Initial Experiments for Building a Guarani WordNet

Luis Chiruzzo
Marvin Agüero-Torales
Aldo Alvarez
Yliana Rodríguez

Global WordNet Conference 2023

UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY

UNI
PARAGUAY

UNIVERSIDAD
DE GRANADA

Introduction

Guarani: South American indigenous language

Between 6 and 10 million speakers

Paraguay, Argentina, Brazil, Bolivia

Contact with Spanish and other European languages for about 500 years

Both indigenous and non-indigenous speakers

NLP for Indigenous Languages

Low resource languages

Very little work in NLP

AmericasNLP

What about WordNet?

NLP for Indigenous Languages

Low resource languages

Very little work in NLP

AmericasNLP

What about WordNet?

- Shipibo-Konibo (Maguiño-Valencia et al., 2018)
- Quechua (Melgarejo et al., 2022)

Guarani Language

Tupi-Guarani family

Agglutinative and polysynthetic

Mostly SVO

Uses postpositions for marking case

Guarani Language

Verbal morphology

ndaguatái

I don’t walk

Guarani Language

Verbal morphology

ndaguatái

I don’t walk

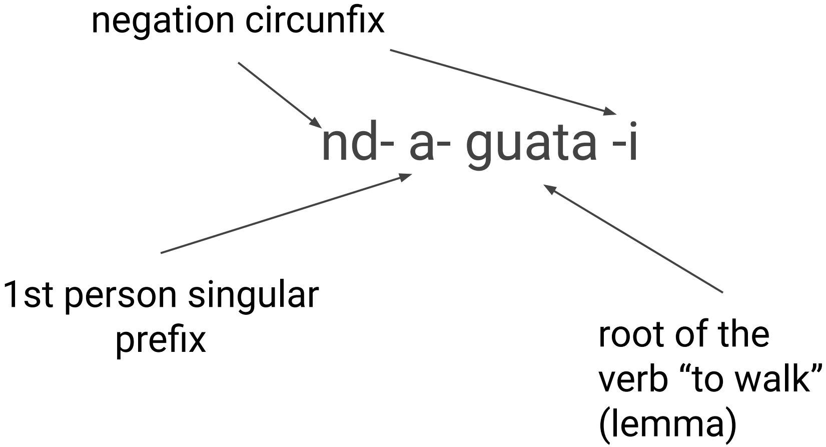
nd- a- guata -i

Guarani Language

Verbal morphology

ndaguatái

I don’t walk



negation circunfix

nd- a- guata -i

1st person singular prefix

root of the verb “to walk” (lemma)

Guarani Language

Noun morphology: triform nouns

tembi’u

food (generic)

hembi’u

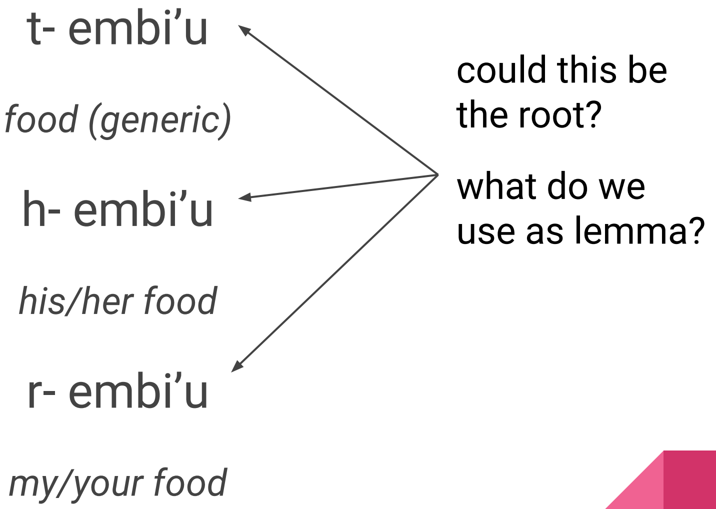
his/her food

rembi’u

my/your food

Guarani Language

Noun morphology: triform nouns



t- embi'u

food (generic)

h- embi'u

his/her food

r- embi'u

my/your food

could this be the root?

what do we use as lemma?

