Universal Dependencies for Mbyá Guaraní

Guillaume Thomas
Department of Linguistics
University of Toronto
guillaume.thomas@utoronto.ca

Abstract

This paper presents the first treebank of Mbyá Guaraní, a Tupí-Guaraní language spoken in Argentina, Brazil and Paraguay. The Mbyá treebank is part of Universal Dependencies, a project that aims to create a set of guidelines for the consistent grammatical annotation of typologically different languages. We describe the composition of the treebank, and non-trivial choices that were made in the adaptation of Universal Dependencies guidelines to the annotation of Mbyá.

1 Introduction

Universal Dependencies (UD) is a cross-linguistic treebank annotation project, which aims to provide guidelines that are consistently applicable to typologically different languages (McDonald et al., 2013). Annotation guidelines are meant to be suitable for computer parsing, while enabling rigorous typological research and linguistic analysis of individual languages. They should also be easily understood by non-linguists. At the time of writing this paper, UD version 2.4 consists of 146 treebanks in 83 languages (Nivre et al., 2019).

This paper discusses the creation of a UD treebank for Mbyá Guaraní, a Tupí Guaraní language (Tupian) spoken in Argentina, Brazil and Paraguay. Work on indigenous American language in Universal Dependencies is still scarce. Previous research on the suitability of Universal Dependencies for the analysis of indigenous American languages include work on Arapaho, an Algonquian language (Wagner et al., 2016) and Shipibo-Konibo, a Panoan language (Vasquez et al., 2018). Outside of UD, Mikkelsen et al. (2014) discuss the development of a dependency treebank of Karuk, a isolate within the Hokan group. Our goals in this paper are to motivate the choices that were made in adapting UD guidelines for the annotation of Mbyá, and to reflect on difficulties that were encountered in this process. In doing so, we hope to contribute to the ongoing debate on the typological foundations of the UD project.

The treebank consists of two parts, each of which has been included in Universal Dependencies v2.4 (Thomas 2019a,b). The paper refers to the latest development version of the UD Mbyá Treebank at the time of writing. We assume familiarity with UD v2 guidelines, as described in UD Guidelines (n.d.).

2 General Information and Treebank Composition

Mbyá is a Guaraní language spoken by approximately 30,000 speakers in Argentina, Brazil and Paraguay (Ladeira, 2018). It belongs to the southern branch (group 1) of the Tupí Guaraní family, together with Nhandeva, Kaiowá and Paraguayan Guaraní, among other languages (Rodrigues, 1986). The main references on the grammar of Mbyá are Robert Dooley's grammatical sketch and lexicon of the language (Dooley, 2015), and Martins (2003)'s doctoral dissertation.

The UD Mbyá treebank consists of two corpora. The largest one is composed of narratives collected by Robert Dooley, written by two Mbyá Guaraní speakers, Nelson Florentino and Darci Pires de Lima, between 1976 and 1990 in Brazil. It contains 11,771 tokens (1,046 sentences). Interlinearized versions of these narratives are archived on the Archive of the Indigenous Languages of Latin America (Dooley, n.d.), and were used with Robert Dooley's authorization. The second corpus is composed of three speeches by Paulina Kerechu Núñez Romero, a Mbyá Guaraní speaker from Ytu community, Caazapá Department, Paraguay, which were recorded by the author. It consists of 1,318 tokens (98 sentences).

There is no standard orthography of Mbyá. Dooley's corpus uses the orthography presented in Dooley (2015), which is popular among Mbyá communities in the south of Brazil. The texts collected in Paraguay uses an adaptation of this orthography to Spanish based spelling conventions adopted in Mbyá communities in Argentina and Paraguay.

The texts were manually interlinearized in SIL Fieldworks Language Explorer (FLEx; Black and Simons (2008)). Robert Dooley's interlinearization of Florentino and Pires de Lima's narratives was imported into FLEx and revised to fit our annotation guidelines. Language specific parts of speech were added manually at this stage of annotation. Interlinearized narratives were exported in the XML FLEx-Text format from FLEx and converted to the CoNLL-U format using a Python script. Morphosyntactic features were automatically created in the conversion stage. Universal POS tags were automatically converted from language specific tags at this stage too, and were later corrected manually. Dependency annotation was semi-automatic. A first set of 500 sentences were annotated manually in Arborator (Gerdes, 2013), and was used to train a parser in UDPipe (Straka et al., 2016). This parser was used to annotate the rest of the corpus, which was manually corrected in Arborator. The annotation team consisted of the author and research assistants at the University of Toronto.[1]

3 Annotation Guidelines

Our annotation guidelines are based on version 2 of Universal Dependencies (UD Guidelines, n.d.). In this section, we describe the adaptation of UD guidelines to Mbyá, focusing on phenomena that are specific to the annotation of Mbyá or that raise interesting issues for the UD annotation scheme.

