Building a Mbyá Treebank

ICLDC 6

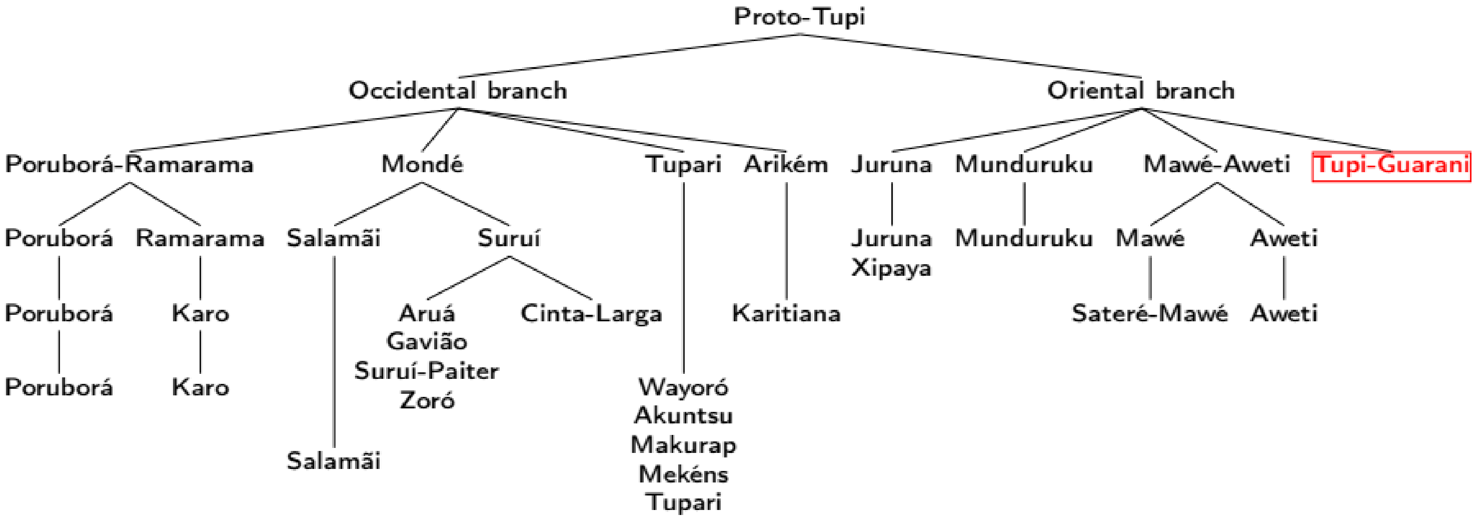
Guillaume Thomas, University of Toronto

Building a Mbyá Treebank

- Mbyá language
- The Mbyá Treebank:
  - What kind of annotation? What kind of uses?
- Annotation Work:
  - Timeline, workflow, resources used, archiving.
- Next steps