Process Overview

- Set of lemmas in the source language sl_i
- Set of lemmas in the target language tl_j
- Set of possible translations between lemmas $<sl_i, tl_j>$

Selectors: heuristic that takes the sl_i lemmas of a synset *s*, and the tl_j translation candidates, and chooses which candidates are suitable for *s*

Sources: bilingual dictionaries (Guarani-Spanish). POS are necessary, too.

Selectors - Monosemy

sl_i only belongs to one synset s (it is monosemic)

⇨

assign all translations tl_j to the synset s

Selectors - Monosemy

sl_i only belongs to one synset s (it is monosemic)

⇒

assign all translations tl_j to the synset s

This assumes that WordNet is *complete*, so that monosemic lemmas are really monosemic

Selectors - Single Translation

For a lemma sl_i there exists only one translation tl_j

sl_i belongs to synsets $s_1 \dots s_n$

⇨

assign tl_j to synsets $s_1 \dots s_n$

Selectors - Single Translation

For a lemma sl_i there exists only one translation tl_j

sl_i belongs to synsets $s_1 \dots s_n$

⇒

assign tl_j to synsets $s_1 \dots s_n$

This assumes the dictionaries are *complete*, so that tl_j is the correct translation in all possible contexts

Selectors - Factorization

Synset s has many lemmas $\{sl_1 \ldots sl_n\}$ (n >= 2)

the lemmas have translations $\{ tl_{1,1} \ldots tl_{1,j} \} \ \ldots \ \{ tl_{n,1} \ldots tl_{n,k} \}$

⟹

assign the intersection of the translation sets to synset *s*

Selectors - Factorization

More restrictive, but at the same time more precise

Synset s has many lemmas $\{sl_1 \ldots sl_n\}$ (n >= 2)

the lemmas have translations $\{ tl_{1,1} \ldots tl_{1,j} \} \ \ldots \ \{ tl_{n,1} \ldots tl_{n,k} \}$

$\Longrightarrow$

assign the intersection of the translation sets to synset *s*

Dictionaries

Avalos: *Ñe'eryruguasu* Guarani-Spanish bilingual dictionary by Celso Ávalos Ocampos

