Proceedings of SIGUL2022 @LREC2022, pages 48–58
Marseille, 24-25 June 2022

Tupian Language Resources

Data, Tools, Analyses

Lorena Martín Rodríguez[*]
Tatiana Merzhevich[*]
Wellington Silva[O]
Tiago Tresoldi[‡]
Carolina Aragon[†]
Fabrício Ferraz Gerardi[*]
[*]Universität Tübingen
lorena.martin-rodriguez, tatiana.merzhevich, fabricio.gerardi@uni-tuebingen.de

[O]Fundação Getúlio Vargas-RJ
wellington.71319@gmail.com

[‡]Uppsala Universitet
tiago.tresoldi@lingfil.uu.se

[†]Universidade Federal da Paraíba
carolinac.aragon@gmail.com

Abstract

TuLaR (Tupian Language Resources) is a project for collecting, documenting, analyzing, and developing computational and pedagogical material for low-resource Brazilian indigenous languages. It provides valuable data for language research regarding typological, syntactic, morphological, and phonological aspects. Here we present TuLaR's databases, with special consideration to TuDeT (Tupian Dependency Treebanks), an annotated corpus under development for nine languages of the Tupian family, built upon the Universal Dependencies framework. The annotation within such a framework serves a twofold goal: enriching the linguistic documentation of the Tupian languages due to the rapid and consistent annotation, and providing computational resources for those languages, thanks to the suitability of our framework for developing NLP tools. We likewise present a related lexical database, some tools developed by the project, and examine future goals for our initiative.

Keywords: Tupian Languages, NLP, Amazonian Languages, Historical Linguistics, Treebanks, Morphology, Finite-State

1. Introduction

The Tupian Language Resources (TuLaR) project follows the precept of promoting linguistic resource development for minority or under-studied languages (Hinton, 2003; Pine and Turin, 2017), especially considering how limited availability interferes with the subsequent production of scientific knowledge and commercial support (Mager et al., 2018; Hedderich et al., 2021). In many scenarios, the lack of such resources leads scientific and commercial initiatives for computational linguistics to only engage with majority or dominant languages, even when there are multi- and cross-linguistic concerns. Such an effect builds up hidden biases against low-resourced languages, even from their own speakers, and, as such, our effort is in line with the objectives of the conference's call: by providing the computational foundations and facilitating the production of teaching material, we aim at fostering the direct participation of minority language communities in the development of computational resources and theoretical knowledge.

The goal of TuLaR is to contribute to the production of computational resources and linguistic knowledge for research and for cooperative work with indigenous communities, especially for those whose languages are categorized as threatened or vulnerable (Eberhard et al., 2021; Languages Project, 2020). It aims to improve the understanding of its morphology and syntax interrelations, thus facilitating their use in natural language processing tasks. For this, we are building different databases (lexical, syntactic, morphological, and fauna-flora) that also aim to consider the historical relations among Tupian languages, as well as its contemporary use, in order to support multilingual tasks that can contribute in increasing the linguistic and cultural knowledge of South American indigenous languages.

TuLaR comprises four databases: TuLeD (Tupian

Lexical Database) (Gerardi et al., 2021b; Gerardi et al., 2021a) with 90 languages (upcoming release), TuMoD (Tupian Morphological Database) (Gerardi, 2022a) with 51 languages, TuPAn (Tupian Plants and Animals) (Gerardi, 2022b) with 25 languages, and TuDeT (Tupian Dependency Treebanks) (Gerardi et al., 2022) with 9 languages. All databases are work-in-progress in different stages of completion.

Among this project databases, this work focuses on the specifications of TuDeT in view of its applicability and results (current and future outcomes). On the scientific side we are concentrating on measuring syntactic complexity of the languages, but we extend our tools used so that we can apply them for all treebanks in Universal Dependencies (UD) (De Marneffe et al., 2021).

On the practical side, we also intend to use the collection of sentences in TuDeT to create educational materials for the communities. One of the main goals of TuDeT is to raise literacy by promoting new teaching materials in indigenous context, to help the communities in stand against language domination.[1]

It would not be out of place at this point to discuss available tools or corpora for Tupian languages, but none exists. TuDeT is the first collection of sentences open-access, despite its inceptive state, as are the tools being built within, such as the Guajajara morphological analyzer (see Section 4.3.). One almost obvious exception is Paraguayan Guarani, a language that enjoys official status and spoken by about six million people. We are aware of a morphological analyzer (Kuznetsova and Tyers, 2021), but not of annotated or tokenized corpora. A parallel corpus Guarani-Spanish is being developed (Chiruzzo et al., 2020). Additional documentation data exists for Aweti (Drude and Reiter, 2005) and Ache (Thompson et al., 2012), but their access is restricted.

Here we introduce our project and discuss its purpose (this section), before describing its main components: the dependency treebanks in terms of their basis and process and annotation (Section 2.) and the lexical database (Section 3.). We address the incipient development of related NLP tools (Section 4.) before concluding remarks that discuss the relevance and potential outcomes of the project's output (Section 5.).

