A finite-state morphological analyser for Paraguayan Guaraní

Anastasia Kuznetsova[♢†]
♢ Department of Computer Science
Indiana University
Bloomington, IN
anakuzne@iu.edu

Francis M. Tyers[†]
† Department of Linguistics
Indiana University
Bloomington, IN
ftyers@iu.edu

Abstract

This article describes the development of morphological analyser for Paraguayan Guaraní, an agglutinative indigenous language spoken by nearly 6 million people in South America. The implementation of our analyser uses HFST (Helsiki Finite State Technology) to model morphotactics and phonological processes occurring in Guaraní. We assess the efficacy of the approach on publically available corpora and find that the naïve coverage of analyser is between 86% and 91%.

1 Introduction

Morphological modelling, under which we subsume both morphological analysis and morphological generation is one of the core tasks in the field of natural language processing. It is used in a wide variety of areas, including but not limited to: orthographic correction (Pirinen and Lindén, 2014), electronic dictionaries (Johnson et al., 2013), morphological segmentation for machine translation (Tiedemann et al., 2015; Forcada et al., 2011), as an additional knowledge source for parsing languages with non-trivial morphology (Gökırmak and Tyers, 2017; Tyers and Ravishankar, 2018), and in computer-assisted language-learning applications (Ledbetter and Dickinson, 2016).

In this article we describe a morphological analyser for Paraguayan Guaraní (in Guaraní: *Avanẽ'e*, ISO-639: gn, grn), one of the official languages of Paraguay. Although Guaraní is an official language and spoken by over six million people throughout the South American continent (Eberhard et al., 2018), it does not benefit from a wide range of freely-available data and tools for building natural language processing systems. If we use Wikipedia as a proxy for viability of crowdsourcing linguistic data, as in (Moshagen et al., 2014), we see that although Guaraní has a large speaker population, the potential for crowdsourcing and big freely-available data is limited.[1]

The absence of large amounts of textual data means that data-driven approaches are hard to apply. In addition, supervised approaches, including neural networks which have become increasingly popular, require large amounts of annotated data to be trained. This in turn requires large numbers of trained annotators to annotate it. Given that neither of these are available, we apply tried-and-tested technique relying on formal linguistic description by means of finite-state transducers. Finite-state techniques have been widely applied to morphological modelling of many languages and are state of the art for many languages, especially those with non-trivial morphology such as languages described as agglutinative (Çöltekin, 2010; Pirinen, 2015) or polysynthetic (Schwartz et al., 2020; Andriyanets and Tyers, 2018).

The remainder of the article is laid out as follows: In Section 2 we give an overview of Guaraní, paying special attention to aspects of morphology and morphosyntax. Section 3 reviews the prior work, Section 4 describes the implementation of the analyser, including information about the linguistic data and tools used. We evaluate our analyser in Section 5, giving both a qualitative, quantitative and comparative evaluation. And finally in Section 6 we give some final remarks and comment on potential future work for Guaraní.

2 Language

Guaraní (Native name: *Avanẽ'e*) is one of the most spoken indigenous languages of South America that belongs to Tupi-Guaraní stock. It is divided into dialects or even languages such as Paraguayan

[1] We note that the Guaraní *Vikipetã*, https://gn.wikipedia.org/ currently has a total of 3,767 articles (as of the 15th July 2020), while the English Wikipedia, http://en.wikipedia.org/ has 6,122,333 as of the same date. If we compare with a language with a similar number of speakers and official status, for example Catalan, we see that the Catalan *Viquipèdia* has vastly more articles 652,079.

Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, pages 81–89

Guaraní, Bolivian Guaraní and some other dialects spoken in Brazil (Ava, Kaiowá, Nhandeva, Mbyá etc.). According to Ethnologue[2] population that speaks all the varieties of Guaraní is 6.162.840 people. The majority of Guaraní speaking population is located in Paraguay where Guaraní is considered the official language and consists of 5.850.000 monolinguals and bilinguals. See Figure 1 for Guaraní speaking area.

