 Lemmatisation

As the objective of the corpus is to provide a view into the synchrony and diachrony of lexical, morphological and phonological features, texts were parsed at all three of these levels. The first stage of the process — lemmatisation — breaks up the texts into individual lexical elements, assigning them a part-of-speech (POS) category and a single identifiable, orthographically-consistent label across texts. Spellings follow the conventions of the Catrileo Alphabet (CA – Croese *et al* 1978)[7] and are listed alongside English and Spanish translations. Throughout, however, the original, non-standardised spelling is kept as the main, marked-up text, allowing our analysis to be fully transparent, while at the same time preserving the surface appearance of these heritage materials.

Further homogeneity and comparability across sources is achieved by using an uninflected version of each word as their key label: the lemma. These lemmas are given in a shape that matches the entries in Augusta's *Diccionario Araucano* (1916), if available. Such labels can be seen in (2), representing a schematic of the tags, and (3) representing a sample word-level (<w>) XML implementation. Lemmas for finite verbs (see 2a,b) are given with the ending *-(ü)n* representing a multifunctional element that is akin to an infinitive or to a gerund, like English *-ing* (cf. Zúñiga 2006:141-3). This is because such verbs cannot stand alone without a finite inflection. Non-verbal categories are mostly uninflected, so they are presented unmodified (the main exception to this is the adjectival pluraliser *-ke* – see 8c below).

(2)

| | FORM | CA | LEMMA | POS | ENGLISH | SPANISH |
|---|---|---|---|---|---|---|
| a. | ⟨kimaqen⟩ | kimagen | *kimün* | V | 'know' | 'saber/conocer' |
| b. | ⟨kimdəŋulai⟩ | kimdüngulay | *kimdüngun* | V | 'know to speak' | 'saber hablar' |
| c. | ⟨quimn⟩ | kimün | *kimün* | N | 'knowledge' | 'sabiduría/conocimiento' |

(3) Sample XML:

[5] Overwhelmingly, the CHM uses the standard TEI attributes as recommended. The main exceptions to this are @corresp and @sameAs, which are normally pointer elements. Here, they represent English and Spanish translations of the lemma (see examples 2 and 3).

[6] A full set of the HTML versions of the texts, alongside descriptions of the materials and the image-based PDFs are available in the CHM Source Material website at: https://benmolineaux.github.io/bookshelf/en/.

[7] This alphabet is known elsewhere as the 'Unified' alphabet. Here we choose to acknowledge the Mapuche linguist who did most to develop and promote it: Dr. María Catrileo.

```
<w lemma="kimün" lemmaRef="kimagen" pos="V" corresp="know" sameAs="saber/conocer"> kimaqen</w>
```

In practical terms, the original spelling forms were first fed through a simple search-and-replace orthographic CA-remapping XSLT, which took into account the spelling conventions of each text. The output of this process (XML: *lemmaRef*) was used as a temporary lemma on which inflected forms could be matched throughout the corpus texts. The lemma proper was added after morphological parsing (see §3.3) by copying over the base forms of the root and derivational suffixes and adding the multifunctional *-(ü)n* to finite verbs. POS and translation information was encoded independently, by hand. Having done this for one form, the POS and translation tags were expanded to matching CA-remapped items in successive texts, which in turn were subject to manual annotation for the unmatched items. Iterations of the procedure — conducted via bespoke XSLT scripts — led to nearly 70% of words in a given text being automatically lemmatised.

### 3.3 Morphological parsing:

In this second stage of tagging, words were broken up into individual morphemes. These were tagged with an invariant base form, a type (root, prefix or suffix) and a meaning (for roots) or function (for prefixes/suffixes). Here the process was facilitated by the generally agglutinative morphology of the language, where each morpheme tends to be both phonologically distinct and carry a single meaning. Of course, the polysynthetic nature of the language (cf. Zúñiga 2017) does create some extreme morphological complexity word-internally, with individual words in the corpus containing up to eight distinct morphemes.

As in the case of lemmatisation, morphological parsing was conducted by hand for the first text, and then expanded by XSLT scripts onto successive texts, with intervening rounds of hand-correction. The output of the process is schematised in the samples in (4), with morpheme-level (<m>) XML tags provided in (5). Where possible, the gloss labels follow the Leipzig Glossing Rules (https://www.eva.mpg.de/lingua/resources/glossing-rules.php), with a view to facilitating consistency and comparability.