2. The Tupian Dependency Treebanks (TuDeT)

All languages in TuDeT belong to the Tupian family, one of the largest language families in South-America (Rodrigues and Cabral, 2002; Rodrigues and Cabral, 2012; Galucio et al., 2015). The vitality level of these languages varies significantly. A sociolinguistic fact about them is the non-correlation between the amount of speakers and the status of the languages. Some languages with only a few hundred speakers each (such as Ka'apor and Karo) are less threatened than others with thousands of speakers (such as Guajajara and Munduruku) which, however, are in an alarmingly rapid process of shifting to Portuguese and abandoning native languages. The nine languages in TuDeT are shown with their respective number of speakers and status from (Eberhard et al., 2021) in Table 1. The presence of two extinct languages, Tupinamba and Old Guarani, plays an important role in understanding diachronic aspects of this language family. The geographic distribution of the languages in TuDeT is shown in Figure 1.

Annotated sentences in TuDeT stem from various sources. For the extinct languages, Tupinamba and Old Guarani, all texts known for these languages are being annotated: grammatical descriptions, e.g. (de Anchieta, 19331595; de Montoya, 1876a), religious texts, e.g. (Araújo, 19521618; de Montoya, 1876b), poetry and theater plays. For the modern languages, we took sentences from grammatical descriptions e.g. (Gabas Jr, 1999; Braga, 2005; Rose, 2011; Aragon, 2014), fieldwork data collection, articles describing aspects of the languages and stories told by native speakers, e.g. (Castro and Guajajara, 2020; Campos Castro and Gervason Defilippo, 2021). The current state of TuDeT treebanks is given in Table 2.

Language	Glottocode	Speakers	Status
Akuntsu	akun1241	3	Nearly extinct
Guajajara	guaj1255	12000	Vigorous
Ka'apor	urb1250	600	Developing
Karo	karo1305	200	Vigorous
Makurap	maku1278	40	Moribund
Munduruku	mund1330	5000	Threatened
Old Guarani	oldp1258	0	Extinct
Teko	emer1243	400	Vigorous
Tupinamba	tupi1273	0	Extinct

Table 1: Languages in TuDeT.

A relevant feature of TuDeT is its unified terminology for the morphological annotations. Having consulted various language descriptions, we have arrived at a general terminology so that the morphological features and their values are the same, as far as possible, for all languages (in TuDeT). Since different descriptions often treat the same constructions in different ways and using different terminology, we have adapted these observations to the framework of Universal Dependencies considering diachronic and synchronic aspects of the

[1]The project is about to publish a book for the alphabetization of Makurap children (Tupi, Tupari) (Aragon and Makurap, 2022).

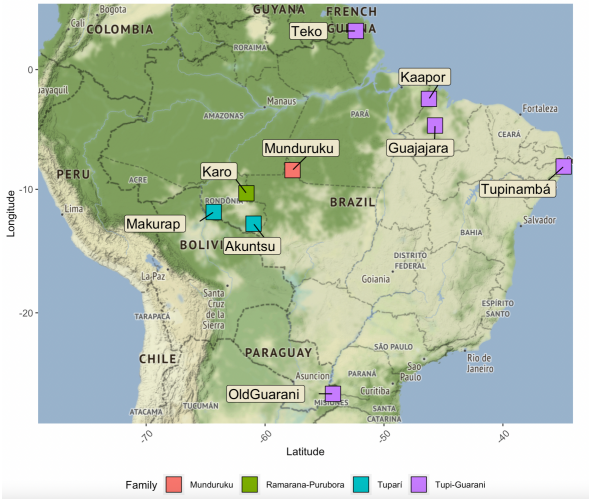

Figure 1: Languages in TuDeT.

languages.

2.1. The Universal Dependencies Framework

Universal Dependencies (De Marneffe et al., 2021) is a multilingual formalism for treebanking, including annotation guidelines[2] for dependency relations, morphological analysis, part-of-speech tagging, and other linguistic features. Besides the languages in TuDeT, one more Tupian language is present in UD, Mbya Guarani, so that ten languages represent the Tupian family in UD. Although we acknowledge some drawbacks of UD, e.g., (Osborne and Gerdes, 2019), it is still the best open-access possibility available. The annotations use the standard UD style POS tag inventory, morphological features and universal dependency relations from Universal Dependencies v2 (Nivre et al., 2020), and are encoded using the CoNLL-U format[3]. They are enriched with additional dependency subtypes and language-specific morphological features to reflect specific traits of Tupian languages.

This combination of standard annotations with specification through subtypes makes UD a satisfactory annotation framework for the analysis of individual languages and for the study of linguistic typology. Each of the treebanks is accompanied by a documentation for all features, syntactic, morphological and POS.

The adaptability of the UD framework to language-specific features is relevant to treat characteristic features of Tupian languages and facilitating NLP tasks. One example of specific values that characterize these languages are ideophones, which show unique syntactic patterns as they can co-occur with certain lexical items in the sentence (restricted collocations) and they are usually exposed to different reduplication processes (Voeltz and Kilian-Hatz, 2001). In UD, ideophones are not part of the POS tag-set, therefore their description in our treebanks requires special treatment. Another case concerns the so called relational prefixes (Rodrigues, 2009), a feature described uniquely for some Brazilian indigenous languages, which mark syntactic contiguity or non-contiguity of heads and their dependents.

