Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6540–6547
Marseille, 20-25 June 2022

A *Mapudüngun* FST Morphological Analyser and its Web Interface

Andrés Chandía
Universitat Pompeu Fabra
Translation and Language Sciences Department
andres@chandia.net / andres.chandia@upf.edu

Abstract

This paper describes the development and evaluation of an FST analyser-generator for the *Mapuche* language, which is publicly available through a web interface. As far as we know it is the first system of this kind for *Mapudüngun*. Following the *Mapuche* grammar by Smeets, we have developed a machine including the main morphological and phonological aspects of *Mapudüngun*. Through this computational approach we have produced an FST analyser-generator capable of classifying and appropriately tagging all the components (roots and suffixes) that interact in a *Mapuche* word-form. A double evaluation has been carried out showing a good level of reliability. In order to face the lack of standardization of the language, additional components (an enhanced analyser, a spelling unifier and a root guesser) have been integrated into the tool. The generated corpora, the lexicons and the FST grammars are available for further development and comparison of results. The *Mapuche* people inhabits southern Chile and Argentina and they speak *Mapudüngun*.

Keywords: Mapudüngun, Mapuche, Morphology, Morphological Analyser, Finite-State Technology, FST, FOMA

1. Introduction

We have defined and developed an FST analyser and generator for *Mapudüngun* using the finite state technology. After the development and compilation of the transducer, the system has been evaluated and implemented as a web service for public free access.

This paper introduces the basics of the system development, the evaluation and the codes and files comprised in the machine, as well as a brief description of the web interface. A long description of the grammar is FST Morphological Analyser and Generator for *Mapudüngun* (Chandía, 2021) that may be consulted for details.

Düngupeyem is a computational tool (a rule-based morphological analyser and generator) for *Mapudüngun*, built with FST using foma (Hulden, 2009), an open source finite state compiler. Foma language descriptions exist for several languages including morphologically complex ones as Quechua (Ríos, 2015), Basque (Alegría et al., 2009) and others.

The key component is a set of regular expressions encoding the grammar. The main taking parts are declared in several scripts, lexicon (roots, suffixes, particles, etc.) and rules for morphotactics and morphophonology.

Regarding other computational projects on *Mapudüngun*, there are really few; the most remarkable one is a data collection and an attempt to develop a morphological analyser and a spellchecker (Levin et al., 2002) at the Carnegie Mellon University. There is also a basic Spanish-*Chesüngun*[1] translator (Rumian, 2011). Our own related projects include CORLEXIM (Chandía, 2014a), a collection of 8 digitized dictionaries of *Mapudüngun*. A spellchecker prototype (Chandía, 2016), online and as a LibreOffice plug-in. IECMap (Chandía, 2017), a morphologically annotated corpus and interface to exploit it. And a stand-alone spelling unifier (Chandía, 2014b), which is also integrated into the FST analyser[2].

2. Verb morphology and encoding

The *Mapuche* verb has a quite complex morphology, and it is the main aim of this implementation. *Mapudüngun* is a polysynthetic and agglutinative language. Polysynthesis means that there are many elements or morphemes in (verb) forms, which is typical of the Native American languages. In agglutinative languages a series of concepts are distributed in several morphemes (Zúñiga, 2006). Agglutination is when morphemes are inside words being identifiable in different contexts; as in Basque, Turkish and Quechua. The original meaning of the stem is modified by the affixes attached to it. Verbs may contain many morphemes, e.g., *di-tu-l-me-tu-a-fi-ñ* 'I will reach it, I will find it'. This word has 8 significant elements. Nominal forms are simple, while verbal ones are extremely complex, presenting a good number of derivative and inflectional morphemes. There is a complete description of the language in the book we base our analyser upon: "A Grammar of *Mapuche*" (Smeets, 2008).

2.1. Verb morphology

After a verbal root, suffixes "occur in a more or less fixed position relative to one another" (Smeets, 2008). Verb suffixes are located on one of the 36 slots assigned to the verb form on the basis of their relative position and function. Slot 1 occupies word final position and slot 36 is next to the root. The order of these slots determines the morphotactics of verb forms. Some slots

[1] A *Williche* (southern) variant of *Mapudüngun*.

[2] All resources available on http://www.chandia.net

host a few mutually excluding affixes, some of them may include variation in their form and some others may be zero markers. Some suffixes may exclude others from different slots due to grammatical or semantic reasons. Even though it is not rare to find up to 7 or 8 suffixes following the root; see Ex.1; verbs usually contain between 4 and 6 suffixes in spontaneous speech.

Below, we graphically represent 2 verb forms. **S** is stem. Every dot represents a slot, leftmost dot is slot 36, rightmost one is slot 1. **X** is a suffix occurrence in a slot. **Ø** is also a suffix occurrence but with a null morpheme, which is a morpheme with no phonemic or graphic realization:

Minimal intransitive verb 2^{nd} person plural
S . . . . X X X . . .

Minimal transitive verb 2^{nd} → 1^{st} persons plural
S . . . . X . . . . X Ø X . . .