The Mbyá Language

Classification


Proto-Tupi

Occidental branch

Oriental branch

Puruborá-Ramarama

Mondé

Tupari

Arikém

Juruna

Munduruku

Mawé-Aweti

Tupi-Guarani

Puruborá

Ramarama

Salamãi

Suruí

Juruna
Xipaya

Munduruku

Mawé

Aweti

Puruborá

Karo

Aruá
Gavião
Suruí-Paiter
Zoró

Cinta-Larga

Karitiana

Sateré-Mawé

Aweti

Puruborá

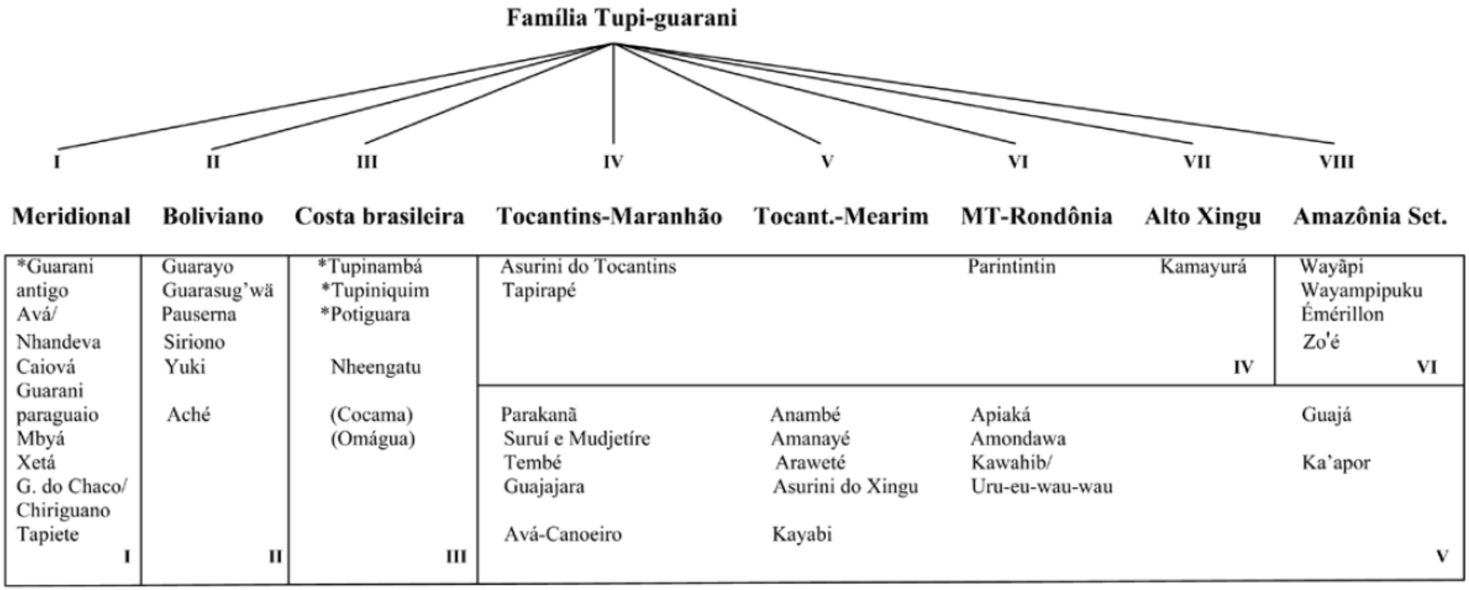
Karo

Wayoró
Akuntsu
Makurap
Mekéns
Tupari

Salamãi

(Dietrich 2010)