(4)

| | FORM | MORPHEMES |
|---|---|---|
| a. | ⟨kim-a-qen ⟩ | know-FUT-IND.2.S>1.S |
| b. | ⟨kim-dəŋu-la-i⟩ | know- thing/news/word/thought/speech-NEG-IND.3 |
| c. | ⟨quim-n⟩ | know-NMLZ |

(5) Sample XML:

```
<w lemma="kimdüngun" lemmaRef="kimdüngulay" pos="V" corresp="know to say" sameAs="saber decir">
  <m baseForm="kim" type=”root” corresp="know" sameAs="saber/conocer">kim</m>
  <m baseForm="düngu" type="root" corresp="thing/news/word/thought/speech"
       sameAs="cosa/noticia/palabra/pensamiento/habla">dəŋu</m>
  <m baseForm="la" type="sfx" corresp="neg" sameAs="neg">la</m>
  <m baseForm="iy" type="sfx" corresp="ind3" sameAs="ind3">qen</m>
</w>
```

### 3.4 Grapho-phonological parsing

The final stage of tagging is grapho-phonological parsing (Kopaczyk et al. 2018), which entails providing IPA-based sound values for each word (as in 7), following a list of spelling-based rules

for each text. Reconstructions of the phonic equivalences of the spellings in the CHM texts are based on a linguistically-informed interpretation of the authors' and compilers' own declared spelling-sound mappings, a process which is not devoid of challenges and pitfalls (see Molineaux *in press a*). The results effectively reconstruct the phonic structure for each text, such that it can be compared with others from different periods and locations, helping to map phonological variation and change from the bottom up. The XML structure for spelling-sound mappings is given in (8), detailing the character (<c>) element and its attributes.

(8)

| | FORM | SOUND | LEMMA | ENGLISH | SPANISH | SOURCE |
|---|---|---|---|---|---|---|
| a. | ⟨huera⟩ | [weɻa] | weda | 'bad' | 'malo(a)/mal' | 1621 Sermones |
| b. | ⟨wedake⟩ | [weθake] | weda | 'bad-P' | 'malo(a)/mal-P' | 1910 Lecturas |
| c. | ⟨weshá⟩ | [weˈʃa] | weda | 'bad' | 'malo(a)/mal' | 1930 Coña |

(9) XML: <m><c corresp="w">hu</c><c corresp="e">e</c><c corresp="ɻ">r</c><c corresp="ˈa">a</c></m>

## 4. Searching the corpus

### 4.1 Search domains

The CHM's front end is a web-based interface, tailor-made using the jQuery JavaScript library. It provides a number of search options (in both English and Spanish) performed directly across the tagged XML corpus documents. These searches can be conducted at the main three levels of tagging (word, morpheme and sound/spelling), as well as allowing users to correlate these features across texts and with relevant non-linguistic metadata (see Fig 3).