Ex.1 Verb with 10 suffixes (Smeets, 2008): 443 .76
nü-nie-ñma-r-pu-tu-e-y-iñ-mu
’they continued to take it away from us’

Root[3]: *nü*- `-IV`[4]`.nü_tomar` ’to take’
Suffixes:

1. *-nie-* Progressive persistent (`+PRPS.nie32`)
2. *-ñma-* Indirect object (`+IO.ñma26`)
3. *-r-* Interruptive (`+ITR.r18`)
4. *-pu-* Locative (`+LOC.pu17`)
5. *-tu-* Iterative/restorative (`+RE.tu16`)
6. *-e-* Internal direct object (`+IDO.e6`)
7. *-y-* Indicative (`+IND.y4`)
8. *-Ø-* First-person (`+1.Ø3`)
9. *-iñ-* Plural (`+PL.iñ2`)
10. *-mu* Dative subject (`+DS3A.mew1`)[5]

Tags between parenthesis express, starting from the left, the abbreviated name of the part of speech (POS) or suffix. POS are introduced by a **–** sign, suffixes, by a **+** sign. `-IV` is ’intransitive verb’, `-N` is ’noun’, etc. For suffixes, `+PRPS` is ’progressive persistent’, `+IDO` is ’internal direct object’, `+PL` is ’plural’, etc.

After the abbreviated POS or suffix name, separated by a dot, is the root or suffix standard form. Roots are followed by their meaning in Spanish with an underscore as separator: `.nü_tomar`. The analyser lexicon is collected with the Spanish translation, that’s why all the examples in this article present the root meaning in Spanish.

The number at the end of each tag indicates the slot where the verb suffix is located. For instance:
`-IV.nü_tomar`: intransitive verb root *nü* which means ’tomar’ in Spanish (’to take’ in English).
`+PRPS.nie32`: progressive persistent suffix, which form is *nie*, is located in slot 32.
`+IDO.e6`: internal direct object suffix, which form is *e*, is located in slot 6, etc.

These items are shown at the analyser results, and similar structure and information is in all the examples.

Slots 1 to 15 hold inflectional suffixes in fixed positions. Slots 16 to 27 hold derivational suffixes, some of which are mobile. Slots 28 to 36 hold derivational suffixes in fixed positions. Mobile suffixes are assigned to their most usual position.

2.2. Root encoding

The system reads files containing roots and incorporate them into the analyser. Lists are collected by POS.

Df.1 `NROOT @re"roots/nroot.lex";`

Definition Df.1 introduces noun roots, so any listed noun is found throughout `NROOT` regular expression. There is a file per category (verbs, adjectives, particles, etc.), e.g., `ajroot.lex` for adjectives. The lexicon is divided in two groups: verbalizable and non-verbalizable forms.

2.3. Suffix encoding

Every suffix is assigned to a slot, encoded in a file, the addition of such information to the main script is made by calling these files which contain the slots fillers:

Df.2 `define NEG @re"slots/1-15-Inflectional-Suffixes/slot-10.aff";`

This file contains negation suffixes filling slot 10:

Df.3
```
["+NEG".la10]:{la}
| ["+NEG"{.ki10}]:"@NK"
| ["+NEG"{.no10}]:[{no}|{nu}]
| ["+NEG"{.kino10}]:[{kino}|{kinu}];
```

Four suffixes fill slot 10. Every suffix section indicates its upper or abstract level (left of the colon) and lower or surface level (right of the colon). Every abstract form is returned on analysing, giving the category tag, the suffix form and the slot number: `+NEG.no10`. These forms are also used to apply morphotactics in the abstract level at analysis and generation processes. At surface level are the forms analysis is referring to, and/or the tags that trigger a process of replacement, like `@NK` (in Df.3) which has a rule replacing it for certain form in a defined context.

The number of roots, suffixes, rules and so are detailed in section 5.2.

[3] This analysis representation highlights the amount of suffixes, it is not the standard output, fig.4 (p.7) shows resulting analyses on the web interface, which is the standard.

[4] *Mapuche* roots seem to be focused on the action or state (or the idea of) only, without taking into account the actants involved in the action. The incorporation of specific suffixes in complete verbal forms lead us to this conclusion, otherwise it would not be necessary such incorporation of suffixes, e.g.:

1. *iñche allkün* ’I hear’ (I have the capability of hearing)
2. *iñche allkütun* ’I hear it’ (there is *something* I hear)
3. *iñche allkütulen* ’I’m hearing it’ (It has my attention)

In 2, the suffix *-tu-* is a **transitivizer**, it means that this suffix is necessary to understand the implication of transitivity and, obviously, of a patient the verb ’to hear’ has. In 3, the suffix *-küle-* ∼ *-le-* is added to express the state of doing the action the verb is referring to. On the contrary, in 1, there is no other suffix than *-n*, which expresses the first-person singular.

[5] All tags meaning are available in (Chandía, 2021), p. 55.

3. Morphophonological rules

The co-occurrence of roots and suffixes in verb forms generates phonological changes. In *Mapudüngun*, this interaction may be between suffixes, the root and a consecutive suffix, roots inside a compound stem, or between all of them when forming complex stems. These changes are encoded in the *lower* or *surface* side of the language, and they affect the word form.