Another advantage of the UD framework is that its extended documentation and highly standardized annotations make it suitable for rapid, consistent annotation as well as easily comprehended by non-linguist audiences. This contributes to our goal of increasing the linguistic documentation and understanding of the Tupian languages.

Moreover, the competitive scores reached in the ConLL 2017 and 2018 Shared Tasks, illustrate the suitability of the framework in developing high-accuracy computer parsers and other downstream NLP tasks (Zeman et al., 2018). Thanks to this, we can develop NLP tools employing the annotated data (see Section 4.), such as the morphological analyzers that are being built for Guajajara and Munduruku, which rely almost exclusively on the respective treebanks.

Alternatives such as SUD (Gerdes et al., 2018) are worth consideration and a future conversion to a surface-syntactic annotation schema and parallel maintenance is planned.

2.2. The Annotation Process

Initially, all annotations were/are being carried out manually by linguists and computational linguists with a strong background knowledge of Tupian languages. Each treebank has one main annotator and all annotations are revised by the two Tupian specialists in the team.

2.2.1. Data standardization

Most of the languages present in TuDeT either lack a standardized orthography or have only recently acquired one. Therefore, we employ rule-based approaches to unify the orthographic differences found in the texts to be annotated. This is done with a two-fold approach:

Phonetic representation: the different sources annotated employ different symbols for certain sounds. We unify the texts in a single orthographic representation of the phonemes. For example, the glottal stop /ʔ/ is generally represented by an apostrophe (’), but we represent it using its IPA symbol (ʔ).

Word boundaries: the sources do not agree whether or not certain morphemes are bound. This affects mainly affixes, clitics, and certain particles.

[2]https://universaldependencies.org/guidelines.html
[3]https://universaldependencies.org/format.html

We decide the status of these morphemes based on diachronic, typological, and syntactic criteria.

2.2.2. Manual annotation

We combine manual approaches with supervised computational methods for the annotation of the linguistic corpora. We start by manually annotating a subset of the linguistic data according to the UD framework described above. The morphosyntactic features of the sentences are encoded using three main linguistic aspects: POS tags, morphological features, and dependency relations.

POS tags: Parts-of-speeches in UD are a predefined tag-set, but it allows for a language-specific tag-set as well. Tupian languages are challenging for theories of word-classes as also are native American languages or languages of Southeast-Asia (Mithun, 2001; Van Valin Jr, 2008; Enfield, 2021). In establishing word-classes for the languages in TuDeT, we adopt an approach suggested by the literature (Croft, 1991; Croft, 2001; Croft, 2022a; Haspelmath, 2021) which avoids the splitting and lumping of word-classes (Croft, 2022b; Croft, 2022a) and accounts for the fact that all lexical roots in many Tupian languages are (existential) predicates, which require additional morphology for functioning as arguments, even roots that are semantically "things or objects". Some treebanks lack the adjective label (ADJ) as a POS, since this label is not relevant – a feature already noticed in the early Jesuitic descriptions of (Old) Guarani and Tupinamba (Alexander-Bakkerus et al., 2020).

Features: The morphological information of each token also stems from a predefined tag-set expanded with language-specific features and values. All features and values are explained in the standardized UD documentation style.

Based on the experience of some team members with Tupian languages, as linguists and field workers, we have adopted some unified terminology for morphological features which often contradicts descriptions of these languages. One example is the controversial status of the so called relational morpheme (R_2), which marks the non-contiguity of head and its dependent (Meira and Drude, 2013; Cabral, 2000). Many authors (Rose, 2011; Harrison and Harrison, 2013) treat it as a third person marker, but in the TuDeT treebanks, similar constructions are marked with the same features and values.

(1) a. Mari **i**-purag
Mari $\mathbf{r_2}$-beauty
"Mari is beautiful"

b. Kujã **i**-poraŋ
Woman $\mathbf{r_2}$-beauty
"The woman is beautiful"

c. Wãĩwĩ **i**-puruʔa
woman $\mathbf{r_2}$-pregnant
"The woman is pregnant"

Dependency relations: We use the dependency relations from the UD guidelines along with certain language-specific subtypes, e.g. the relations *obl:subj* and *obl:obj* are employed in strictly head-marking languages such as Tupinamba, where the core arguments are bound to the predicate as a single phonological word, so that NPs related to these arguments cannot be the argument themselves and thus must be in a different dependency relation. This can be seen in Figures 2 and 3, where the strictly head-marking character is considered by the subtypes of the oblique relation, since the root contains the predicate and two core-arguments.

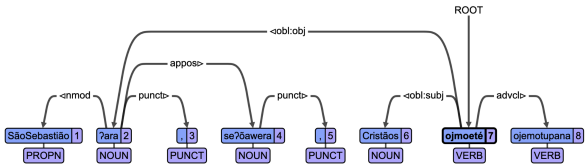

Figure 2: Example of dependency annotation from the Tupinamba UD-treebank.