Mbyá Quick Facts



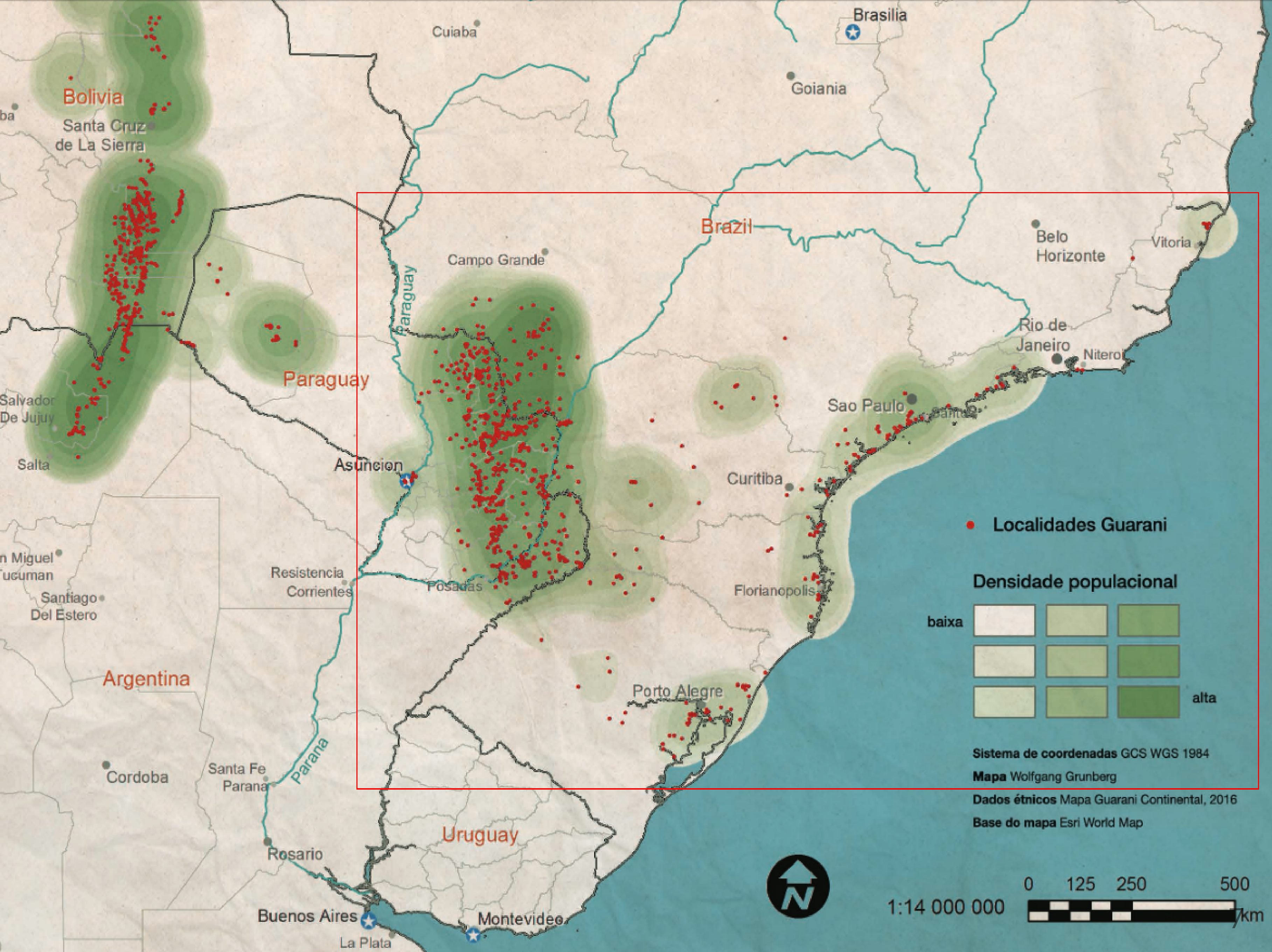
Família Tupi-guarani

I	II	III	IV	V	VI	VII	VIII
Meridional	Boliviano	Costa brasileira	Tocantins-Maranhão	Tocant.-Mearim	MT-Rondônia	Alto Xingu	Amazônia Set.
*Guarani antigo Avá/ Nhandeva Caiová Guarani paraguaio Mbyá Xetá G. do Chaco/ Chiriguano Tapiete I	Guarayo Guarasug’wä Pauserna Siriono Yuki Aché II	*Tupinambá *Tupiniquim *Potiguara Nheengatu (Cocama) (Omágua) III	Asurini do Tocantins Tapirapé		Parintintin	Kamayurá IV	Wayãpi Wayampipuku Émérillon Zo'é VI
			Parakanã Suruí e Mudjetíre Tembé Guajajara Avá-Canoeiro	Anambé Amanayé Araweté Asurini do Xingu Kayabi	Apiaká Amondawa Kawahib/ Uru-eu-wau-wau		Guajá Ka’apor V

(Dietrich 2010)

Mbyá Quick Facts

- Approx. 30,000 Mbyá speakers (Instituto Socioambiental):
  - Argentina: 2147 (INDEC, 2010)
  - Brazil: 7000 (Funasa, Funai, 2008)
  - Paraguay: 21422 (Censo Nacional, 2012)
- Next slide:
  - Guarani languages, Mapa Guarani Continental (2016)
  - Mbyá spoken inside red rectangle.