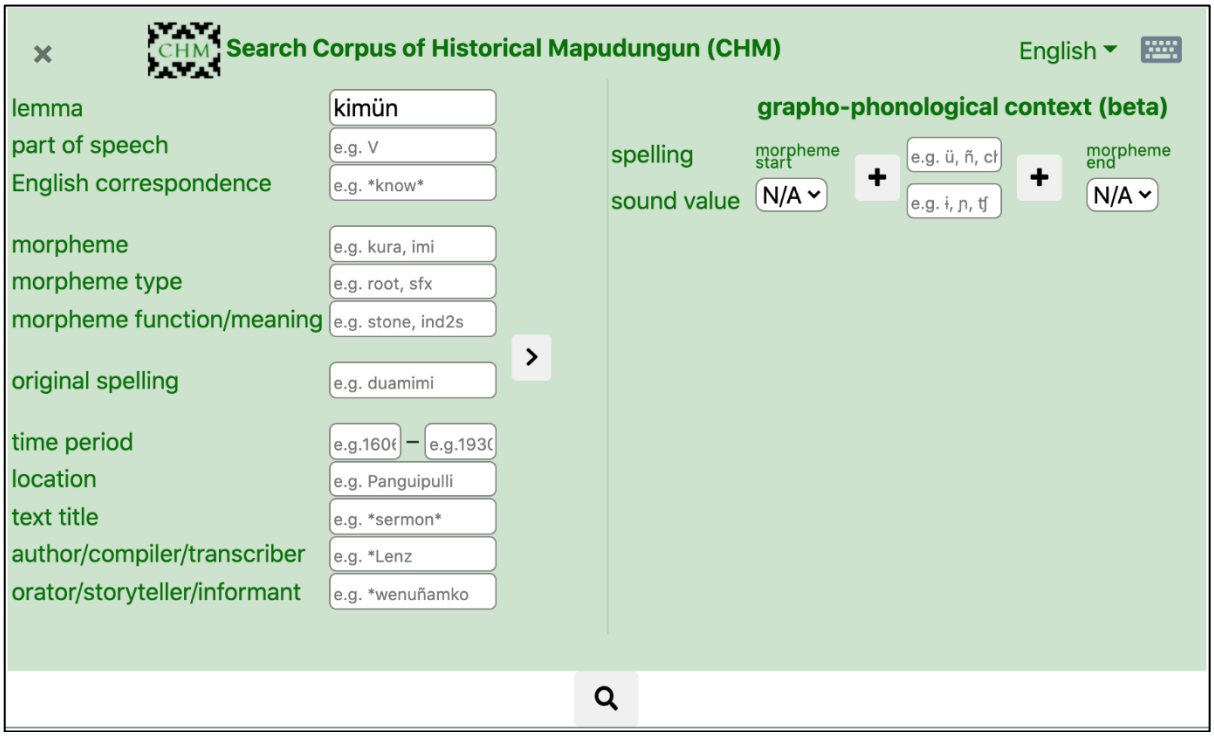



**Figure 3:** CHM search form, displaying lemma, morpheme, spelling, and metadata (left) and grapho-phonological mappings (right).

Searching by lemma and POS: The search box for 'lemma' allows users to find Mapudungun words throughout the available corpus, based on their CA spelling or their English/Spanish translations. It is also possible to restrict these searches to particular parts of speech, or simply to search for all the elements in a particular POS by entering the relevant abbreviation (e.g. V= verb; IJ= interjection).

Searching by morpheme: Since the CHM is morphologically parsed, users can search the texts by individual roots, prefixes and suffixes in the corpus. As with lemmas, the morphemes follow the basic form given in Augusta's dictionary (1916), adapted to CA spellings. It is possible to restrict these searches by morpheme type, that is, by whether they are a root, prefix or suffix. Finally, it is also possible to search by the meaning or function of the morpheme, which in the case of roots is an English/Spanish translation (e.g. *kura* = 'stone'), while in the case of suffixes is a gloss acronym (i.e. *-imi* = IND2S).

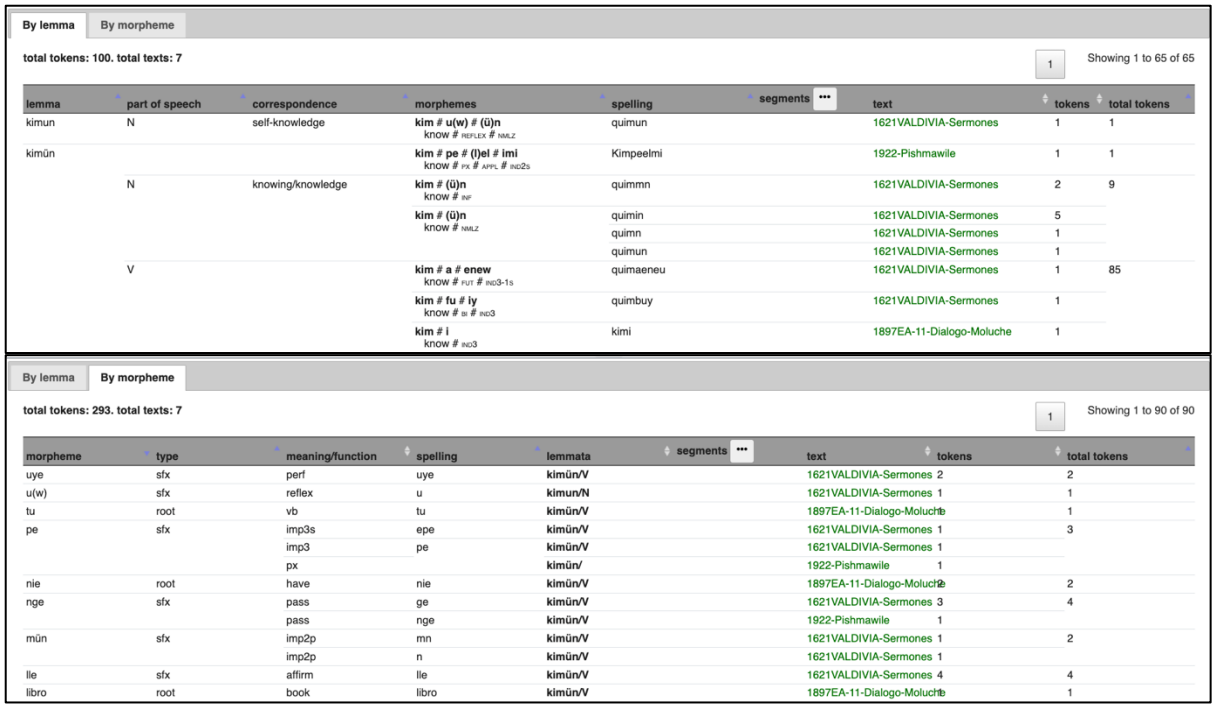
Searching by original spelling: this feature allows for the exploration of particular strings of letters corresponding to a particular texts' spelling. Such searches take into account the actual word in context, so they include all inflectional morphology.