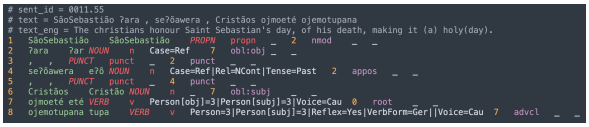

Figure 3: Example of annotation in CoNLL-U format from the Tupinamba UD-treebank.

2.2.3. Supervised annotation

For the supervised annotation, we employ UDPipe 2 (Straka, 2018), a multi-task system for automatic annotation within the UD framework which performs with high accuracy for several languages. We train the model using the manually annotated corpora of sentences available. The resulting annotations are then revised and corrected before their insertion into the treebanks. As expected, the output of the model improves proportionally to the number of annotated sentences. Guajajara is a good example for this approach: the first release of the Guajajara UD-treebank contained 276 sentences. After 500 sentences were reached, this manually annotated dataset served as a training model for automatic dependency parsing. The accuracy of a predictive model has been proven positive, with an accuracy of 99.96%. Currently, the treebank has been enlarged up to 1126 sentences, which should

allow for more precision and consequently better quality of the automatically annotated sentences. Transfer approaches have been implemented for Paraguayan Guarani (Mager et al., 2021), but the performance showed lower automatic scores. Therefore, we initially excluded the possibility of using transfer approaches. However, there has been recent promising work regarding zero-shot methods (Blum, 2022), so transfer approaches could be considered to improve the annotation process.

Table 2 contains the number of sentences and tokens that are part of each TuDeT treebank. It is relevant to mention that not all the treebanks have been created at the same time, which is reflected in the quantity of annotated texts.

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

Table 2: Amount of sentences and tokens in each TuDeT treebank.

3. TuLeD

The Tupian Lexical Database (TuLeD) is the largest online database dedicated to languages of a South-American family. It is an open-source database[4], which provides an extensive list of lexical items with cognate assignment, phonetic alignment (shown in Figure 4), cultural or linguistic notes, and borrowing information. The data is presented in a standardized format according to the CLDF (cross-linguistic data format) standards (Forkel et al., 2018), and corresponds to the main principles of FAIRness (Findability, Accessibility, Interoperability, and Reusability) (Wilkinson et al., 2016), which enables ease of access, straightforward sharing and manipulation. Such word lists can be applied in typological language comparison and other linguistic tasks. This database comprises 78 languages, 404 concepts[5]. The concepts are connected to CONCEPTICON glosses (List et al., 2016), which allow for a network of semantic relationships cross-linguistically. The geographic distribution of the languages and language families presented in TuLeD is shown in Figure 5.

[4] `https://tular.clld.org/contributions/tuled`

[5] The next release of TuLeD will comprise 91 languages and 650 concepts.

Figure 4: Example of phonetic alignment from TuLeD for three different cognate classes.

Figure 5: Map of languages in TuLeD colored according to sub-group.

Although TuLeD cannot yet be considered as a dictionary (it does not supply information about, for example, grammar, usage, and synonym discrimination), it plays an important role in laying out ways to help the process of vocabulary learning besides accommodating the phonetic-phonological profile of the languages. TuLeD, besides containing the traditional items of the Swadesh List (Swadesh, 1955), which are said to be the most borrowing-resistant items of languages, also contains culturally relevant items for the Tupian populations (Ferraz Gerardi et al., 2021).

Two additional databases are part of TuLaR: TuMoD (Tupian morphological database) and TuPAn (Tupian plants and animals). As they are under intensive development and have not been publicly released yet, they are not discussed here.

4. TuDeT Tools

The development of NLP tools is an important part of the project and is still in its initial phase. As of now, two tools are almost ready for release, and are presented below.

4.1. TuDeTstats

In order to track relevant statistics from the treebanks and measure syntactic complexity, which are informative of synchronic and diachronic aspects of the languages, we have built a web application which uses two different approaches. On one side, complexity measures are computed (e.g. MDD: mean dependency distance in a sentence (Gibson, 1998), LEFT: proportion of left dependents (Chen and Gerdes, 2017), NDD: normalized dependency distance (Lei and Jockers, 2020)) along with part-of-speech tags and syntactic dependencies (using code from (van Cranenburgh, 2019))[6]. On the other side, we have added unigrams, and selected bi- and trigrams of POS tags along with a raw count of left dependents [7].

The combination of these complexity measures with n-grams, as we show, performs better than the complexity measures alone. With Linear Discriminant Analysis (LDA), for example, the inclusion of n-grams can account for family membership. The family-cluster is less clear when only complexity measures are used. This is shown in Figure 6 where only complexity measures were used to cluster according to family membership languages of five different families in UD[8]. Figure 7 shows the clusters combining complexity measures and selected n-grams alongside with HeadLeft. Measures such as these are important because they can tell us how structurally different text types are for the family's internal analyzes.

4.2. Visualization

TuDeTstats is built in the R programming language (R Core Team, 2021) with the Shiny package (RStudio, Inc, 2014) for reactive web applications. Together, they provide access to modern analytics and visualization algorithms for linguistic research. Figure 8 shows the TuDeTStats application with selected measures displayed for the Tupian treebanks in UD.