Sistema de coordenadas GCS WGS 1984

Mapa Wolfgang Grunberg

Dados étnicos Mapa Guarani Continental, 2016

Base do mapa Esri World Map

Mbyá Quick Facts

- Word Order:
  - Predominantly SVO in matrix, SOV in subordinated.
  - About 50% arguments in corpus are omitted.
- Head marking, rich agglutinating verbal morphology
- Active Stative Language
- Omnipredicative language

Linguistic Situation

- Inventário da Linguá Mbyá, IPOL (Instituto de Investigação e Desenvolvimento em Política Linguística), 2011:
    - 69 Mbyá communities in 6 Brazilian States
    - 313 families, 596 individuals
    - 98% individuals reported oral production fluency
    - 41% individuals reported writing fluency
    - 94% parents reported transmission to children

Linguistic Situation

- INRC questionnaire: interview of 40 inhabitants of Sapukai community in RJ (11: 6-15, 16: 16-30, 10: 31-60, 4: > 60):
    - 100% were native speakers of Guarani.
    - 95% speak Portuguese as a second language.
    - Portuguese first learned at school (average 7.5 year old).
    - 37% declare that they can ‘speak well’ in Portuguese, 47% that they can ‘speak more or less’ and 17% that they can ‘hardly speak.’
    - 95% never use Portuguese in Sapukai except when a non-indigenous person is visiting.

Publications in Mbyá

- No reliable, exhaustive inventory.
- Personal inventory, 2013, building on:
  - "Valorização do mundo cultural Guarani Mbya", Ladeira et al. 2011
  - "Guia de fontes e bibliografia sobre línguas indígenas e produção associada", Facó Soares, 2010.
  - "O livro indígena e suas múltiplas grafias", Machado Alvez de Lima, 2012.
- Outdated, not exhaustive.
- Excludes audiovisual productions (numerous and popular).

Publications in Mbyá

Genre	Count
Mythological narratives	4
Religious prose and songs	3
Autobiographical narratives & *nhemongeta*	13
Other narratives	3
Language education	21
Evangelical prose	2
Other	5
Total	51

- Argentina: 6%, Brazil: 72%; Paraguay: 22%
- Date range: 1959 - 2010

Language Scholarship

- Relatively well described/documented language:
  - Rich scholarship on language family
  - Advanced grammatical description (Dooley 2015)
  - Relatively large number of speakers, large territory
  - Relatively large number of written and audiovisual documents
- Lacunae:
  - Lack of publicly available, searchable corpora
  - Virtually no documentation of variation

The Mbyá Treebank: composition

Project Goals

- Create a multi-layer corpus of texts composed of:
  - legacy materials
  - original documentation: INRC corpus
- Include documents from various areas and times
- Develop automatic annotation and analysis tools
- Accessibility:
  - Publicly accessible archive
  - Advanced search and visualization tools

Digitized material

Source	Genre	Country	Word Count
Dooley's AILLA collection	Narratives	BRA	12509
INRC corpus	Oratory	BRA	24225
Ayvu Rapyta	Narratives	PRY	3586
El Canto Resplandeciente	Narratives	ARG	4032
Opa Mba'e Re Nhanembo'e Aguã	Alphabetization	BRA	3064
Nhandereko Nhemombe'u Tenonderã	Narratives	BRA	14192
Varai Parai Regua	Narratives	BRA	586

Morphological glosses, lemmas, POS, syntactic dependencies, coreference, animacy.