3.1 Lexical Categories

Universal Parts of Speech (POS) in UD include 6 POS for open class words: Adjectives (`ADJ`), Adverbs (`ADV`), Interjections (`INTJ`), Nouns (`NOUN`), Proper Nouns (`PROPN`) and Verbs (`VERB`). Here, we will focus on the distinction between `ADJ`, `ADV`, `NOUN` and `VERB`. While most scholars of Guaraní languages recognize the existence of a noun/verb distinction, there is less agreement on the existence of a distinction between adjectives and adverbs in these languages (see Dietrich (2017) for a recent discussion). Consequently, we have not included these categories in the language specific tagset for Mbyá. The following subsections describe the mapping from these language specific categories to the universal POS of UD.

Verbs The language specific tagset of Mbyá includes subcategories of verbs that reflect their valency and agreement class. In order to understand this categorization, it is necessary to give some background on agreement in the language. Subjects and objects are cross-referenced on verbs by prefixes that encode person and number. There are two sets of cross-reference prefixes, which I refer to as 'set A' and 'set B' prefixes, following Tonhauser (2017). These two sets distinguish two classes of intransitive verbs. Set A prefixes are used to index the subject of active (dynamic) verbs, while set B prefixes are used with inactive (stative) verbs, as illustrated by examples (1) and (2). Accordingly, we distinguish active (`vi:a`) from inactive (`vi:i`) intransitive verbs in our language specific tagset:[2]

(1) A-vaẽ.
A1.SG-arrive
`VERB`
`vi:a`
'*I arrived.*'

(2) Xe-kane'õ.
B1.SG-tired
`VERB`
`vi:i`
'*I am tired.*'

[1] Gregory Antono, Laurestine Bradford, Vidhyia Elango, Jean-François Juneau, Angelika Kiss, Barbara Peixoto, Darragh Winkelman.

[2] Glosses: A1.SG: first person singular 'active' inflection; A1.PL.INC: first person plural inclusive 'active' inflection; A1.PL.EXCL: first person plural exclusive 'active' inflection; B1.SG: first person singular 'inactive' inflection; BDY; information structure boundary marker; COMP: completive aspect; CONT: continuative aspect; DM discourse marker; DS different subject marker; HSY: hearsay evidential; MIR: mirative evidential; NEG negation; NMLZ nominalization; NPOSSD: non-possessed nominal form; PAST: past tense; R: linker morpheme; REFL: reflexive; REL: relativizer; SS same subject marker.

Words that were categorized as verbs in the language specific tagset were also tagged as VERB when they are used as predicates. In addition, inactive verbs (vi:i) are also attested as modifiers of nouns and of non-nominal heads, in which case they were tagged as (ADJ) or (ADV), respectively:

(3)

Avaxi	o-nhotỹ	r-yxy	porã.
Corn	A3-plant	R-line	good
NOUN	VERB	NOUN	ADJ
n	vt	n	vi:i

'*He planted the corn in a good line*.' (Dooley, 2015)

(4)

Oro-vy'a	porã.
A1.PL.EXCL-happy	good
VERB	ADV
vi:a	vi:i

'*We were very happy*.' (Dooley, 2015)

Alternatively, verbs tagged as ADJ or ADV could have been tagged as VERB, and analyzed as reduced relative or adverbial clauses. However, when used as modifiers, these verbs are typically uninflected (i.e. they do not bear cross-reference prefixes), unlike verbs in fully fledged clausal modifiers. Consequently, we believe that verb roots used as modifiers do not head a clause, which means that they should be tagged as ADJ or ADV in UD v2.4 guidelines.

Nouns Drawing a distinction between nouns and inactive verbs is not trivial in Mbyá, since the B set of cross-reference markers of inactive verbs is also used as possessive markers on nouns, and nouns are productively used as predicates without copula (for a discussion of this issue in Tupí-Guaraní languages, see Queixalós (2001)). Following Dooley (2015), we categorize as nouns those words that can be used as arguments without additional marking (such as nominalizing morphology). In order to preserve the distinction between nominal and verbal predications, these words were tagged as nouns both in their argument uses and in their predicative uses.