4.3. Morphological analyzers

Based on the collected texts and the morphology presented in the treebank, a finite-state transducer for Munduruku is being built using HFST (Lindén et al., 2009) and Xerox functions. The analyzer contains a lexicon of root words, morphological and phonological rules, and composition operators. Another morphological analyzer for Guajajara is in the early stages of development, also using HFST, and we have plans to experiment with FOMA (Hulden, 2009) and OpenFST (Allauzen et al., 2007). The training set for the lexicon was extracted from the Guajajara UD-treebank, which contains 700 unique lemmas. Unfortunately, it is difficult to evaluate the analyzer at an early stage. However, a test-set for accuracy evaluation is being developed as the amount of rules increase.

A significant advantage of these morphological analyzers is that they can be adapted to other languages of the Tupian family. For example, we have already started to build a analyzer of Tupinamba based on the templates available for Guajajara. Rule-based systems of a morphological analyzer can be used for future NLP applications, such as morphological inflection and derivation tasks, automatic annotation of morphological features and machine translation. An example of an output from the Munduruku morphological analyzer is shown in Figure 9[9].

5. Conclusion

TuLaR contributes to expanding the linguistic description, documentation, and computational linguistic resources available for under-researched and low-resource languages of the Tupian family through nine languages following the Universal Dependencies framework and allows developing NLP tools, providing analyzes at different levels (phonology, morphology, and syntax). Future directions may focus on the development of NLP tools such as tokenizers, lemmatizers, morphological analyzers or automatic translation of written texts, as well as web-based systems with new language resources. All these are valuable initiatives to increase linguistic policies regarding endangered languages, rekindling ways to revitalize not only the language and culture, but also the indigenous community identity (Hinton, 2003; Pine and Turin, 2017).

Thus, the creation of linguistic resources presented for the Tupian family in this paper is an example of how computational linguistics products correlate with linguistic research and indigenous communities' necessities in a way to implement efforts to ensure the triad **documentation-conservation-revitalization**, contributing towards a more inclusive computational linguistics.

An important aspect of the work here presented lies is that all tools and the data are available in open access. We are glad to engage in academic cooperation, as well as with the communities. We

[6]We are aware of the controversial topic of complexity in language and measures of syntactic complexity, nonetheless it is appropriate to employ the term for the measures implemented in our application– see (Jiménez, 2018).

[7]The web application can be accessed in its pre-release version from `https://ffgerardi.shinyapps.io/TuDeT-Stats/`.

[8]We have included larger figures in Appendix A.

[9]The Munduruku finite-state morphological analyzer can be accessed from `https://github.com/LanguageStructure/Munduruku_FST`.

look forward to participating in similar projects, but we also welcome collaborators in our projects.

6. Acknowledgements

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

7. Bibliographical References

Alexander-Bakkerus, A., Rebeca, R. F., Zack, L., Zwartjes, O., and Case, J., (2020). *Were there ever any adjectives? The recognition of the absence of an autonomous adjective class in Tupi-Guarani as demonstrated in the earliest missionary grammars.*, page 139–155. Brill.

Allauzen, C., Riley, M., Schalkwyk, J., Skut, W., and Mohri, M. (2007). Openfst: A general and efficient weighted finite-state transducer library. In *International Conference on Implementation and Application of Automata*, pages 11–23. Springer.

Aragon, C. C. and Makurap, A. O. (2022). *Ensinando a língua Makurap*, volume 1. Oikos: São Leopoldo.

Aragon, C. C. (2014). *A grammar of Akuntsú, a Tupían language*. Ph.D. thesis, University of Hawai 'i„ at Mānoa. unpublished PhD thesis.

Araújo, A. d. (1952[1618]). *Catecismo na língua brasílica*. Pontifícia Universidade Católica do Rio de Janeiro.

Blum, F. (2022). Evaluating zero-shot transfers and multilingual models for dependency parsing and POS tagging within the low-resource language family tupían. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 1–9, Dublin, Ireland, may. Association for Computational Linguistics.

Braga, A. d. O. (2005). *Aspects morphosyntaxiques de la langue Makurap/Tupi*. Ph.D. thesis, Toulouse 2. unpublished PhD thesis.

Cabral, A. (2000). Flexão relacional na família tupí-guaraní. *ABRALIN, Boletim da Associação Brasieira de Lingüística*, 25:233–262.

Campos Castro, R. and Gervason Defilippo, J. (2021). Histórias originárias em tenetehára (tupí-guaraní) como estratégia de revitalização linguística. In Patrícia Goulart Tondineli, editor, *(Re)vitalizar línguas minorizadas e/ou ameaçadas: teorias, metodologias, pesquisas e experiências*, pages 109–138. Coleção Pós-Graduação da UNIR - EDUFRO.

Castro, R. C. and Guajajara, P. C. (2020). Izipi mehe: Cibercaminhos linguísticos e literários para a preservação da cultura tenetehára. *Revista Brasileira de Linguística Antropológica*, 12:251–282.

Chen, X. and Gerdes, K. (2017). Classifying languages by dependency structure. typologies of delexicalized universal dependency treebanks. In *Proceedings of the fourth international conference on dependency linguistics (Depling 2017)*, pages 54–63.