POS & syntactic dependencies (under correction), ELAN transcription

POS & syntactic dependencies (under correction)

Dooley's AILLA collection

- Narratives produced between 1970 and 1990.
- SIL "Indigenous Literature Workshops".
- Archived on AILLA:
  - Archive of the Indigenous Languages of Latin America
  - UT-Austin
- 1,000 sentences, two authors from Paraná (Brazil):
  - Nelson Florentino
  - Darci Pires de Lima

Dooley's AILLA collection

- Why start with this collection?
    - Already available in a publicly accessible archive.
    - Dooley authorized archiving the Treebank on AILLA.
    - Convenient starting point:
        - Interlinearized,
        - Spelling normalized,
        - Consistent with Dooley's (2015) grammar.

INRC corpus

- Inventário Nacional de Referências Culturais do Povo Guarani dos Estados do Rio de Janeiro e Espírito Santo.
- Language documentation project, 2013-2015.
    - Funded by IPHAN
    - Hosted by Museu do Índio/FUNAI
    - Scientific coordinator: the author of this talk
- Document verbal arts in Mbyá:
    - oratory,
    - songs,
    - cerimonial discourse.

INRC corpus

- 6 communities from Espirito Santo and Rio de Janeiro:
    - Araponga (RJ), Parati Mirim (RJ), Sapukai (RJ), Tekoa Porã (ES), Piraque Açu (ES), Mboapy Pindo (ES).
- Mbyá documentation team:
    - 12 young adults and 12 elders
    - recording, editing, translation
    - elders recorded in public speech events and cerimonies
- Personal and logistic support from Museu do Índio and FUNAI

INRC corpus

Alexandre Ferreira Benites, Alexandro Karai Benite, Antonio Carvalho, Agostinho da Silva, Diego Escobar, Edison Jecupé, Edmilson Karai da Silva, Francisco Karai Tataendy, Fernandes da Silva, Genilson da Silva, Ilson Fernandes, João da Silva, Jonas Ernesto da Silva, Juninho da Silva, Luiz Almeida, Maikon Brando da Silva Vaz, Maninho Pepe, Marciana Para Mirim de Oliveira, Marcio Kuaray da Silva, Marilza da Silva, Miguel Benites, Nelson Carvalho dos Santos, Nirio da Silva, Pedro da Silva, Rodrigo da Silva, Teresa da Silva, Thiago Vera Benite da Silva, Vanda de Lima Carvalho, Wesley Silveira Carvalho

INRC corpus

- 80 hours of audiovisual recordings, including:
    - 10 hours of audio recording of oratory,
    - 2 hours of audio recording of songs.
- Archived at the Museu do Índio/FUNAI, Rio de Janeiro.
- Audiobook and song recordings to be published by Museu do Índio.

INRC corpus

- Among the oratory documented in the project:
  - 3 hours transcribed and translated in ELAN,
  - 7 speakers, 3 men, 4 women.
  - Included in the Mbyá Treebank.
  - Access restricted (to be archived at Museu do Índio).
- First publicly archived corpus of oral discourse in Mbyá.
- Counterbalance written narratives accessible in legacy materials.

The Mbyá Treebank: Annotation

Annotation framework: desiderata

- Consistency with state of the art scholarship on Mbyá:
  - Dooley's (2015) grammar and dictionary
- Facilitate comparison with other languages:
  - in particular Tupi Guarani languages
- Guidelines should have a soft learning curve.
- Framework should be supported by widespread NLP tools.

Universal Dependencies

> Universal Dependencies (UD) is a framework for cross-linguistically consistent grammatical annotation and an open community effort with over 200 contributors producing more than 100 treebanks in over 70 languages.