Searching by graphemes and sound values: A separate search focusing on individual segments and their sound-spelling mappings is also available. Users can search by individual graphemes or IPA-based phones or by a series of these. It is also possible to specify whether the target segment(s) are at the start or end of the morpheme, helping users explore phonotactic organisation.

Metadata: Searches can also be restricted to texts with specific characteristics following the XML metadata.

### 4.2 Outputs

Results are presented in downloadable tables, organised either by lemma or morpheme (Figure 4). Columns contain the tags for each of the items matching the search terms with counts by text and lemma/morpheme. Individual attestations are hyperlinked to their source texts, available as HTML files allowing quick view of the different tags as pop-up bubbles (Figure 5).



By lemma | By morpheme

total tokens: 100. total texts: 7 — Showing 1 to 65 of 65

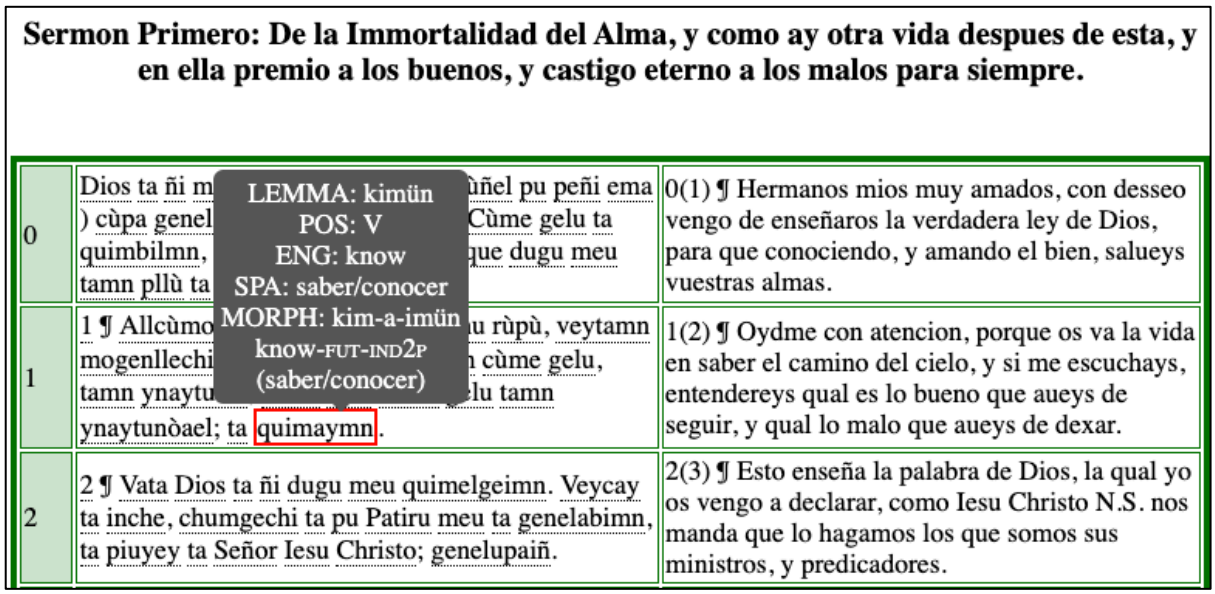
| lemma | part of speech | correspondence | morphemes | spelling | segments | text | tokens | total tokens |
|---|---|---|---|---|---|---|---|---|
| kimun | N | self-knowledge | **kim # u(w) # (ü)n** know # REFLEX # NMLZ | quimun | | 1621VALDIVIA-Sermones | 1 | 1 |
| kimün | | | **kim # pe # (l)el # imi** know # PX # APPL # IND2S | Kimpeelmi | | 1922-Pishmawile | 1 | 1 |
| | N | knowing/knowledge | **kim # (ü)n** know # INF | quimmn | | 1621VALDIVIA-Sermones | 2 | 9 |
| | | | **kim # (ü)n** know # NMLZ | quimin | | 1621VALDIVIA-Sermones | 5 | |
| | | | | quimn | | 1621VALDIVIA-Sermones | 1 | |
| | | | | quimun | | 1621VALDIVIA-Sermones | 1 | |
| | V | | **kim # a # enew** know # FUT # IND3-1S | quimaeneu | | 1621VALDIVIA-Sermones | 1 | 85 |
| | | | **kim # fu # iy** know # SI # IND3 | quimbuy | | 1621VALDIVIA-Sermones | 1 | |
| | | | **kim # i** know # IND3 | kimi | | 1897EA-11-Dialogo-Moluche | 1 | |