Guaraní is an agglutinative concatenative language. It's morphology has both derivational and inflectional traits: it uses suffixes, preffixes and circumfixes for word production. Roots (or stems) affect the phonology of affixes concatenated to the stem and vice-versa, mostly in cases of nasal harmonization or incorporation[3]. The majority of the words in Guaraní are oxytone with some exceptions when accentuation rules apply (Estigarribia, 2017).

Only recently Paraguayan institution *Academía de la Lengua Guaraní* approved current orthographic standard for written Guaraní (Sánchez, 2018). Thus in literature published before 2018 writing standards vary significantly. For example, postposition 'hağua' or 'haguã' can be written with $\tilde{g}$ or $\tilde{a}$ where nasalization is marked graphically by tilde. According to phonological rules, nasalization propagates over the entire syllable if there are any nasal phonemes in it (Krivoshein de Canese, 1983), therefore, both spellings are acceptable. In addition, tilde indicates the stress for nasal vowels and special nasality marking in *haguã* may be considered excessive. In Wikipedia corpus some nasalized phonemes are also marked with diaeresis "¨" (ï, ÿ, ä, etc.). Our transducer handles all the spelling varieties and treats them as orthographic errors.

Despite Guaraní being one of the most spoken low-resource languages of South America grammars thoroughly describing the language are not abundant. Throughout this paper we mostly consult with (Krivoshein de Canese, 1983), (Estigarribia, 2017) and (Dietrich, 2017), although there are earlier reliable grammars available (Gregores and Suarez, 1967).

3 Prior work

Most of the existing computational resources for Guaraní so far are online dictionaries or translators supported by the community. They are based on aligned publicly avaliable corporas such as Wikipedia or Guaraní–Spanish Bible. For example, *iguarani.com* and *glosbe.com* are mostly supported by non-professionals i.e. native speakers or other enthusiasts. Glosbe even has its API (Application Programming Interface). But as textual sources for Guaraní are scarce these translators are not always reliable and lacking words.

At Indiana University Michael Gasser (Gasser, 2018) developed *Mainumby* translation system created mostly for Paraguayan translators with implementation of finite state morphological analyzer *ParaMorfo* embedded into translator. This analyser is very close to what we have done although is focused mostly on the form generation rather than morphological analysis. The analyser discussed in this paper and ParaMorfo were built independently and we will evaluate two transducers for comparison.

4 Development

Transducer-based morphology modelling is essentially the mapping between elementary morphological units (morphemes) to morphological (part of speech) tags or whole lexemes. This mapping reflects the combinatorics and morphological constraints of natural language i.e. which morphemes can combine into a lexeme and which morphemes are incompatible.

FST-based approaches use *continuation lexicons* term to denote the mapping as we will reference them throughout the paper. The implementation of continuation lexicons in our analyser is entirely built on dictionaries publicly available on the web. One of them is L3 project Guaraní dictionary[4] from the *hltdi-l3* GitHub repository.

Our two-level transducer uses two formalisms:

- `lexc` formalism which models morphotactics (morpheme combinatorics);
- `twol` formalism is used for implementing phonological rules.

Both of the formalism use specific syntax following HFST platform conventions. Our analyser is a part of Apertium[5] open-source platform and can be used freely and enhanced by any member of open-source community. In the paper it is referred as Apertium analyser.

[2] `https://www.ethnologue.com/language/grn`

[3] *Incorporation* is a type of word formation that comprises a compound from a verb and an object of that verb i.e. object is incorporated by a verb and becomes a sole lexeme.

[4] `https://github.com/LowResourceLanguages/hltdi-l3/blob/master/dicts/lustig_words_gn_es.txt`