Adjectives and Adverbs Some roots can be used as modifiers but not as predicates. Because of the lack of evidence of a lexical distinction between adjectives and adverbs in the language, they were tagged as modifiers in the language specific tagset, and as ADJ or ADV in the universal tagset. This is illustrated by the use of *guaxu* ('big', 'a lot') in the following examples:

(5)

Ja-j-apo	karu	guaxu.
A1.PL.INCL-B3-do	meal	big
VERB	NOUN	ADJ
vi:a	n	mod

'*We prepared a big meal*.' (Dooley 2015)

(6)

A-karu	guaxu.
A1.SG-eat	a lot
VERB	ADV
vi:a	mod

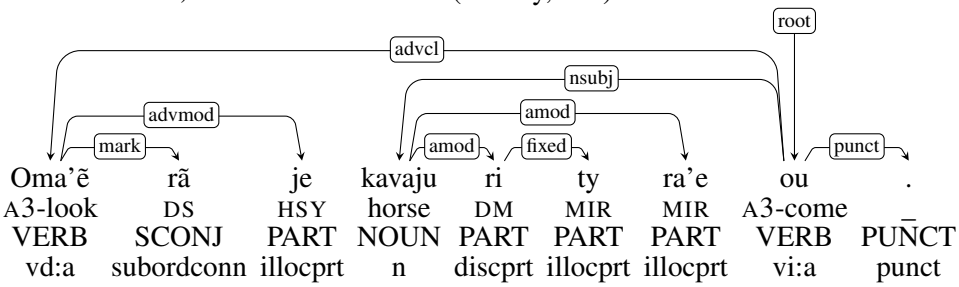
'*I ate a lot*.' (Dooley 2015)

3.2 Particles

Mbyá has a rich system of particles, which have been glossed as PART in the universal tagset. Language specific tags for particles encode their semantic function: aspect particles (aspprt), discourse particles discprt, focus particle (focprt), illocutionary particles (illocprt), intensifiers (intprt), modal particles (modprt), quantificational particles (quantprt), question particles (qprt) and temporal particles (temprt).

Many of these particles can be used as dependents of nouns as well as of verbs. This raises an issue for the UD v2 annotation scheme, since the only functional dependencies of nominal heads admitted by the guidelines are determiners (det), classifiers (clf) and case (case). The solution we have adopted is to default to amod for particles used as modifiers of nouns, as illustrated in example (7), where the mirative particles *ri ty* and *ra'e* modify the noun *kavaju* ('horse'):

(7) '*When he looked, a horse had arrived.*' (Dooley, n.d.)



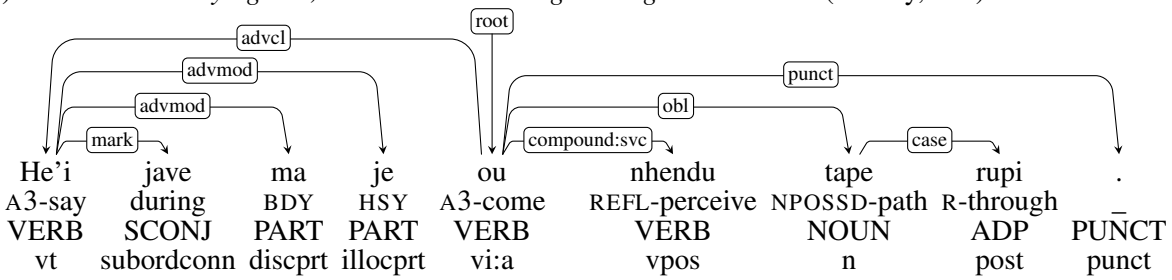
This solution, however, is unsatisfying, since particles are not adjectives, but belong to a closed class of functional items. A more satisfying solution would be to introduce a dependency relation for modifiers that is neutral with respect to the syntactic category of the dependent, as advocated by Croft et al. (2017).