There are frequent required phonological changes in *Mapudüngun* reflected in spelling, but the epenthesis of voiced velar fricative /ɣ/ represented by *g* can be optional in some contexts. In fact, epenthesis of a glottal stop is optional between the ending vowel of a root and the initial vowel of a following root in compounds.

Ex.2 (Smeets, 2008): 52
dew-ma-iyal-la-y → *dew-ma-***g***iyal-la-y* 'he did not prepare food'

```
-IV.dew_terminar+EXP.ma35-NN.iyal
_comida-CR.IV+NRLD.a9+IND.y4+3.Ø3
```

Lexical forms must be collected in an intermediate form with the appropriate tags to later apply the rules transforming them into the final surface forms, e.g., the portmanteau suffix for indicative, 1st p., singular, may occur as *-ün-, -üñ-, -n-* or *-ñ-*. Rules to generate them:

```
Df.4  ["@Ü" -> ü || [CON|SVW|.#.]  _ ]
.o.   ["@Ü" -> 0 || VOW _ ];
```

This applies not only to this case, but wherever the tag `@Ü` is found. First rule indicates that `@Ü` is replaced by *ü* when a consonant or semi-vowel precedes it, or when it is at word beginning (`.#.` *word boundary*). The second rule states that `@Ü` is deleted if preceded by vowel. In combination with the *intermediate representation*, the FST compiles all 4 possible forms.

Tags are created and assigned arbitrarily. For the rules to work, it is necessary to place the corresponding tag in the appropriate position. As the rule is applied to the surface level, the tag is placed at that same level, which is encoded to the right of the colon. Following rule incorporates Pluperfect mark (PLPF) to the system:

```
Df.5 ["+PLPF"{.üwye15}]:["@Ü"{wye}];
```

To the left of the colon is the upper or abstract level, the analysis representation. To the right, the lower level, the surface representation, where the tag is added preceding the suffix as initial character.

Phonological changes comprise elision, epenthesis and interchange; some are optional and reflect affection (or other emotional senses), or register (ceremonial or colloquial); Smeets (2008) mentions 16 cases of consonants (and semi-vowels) alternation, and 6 of vowels. Simplification of geminate sounds is also related to colloquial speech, while pronouncing them is related to ceremonial speech. There are also some special cases of verb roots which realize differently depending on the context; also suffixes that change their form obligatorily in some contexts and optionally in others. The system collects 44 rules, all of them composed by two or more sub-rules; see section 5.2.

4. Morphological rules

The majority of these rules are devoted to manage morphotactics, which is the set of constraints that regulates the co-occurrence of morphemes. Concisely, the verb is a stem followed by a series of suffixes that complete the verb form. Once the lexicon and suffixes interacting in the verb form are declared; see 2.2, 2.3; it is necessary to regulate their interaction.

4.1. Stem codification

Suffixes also appear in verb stems except for stems formed by a single verb root. Adjective, adverb, noun and other roots may become verb stem linking to them a verbalizing suffix. To simplify, we call verb stem to any form, simple or complex, to which suffixes are bonded in order to form a complete verb predication. While a simple stem is made of one root only, a complex stem may be composed by two roots, a root with some suffixes, or a combination of them all.

The simplest stem type is formed by a single verb root. Verb suffixes may be added immediately to this type of stem. Other type of stem that accepts verb suffixes immediately is the simple compound where one of the members is a verb root, the other member may be another verb root, an adjective, adverb, noun, numeral or a question root; see table 1.

Table 1: Simple stem forms

Stem	Suffixes
Verb root	+Suffixes
Verb root + Verb root	+Suffixes
Verb root + Non-verb root	+Suffixes
Non-verb root + Verb root	+Suffixes

Compounds encoding needs to be ruled by applying tag neutralization and verb valence.

Tag neutralization: complex stems and complex compounds are *complex* when incorporating suffixes into the stem, having their own rules of interaction. For this reason, POS and suffixes tags are modified while applying the inner compound rules: Ru.1.

Ru.1 POS and suffixes tags neutralization (sample)

```
define NeutTv ["-tv0" <- "-TV"];
define NeutCa ["+ca0" <- "+CA"];
```

Table 2: More complex stem forms

Stem	Slot 36	Suffixes
Non-verb root	+VRB	+Suffixes
Non-verb compound	+VRB	+Suffixes
Reduplicated root	+SFR	+Suffixes

Table 3: Intransitive and transitive suffixes per slot

Slot	23	10	6	4	3	2	1
Intransitive	-	negation	-	mood	person / portmanteau	number	-
Transitive	agent	negation	object	mood	person / portmanteau	number	dative subject

Valence in compounds: when one of the compound roots is a verb and the other is not, the resulting compound gets the valence from the verb. When both members are verbs, valence is derived from the second.

One more degree of complexity is the addition of a verbalizer (s36) to some single roots or compounds to become verb stems. Adjectives, adverbs, nouns, numerals, onomatopoeia, proper nouns, question forms, and compounds with no verb root, need a verbalizer in slot 36. Reduplicated roots need a *stem formative*. Table 2 summarizes previous description.