- https://universaldependencies.org/
- Coordinated by Joakim Nivre (Uppsala University)

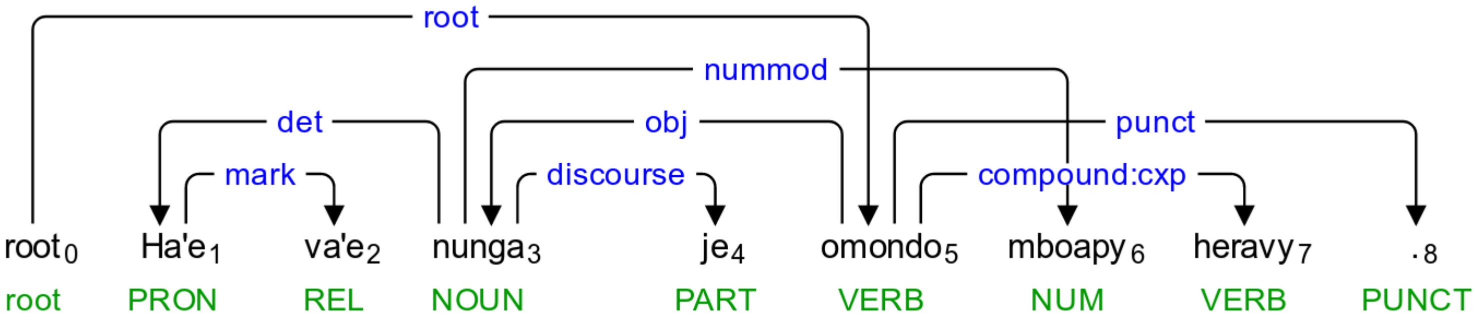
Universal Dependencies

- Cross-linguistic framework:
  - 17 universal POS tags
  - 37 universal syntactic relations
  - 49 grammatical features
- Cross-linguistic flexibility:
  - Language specific POS tags
  - Language specific subtyping of syntactic relations
  - Typologically diverse projects: Ainu, Amharic, Japanese, Vietnamese, Yoruba…

Universal Dependencies

- Dependency grammar:
  - Grammatical relations between words: dependencies.
  - Surface oriented, no transformation.
- Universal Dependencies:
  - Content word are heads of function words.

Universal Dependencies



"*Of this type he removed three, and he took them with him.*"