[5] `https://github.com/apertium/apertium-grn`

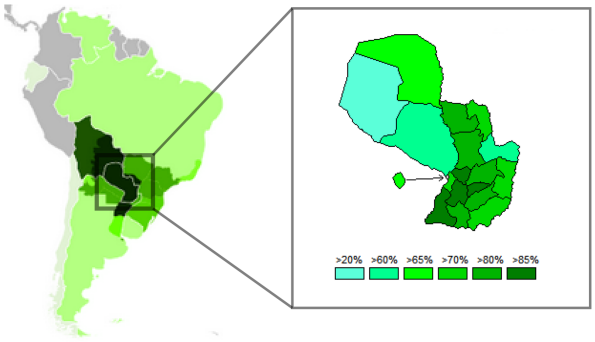

Figure 1: Areas where Guaraní is spoken in South America (including language varieties). The very dark green shows areas where the language has official status, dark green shows areas where there are a considerable number of speakers, while the light green shows areas where the language is official by virtue of its recognition by the Mercosur trade block. The box zooms in on Paraguay and shows the percentage of people having Guaraní as a native language by department according to the 2002 census.

```
LEXICON Nouns
achegety:achegety N ; ! "abecedario"
aguara:aguara N ; ! "zorro"
aguyjevete:aguyjevete N ; ! "gratitud"
ahoja:ahoja N ; ! "manta"
aho iha:aho iha N ; ! "carpa"
```

Figure 2: Lexicon for noun stems from `lexc` file. The first element before colon is an underlying form, the second element stands for surface form of the nouns adding further lexicons to the surface stem (N-lexicon). After exclamation mark follows the comment with translation to the word.

4.1 Morphotactics

4.1.1 Ambiguity of classes

The nature of stems in Guaraní is ambiguous. Those may pertain either to nominal or verbal classes. The same root may represent either a verb or a noun and even an adjective depending on the syntactic role, position in a sentence and morphological units attached to the root. Both nouns and verbs can serve as predicates: verbs express an action and nouns define qualities, states and notions (Dietrich, 2017). As a convention we group nouns, adjectives, adverbs as into a nominal class and refer to them as *nominals* and call verbal stems as *verbals*. Notably 'adjectives' and 'adverbs' are not always distinguished by the researchers in the literature. A lot of roots in these classes take comparative suffixes to form degree constructions at the same time verbs show similar behaviour so we cannot call them adjectives in its' full sense (Dietrich, 2017).

Because of the ambiguity the same stems appear in various basic lexicons (nouns, verbs, adjectives). In Table 1 we illustrate possible analyses for *arandu* root. As a noun *arandu* means 'intelligence' and as an adjective 'wise, educated'.

Our transducer consists of several lexicons: Nouns (4455), Verbs, divided in two groups by transitivity (2537), Adjectives (1668), Adverbs (457) and other morphological categories including pronouns, determiners, toponyms, anthroponyms, barbarisms (in their majority Spanish loanwords), etc. In the following sections we discuss concrete linguistic phenomena in Guaraní as well as present our implementation decisions for them.

4.1.2 Nouns

Nouns in Guaraní can attach various suffixes and prefixes with pronominal, spacial, temporal meaning. They can serve as predicates and incorporate other nouns. Some nouns are so called multi-roots as they have several initial forms expressing different kinds of relations. Figure 2 shows an example of Nouns lexicon. A simplified version of non-deterministic FST for noun derivation is shown on the Figure 3. The figure shows two branches of prefixing possible for nominal stems in Gurarani followed by case inflection, diverse types of derivation (pluralization, degree suffix attachment) and incorporation.

Case affixes Nouns in Guaraní can attach case affixes which sometimes behave as postpositions. The nouns and postpositions often times are written separately as the analysis of Wiki-corpus shows. Such behaviour of affixes is also described in grammar books (Estigarribia and Pinta, 2017). The exam-

Form	Translation
arandu<n>	‘intelligence’
arandu<adj>	‘wise, educated’

Table 1: Possible analyses for ‘arandu’