    18k lemma pairs

DC: Online dictionary from the web portal Descubriendo Corrientes

    14k lemma pairs

Wiktionary: Multilingual collaborative online dictionary

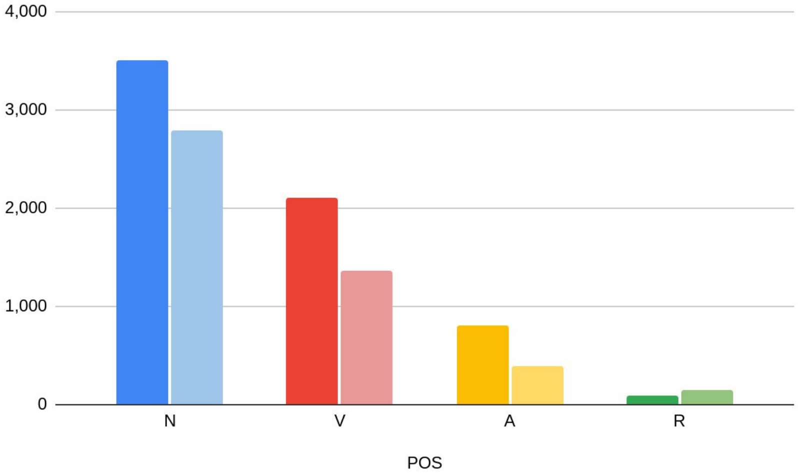
    2k lemma pairs

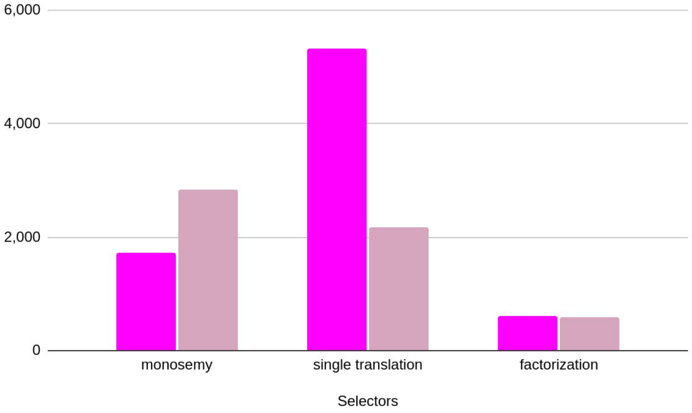
Overall Results

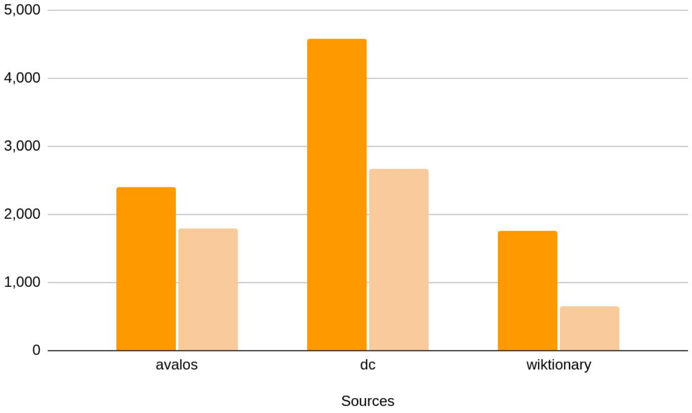
12k <synset, lemma> pairs

6.5k unique synsets

4.3k unique lemmas







Evaluation

Two native speakers annotated a sample of results

Annotators were given this context:

- synset ID
- Spanish translation of the synset definition
- known Spanish lemmas (from Spanish WordNet)

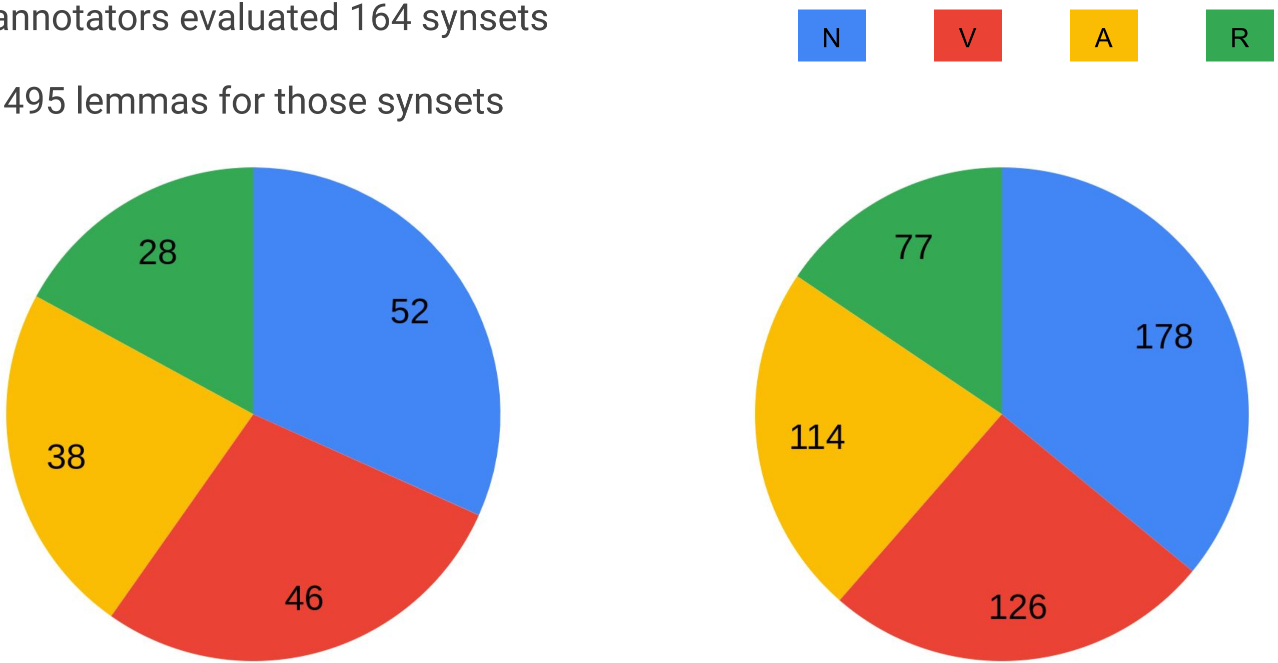
They labeled if the lemmas extracted by the selectors for that synset are valid