Chiruzzo, L., Amarilla, P., Ríos, A., and Giménez Lugo, G. (2020). Development of a Guarani - Spanish parallel corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2629–2633, Marseille, France, May. European Language Resources Association.

Croft, W. (1991). *Syntactic categories and grammatical relations: The cognitive organization of information*. University of Chicago Press.

Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press on Demand.

Croft, W. (2022a). *Morphosyntax: Constructions of the World's Languages*. Cambridge University Press. Draft version of 2021.

Croft, W. (2022b). Word classes in radical construction grammar. In Eva van Lier, editor, *Oxford handbook of word classes*. Oxford University Press.

de Anchieta, J. ((1933)[1595]). *Arte de gramática da língua mais usada na costa do Brasil*. Imprensa Nacional.

De Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal dependencies. *Computational linguistics*, 47(2):255–308.

de Montoya, A. R. (1876a). *Arte de la lengua guarani o mas bien tupi*. Faesy & Frick.

de Montoya, A. R. (1876b). *Catecismo de la lengua guaraní*, volume 4. BG Teubner.

Drude, S. and Reiter, S. (2005). Collection "awetí".

Eberhard, D. M., Simons, G. F., and Fennig, C. D. (2021). *Ethnologue: Languages of the World. Twenty-fourth edition*, volume 16. SIL international, Dallas, TX.

Enfield, N. J. (2021). *The Languages of Mainland Southeast Asia*. Cambridge Language Surveys. Cambridge University Press.

Ferraz Gerardi, F., Aragon, C. C., and Reichert, S. (2021). When the macaw teaches you to eat the brazil nut: Introducing a concept list of tupían languages. Computer-Assisted Language Comparison in Practice, 01/12/2021, https://calc.hypotheses.org/2988, 12.

Forkel, R., List, J.-M., Greenhill, S., Rzymski, C., Bank, S., Cysouw, M., Hammarström, H., Haspelmath, M., Kaiping, G., and Gray, R. (2018). Cross-linguistic data formats, advancing

data sharing and re-use in comparative linguistics. *Scientific Data*, 5:180205, 10.

Gabas Jr, N. (1999). *A Grammar of Karo*. Ph.D. thesis, University of California, Santa Barbara. unpublished PhD thesis.

Galucio, A. V., Meira, S., Birchall, J., Moore, D., Gabas Júnior, N., Drude, S., Storto, L., Picanço, G., and Rodrigues, C. R. (2015). Genealogical relations and lexical distances within the tupian linguistic family. *Boletim do Museu Paraense Emílio Goeldi. Ciências Humanas*, 10(2):229–274.

Gerardi, F. F., Reichert, S., and Aragon, C. C. (2021a). Tuled (tupían lexical database): introducing a database of a south american language family. *Language Resources and Evaluation*, 55(4):997–1015.

Gerardi, F. F., Reichert, S., Aragon, C., List, J.-M., and Wientzek, T. (2021b). Tuled: Tupían lexical database, 03.

Gerardi, F. F., Reichert, S., Aragon, C., Martín-Rodríguez, L., Godoy, G., and Merzhevich, T. (2022). Tudet: Tupían dependency treebank, 05.

Gerardi, F. F. (2022a). Tumod: Tupían morphological database. Forthcoming.

Gerardi, F. F. (2022b). Tupan: Tupían plants and animals. Forthcoming.

Gerdes, K., Guillaume, B., Kahane, S., and Perrier, G. (2018). SUD or surface-syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 66–74, Brussels, Belgium, November. Association for Computational Linguistics.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Harrison, C. and Harrison, C. (2013). *Dicionário guajajara-português*. Anápolis: International Linguistic Association (SIL).

Haspelmath, M. (2021). Word class universals and language-particular analysis. In Fulano, editor, *Oxford handbook of word classes*. Oxford University Press.

Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., and Klakow, D. (2021). A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Hinton, L. (2003). Language revitalization. *Annual Review of Applied Linguistics*, 23:44–57.

Hulden, M. (2009). Foma: a finite-state compiler and library. In *Proceedings of the Demonstrations Session at EACL 2009*, pages 29–32.

Jiménez, C. C. (2018). *Complejidad lingüística: orígenes y revisión crítica del concepto de lengua compleja*. Peter Lang.

Kuznetsova, A. and Tyers, F. M. (2021). A finite-state morphological analyser for paraguayan guaraní. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 81–89.

Languages Project, E. (2020). Catalogue of endangered languages.

Lei, L. and Jockers, M. L. (2020). Normalized dependency distance: Proposing a new measure. *Journal of Quantitative Linguistics*, 27(1):62–79.

Lindén, K., Silfverberg, M., and Pirinen, T. (2009). Hfst tools for morphology – an efficient open-source package for construction of morphological analyzers. In Cerstin Mahlow et al., editors, *State of the Art in Computational Morphology*, volume 41, pages 28–47, 08.

List, J.-M., Cysouw, M., and Forkel, R. (2016). Concepticon: A resource for the linking of concept lists. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2393–2400, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Mager, M., Gutierrez-Vasques, X., Sieera, G., and Meza, I. (2018). Challenges of language technologies for the indigenous languages of the americas. In *Proceedings of the 27th International Con- ference on Computational Linguistics*, pages 55–69.

