UPPSALA
UNIVERSITET

Max-Planck-Institut
für evolutionäre Anthropologie

UFPB

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Tupian Language Resources

Data, Tools, Analyses

24. Juni 2022

Introduction

What is TuLaR?

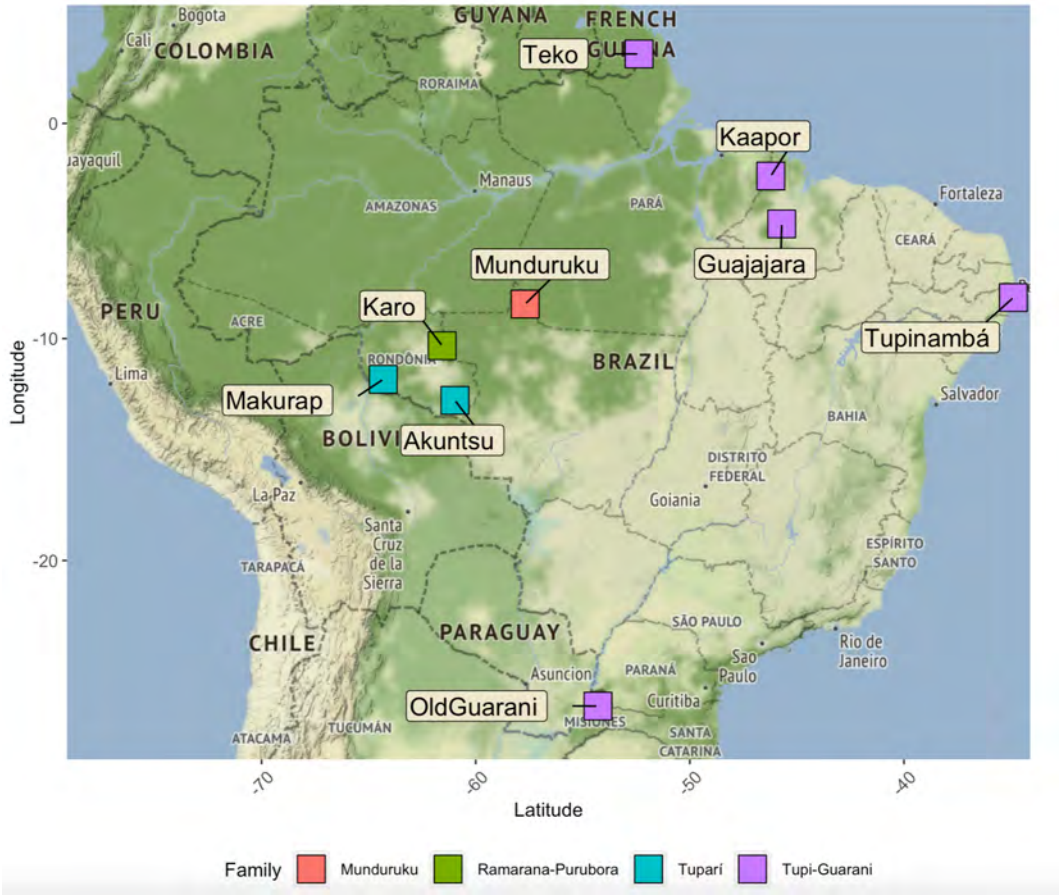
Database	Languages
TuLeD (Tupian Lexical Database)	90
TuMoD (Tupian Morphological Database)	51
TuPAn (Tupian Plants and Animals)	26
TuDeT (Tupian Dependency Treebanks)	9

Our Goals

- Production of computational resources
- Production of linguistic knowledge
- Analysis of syntax and morphology
- Creation of resources for threatened languages
- Collaboration with indigenous communities
- Increasing the linguistic and cultural knowledge of South American indigenous languages.

Example of collaboration with the communities

The Tupian Dependency Treebanks (TuDeT)