Particles that modify non-nominal heads were annotated with the `advmod` relation, as illustrated by the hearsay evidential *je* in example (7). Alternatively, particles expressing tense, aspect, mood and evidentiality (TAME) might have been tagged as `AUX` when they modify a verb, in which case the relation `aux` would have been used. However, since TAME particles have no morphological verbal features, and are so flexible in their distribution, we are reluctant to tag them as auxiliaries. This being said, the semantic subcategorization of particles in the language specific tagset should make it trivial to map TAME particles to auxiliaries, when they are used as modifiers of verbs.

3.3 Complex Predicates

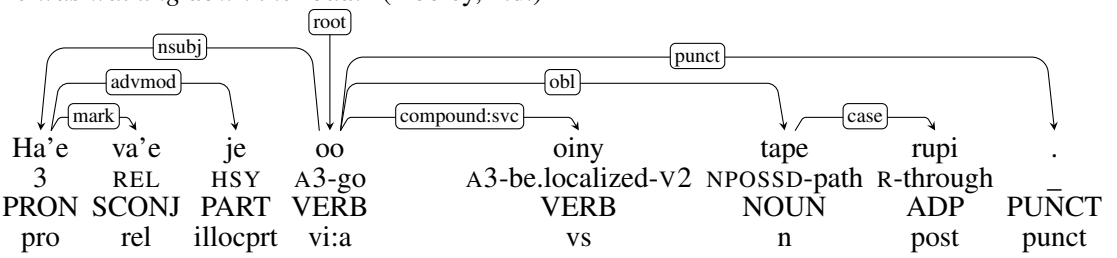
There are at least two types of complex predicates in Mbyá. The first one is formed by combining the main verb with a bare uninflected root, which we glossed `vpos` in the language specific POS tagset. Postposed roots are used in the expression of agent oriented modality (e.g. *pota*, 'try to'), or sensory evidentiality, like *nhendu* ('audibly') in the following example:

(8) '*As she was saying this, she heard something coming on the road.*' (Dooley, n.d.)



A second type of complex predicates is formed by using one of a limited number of verbs in the so-called 'gerund' form common in Tupí-Guaraní languages (Rodrigues, 1953; Jensen, 1989), which we glossed `vs` in the language specific tagset. The dependent verb can be interpreted literally as expressing the position in which the event described by the main verb is realized, but it can also have an aspectual value. This construction was described by Dooley (1991):

(9) '*He was walking down the road.*' (Dooley, n.d.)



Both types of complex predicates were annotated with the relation `compound:svc` in the treebank.

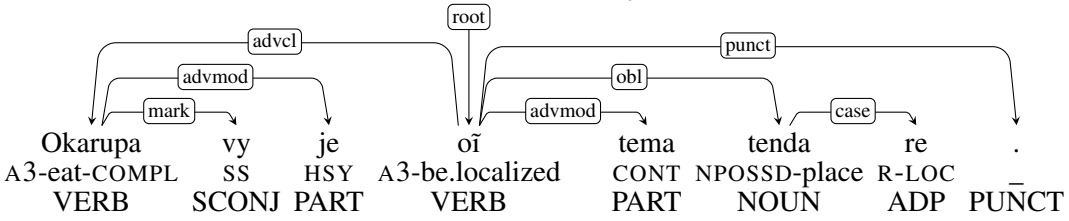
3.4 Strategies of Subordination

The UD v2 annotation scheme recognizes three major types of subordinate clauses: core clausal arguments, adverbial clauses and relative clauses. We describe each of these in turn in this section.

3.4.1 Adverbial Clauses

Adverbial clauses are introduced by a variety of subordinating conjunctions (SCONJ). Among these, it is useful to distinguish plain subordinating conjunctions from switch reference markers. The former, such as *jave* ('when'), only express temporal or causal relations between clauses. The switch reference markers *vy* and *ramo*/*rã* also function as subordinating conjunctions, but do not encode a specific temporal or causal relation. Instead, they indicate whether the subject of the subordinated clause is the same as that of the superordinate clause. In example (10), the same subject marker *vy* indicates that the subject of the eating is the same as the subject of the sitting:

(10) '*After he had eaten, he was sitting on a bench.*' (Dooley, n.d.)

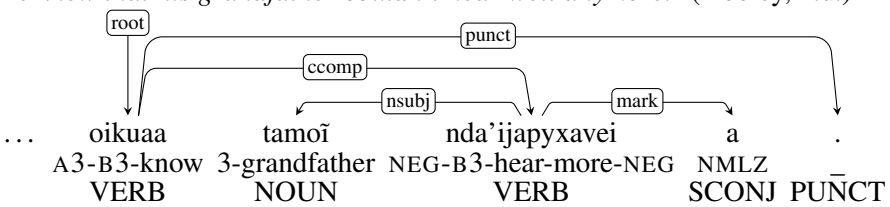


Our annotation guidelines treat switch reference markers as subordinating conjunctions, which are introduced by the relation mark.