Also, they could suggest lemmas not chosen by the selectors

Evaluation

In total the annotators evaluated 164 synsets

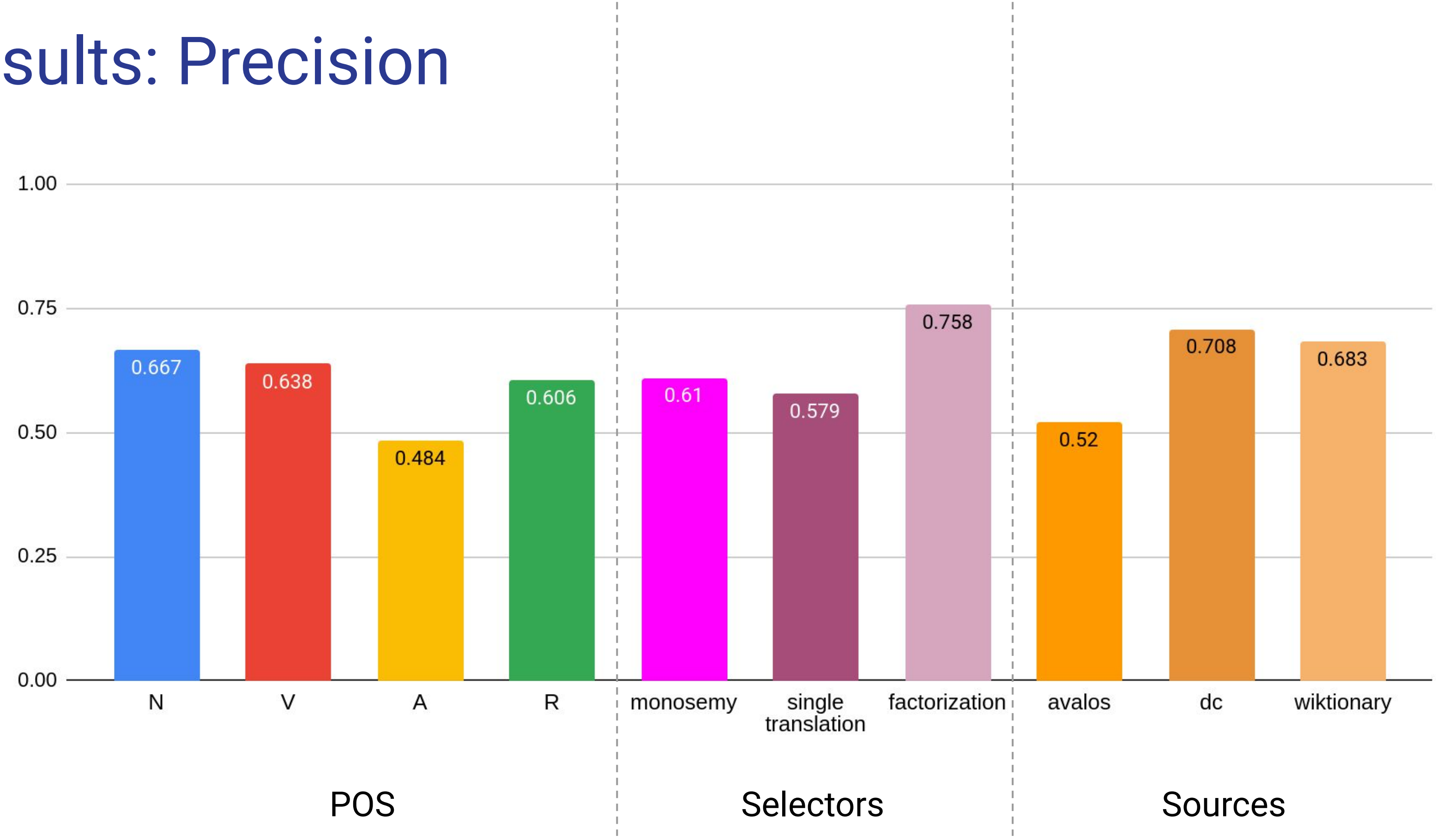
They chose 495 lemmas for those synsets



Synsets

Lemmas

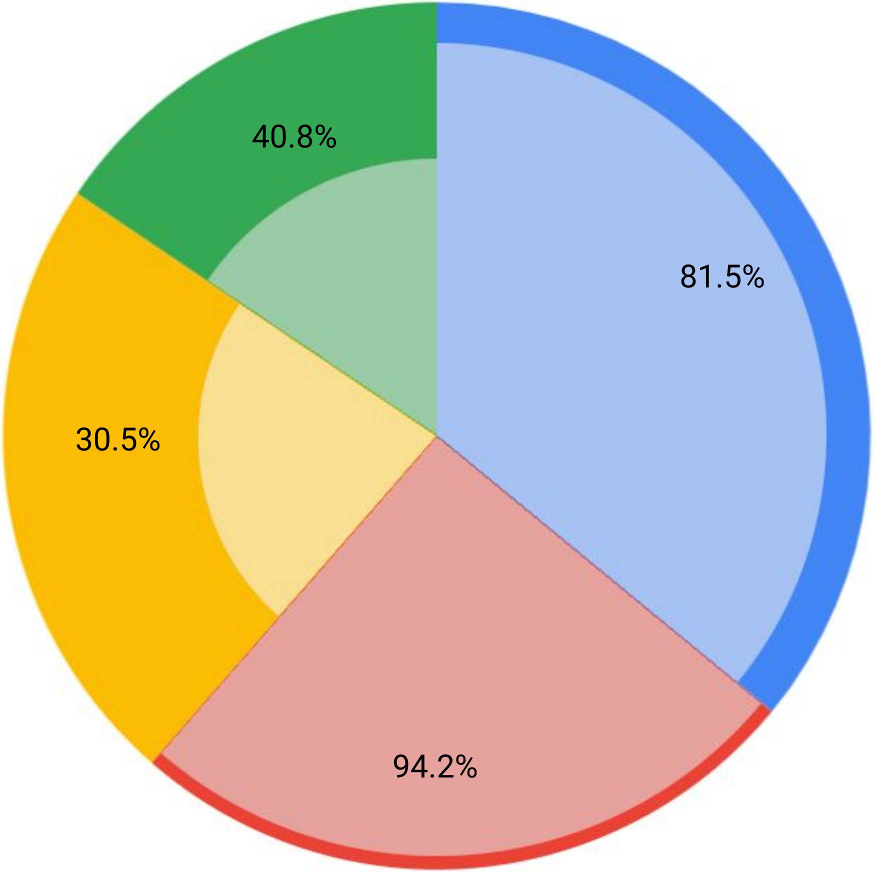
Results: Precision



Results: Coverage

Coverage of the union of sources

🟦 N 🟥 V 🟨 A 🟩 R



Results: Discussion

Lower precision

Single Translation → Incomplete dictionaries

Monosemy → Incomplete Spanish WordNet

Results: Discussion

Lower precision

Single Translation → Incomplete dictionaries

Monosemy → Incomplete Spanish WordNet

Higher precision

Factorization → But fewer lemmas found

Results: Discussion

Annotators consistently chose all forms of a triform noun

branch.n.02

Results: Discussion

Annotators consistently chose all forms of a triform noun

branch.n.02

- takã
- hakã
- rakã
- takãmby
- hakãmby
- rakãmby

Results: Discussion

Annotators consistently chose all forms of a triform noun

branch.n.02	- takã	✅
	- hakã	✅
	- rakã	✅
	- takãmby	❌
	- hakãmby	❌
	- rakãmby	❌

Results: Discussion

Annotators consistently chose all forms of a triform noun

branch.n.02	- takã	✅
	- hakã	✅
	- rakã	✅
	- takãmby	❌
	- hakãmby	❌
	- rakãmby	❌

So all forms of the triform nouns are added to the synset

Results: Discussion

Inconsistencies in orthography

Especially in Wiktionary:

Results: Discussion

Inconsistencies in orthography

Especially in Wiktionary:

Spanish		Guarani
rama	-	rakä
rama	-	hakã
rama	-	taka

Results: Discussion

Inconsistencies in orthography

Especially in Wiktionary:

Spanish		Guarani	
rama	-	rakä	non-standard diacritic
rama	-	hakã	standard orthography
rama	-	taka	no diacritic at all!

Conclusions

Three selectors, three bilingual dictionaries

11,967 lemmas extracted for 6,519 synsets

Best precision for factorization (76%), and DC dictionary (71%)

Conclusions

Three selectors, three bilingual dictionaries

11,967 lemmas extracted for 6,519 synsets

Best precision for factorization (76%), and DC dictionary (71%)

Create new selectors

Expand the evaluation

Expand the sources, for example use larger corpora or MT systems

Aguyje!

Thank you!