Single non-verb roots need a verbalizer to become verb stems.

Verbalizers (slot 36): *-Ø-, -nge-, -tu-, -l-*.

There is a rule stating that a single non-verb root must be followed by a verbalizer in order to form a verb.

Non-verb compounds may be verbalized by a suffix of slot 36. Suffixes verbalizing single adjectives, also verbalize adjective compounds, and so on.

Reduplicated forms need a stem formative (s36) to further attach verb suffixes.

Complex single root stems made up by a single root, a compound or a reduplicated root that incorporates at least one suffix (rarely more than 3) into the structure are considered *complex stems*. Encoded complex single root stems are adjectival, adverbial, nominal, numeral, questions and nominalized verbs.

Complex nominalized verb stems. This type of stem is treated in 3 steps. The first step defines the form, including all the optional and obligatory normalizing suffixes, the second step sets the neutralization of all tags identifying the suffixes occurring in the form, and the third step regulates the interaction of all possible suffixes in the stem, including verbalizers.

4.2. Verb paradigms and morphotactics among verb suffixes

There are about 80 verb suffixes spread in 36 slots. Suffixes belonging to the same slot exclude one another. Some suffixes exclude others for grammatical or semantic reasons, for example, once a verb has taken an inflectional nominalizer (s4), it may not take suffixes of mood (s4), person (s3) and number (s2).

To treat suffixes co-occurrence, first, the suffix sequence is established including all the possible variants generated by suffix mobility[6]. Almost all suffixes are marked as optional except for `PERSON`, `PTMT`, `FLECNOM` and `NMZ`. Because the *Mapuche* verb is either finite (`PERSON` and `PTMT`) or nominalized (`FLECNOM` or `NMZ`), these are the obligatory suffixes for those forms.

To rule suffixes co-occurrence, first, the minimal verb form is incorporated, i.e., an intransitive verb root plus suffixes expressing mood, person and number, obligatory in a finite verb form, then, the transitive verb related suffixes, to finally set the rules dealing with these minimal verb forms.

Verbs may take 2 possible forms, nominalized or finite, the last one presents mood (s4), person or portmanteau morphs (s3), number (s2), dative subject (s1, transitive forms).

Suffixes from slots 23 and 6 are incorporated in order to complete the transitive verb paradigm, and negation suffixes (s10) which complement with mood markers and have a particular incidence in imperative negative forms.

Table 3 shows suffixes per slot implied in transitive and intransitive verb paradigms. Not all of them co-occur in a transitive form, e.g., agent markers (s23) do not co-occur with direct objects (s6) or dative subjects (s1).

Thirty-three rules regulate verb paradigms, some containing sub-rules, and some including the interaction with inflectional nominalizers (s4). No reference to mood, person or number may be made when a verb takes one of the nominalizers; but nominalized verbs may include agents or objects with the corresponding dative subject. For example:

Ru.2 **Constraint rule**

```
define RuPr10 [~$["+CND" ?*
[["+1"{.Ø3}]|["+3"[{.Ø3}|{.ng3}]]]]];
```

Symbols combination ∼`$` says that *the conditional may not be followed by a* 1^{st} *person suffix in its null form -Ø-, or the* 3^{st} *person suffix in its forms null or* -ng-.

In this way 12 rules regulate verb nominalization: co-occurrence of inflectional or derivational nominalizers with dative subject, non class-changing, class-changing and instrumental suffixes.

Additionally, 30 rules regulate the occurrence of suffixes that are not obligatory in transitive or intransitive forms. For example: ”The reflexive/reciprocal suffix *-w-* does not combine with a suffix in slot 23, 6 or 1. The reflexive morpheme *-w-* may occur in intransitive verbs, i.e., in verbs which do not take a suffix in slot 6”, (Smeets, 2008).

5. Compilation and extended analyser

The flow chart deployed in figures 1, 2 and 3 shows the interconnection of all the related elements along the

[6] Suffixes are assigned to their most common position (slot), but some of them may appear displaced from their usual position with respect to co-occurring suffixes.