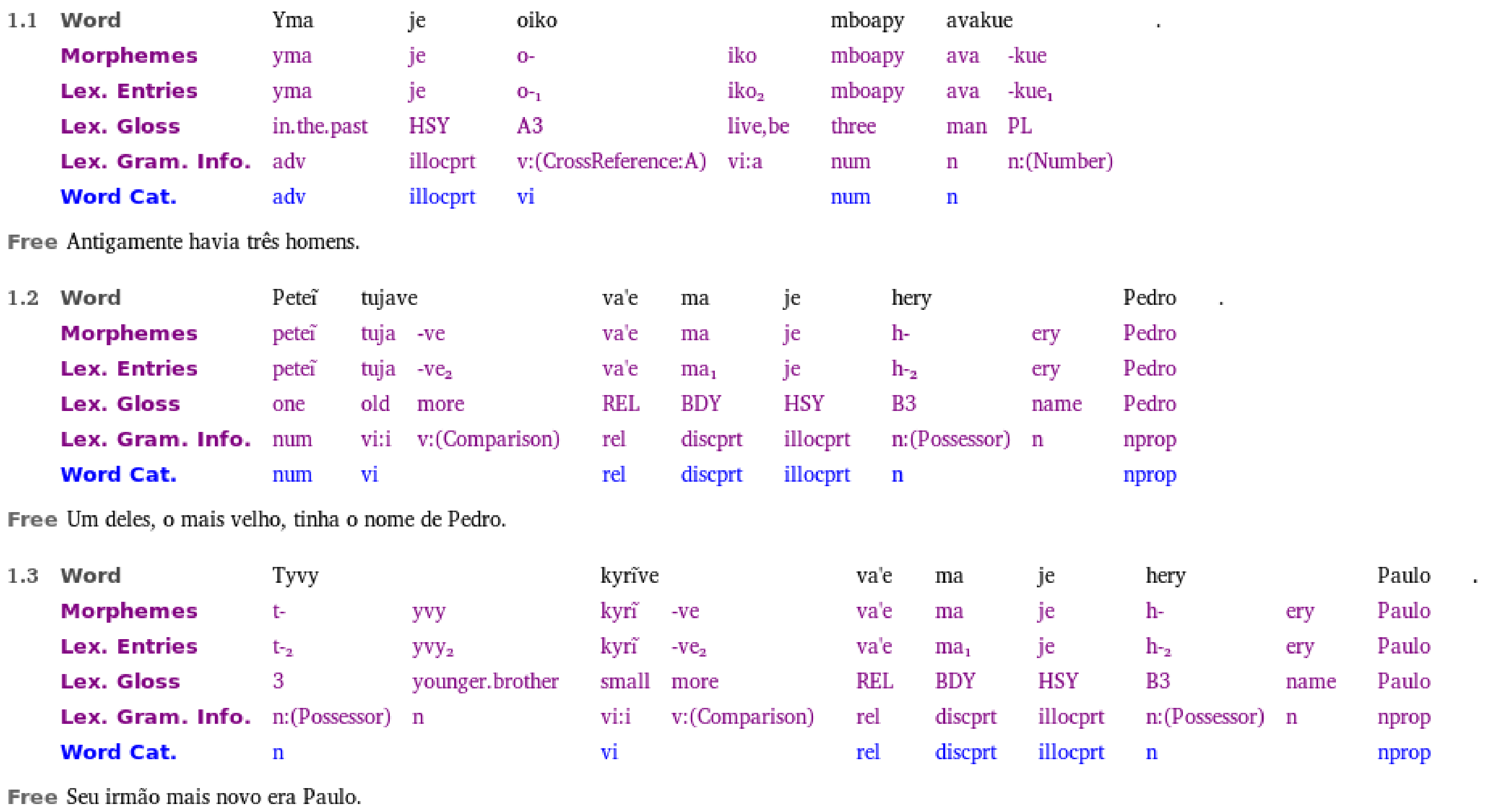
Interlinearization & POS tagging

- Spelling and word tokenization:
    - No generally accepted spelling and word tokenization conventions in Mbyá.
    - Follow Dooley's (2015) conventions.
    - Popular in the South/South-East of Brazil.
    - In use in the INRC communities.

Mbyá specific POS tagging

- Split-S cross-referencing system:
  - inactive intransitive
  - active intransitive
- Most 'adjectives' are inactive verbs
- Nouns are productively used as possessive/attributive predicates
- Rich particle inventory:
  - Tense/Aspect/Modality, quantification, focus, information structure, illocutionary modifiers, ...

Interlinearization & POS tagging: SIL FLEX



1.1	Word	Yma	je	oiko		mboapy	avakue		.
	Morphemes	yma	je	o-	iko	mboapy	ava	-kue	
	Lex. Entries	yma	je	o-$_1$	iko$_2$	mboapy	ava	-kue$_1$	
	Lex. Gloss	in.the.past	HSY	A3	live,be	three	man	PL	
	Lex. Gram. Info.	adv	illocprt	v:(CrossReference:A)	vi:a	num	n	n:(Number)	
	Word Cat.	adv	illocprt	vi		num	n		

Free Antigamente havia três homens.

1.2	Word	Peteĩ	tujave		va'e	ma	je	hery		Pedro	.
	Morphemes	peteĩ	tuja	-ve	va'e	ma	je	h-	ery	Pedro	
	Lex. Entries	peteĩ	tuja	-ve$_2$	va'e	ma$_1$	je	h-$_2$	ery	Pedro	
	Lex. Gloss	one	old	more	REL	BDY	HSY	B3	name	Pedro	
	Lex. Gram. Info.	num	vi:i	v:(Comparison)	rel	discprt	illocprt	n:(Possessor)	n	nprop	
	Word Cat.	num	vi		rel	discprt	illocprt	n		nprop	