Mager, M., Oncevay, A., Ebrahimi, A., Ortega, J., Rios, A., Fan, A., Gutierrez-Vasques, X., Chiruzzo, L., Giménez-Lugo, G., Ramos, R., Meza Ruiz, I. V., Coto-Solano, R., Palmer, A., Mager-Hois, E., Chaudhary, V., Neubig, G., Vu, N. T., and Kann, K. (2021). Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online, June. Association for Computational Linguistics.

Meira, S. and Drude, S. (2013). Sobre a origem histórica dos "prefixos relacionais" das línguas tupí-guaraní. *Cadernos de Etnolingüística*, 5(1):1–30.

Mithun, M. (2001). *The languages of native North America*. Cambridge University Press.

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., and Zeman, D. (2020). Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection.

Osborne, T. and Gerdes, K. (2019). The status of function words in dependency grammar: A

critique of universal dependencies (ud). *Glossa: a journal of general linguistics*, 4(1):1–28.

Pine, A. and Turin, M. (2017). *Language revitalization*. Oxford University Press.

R Core Team, (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rodrigues, A. D. and Cabral, A. (2002). Revendo a classificação interna da família Tupí-Guaraní. *Línguas Indígenas Brasileiras. Fonologia, Gramática e História, Atas do I Encontro Internacional do GTLI da ANPOLL*, 1.

Rodrigues, A. D. and Cabral, A. S. (2012). Tupían. In Lyle Campbell et al., editors, *The Indigenous Languages of South America*, volume 2, pages 495–574. de Gruyter, Berlin.

Rodrigues, A. D. (2009). A case of affinity among tupí, karíb, and macro-jê. *Revista Brasileira de Linguística Antropológica*, 1(1):137–162.

Rose, F. (2011). *Grammaire del L'Émérillon Teko, une langue Tupi-Guarani de Guyane Française*. Peeters.

RStudio, Inc, (2014). *shiny: Easy web applications in R*. URL: `http://shiny.rstudio.com`.

Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium, October. Association for Computational Linguistics.

Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *International journal of American linguistics*, 21(2):121–137.

Thompson, W. M., Roessler, E.-M., Hauck, J. D., Susnik, B., and Sousa, L. T. (2012). Collection "aché".

van Cranenburgh, A. (2019). udstyle. `https://github.com/andreasvc/udstyle`. unpublished code.

Van Valin Jr, R. D. (2008). Rps and the nature of lexical and syntactic categories in role and reference grammar. In Robert D Van Valin Jr, editor, *Investigations of the syntax-semantics-pragmatics interface*, pages 161–78. John Benjamins, Amsterdam.

Voeltz, F. E. and Kilian-Hatz, C. (2001). *Ideophones*. John Benjamins Publishing.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

Zeman, D., Hajič, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., and Petrov, S. (2018). CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium, October. Association for Computational Linguistics.

A Appendix

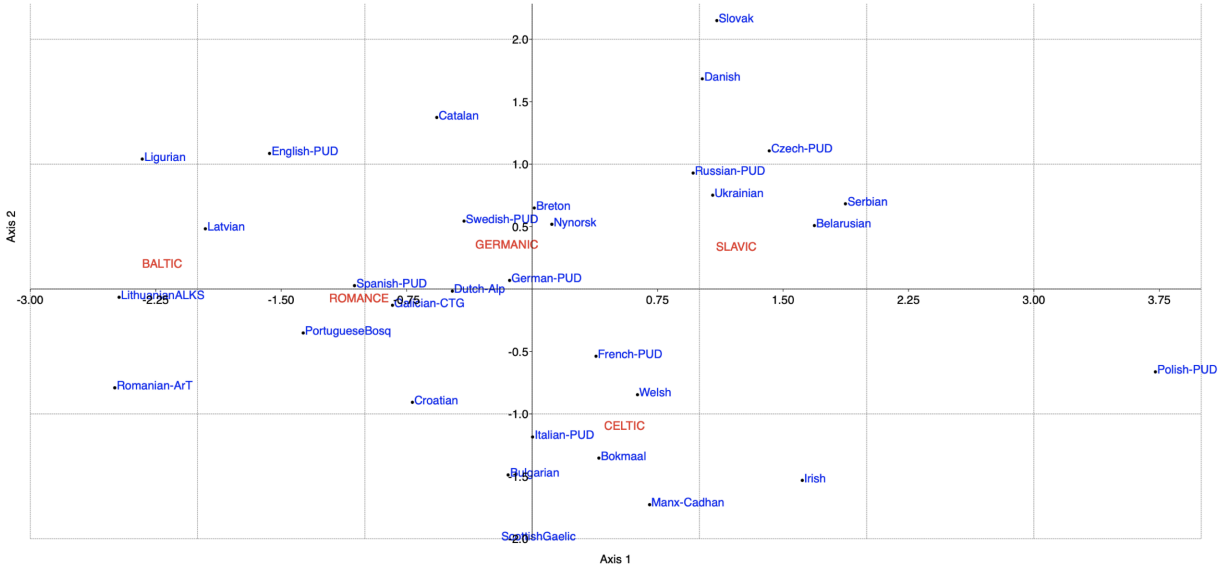

Figure 6: LDA using complexity measures.