3.4.2 Nominalized Complement Clauses

There are no subject clauses in the treebank. Complement clauses are attested and are formed by nominalizing the dependent verb with the morpheme *a*. While Dooley (2015) analyzes this morpheme as a suffix, we might argue that it is a clitic, since its host is not necessarily the verb, but can be one of its adverbial modifiers. This makes it unsatisfying to analyze the nominalizer as a suffix and to represent its function as a verbal feature, since in some cases this feature would have to appear on a word other than that to which the nominalizer is affixed. Consequently, we decided to represent this nominalizer as a token in the dependency annotation, where it is tagged as SCONJ and related to its head by a mark dependency, as illustrated in example (11). This decision is consistent with some writing conventions for Mbyá, such as those adopted by Cadogan (1959).

(11) '*He knew that his grandfather couldn't hear well anymore.*' (Dooley, n.d.)



Nominalized clauses have several morphosyntactic features that are characteristic of noun phrases. They are compatible with temporal suffixes, which in Guaraní languages are nominal markers, and they may be used as complements of post-positions. Nevertheless, they preserve their full clausal structure. In particular, the verbs that head clausal nominalizations project their regular argument structure, bear cross-reference markers, may be modified by adverbs and may be part of complex predicates.

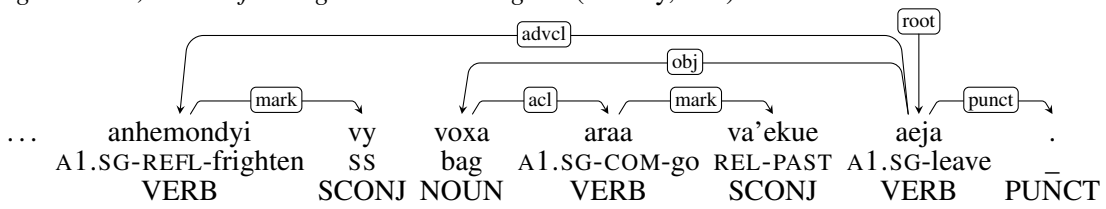
The mixed categorial status of clausal nominalizations in Mbyá raises the question of which dependency relation should be used to relate them to their head. In particular, should nominalized complements of verbs be introduced by the ccomp relation, or by the obj relation? Using the ccomp relation allows us to capture the fact that these complements are propositional. In this case, we take ccomp to indicate the

semantic status of its dependent, a proposition rather than an individual. In addition, the use of `ccomp` indicates the fact that the complement has a full clausal structure. On the other hand, using `obj` is consistent with the nominal category of the complement, in particular its compatibility with adpositions, which UD represents as case marking of nominal dependents. In devising annotation guidelines for the Mbyá treebank, we have decided to use the `ccomp` relation with nominalized clauses that denote propositions.

3.4.3 Relative Clauses

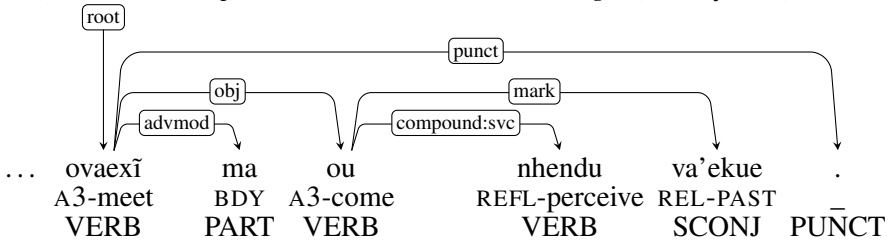
Relative clauses are formed with the enclitic *va'e*, as illustrated in example (12).

(12) '*I got scared, and I left a bag that I had brought.*' (Dooley, n.d.)



Like the clitic *a*, *va'e* is a nominalizer, and relative clauses have a mixed syntactic category. This create an issue for the annotation of free relative clauses used as arguments of verbs, like *ou nhendu va'ekue* (lit. 'that he had heard coming') in the following example:

(13) '*He met the person that he had heard coming.*' (Dooley, n.d.)