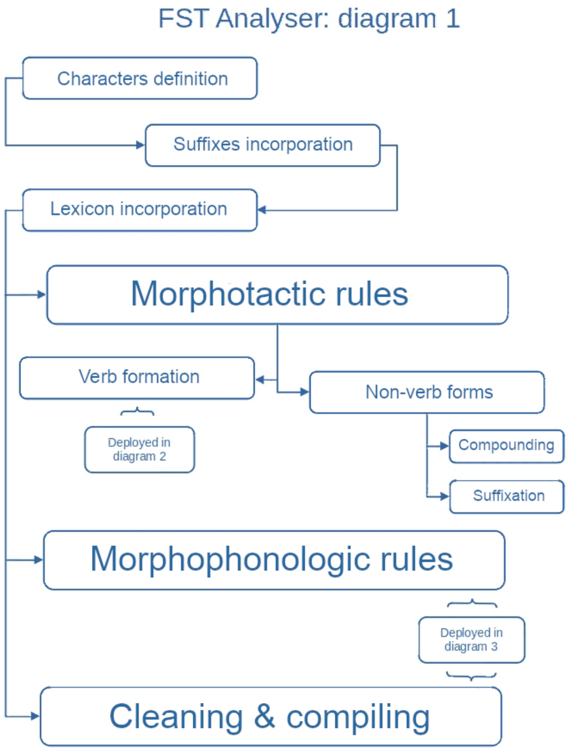


Figure 1: General view

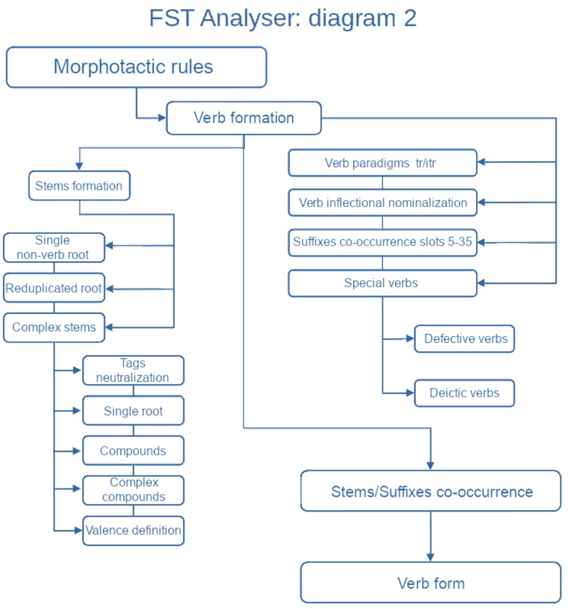


Figure 2: Verb formation

process. Figure 1 explains the main blocks of the FST grammar. Fig. 2 describes the main sections of the morphotactics, and fig. 3 the sections inside the phonological grammar and the cleaning process.

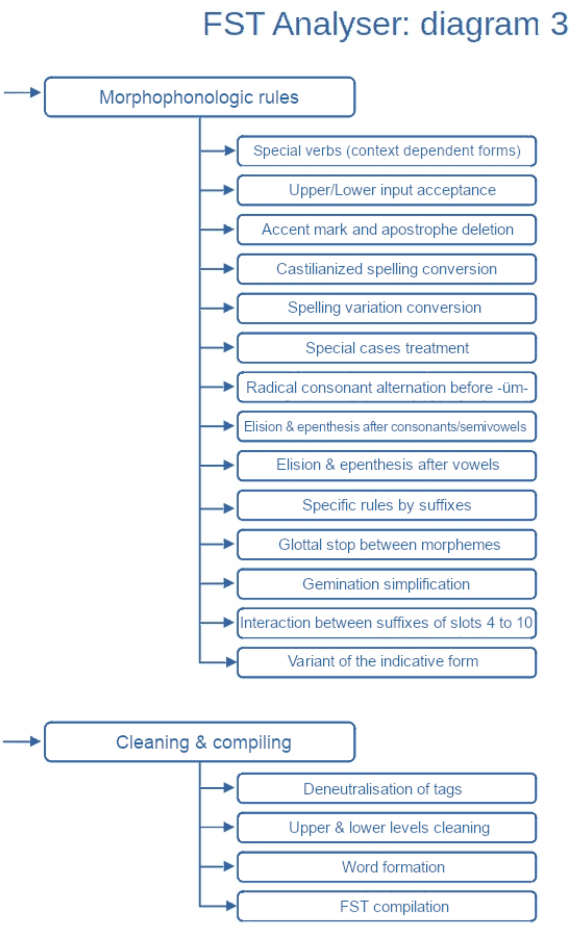

FST Analyser: diagram 3
Morphophonologic rules
Special verbs (context dependent forms)
Upper/Lower input acceptance
Accent mark and apostrophe deletion
Castilianized spelling conversion
Spelling variation conversion
Special cases treatment
Radical consonant alternation before -üm-
Elision & epenthesis after consonants/semivowels
Elision & epenthesis after vowels
Specific rules by suffixes
Glottal stop between morphemes
Gemination simplification
Interaction between suffixes of slots 4 to 10
Variant of the indicative form
Cleaning & compiling
Deneutralisation of tags
Upper & lower levels cleaning
Word formation
FST compilation

Figure 3: Morphotactics, cleaning & compiling

5.1. Extended analyser

There is a significant variation in *Mapudüngun* spelling, mainly due to the existence of different spelling proposals, together with the strong influence of Spanish orthography. Some texts may present a mixture of these proposals along with Castilianized spelling. We have developed an extended analyser, in lexicon and rules, that fit another variants of *Mapudüngun*. We have named it *Düngupeyem*, keeping the basic one as *SmeetsAnalyser*.

A function included in the extended engine is spelling unification, which mainly manages the influence of Spanish orthography. This process includes deleting accentuation marks á → *a*, é → *e*, etc., conversion into lowercase characters and phonological conversion, e.g.:

b → *f*;
ca → *ka*, co → *ko*, cu → *ku*, ce → *se*, ci → *si*;
gue → *ge*, gui → *gi*;
hua → *wa*, hue → *we*, hui → *wi*, huo → *wo*, huu → *wu*;
ha → *a*, he → *e*, hi → *i*, ho → *o*, hu → *u*;
j → *k*, qu → *k*;
v → *f*;
tx → *tr*, x → *tr*;
z → *d*

5.2. Dimensions of the grammar

Below, we specify the dimensions of elements within the FST grammar. The included data refers to the length of the elements interacting in the system: lexicon, suffixes, rules, etc.