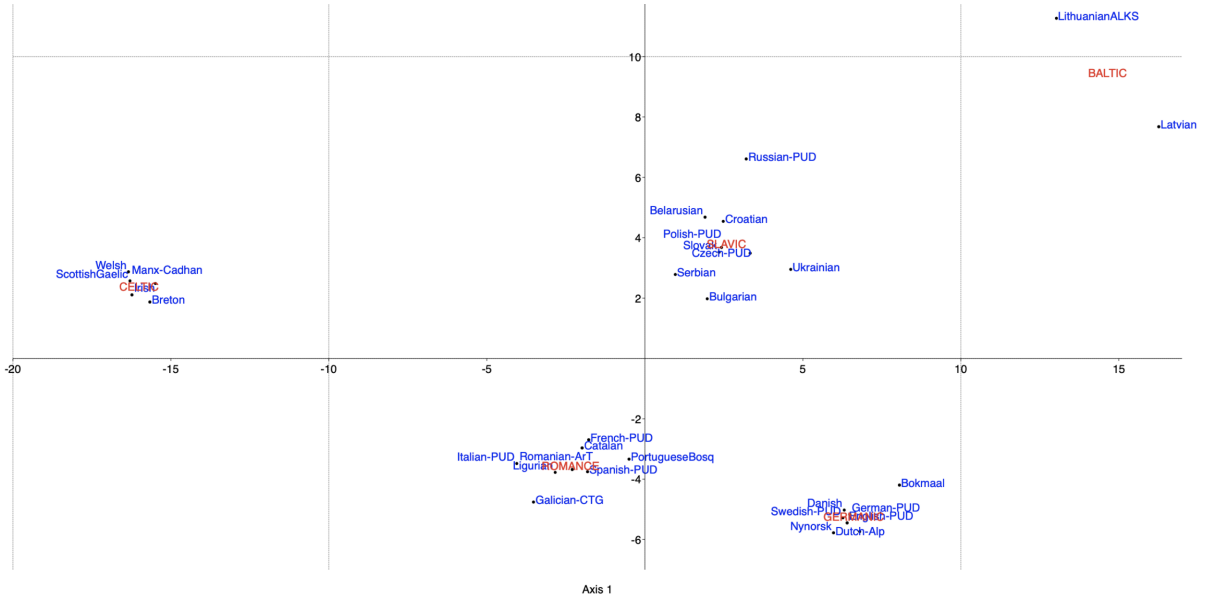

Figure 7: LDA combining complexity measures with n-grams.

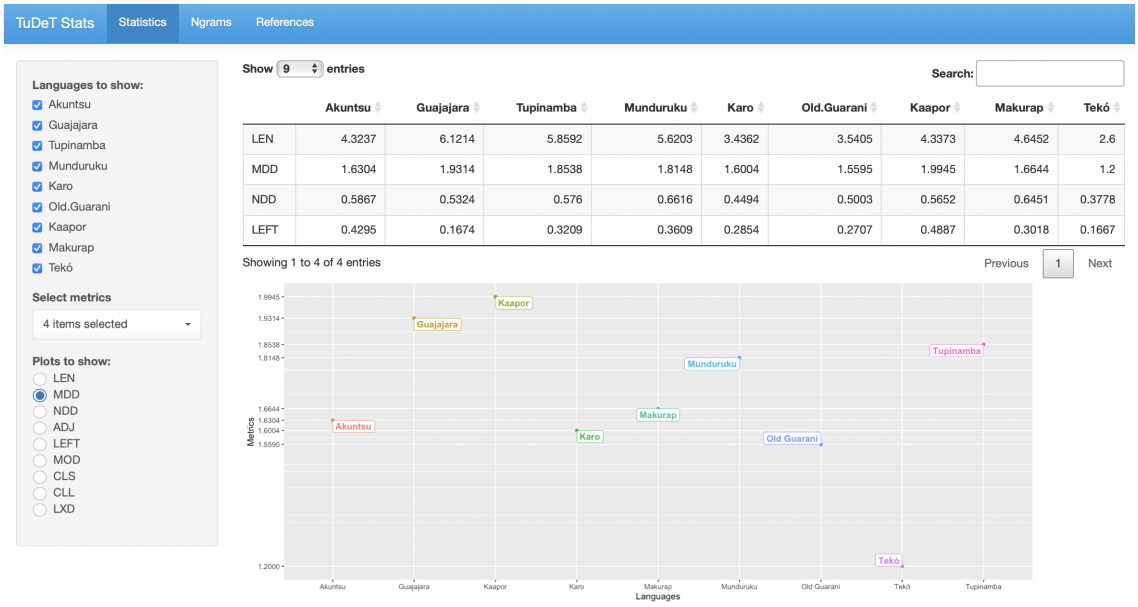

Figure 8: Example of Mean Dependency Distance for languages in TuDeT.

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

apply up oxi
1SG+R1+mother

apply down 1SG+R1+arrow
odop

apply up tao
R2+leg
```

Figure 9: Output examples of the Munduruku finite-state morphological analyzer.