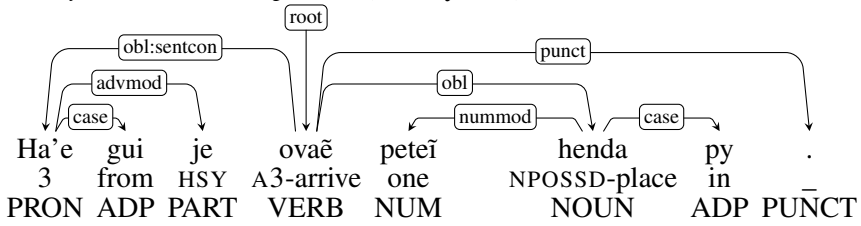


In this example, the complement of *ovaexĩ* ('meet') is syntactically clausal, yet it denotes an individual. The first property motivates the use of a `ccomp` relation, while the second supports the use of an `obj` relation. We have opted for the latter, in order to capture the contrast between clausal nominalizations that denote individuals, introduced by `obj`, and those that denote propositions, introduced by `ccomp`.

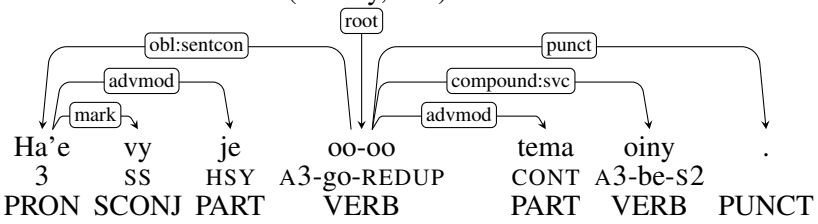
3.5 Discourse Connectives

Most sentences in narratives start with a sentence initial discourse connective. These connectives are composed of the pronoun *ha'e*, which is generally followed by an adposition or a subordinating conjunction, as illustrated in examples (14) and (15). Following Dooley (2015), we assume that occurrences of *ha'e* in discourse connectives denote propositions or situations that are described or made salient by a preceding discourse unit, much like the demonstrative *this* in the English connective *contrary to this*. Since their head is pronominal, we analyze sentence initial discourse connectives as oblique modifiers of the root:

(14) '*Finally, he arrived at a place.*' (Dooley, n.d.)

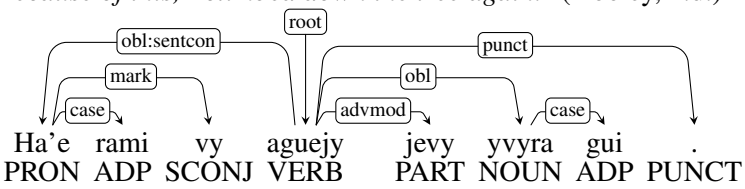


(15) '*He walked and walked.*' (Dooley, n.d.)



Note that these propositional pronouns can be modified by an adposition and a subordinating expression simultaneously, as illustrated in (16):

(16) '*Because of this, I climbed down the tree again.*' (Dooley, n.d.)



Sentence initial discourse connectives are another manifestation of the blurring of the distinction between clausal and nominal categories in Mbyá. Their heads are pronominal. As such, they are compatible with plural marking by the particle *kuery*, and they can be introduced by post-positions, which are characteristic features of nouns. On the other hand, their heads denote propositions, and are compatible with subordinating conjunctions like the same subject marker *vy* in example (15), which normally introduces adverbial clauses. In this example, *vy* indicates that the subject of the verb *oo* is the same as that of the previous sentence, which provides an antecedent to the pronoun *ha'e*.

We have decided to annotate sentence initial discourse connectives as obliques (`obl`) rather than adverbial clauses (`advcl`), thereby giving more weight to the form of their heads (pronominal) than to their interpretation (propositional).

4 Conclusion

We presented the Mbyá treebank, a syntactically annotated corpus of Mbyá in Universal Dependencies. We discussed the adaptation of UD guidelines to the annotation of Mbyá, highlighting questions raised by mixed categories (nominalizations) and the use of functional particles as adnominal modifiers.

Acknowledgements

Initial work on the treebank was made possible by a research award to the author by the Connaught Fund at the University of Toronto.

References

Andrew Black and Gary Simons. 2008. The SIL Fieldworks Language Explorer Approach to Morphological Parsing. In *Computational Linguistics for Less Studied Languages: Texas Linguistics Society, 10*, pages 37–55. CSLI Publications.