The FST system counts 33,668,316 arcs and 8,401,431 states. Its size is 513.6 MB and it is of a cyclic type. Bellow, table 4 shows the amount of verbalizable lexicon. Table 5, the amount of non verbalizable lexicon. Table 6 shows the amount of suffixes separating them in two groups, suffixes that form part of a conjugated verb, and suffixes that nominalize verb forms or are added to nominalized verbs. Finally, table 7 shows the amount of generated rules for the system to work.

Table 4: Roots (verbalizable lexicon): 2,075

Adjectives	130	Onomatopoeia	12
Adverbs	21	Proper nouns	68
Intransitive vbs.	293	Questions	4
Nouns	1,318	Transitive verbs	215
Numerals	14		

Table 5: Non-verbalizable lexicon: 252

Adverbs	76	Negations	1
Anaphoric pron.	6	Numbers	10
Auxiliaries	8	Particles	19
Conjunctions	8	Personal pron.	9
Demonstratives	6	Possessive pron.	6
Foreign express.	7	Prepositions	5
Interrogative pron.	7	Punctuation marks	58
Interjections	26		

Table 6: Suffixes: 108

Verb suffixes: 93

Fix derivational	27	Mobile derivational	15
Inflectional	50	Non-slot	1

Non-verb suffixes: 15

Class-changing	3	Nominalizers	5
Instrumental	1	Non class-changing	6

Table 7: Rules: 522

Character definition	3	Phonological	89
Cleaning	5	RegEx files	75
Morphological	350		

6. Evaluation

In this section we will show the coverage of the standard and extended analysers and the ambiguity of the results.

6.1. Corpora

Three corpora have been collected in order to calculate the coverage of the analysers and the ambiguity of the results (see (Chandía, 2021), p. 48 for details on corpora).

Gold standard. We have compiled a corpus including the sentences from *A grammar of Mapuche*, which words were analysed and disambiguated by Smeets (2008). This corpus includes all the sentences from chapters 10 to 18, and 21, which deal with nouns, adjectives, adverbs, numerals, demonstratives and anaphoric pronouns, personal, possessive, interrogative pronouns, suffixation and verbalization. Also, all the 17 texts of *Part VIII-Texts*, (Smeets, 2008).

Control corpus. 240 sentences extracted out of the Gold standard, and containing a total of 1,671 words, which correspond to 650 forms. It is used to check the results from the analyser. A correct analysis for every word of this corpus must appear in the output.

Comparison corpus. Made of 170 sentences containing 1,256 words (511 forms). Texts are extracted from *Mapudüngun. El habla Mapuche, chapter V. Textos en mapudüngun*, (Zúñiga, 2006).

6.2. Coverage and ambiguity using the standard analyser

An analyser that strictly fits with Smeets (2008) work has been compiled as the *SmeetsAnalyser*, and an extended analyser as *Düngupeyem*, which includes another variants of *Mapudüngun*.

Firstly, we have analysed the control corpus with the *SmeetsAnalyser* and have obtained the following results:

Word-forms: 650; produced analyses: 2,232; unknown words: 2

This means that the coverage is very good, 99.7%, and the ambiguity rate (number of analyses per word) reaches 3.59.

This ambiguity rate is not strange because ambiguity in *Mapudüngun* is high. There are some ambiguous words that appear very often in texts, like *ka*: `AJ` 'Other', `CJ` 'and', `PT` 'certainty, indignation, cynicism'; *fey*: `AV` 'then', `DP` 'which, that', `IV` 'to be that', `PP` 'he, she, them'; and most of the words ending in *n*, the shared form between the 1st person, singular suffix and the plain verbal noun mark:

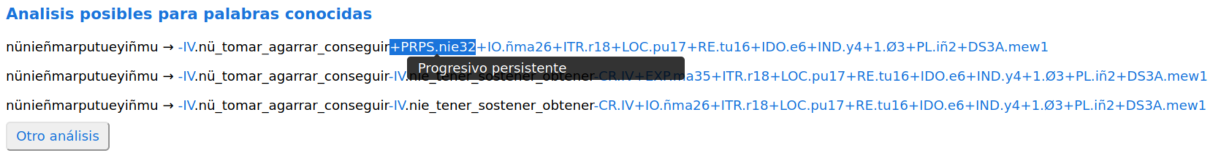
küdawün → `-IV.küdaw_trabajar+IND1SG.n3`
küdawün → `-IV.küdaw_trabajar+PVN.n4`

Previous 9 analyses divided by the 3 given examples establish the *average Mapudüngun ambiguity* (**ama**) for these words.

In order to evaluate the *SmeetsAnalyser* on another type of text, we test the comparison corpus, and we obtained the following results:

Word-forms: 511; produced analyses; 1,368; unknown words: 120

This means that the coverage is worse, 76,5%, and the ambiguity rate similar, 3.49.