By lemma | By morpheme

total tokens: 293. total texts: 7 — Showing 1 to 90 of 90

| morpheme | type | meaning/function | spelling | lemmata | segments | text | tokens | total tokens |
|---|---|---|---|---|---|---|---|---|
| uye | sfx | perf | uye | **kimün/V** | | 1621VALDIVIA-Sermones | 2 | 2 |
| u(w) | sfx | reflex | u | **kimun/N** | | 1621VALDIVIA-Sermones | 1 | 1 |
| tu | root | vb | tu | **kimün/V** | | 1897EA-11-Dialogo-Moluche | 1 | 1 |
| pe | sfx | imp3s | epe | **kimün/V** | | 1621VALDIVIA-Sermones | 1 | 3 |
| | | imp3 | pe | **kimün/V** | | 1621VALDIVIA-Sermones | 1 | |
| | | px | | **kimün/** | | 1922-Pishmawile | 1 | |
| nie | root | have | nie | **kimün/V** | | 1897EA-11-Dialogo-Moluche | 2 | 2 |
| nge | sfx | pass | ge | **kimün/V** | | 1621VALDIVIA-Sermones | 3 | 4 |
| | | pass | nge | **kimün/V** | | 1922-Pishmawile | 1 | |
| mün | sfx | imp2p | mn | **kimün/V** | | 1621VALDIVIA-Sermones | 1 | 2 |
| | | imp2p | n | **kimün/V** | | 1621VALDIVIA-Sermones | 1 | |
| lle | sfx | affirm | lle | **kimün/V** | | 1621VALDIVIA-Sermones | 4 | 4 |
| libro | root | book | libro | **kimün/V** | | 1897EA-11-Dialogo-Moluche | 1 | 1 |

Figure 4: Partial results (in table form) organised by lemma (above) and morpheme (below) using the bespoke CHM search facilities



**Sermon Primero: De la Immortalidad del Alma, y como ay otra vida despues de esta, y en ella premio a los buenos, y castigo eterno a los malos para siempre.**

| | | |
|---|---|---|
| 0 | Dios ta ñi m… …ùñel pu peñi ema ) cùpa genel… …Cùme gelu ta quimbilmn,… …que dugu meu tamn pllù ta… | 0(1) ¶ Hermanos mios muy amados, con desseo vengo de enseñaros la verdadera ley de Dios, para que conociendo, y amando el bien, salueys vuestras almas. |
| 1 | 1 ¶ Allcùmo… …u rùpù, veytamn mogenllechi… …n cùme gelu, tamn ynaytu… …lu tamn ynaytunòael; ta quimaymn. | 1(2) ¶ Oydme con atencion, porque os va la vida en saber el camino del cielo, y si me escuchays, entendereys qual es lo bueno que aueys de seguir, y qual lo malo que aueys de dexar. |
| 2 | 2 ¶ Vata Dios ta ñi dugu meu quimelgeimn. Veycay ta inche, chumgechi ta pu Patiru meu ta genelabimn, ta piuyey ta Señor Iesu Christo; genelupaiñ. | 2(3) ¶ Esto enseña la palabra de Dios, la qual yo os vengo a declarar, como Iesu Christo N.S. nos manda que lo hagamos los que somos sus ministros, y predicadores. |

LEMMA: kimün
POS: V
ENG: know
SPA: saber/conocer
MORPH: kim-a-imün
know-FUT-IND2P
(saber/conocer)

Figure 5: Sample HTML text from the CHM (Valdivia 1621) with pop-up tagging

In the case of the grapho-phonological mapping, results are available in table form as well as in a specialised visualisation tool originally developed for Medieval Scots sound-spelling mappings (see Kopaczyk et al. 2018). This resource allows a view of the one-to-many and many-to-one relationship between spellings and sounds across the corpus (cf. Figure 5), thus helping to pinpoint its variation across time and space.