León Cadogan. 1992. *Ayvu Rapyta; Textos Míticos de los Mbyá Guaraní del Guairá*. Boletim n^o 227 – Antropología n^o 5. São Paulo: Universidade de São Paulo.

William Croft, Dawn Nordquist, Katherine Looney and Michael Regan. 2017. Linguistic typology meets Universal Dependencies. In Markus Dickinson, Jan Hajic, Sandra Kübler and Adam Przepiórkowski (eds), *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories* (TLT15), pages 63–75. CEUR Workshop Proceedings

Wolf Dietrich. 2017. Word Classes and Word Class Switching in Guaraní Syntax. In Bruno Estigarribia and Justin Pinta (eds), *Guaraní Linguistics in the 21st century*, pages 158–193. Leiden: Brill.

Robert A. Dooley. 1991. A double-verb construction in Mbyá Guaraní. In *Work Papers of the Summer Institute of Linguistics, University of North Dakota Session* 35:31–66.

Robert A. Dooley. 2015. *Léxico guarani, dialeto mbyá*. Summer Institute of Linguistics.

Robert A. Dooley. n.d. *Mbyá Guaraní collection of Robert Dooley*. The Archive of the Indigenous Languages of Latin America: www.ailla.utexas.org. Media: text. Access: 100

Kim Gerdes. 2013. Collaborative dependency annotation. In *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, pages 88–97.

Cheryl Jensen. 1989. *O desenvolvimento histórico da língua Wayampi.* Campinas: Editora da UNICAMP.

Maria Inês Ladeira. 2018. Guarani Mbya. Em Fany Pantaleoni Ricardo (ed.), *Povos Indígenas no Brasil*, Instituto Socioambiental. `https://pib.socioambiental.org/pt/Povo:Guarani_Mbya`

Marci Fileti Martins. 2003. *Descrição e análise de aspectos de gramática do guarani mbyá [Description and analysis of some grammatical aspects of Guaraní Mbyá]*. Ph.D. thesis, State University of Campinas.

Line Mikkelsen, Andrew Garrett, Erik Maier and Clare Sandy. 2014. Developing a syntactically parsed corpus of Karuk. Talk presented at the Annual Meeting of the Society for the Study of the Indigenous Languages of the Americas, Minneapolis, MN. January 3rd.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 92–97.

Joakim Nivre, Mitchell Abrams, Agić Željko et al. 2019. *Universal dependencies 2.4*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Francesc Queixalós (ed.) 2001. Des noms et des verbes en tupi-guarani: état de la question. LINCOM Studies in Native American Linguistics, vol. 37. Munich: LINCOM Europa.

Aryon Dall'Igna Rodrigues. 1953 Morfologia do verbo tupí. *Letras*, 1:121–152.

Aryon Dall'Igna Rodrigues. 1986. *Línguas brasileiras*. São Paulo: Loyola.

Milan Straka, Jan Hajič and Jana Straková. 2016. UD-Pipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC'16).

Guillaume Thomas. 2019a. UD Mbya_Guarani_Dooley, Mbyá Guaraní treebank based on narratives collected by Robert Dooley. In Nivre et al. 2019.

Guillaume Thomas. 2019b. UD Mbya_Guarani_Thomas, Mbyá Guaraní treebank based on narratives collected by Guillaume Thomas. In Nivre et al. 2019.

Judith Tonhauser. 2017. The Distribution of Implicit Arguments in Paraguayan Guaraní. In Bruno Estigarribia and Justin Pinta (eds), *Guaraní Linguistics in the 21st century*, pages 231–258. Leiden: Brill.

Alonso Vasquez, Renzo Ego Aguirre, Candy Angulo, John Miller, Claudia Villanueva, Željko Agić, Roberto Zariquiey and Arturo Oncevay. 2018. Toward Universal Dependencies for Shipibo-Konibo. Proceedings of the Second Workshop on Universal Dependencies (UDW 2018), pages 151–161. `https://www.aclweb.org/anthology/W18-6018`

Irina Wagner, Andrew Cowell and Jena Hwang. 2016. Applying Universal Dependency to the Arapaho Language. Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016), pages 171–179. DOI: 10.18653/v1/W16-1719.

Universal Dependencies n.d. Universal Dependencies Guidelines. `https://universaldependencies.org/guidelines.html`