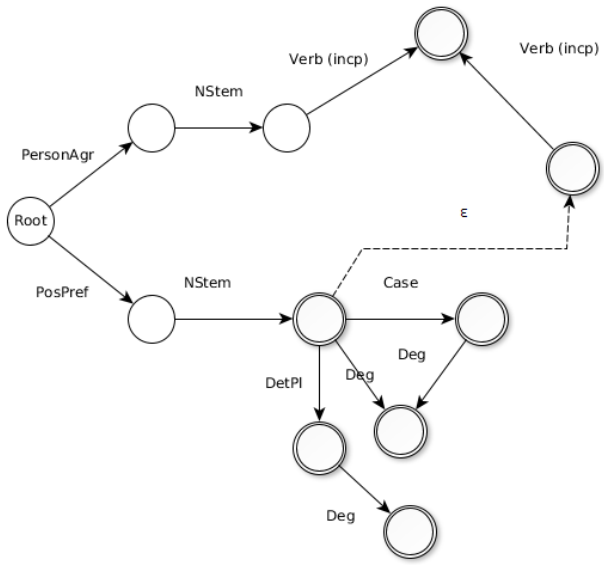

Figure 3: Reduced FST for Guaraní noun derivation/inflection. Labels used: *PersonAgr* for personal agreement prefixes, *PosPref* for possessive prefixes, *NStem* for nominal stems, *Verb* is used for marking incorporation of the noun by verb, *Deg* for degree, *DetPl* for plural determiner.

ples below illustrate the difference between usages of those segments. In (1), the suffix *-pe* (nasal variant of *-me*) expresses locative and in (2), the postposition *hagũa* expresses direction.

(1) tetã-me
country-LOC

‘in the country’

(2) *Ou* *o-mba’apo* *hagũa.*
come-SG3 POSS.SG3-work to

‘S/he comes to work’

The morphotactic transducer (*lexc* file) contains a CASE lexicon with postpositional tag <post> and inflects nominal lexicons.

Incorporation is a morphological process that fuses nouns into a verbal form as a direct object. Normally the object referring to a human being follows the verb. In case of incorporation the object is inserted between personal agreement marker and verbal stem. The verb itself remains intransitive while incorporating a noun. Compare examples (3) and (4) from (Dietrich, 2017).

(3) Iñirũ katu he’i ichupe.
3SG-friend but answer-3SG 3SG.DAT

‘But his friend answered him.’

(4) a. *a-johéi*
SG1-wash

‘I wash (it)’

b. *a-**py**-héi*
SG1-feet-wash

‘I wash my feet.’

Noun incorporation in Guaraní transducer is modelled as follows: verbal stems are attached to ṡtems in NOUNS lexicon.

Multi-form roots A challenging aspect of Guaraní nominals is that some of them have two or three initial forms (they are called biforms and triforms by Krivoshein and multiform roots by Estigarribia). They alter the first allomorph consonant of the word depending on the semantics a speaker wants to express. Most of these forms begin with /t-/ (biforms predominantly express the terms of kinship). Representations of biforms and triforms are distinguished by possessiveness. Absolute form generally begins with /t-/, the second form is relational where the possessor is not a 3P pronominal and starts with /r-/. The third form begins with /h-/ where there is a 3P pronominal possessor (see Table 2, examples are taken from (Estigarribia and Pinta, 2017)).

The transducer handles these allomorphs as determiners or possessive pronouns. The initial form marker /t-/ is eliminated by the rule and then triform nominal stems are appended to /r-/ and /h-/ initial segments (see Figure 4).

4.1.3 Verbs

Verbal classification The most complex part of morphological combinatorics is verbal modelling that could be completed in multiple ways depending on classification strategy. Verbal forms can be

Example	Gloss	Translation	Form
tembiapo	**t**embiapo	'work'	Absolute
Huã rembiapo	Huã **r**-embiapo	'Juan's work'	Relational
hembiapo	**h**-embiapo	'his/her work'	POSS.3-possessor

Table 2: Representations of *tembiapo* noun with it's three forms where the first form is absolute, second is relational with non-pronominal possessor and the third form with the pronominal possessor.

```
LEXICON DetTriformes
r%<det%>%+:r%{t%} Triformes ;
h%<prn%>%<pos%>%+:h{t%} Triformes ;
```

Figure 4: Lexicon defining triforms in `lexc` file. Special character `%{t%}` works here as archiphoneme and is a part of morphophonological module. It is always implied in underlying representation of the word and it actualizes on the surface only when 'r' or 'h' sounds are not around in the context i.e. in absolute forms.

divided by transitivity, areales a(i)reales and chendales. We give the definition for all the subclasses below.

According to (Estigarribia, 2017) *aireales* are the verbs that take /-i-/ sound between personal agreement suffixes and the root. /-i-/ vowel is a phonetic segment that does not carry any morphological load but it can significantly change semantics of the word. For example, areal verb *ke* "to enter" acquires a new meaning "to sleep" when /-i-/ is added. So *a-i-ke* means "I sleep" instead of *a-ke* "I enter".