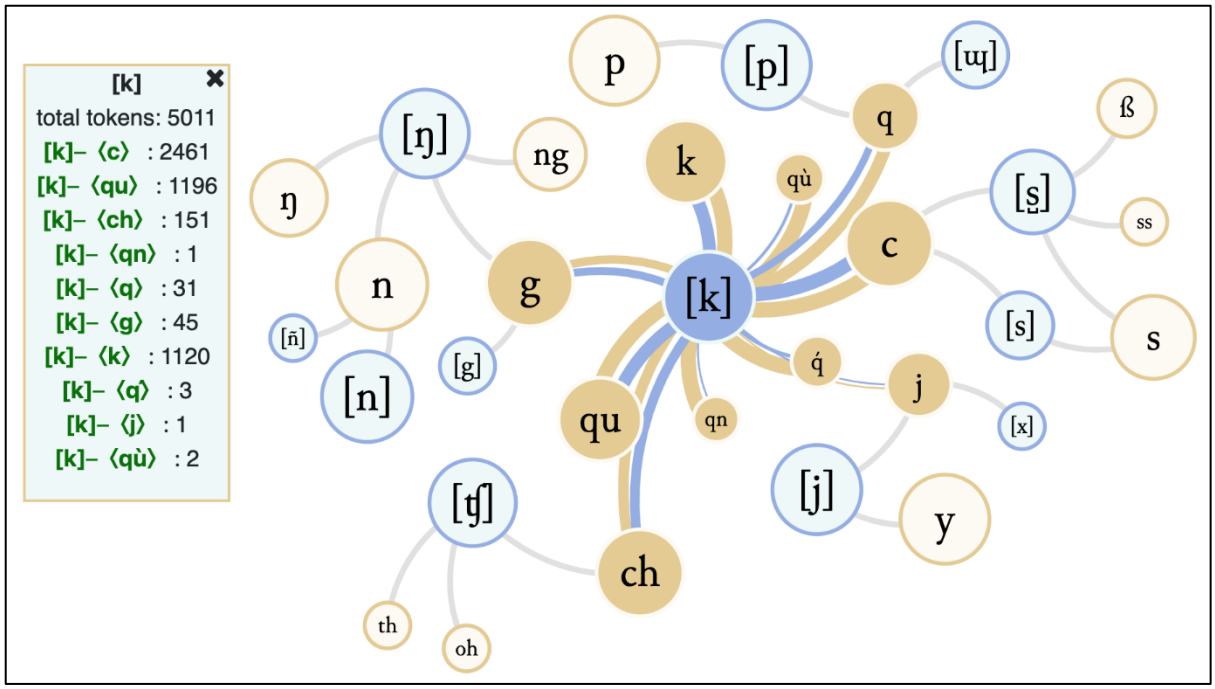



**Figure 6:** Graphemic substitution set for [k] across the CHM in the grapho-phonological visualisation tool

## 5. Conclusions:

The CHM represents a first richly-tagged, freely-available corpus of historical texts from a Native American language. Its construction and search capabilities make extensive use of the TEI linguistic tags, which show themselves to be particularly well suited for representing certain typological features of Mapudungun, such as the language's concatenative, polysynthetic, agglutinative structure. The tagging structure also allows users access to digital versions of texts that are, on the surface, true to the original, with no spelling or grammatical standardisation, while at the same time offering a unifying orthography (through lemmatisation) and careful analysis of the morphology and phonology of the language (through morpho-phonological parsing). The approach thus combines fidelity (and respect for Mapuche heritage materials) and a flexible, powerful search capacity. While the front end of the CHM is a bespoke tool, its underlying XML structure can be easily reused for further purposes (e.g. Syntactic parsing) or reproduced for application to other minority languages.

The design features and usability of the CHM should allow studies of particular patterns in the word and sound structure of the language to be analysed both from a qualitative and quantitative perspective. An example of this is my own recent (in press a) analysis of the diachrony of the Mapudungun dental-alveolar contrasts. There I extract all clear instances of the two categories and show that their incidence across individual lexemes, morphemes and periods is relatively stable, despite the contrast's low functional load – a fact that is surprising, given the relative typological rarity of the contrast. The same kind of analysis could be performed for particular sound sequences, for the productive combination and relative order of morphemes, or for the incidence of particular lexical items across, contexts, periods and genre.

While the CHM texts are gathered from all major language areas (Figure 2) and relevant periods (Figure 1), they inevitably remain imbalanced in nature — both temporally and spatially

—, as tends to be the nature of historical corpora. These imbalances, as well as the second-language nature of many of the texts, must be taken into account for both quantitative and qualitative studies.