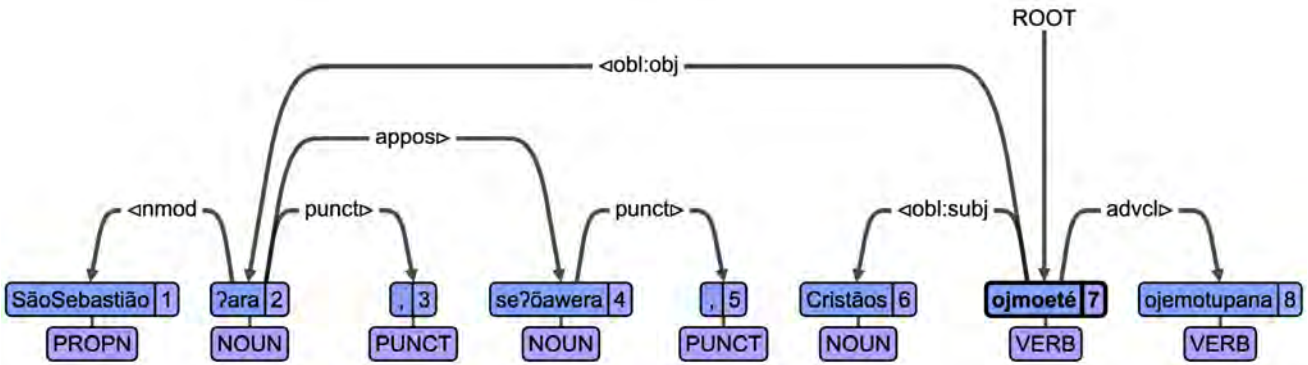


The Languages in TuDeT

Doculect	Glottocode	Number of speakers	Status
Akuntsú	akun1241	3	Nearly extinct
Guajajara	guaj1255	12.000	Vigorous
Ka'apor	urub1250	600	Developing
Karo	karo1305	200	Vigorous
Makurap	maku1278	40	Moribund
Mundurukú	mund1330	5000	Threatened
Old Guaraní	oldp1258	0	Extinct
Tekó	emer1243	400	Vigorous
Tupinambá	tupi1273	0	Extinct

The Universal Dependencies Framework

```
# sent_id = 0011.55
# text = SãoSebastião ʔara , seʔõawera , Cristãos ojmoeté ojemotupana
# text_eng = The christians honour Saint Sebastian's day, of his death, making it (a) holy(day).
1   SãoSebastião    SãoSebastião    PROPN   propn   _   2   nmod    _   _
2   ʔara    ʔar NOUN    n   Case=Ref    7   obl:obj _   _
3   ,   ,   PUNCT   punct   _   2   punct   _   _
4   seʔõawera   eʔõ NOUN    n   Case=Ref|Rel=NCont|Tense=Past   2   appos   _   _
5   ,   ,   PUNCT   punct   _   4   punct   _   _
6   Cristãos    Cristão NOUN    n   _   7   obl:subj    _   _
7   ojmoeté eté VERB    v   Person[obj]=3|Person[subj]=3|Voice=Cau  0   root    _   _
8   ojemotupana tupa    VERB    v   Person=3|Person[subj]=3|Reflex=Yes|VerbForm=Ger||Voice=Cau  7   advcl   _   _
```



The Annotation Process

Sources:

- Grammatical descriptions
- Religious texts
- Fieldwork data collection
- Cultural texts

Data Standardization:

- Phonetic representation
- Word boundaries

Annotation

Manual annotation:

- Part-of-speech tagging
- Features
- Dependency relations

Supervised annotation:
UDPipe 1

Language	Sentences	Tokens
Akuntsu	243	1056
Guajajara	1126	8702
Ka'apor	83	366
Karo	674	2319
Makurap	31	146
Munduruku	158	1016
Old Guarani	59	212
Teko	100	232
Tupinamba	546	4089

Amount of sentences and tokens in each TuDeT treebank

Tupian Lexical Database (TuLeD)

3402	Arua	BAT	dʒĩip	dʒĩip	dʒ ĩ i p	dʒ ĩ i p - -	173[8]	1146[8]	BAT
2569	Cinta-Larga	BAT	ʒip	ʒip	ʒ i p	ʒ i - p - -	173[8]	1146[8]	BAT
2733	Gavião	BAT	dʒip	dʒip	dʒ i p	dʒ i - p - -	173[8]	1146[8]	BAT
21952	Kepkiriwat	BAT	iep	iep	ie p	- ie - p	173[8]	1146[8]	BAT
3135	Monde	BAT	ʒip	ʒip	ʒ i p	ʒ i - p - -	173[8]	1146[8]	BAT
385	Purubora	BAT	ʃipẽj / motaĩ	ʃipẽj	ʃ i p ẽ j	ʃ i - p ẽ j	173[8]	1146[8]	BAT
3910	Surui-Paiter	BAT	lĩip	lĩip	l ĩ p	l ĩ - p - -	173[8]	1146[8]	BAT
3725	Zoro	BAT	dʒip	dʒip	dʒ i p	dʒ i - p - -	173[8]	1146[8]	BAT