Analisis posibles para palabras conocidas

nüinieñmarputueyiñmu → -IV.nü_tomar_agarrar_conseguir+PRPS.nie32+IO.ñma26+ITR.r18+LOC.pu17+RE.tu16+IDO.e6+IND.y4+1.Ø3+PL.iñ2+DS3A.mew1

Progresivo persistente

nüinieñmarputueyiñmu → -IV.nü_tomar_agarrar_conseguir-IV.nie_tener_sostener_obtener-CR.IV+EXP.ma35+ITR.r18+LOC.pu17+RE.tu16+IDO.e6+IND.y4+1.Ø3+PL.iñ2+DS3A.mew1

nüinieñmarputueyiñmu → -IV.nü_tomar_agarrar_conseguir-IV.nie_tener_sostener_obtener-CR.IV+IO.ñma26+ITR.r18+LOC.pu17+RE.tu16+IDO.e6+IND.y4+1.Ø3+PL.iñ2+DS3A.mew1

Figure 4: Analysis results on the web

6.3. Coverage and ambiguity using the Düngupeyem analyser

The low coverage obtained using the standard analyser was the reason because we developed the extended analyser (*Düngupeyem*). So, in a second step we have evaluated this enhanced analyser on both corpora. The results for the control corpus are:

Word-forms: 650; produced analyses: 2,477; unknown words: 2

The coverage remains (it was very high) and ambiguity rises less than 0.24 points per word-form, indicating that reliability of the analyser is quite good.

Using the comparison corpus results are:

Word-forms: 511; produced analyses: 1,828; unknown words: 10

The coverage is improved from 76,5% to 98,0% with a low increment of the ambiguity rate (from 3.49 to 3.64)

For cases where the highest coverage is necessary, we have developed a *roots' guesser*[7].

Using the guesser for unknown words after the *Düngupeyem* only remain three unknown words, and the coverage is 99.4%.

These calculi confirm the reliability of the analyser, even more after adding considerations for other dialects and extending lexicon.

These results are comparable to those reported by Ríos (2015) who reaches 95.2% of correctness with her system for Quechua.

7. Web interface

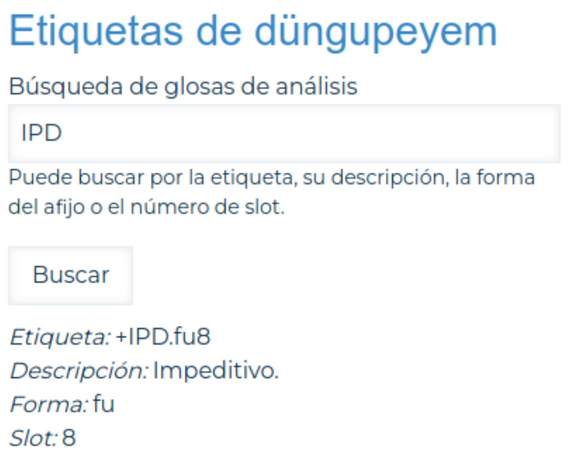
This section introduces the web interface for the analyser and generator. Both share a page with a form for each one to enter text and start the selected process.

The system flow for analysis is explained in ten steps: **1**) input of text, direct or file upload. **2**) Text tokenizing. **3**) Sending text to the analyser, a word by word processing of the *flookup* (Hulden, 2009) utility that reads a line, passes the string to the analyser and formats the output, directing it to a new file. **4**) The output generated by the analyser is stored and cleaned. The analyser stores two types of output in the same file, analysed and non-analysed forms; **5**) Non-analysed forms are moved to a new file, analysed ones remain. **6**) Only performed if analysed forms file has content. Final file is created with the corresponding output strings. **7**) The file with non-analysed forms is sent to the roots' guesser. **8**) Non-analysed forms are added to the output. **9**) Results are displayed to the web user, and if there are unknown or unrecognized forms, **10**) an email is sent to the administrator.

In the interaction interface for analysis and generation there is as little information as possible so the user is not distracted. The email field is worth highlighting. For feedback on the unrecognized or non-analysed forms, the user may add an email address. This way, the user may correct a possible mistake or the administrator may improve the system.

Figure 4 displays the analysis for *nüinieñmarputueyiñmu* 'they continued to take it away from us' as seen on the web interface. First analysis corresponds to Ex1 explained in 2.1, CSS sets tags in blue and shows tag name on hovering them.

There is also a search block on the right side of the page to look for analyses tags (fig. 5).



Etiquetas de düngupeyem

Búsqueda de glosas de análisis

IPD

Puede buscar por la etiqueta, su descripción, la forma del afijo o el número de slot.

Etiqueta: +IPD.fu8

Descripción: Impeditivo.

Forma: fu

Slot: 8

Figure 5: Morphological tags search block

http://www.chandia.net/dungupeyem is the URL to access these services. There is a web page containing the complete set of tags used in analysis, and an explanation of how to read the generated glosses on http://www.chandia.net/glosas-del-dungupeyem.