Another important limitation that must be considered when using the CHM or planning the building of further resources of the type, is that the tagging process is neither neutral nor devoid of technical challenges. While we have attempted a relatively impartial reading of the facts of Mapudungun, these analyses are usually based on contemporary varieties, and researchers do not always agree on their interpretation. These disagreements may be either on the function of a particular suffix or particle, or the sound value of a particular grapheme. It is with these discrepancies in mind that we are keen for the scholarly community to engage with the materials. As for the building of the resource, we must emphasize that this is not a pipeline that is particularly easy to automate. Indeed, attempts at using NLP tools to facilitate parsing of languages with complex morphology like that of Mapudungun may be successful, but if we add in variation in orthography, region and time, results are fairly poor, yielding little accuracy for new, unseen lexical items (Mager et al 2018).

These caveats aside, the CHM paves the way for the application of digital methods to the history of minority, non-standard languages, creating transferable tools, and foregrounding under-studied typological features. Such outcomes will broaden our understanding of language change overall. Locally, we believe the materials to be particularly relevant given the renewed interest in the language by ethnically Mapuche people (Rojas et al 2016), who are often returning to their ancestral tongue after one or more generations of dormancy. We respectfully put forth these materials in the hope that they will provide teachers and learners of Mapudungun with a repository of words in their historical usage and forms, thus supporting revitalisation efforts.

## Primary sources:


Augusta, F. de. 1903. *Gramática araucana*. Valdivia: Imprenta Central J. Lampert. http://www.memoriachilena.gob.cl/602/w3-article-8186.html

Augusta, F. de 1910. *Lecturas araucanas*. Padre Las Casas: San Francisco. http://www.memoriachilena.gob.cl/602/w3-article-9601.html

Augusta, F. de 1916. *Diccionario Araucano-Español y Español-Araucano*. Santiago: Imprenta Universitaria. https://archive.org/details/diccionarioarauc01fluoft

Barbará, F. 1876. *Manual ó vocabulario de la lengua pampa*. Buenos Aires: Imprenta de Mayo. https://archive.org/details/manualvocabulari00barb

Faulkner, T. 1774. *A description of Patagonia*. Hereford: C. Pugh. https://archive.org/details/descriptionofpat01falk

Febrés, A. 1765. *Arte de la Lengua General del Reyno de Chile*. Lima: Calle de la Encarnación. http://www.memoriachilena.gob.cl/602/w3-article-8486.html

Guevara, T. 1911. *Folklore araucano: refranes, cuentos, cantos, procedimientos, costumbres prehispánicas*. Santiago: Imprenta Cervantes. http://www.memoriachilena.gob.cl/602/w3-article-8188.html

Guevara, T. 1913. *Las últimas familias y costumbres araucanas*. Santiago: Imprenta Cervantes. http://www.memoriachilena.gob.cl/602/w3-article-8187.html

Havestadt, B. 1777. *Chilidúgu: sieve tractatus linguæ Chilensis.* Aschendorf. https://archive.org/details/chilidusiveresch01have/page/n5/mode/2up

Lenz, Rodolfo. 1897. *Estudios araucanos.* Santiago: Anales de la Universidad de Chile. http://www.memoriachilena.gob.cl/602/w3-article-7925.html

Manquilef, M. 1911. *Comentarios del pueblo araucano I. La faz social*. Santiago: Imprenta Cervantes. http://www.memoriachilena.gob.cl/602/w3-article-8192.html

Manquilef, M. 1914. *Comentarios del pueblo araucano II. La Jimnasia nacional (juegos, ejercicios y bailes)*. Santiago: Imprenta Barcelona. http://www.memoriachilena.gob.cl/602/w3-article-8193.html

Medina, J. T. 1894. *Noticia biográfica.* In *Doctrina cristiana y catecismo con un confesionario, arte y vocabulario breves en lengua Allentiac por el Padre Luis de Valdivia*, ed. J.T. Medina, Seville: Imprenta de E. Rasco, p. 1–42. http://www.memoriachilena.gob.cl/602/w3-article-9565.html

Mösbach, E. de. 1930. *Vida y costumbres de los indígenas araucanos en la segunda mitad del siglo XIX.* Santiago: Cervantes. http://www.memoriachilena.gob.cl/602/w3-article-8190.html

Schuller, R. R. 1907. *El Vocabulario Araucano de 1642-1643.* Santiago de Chile: Imprenta Cervantes. https://archive.org/details/elvocabularioar00hercgoog

Valdivia, L. de 1606. *Arte, y gramática general de la lengua que corre en todo el Reyno de Chile, con un vocabulario y confessionario*. Lima: Francisco del Canto. https://uvadoc.uva.es/handle/10324/701

Valdivia, L. de 1621. *Sermón en la lengua de Chile: de los mysterios de nuestra santa fe catholica, para predicarla a los indios infieles del reyno de Chile, dividido en nueve partes pequeñas, acomodadas a su capacidad*, Valladolid. http://www.memoriachilena.gob.cl/602/w3-article-8484.html

## References:

Caniuqueo, S. 2006. Siglo XX en Gulumapu. In Mariman, P, Canique, S, Millalén, J. & Levil, R. *¡…Escucha Wingka…!* Santiago: LOM Ediciones, pp. 129-218.