TuLeD

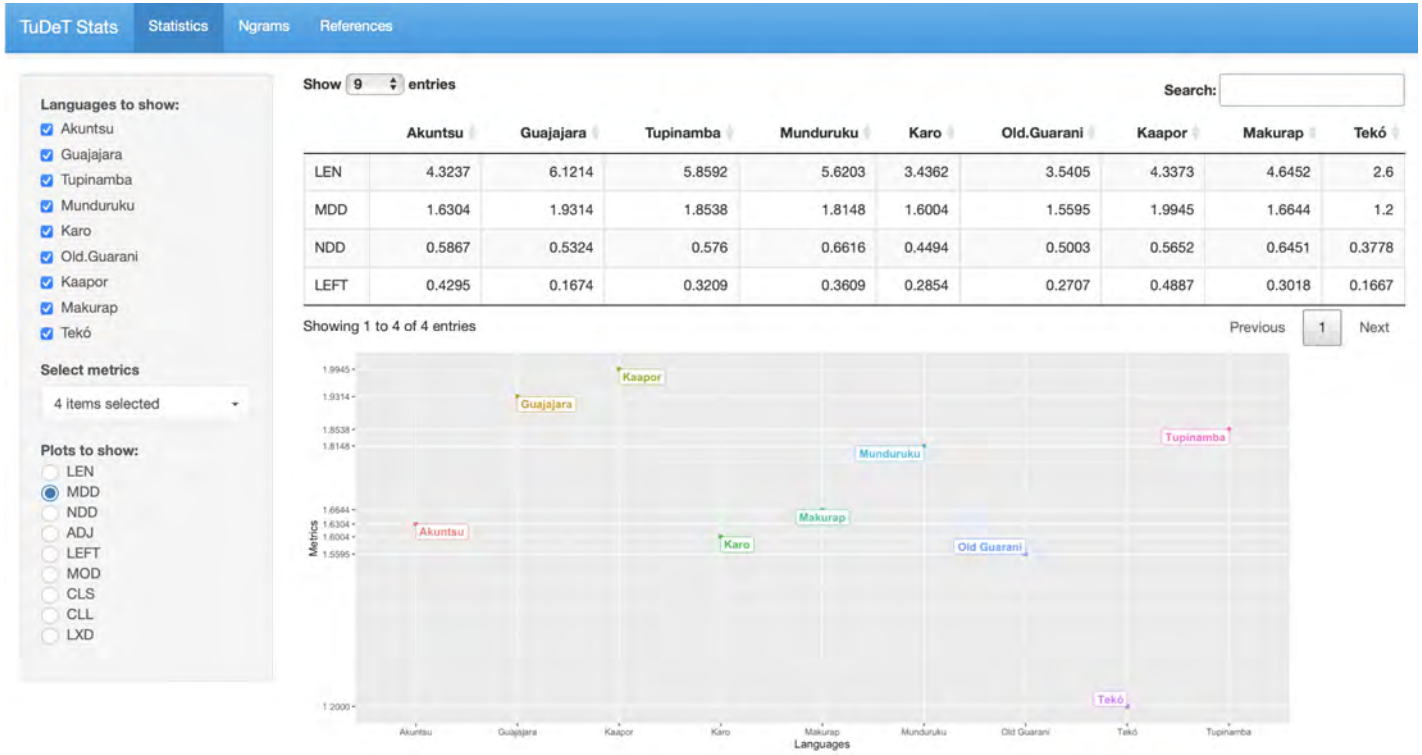
Phonetic alignment in TuLeD

TuDeT Tools

TuDeT Stats

- Complexity measures
  - Mean Dependency Distance
  - Left dependents proportion
  - Normalized Dependency Distance
  - POS tags
  - Syntactic dependencies
- Unigrams
- POS n-grams
- Left dependents count

TuDeT Stats



Morphological analyzers

```
apply up ooroğ
NUMBER=SING|PERSON=1+Perfective+(hunt)

apply up oxi
1SG+R1+mother

apply down 1SG+R1+arrow
odop

apply up tao
R2+leg
```

Example from the Munduruku analyzer

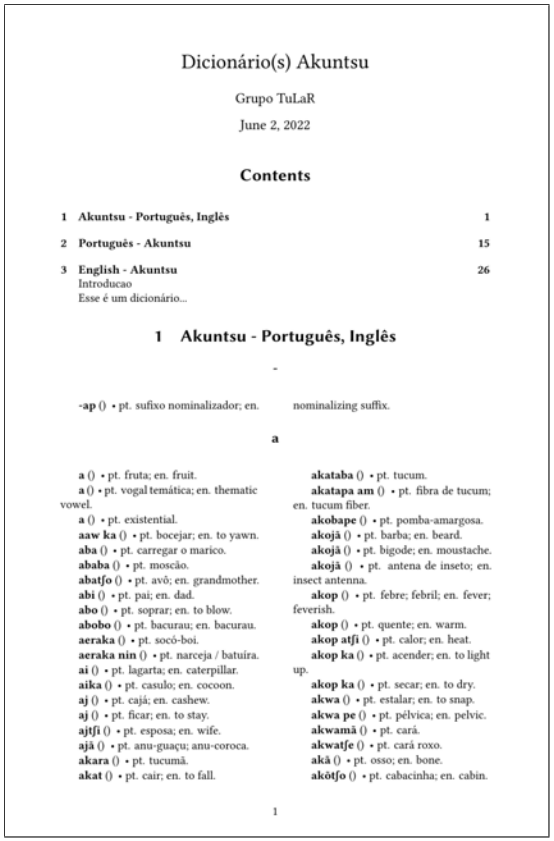
TuLaR in the context of under-resourced NLP