Chendales is a subclass of verbal stems which attach possessive pronouns as prefixes. Possessive prefixes alter active verbs to states. The example below borrowed from (Estigarribia, 2017) shows the difference of active and stative forms:

(5) **a**-monda
SG1.ACT-steal

'I steal = I am stealing'

(6) **che**-monda
SG1.INACT-steal

'I steal=I am a thief'

Possessive prefixes in chendales behave like a subject of the predicate whereas can be interpreted as objects when attached to a(i)real verbs.

(7) Nde **che**-juhu.
SG2.NOMSG1.ACC-meet

'You meet me.'

Excessive splitting of verbal stems into separate verbal classes (transitive/intransitive, areales/aireales, chendales) can result in overgeneration of non-existing forms. Thus we segregate verbal stems in two lexicons by transitivity and then implement specific morphological alterations for each of the subclasses. For example, aireal verbs acquire /-i-/ phoneme between stem and prefix by using special character `%{i%}` called *archiphoneme* in HFST terminology. It allows to specify the context in the rule for a representation of a phoneme's underlying form. Archiphome is mapped to a set of surface representations of the sound and the context is specified for every surface form including 'zero sound'. Thus `%{i%}` appears as 'zero sound' in areal verbs and /-i-/ in aireales.

Verbal affixes Guaraní verbs undergo personal agreement (see example for verb *ke* – 'to enter' in Table 3) as well can attach tense, aspect and mood markers. A general model of verbal strategy can be found on Figure 5.

Tense, aspect and mood markers attach to the predicate but they are not obligatory unless mark future tense. In case a verb does not take any suffix it may be preceded by an adverbial or postpositional tense marker. Compare the two examples where *-akue* is a past tense marker and *va'ekue* is an adverb:

(8) Aha **va'ekue** nde rógape.
1SG-go ADV.PAST 2SG.POS house-LOC

'I went to your house.'

(9) Ou'**akue** che sy rógape.
Come-PAST 1SG.POS mother house-LOC.

'Came yesterday to my mother's house'.

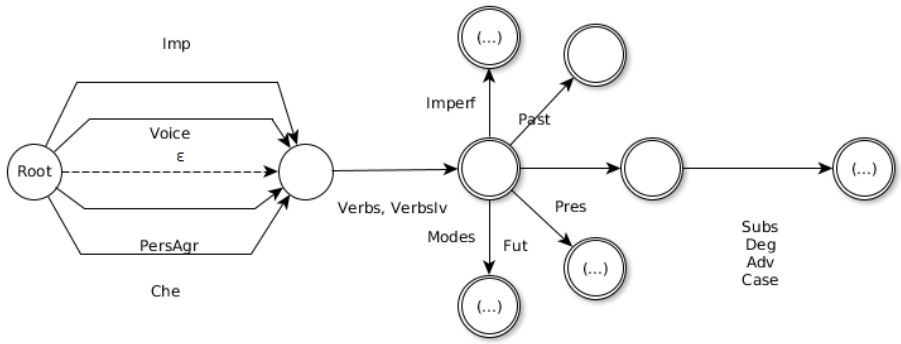

Figure 5: Reduced FST for Guaraní verb strategy. Labels used: *PersAgr* for personal agreement prefixes, *Che* for chendales, *Imp* for imperative, *Deg* for degree, *Imperf* for imperfect. Most of the finite states can be extended further by suffix combinations.

Form	Gloss	Translation
ake	a<prn><p1><sg><nom>+ke	I enter
reke	re<prn><p2><sg><nom>+ke	You enter
oke	o<prn><p3><sg><nom>+ke	S/he enters
jake	ja<prn><p1><pl><nom>+ke	We enter (inclusive)
roke	ro<prn><p1><pl><nom>+ke	We enter (exclusive)
peke	pe<prn><p2><pl><nom>+ke	You enter
oke	o<prn><p3><pl><nom>+ke	They enter

Table 3: Personal agreement for the verb *ke* (enter)

Orthographically there is no agreement in using some of tense markers as affixes or as adverbs. In literature and corpora we can find both interpretations so our analyzer handles it in both ways.