Chiodi, F. & E. Loncon. 1999. *Crear nuevas palabras: Innovación y expansión de los recursos lexicales de la lengua mapuche*, Temuco: Editorial Pillan.

Croese, R., A. Salas, and G. Sepúlveda. 1978. Proposición de un sistema unificado de transcripción fonémica para el mapuche. *Revista de Lingüística Teórica y Aplicada*, 16, 151-159.

Gundermann, H. 2014. Orgullo cultural y ambivalencia: Actitudes ante la lengua originaria en la sociedad mapuche contemporánea. *Revista de Lingüística Teórica y Aplicada* 52:1. Pp. 105-132.

Hanks, W. 2010. *Converting Words: Maya in the Age of the Cross*. Berkley: University of California Press.

Kopaczyk, J, B. Molineaux, V. Karaiskos, R. Alcorn, B. Los. and Maguire, W. 2018 Towards a grapho-phonologically parsed corpus of medieval Scots: Database design and technical solutions. *Corpora* 13(2):255–269.

Mager, M., X. Gutierrez-Vasques, G. Sierra and I. Meza-Ruiz 2018. Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, NM, 55–69.

Molineaux, B. (in press a) The dental-alveolar contrast in Mapudungun: Loss, preservation and extension. *Linguistics Vanguard*.

Molineaux, B. (in press b) A reassessment of word prominence in Mapudungun: phonological vs. morphological activation. In Bogomolets K. & van der Hulst H. (eds.), *Word accent in languages with complex morphology*. Oxford: Oxford University Press.

Molineaux, B. and V. Karaiskos 2021. Corpus of Historical Mapudungun (Version 1.0), © The University of Edinburgh, URL: http://www.amc-resources.lel.ed.ac.uk/CHM/.

Pache, M. 2014. Lexical evidence for Pre-Inca language contact of Mapudungun with Quechuan and Aymaran. *Journal of Language Contact 7*. 345–379.

Rojas, D., C. Lagos and M. Espinoza. 2016. Ideologías lingüísticas acerca del mapudungun en la urbe chilena: el saber tradicional y su aplicación a la revitalización lingüística. *Chungará: Revista de Antropología Chilena* 48:1, 115–125.

Salas, A. 1992. *El mapuche o araucano.* Madrid: MAPFRE.

Smeets, I. 2008. *A Grammar of Mapuche*. Berlin: Mouton de Gruyter.

Tarpley, B. 2018. *Digital humanities dashboard*. Center of Digital Humanities Research, Texas A&M University. http://codhr.dh.tamu.edu/2018/04/24/the-early-modern-ocr-project/.

TEI Consortium, eds. 2021 *Guidelines for Electronic Text Encoding and Interchange v.4.2.2.* 09 April 2021. http://www.tei-c.org/P5/.

Villena, B. 2017. Fuentes para el estudio del mapudungún: propuesta de periodización. *Lenguas y literaturas indoamericanas*, 19(1), pp. 141-167.

Villena, B., Cabré, M. T., & Fernández, S. (2019). Formación de nombres en mapudungún: Productividad, genuinidad y planificación. *Revista signos*, 52(100), 615-638.

Zúñiga, F. 2006. *Mapudungun: El habla mapuche*. Santiago: Centro de Estudios Públicos.

Zúñiga, F. 2017. Mapudungun. In M. Fortescue, M. Mithune and N. Evans (eds.), *The Oxford handbook of polysynthesis*, 696–712. Oxford: Oxford University Press.

Zúñiga, F. and A. Olate. 2017. El estado de la lengua mapuche, diez años después. In I. Aninat, V. Figueroa & R. González (eds.), *El pueblo mapuche en el siglo XXI: propuestas para un nuevo entendimiento entre culturas en Chile*, 342–374. Santiago: Centro de Estudios Públicos.

Zwartjes, O. ed. 2000. *Las gramáticas misioneras de tradición hispánica (siglos XVI y XVII).* Amsterdam/Atlanta: Rodopi.