The interface was already online at the early stages of the project, back in 2013 when it was just a prototype. In December 2020 we have added an analyses counter, and along this last year there have been 178 performed

[7] The roots' guesser is a tool derived from the analyser, which have no lexicon of roots, but the possible root structures in terms of consonants, vowels and semi-vowels, e.g., CVC, CVSV, CVCVCVC are valid *Mapuche* root structures. Suffixes are included, and their combination rules. This machine first check the possible root structures and then the possible suffixes combinations.

analyses by anonymous users.

Taking into account that this is a very specialized tool, we allow ourselves to think that this system is a valuable contribution for the linguistic community in particular, and a good service for spreading the knowledge of the *Mapuche* language and culture.

8. Conclusions

A morphological analyser-generator for *Mapudüngun* has been developed defining a finite state grammar including lexicon, morphotactics and phonological and morphological rules. The tool has been evaluated obtaining good results for "normalized" words. As far as we know it is the first morphological analyser-generator for *Mapudüngun*.

Due to the significant variation in *Mapudüngun* spelling, mainly due to the existence of different spelling proposals and the influence of the Spanish spelling, we have extended the analyser in order to increase the coverage for a wider range of texts.

We have obtained good results, achieving rates comparable to those of other languages with similar topology.

All the resources and related information are available on http://www.chandia.net website; in the menu bar, under *Küdawkawe* 'tools' are *Düngupeyem* (Chandía, 2013), interface for the analyser-generator system; *Ngülamwirinten* (Chandía, 2016), online spellchecker and LibreOffice plug-in; *IECMap* (Chandía, 2017), corpus exploitation interface (indexed corpora are tagged with *düngupeyem*); *Kümewirin Mapudüngun Trapümfe* (Chandía, 2014b), spelling unifier. A repository of the FST tools and the corpora is available on http://www.chandia.net/dungupeyem/repositorio.

We continue working on tools derived from the analyser, and we hope this work will be the basis for current and future work on automatic text processing for *Mapudüngun*. We are collaborating in a couple of projects: 1) the Master thesis "Design and development of a basic computational infrastructure for *Mapudüngun* learning" by Cristian Ahumada (Ahumada, 2022); and 2) a GitHub repository for *Mapudüngun* resources being collected by Taco de Wolff from Centro de Investigación INRIA Chile.

9. Copyrights

We publish every document and resource under the Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) licence.

10. Acknowledgements

As always, I deeply and sincerely thank Iñaki Alegria who have guided me along the confection of this and other articles. He has spent countless hours of revision to improve and present this work in a clear, concise and understandable manner.

11. Bibliographical References

Ahumada, C. (2022). *Diseño y Desarrollo de una Infraestructura Computacional Básica para el Aprendizaje del Mapuzugun.* Tesis para optar al grado de magíster en ciencias, mención computación, Departamento de Ciencias de la Computación. Facultad de Ciencias Físicas y Matemáticas. Universidad de Chile, Santiago, Chile.

Alegría, I., Etxeberria, I., Hulden, M., and Maritxalar, M. (2009). Porting Basque Morphological Grammars to foma, an Open-Source Tool. Technical report, Euskal Herriko Unibertsitatea, Basc Country. Ixa taldea.

Chandía, A. (2021). FST Morphological Analyser and Generator for Mapudüngun. 2109.09176, arXiv, cs.CL.

Hulden, M. (2009). Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 29–32. Association for Computational Linguistics.

Levin, L., Vega, R., Carbonell, J., Brown, R., Lavie, A., Y, E. C., and Huenchullan, C. (2002). Data collection and language technologies for mapudungun. In *In Proceedings of the International Workshop on Resources and Tools in Field Linguistics, LREC.*

Ríos, A. (2015). *A Basic Language Technology Toolkit for Quechua.* A thesis submitted in fulfilment of the requirements for the degree of doctor of philosophy, Faculty of Arts of the University of Zurich. Zurich, Switzerland.

Smeets, I. (2008). *A Grammar of Mapuche*. Mouton de Gruyter, Berlin, Germany. New York, U. S. A., 1st edition.

Zúñiga, F. (2006). *Mapudüngun. El habla mapuche*. Centro de Estudios Públicos, Santiago, Chile, 1st edition.

12. Language Resource References

Andrés Chandía. (2013). *Düngupeyem. Morphological FST analyser-generator for Mapudüngun.* http://www.chandia.net/dungupeyem.

Andrés Chandía. (2014a). *CORLEXIM. Corpus lexicográfico del mapudüngun.* http://corlexim.cl.

Andrés Chandía. (2014b). *Kümewirin Mapudüngun Trapümwe. Spelling unifier for Mapudüngun.* http://www.chandia.net/kmt.

Andrés Chandía. (2016). *Ngülamwirinten. Spellcheker for Mapudüngun [prototype].* http://www.chandia.net/ngülamwirinten.

Andrés Chandía. (2017). *IECMap. Corpus Exploitation Interface for Mapudüngun.* http://www.chandia.net/iecmap.

Salvador Rumian. (2011). *Srupal Süngufü - Traductor Español-Che Süngun (programa experimental).* http://millalikan.blogspot.com/2011/02/cada-vez-mas-cerca-del-traductor.html.