4.2 Morphophonology

Phonological aspects of Guaraní in Apertium analyser are modelled by HFST `twol` formalism and a set of archiphonemes in `lexc` file. `twol` file contains 30 rules that impose constraints on phonological alterations.

As we mentioned Guaraní is oxytone language i.e. the end of the word is always stressed. Accents are used for marking exceptions from this rule. Suffixes (or postpositions) that can attach to the stem may be tonal or atonal. As the stress is generally not marked it causes the shift of the accents in writing. If the suffix is tonal and it is attached to the root the stress should be removed from the stem and shifted to the tonal affix as in plural form of *óga* ('house') – *ogakuéra*. The case of tonal suffixes is solved by a phonological rule specifying contexts where the corresponding characters must change (Figure 6).

One more specific feature of Guaraní phonology is nasalization. Both vowels and consonants can be nasal/nasalized. A special character used for indicating nasalization is tilde. If a syllable contains nasal phoneme it automatically becomes nasal so there is no need to mark the rest of the phonemes of the syllable with the tilde. Although, if the word is a compound and/or incorporates two (or more) nasal roots both tildes remain. The same rule applies to nasal morphemes attached to the root as in examples below.

(10) akãperõ
akã-perõ
'bold-headed'

(11) omitãmohavõ
o-mitã-mohavõ
SG3-child-soap
'She soaps the child'

Nasalization affects suffixes and prefixes with a consonant adjacent to the root if the root is nasal. Consonants change to their nasal equivalents with the same place of articulation i.e. j → n, k → ng, etc. (see Figure 7).

```
"Change tonal vowel to atonal if tonal in affix"
Vt:Va <=> •:_ [Cns:|ArchiCns:|Nas:|VowsAton:|%>: ]+  VowsTon: ;
          •:_ [Cns:|%>:|ArchiCns:|Nas:|VowsAton:]+ %{E%}: ;
          •:_ [Cns:|%>:|ArchiCns:|Nas:|VowsAton:]+ [%{Y%}: g:u:a:] ;


"Delete ending -[i] before comparative -icha"
Vx:0 <=> _ %>: [ i: c: h: a: | %r%:0 i: ] ;
            where Vx in ( ĩ i í ) ;
```

Figure 6: The first `twol` rule handles alteration of tonal vowel (`Vt`) after a special character "•" that we added to each word form in `lexc` file containing accents to indicate tonal vowel if in the following context there are any tonal vowels (`VowsTon`). The second rule executes vowel deletion when preceding *icha* suffix or zero surface `%r%` suffix followed by `-i` which appears in negative circumfix if the stem ended in vowel.

```
•:0 ó:o g:g a:a >:0 {N}:0 {K}:k u:u é:é r:r a:a
i:i r:r ũ:ũ >:0 {N}:n {K}:g u:u é:é r:r a:a
```

Figure 7: Example of transducer's output for nasalization of *kuéra* plural suffix. Archiphonemes `{N}` and `{K}` actualize in a surface form preceded by nasal vowel `ũ` (`>` is a special symbol used for morpheme boundary).

Except nasalization our analyser handles phoneme deletion, vowel alteration, phoneme insertion (including glottal stop between two vowels). Transition of tonal vowel to atonal is showed on Figure 6. This rule applies to the words having a tonal vowel in the stem marked with tonal accent as in Spanish. Vowel 'é' in verb 'wash' 'johéi' changes to 'e' when suffix 'hína' indicating imperfect is added. As a result we receive `joheihína`. The other rule handles vowel deletion to avoid duplicate `i` sound on the morphemes' boundary. This can occur when comparative suffix *-icha* is added to a stem ending with `-i`. As in `morotĩ` 'white' underlying form would result in vowel duplication `morotĩ<adj>+icha<comp>` → `morotĩicha`. The rule deletes the duplicate `i` and we receive `moroticha` '(equally) white' on the surface.

5 Evaluation

To evaluate the analyser we estimate naive coverage metric and compare it to *ParaMorfo* system. *Naive coverage*[6] is the ratio of tokens that receive at least one morphological analysis to the total number of tokens in the corpus.