- NLP for under-resourced, endangered, minority and minoritized languages is technically and ethically different
- Lack of NLP support endangers these communities even more
- We follow both FAIR (findable, accessible, interoperable, reusable) and CARE (collective benefit, authority to control, responsibility, ethics) principles
- Tools and data should empower the communities, also in non-immediately academic output

Picture: Rodolfo Oliveira, Agência Pará

First example: learning material

- Along with data validation: automatic generation of learning material
- Output with minimal technological dependencies and an immediate impact on the communities
- Tools for continuous integration: generates both static HTML and PDFs (via LaTeX) with a template language (Jinja2)



Dicionário(s) Akuntsu

Grupo TuLaR

June 2, 2022

Contents

1 Akuntsu - Português, Inglês

-

-ap () • pt. sufixo nominalizador; en. nominalizing suffix.

a

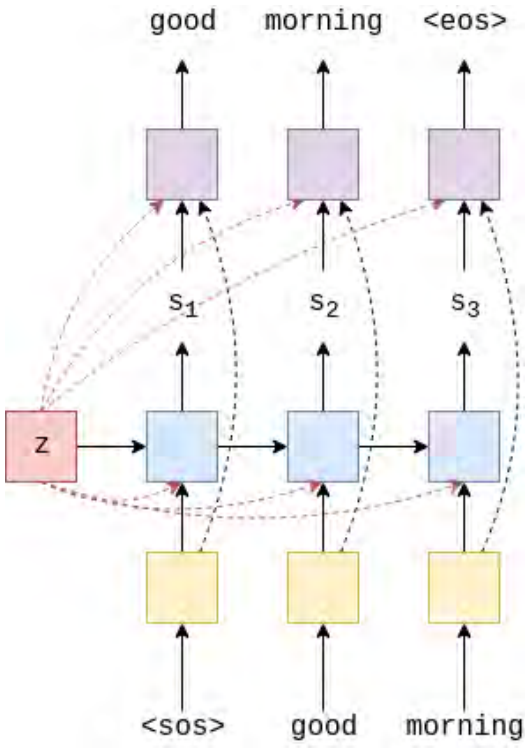
a () • pt. fruta; en. fruit.
a () • pt. vogal temática; en. thematic vowel.
a () • pt. existential.
aaw ka () • pt. bocejar; en. to yawn.
aba () • pt. carregar o marico.
ababa () • pt. moscão.
abatʃo () • pt. avó; en. grandmother.
abi () • pt. pai; en. dad.
abo () • pt. soprar; en. to blow.
abobo () • pt. bacurau; en. bacurau.
aeraka () • pt. socó-boi.
aeraka nin () • pt. narceja / batuíra.
ai () • pt. lagarta; en. caterpillar.
aika () • pt. casulo; en. cocoon.
aj () • pt. cajá; en. cashew.
aj () • pt. ficar; en. to stay.
ajtʃi () • pt. esposa; en. wife.
ajã () • pt. anu-guaçu; anu-coroca.
akara () • pt. tucumã.
akat () • pt. cair; en. to fall.
akataba () • pt. tucum.
akatapa am () • pt. fibra de tucum; en. tucum fiber.
akobape () • pt. pomba-amargosa.
akojã () • pt. barba; en. beard.
akojã () • pt. bigode; en. moustache.
akojã () • pt. antena de inseto; en. insect antenna.
akop () • pt. febre; febril; en. fever; feverish.
akop () • pt. quente; en. warm.
akop atʃi () • pt. calor; en. heat.
akop ka () • pt. acender; en. to light up.
akop ka () • pt. secar; en. to dry.
akwa () • pt. estalar; en. to snap.
akwa pe () • pt. pélvica; en. pelvic.
akwamã () • pt. cará.
akwatʃe () • pt. cará roxo.
akã () • pt. osso; en. bone.
akõtʃo () • pt. cabacinha; en. cabin.

1

First page of an Akuntsu/Portuguese/English dictionary

Second example: machine translation

- Tupían and other under-resourced languages are hard to tackle for automatic machine translation
- Following the recent success of Bapna et al. (2022), we are exploring neural machine translation with different techniques (transfer learning, zero-shot, multi-way translation, enriched data, etc.)
- Very open to collaboration!



Conclusion

- TuLaR: Linguistic description, documentation and creation of NLP resources
- Future projects: NLP resources to support revitalization
- Triad documentation-conservation-revitalization

This research is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 834050).

lorena.martin-rodriguez@uni-tuebingen.de