We estimate performance of our transducer on two publically available corpora: the Bible and the Guaraní Wikipedia. An example of a fully analysed Guaraní sentence is presented in Table 4. The asterisk, *, marks the example of erroneous output. Pronouns like *che* can serve as possessive and personal pronouns. The morphological analyser did not solve the case correctly, as we initially presumed that only personal pronouns will be written separately. Correct analysis of this lexeme is `che<prn><pos><p1><sg>`. Cases like this require enhancement so that any orthographical inconsistencies could be parsed. Table 5 shows the results of naive coverage evaluation as compared to *ParaMorfo*.

For fair comparison we ran Wikipedia texts and the Bible through Apertium analyser and dropped all the tokens that did not belong to open-class category because ParaMorfo does not recognize closed-class words (adverbs, conjunctions, numbers, etc.) and punctuation marks. ParaMorfo segments tokens differently than our analyser so at the end of processing we received different quantity of entry tokens for each analyser.

According to Table 5 the naive coverage of Apertium analyser is significantly higher than of ParaMorfo. One reason is that the latter does not cover Spanish barbarisms present in the corpora in increased proportion after closed class tokens are excluded. Moreover, ParaMorfo does not recognize proper names such as toponyms and anthroponyms.

We also evaluate conventional quality metrics for the analyser such as precision, recall and F-measure. To estimate precision, recall and F-measure we have annotated 8308 tokens from different sources where each tokens has a corresponding valid analysis in the context. Note that this estimate is only the approximation of the scores because in order to have true scores each form should be annotated with *all* valid

[6]The metric is called *naive* coverage because even if the word received an analysis it may not be grammatically correct e.g. in cases of over-generation or some grammatically correct analyses may not be delivered by the transducer.

Surface form	Analysis
Ojapo	o<prn><p3><pl><nom>+japo<v><tv><pres>
oréve	ore<adj>+ve<adj><dist>
guarã	guarã<post>
kuehe	kuehe<adv>
chipa	chipa<n>
che	*ché<prn><pers><p1><sg>
sy	sy<n>
.	.<sent>

Table 4: Example morphological analysis of Guaraní sentence with Apertium tag style. Note that morphological ambiguity in the example was manually solved for illustration purposes.

Corpus	Coverage	Tokens
Apertium:		
Wikipedia	0.86	375989
Bible	0.91	482941
ParaMorfo		
Wikipedia	0.54	379736
Bible	0.64	631724

Table 5: Naive coverage evaluation

analyses of the words instead of a single analysis per word. This is not an easy task to complete without a native speaker.

We define *true positives* as the list of the analyses present both in the gold standard and the transducer's output, *false positives* as those analyses in the transducer's output but not in the gold standard. Finally, *false negatives* are the analyses found in the gold standard but not the analyser's output. This evaluation method was previously used by Richardson and Tyers (2021). Apertium analyser yields the following scores: *precision* 0.30, *recall* 0.86 and *F1-score* 0.45.

Precision here reflects the likelihood of the form produced by the analyser to be in the gold standard, which is high in our case. Precision shows low score because the annotated data only contains one valid morphological analysis per word. Thus, overall we can conclude that the likelihood of the word being analysed correctly is fairly high. It does not possible to compare our results with *ParaMorfo* in this case because of the differences in morpheme mapping between the analysers, this way to do fair comparison additional effort is needed to annotate data using *ParaMorfo*'s tag convention.

Another metric we asses is *average ambiguity rate*, the average number of morphological analyses given by the transducer per word. Average ambiguity rate for Wikipedia corpus is 3.018 analyses per token and for Bible – 3.450 analyses. This fact gives us an interesting observation that Guaraní language is *moderately* polisynthetic as compared to other languages that according to (Estigarribia and Pinta, 2017) may have 5-6 analyses per word.

To briefly summarize our contributions in comparison with ParaMorfo analyser:

- Apertium analyser recognises closed class forms (adverbs, conjunctions, numerals) and punctuation;
- Handles Spanish barbarisms and Proper nouns;
- Flexible with orthographic variation.

6 Conclusions

We presented a finite-state morphological analyser for one of the indigenous polysynthetic languages of South America – Paraguayan Guaraní. Further work implies the expansion of the existing lexicons to reach most possible coverage mainly by adding more stems to continuation lexicons (verbs, nouns, proper names). Currently the analyser provides all possible analyses for a token and it requires further work on morphological disambiguation.

Acknowledgements

We would like to thank Michael Gasser, emeritus professor of Informatics and Computing at Indiana University, Dmitry Gerasimov, from the Institute for Linguistic Studies of the Russian Academy of Sciences, as well as the reviewers for the thorough evaluation of our work and suggestions.

References

Vasilisa Andriyanets and Francis M. Tyers. 2018. A prototype finite-state morphological analyser for Chukchi. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 31–40, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Wolf Dietrich. 2017. Word classes and word class switching in Guaraní syntax. In *Guarani Linguistics in the 21st Century*, pages 101–137. Brill.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2018. Guaraní. `http://www.ethnologue.com`.

Bruno Estigarribia. 2017. A grammar sketch of Paraguayan Guarani. In *Guarani Linguistics in the 21st Century*, pages 7–85. Brill.

Bruno Estigarribia and Justin Pinta, editors. 2017. *Guarani Linguistics in the 21st Century*. Brill.

Mikel Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25:127–144.

Michael Gasser. 2018. Mainumby: un ayudante para la traducción Castellano-Guaraní. In *Tercer Seminario International sobre Traducción, Terminología y Lenguas Minorizadas*.

Emma Gregores and Jorge Suarez, editors. 1967. *A Description of Colloquial Guaraní*. Mouton & Co.

M. Gökırmak and F. M. Tyers. 2017. A dependency treebank for kurmanji kurdish. In *Proceedings of the the International Conference on Dependency Linguistics, Depling 2017*.

Ryan Johnson, Lene Antonsen, and Trond Trosterud. 2013. Using finite state transducers for making efficient reading comprehension dictionaries. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, volume 85, pages 59–71.

Natalia Krivoshein de Canese. 1983. *Gramática de la lengua guaraní*. Asunción.

Scott Ledbetter and Marcus Dickinson. 2016. Cost-effectiveness in building a low-resource morphological analyzer for learner language. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 206–216.

S. Moshagen, T. Trosterud, J. Rueter, F. M. Tyers, and T. A. Pirinen. 2014. Open-source infrastructures for collaborative work on under-resourced languages. In *Proceedings of CCURL workshop 2014 organised in conjunction with LREC2014*.

Tommi A Pirinen. 2015. Development and use of computational morphology of Finnish in the open source and open science era: Notes on experiences with Omorfi development. *SKY Journal of Linguistics*, 28:381–393.

Tommi A. Pirinen and Krister Lindén. 2014. State-of-the-art in weighted finite-state spell-checking. In *Computational Linguistics and Intelligent Text Processing*, pages 519–532, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ivy Richardson and Francis M. Tyers. 2021. A morphological analyser for k'iche'.

Lane Schwartz, Francis Tyers, Lori Levin, Christo Kirov, Patrick Littell, Chi kiu Lo, Emily Prud'hommeaux, Hyunji Hayley Park, Kenneth Steimel, Rebecca Knowles, Jeffrey Micher, Lonny Strunk, Han Liu, Coleman Haley, Katherine J. Zhang, Robbie Jimmerson, Vasilisa Andriyanets, Aldrian Obaja Muis, Naoki Otani, Jong Hyuk Park, and Zhisong Zhang. 2020. Neural polysynthetic language modelling.

Vidalia Sánchez, editor. 2018. *Guarani Ñe'ẽ Rerekuapavẽ*. Editorial Servilibro.

Jörg Tiedemann, Filip Ginter, and Jenna Kanerva. 2015. Morphological Segmentation and OPUS for Finnish-English Machine Translation. Technical report, University of Turku.

F. M. Tyers and V. Ravishankar. 2018. A prototype dependency treebank for breton. In *Actes de la conférence Traitement Automatique de la Langue Naturelle, TALN 2018*, pages 197–204.

Çağrı Çöltekin. 2010. A freely available morphological analyzer for Turkish. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC2010)*.