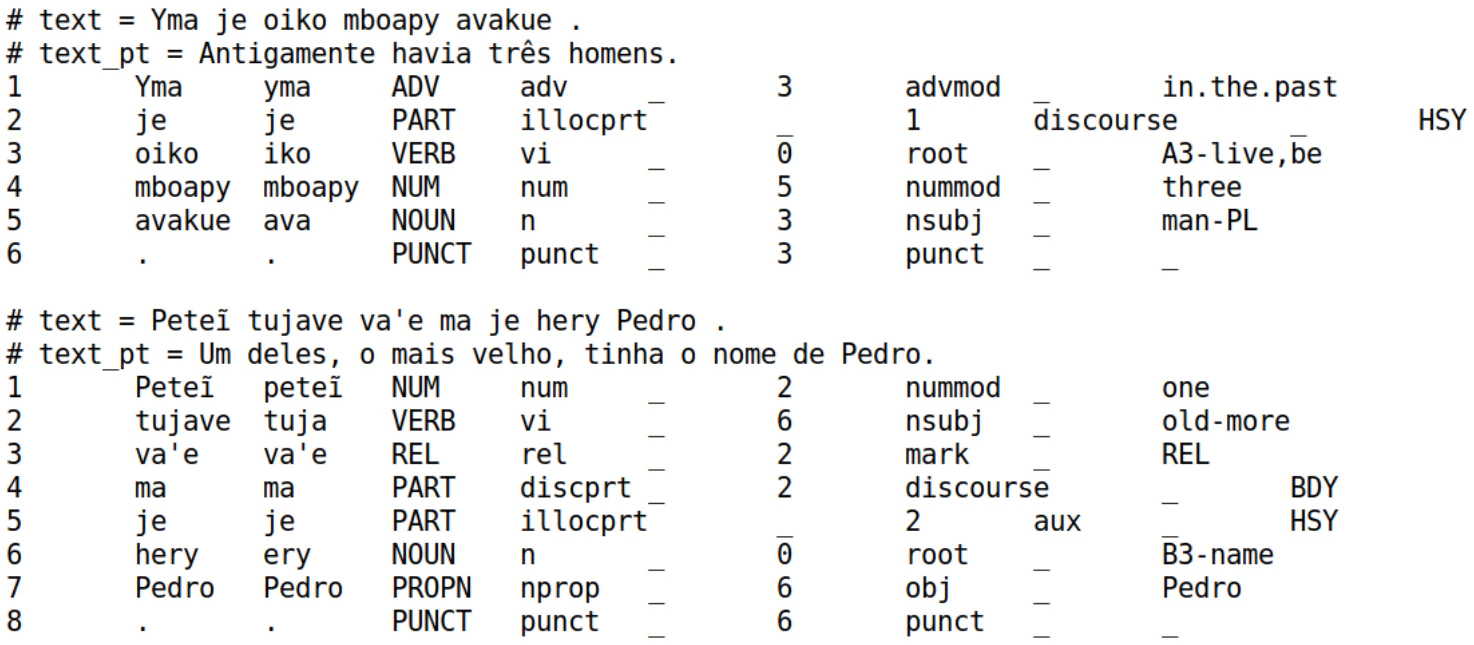
Free Um deles, o mais velho, tinha o nome de Pedro.

1.3	Word	Tyvy		kyrĩve		va'e	ma	je	hery		Paulo	.
	Morphemes	t-	yvy	kyrĩ	-ve	va'e	ma	je	h-	ery	Paulo	
	Lex. Entries	t-$_2$	yvy$_2$	kyrĩ	-ve$_2$	va'e	ma$_1$	je	h-$_2$	ery	Paulo	
	Lex. Gloss	3	younger.brother	small	more	REL	BDY	HSY	B3	name	Paulo	
	Lex. Gram. Info.	n:(Possessor)	n	vi:i	v:(Comparison)	rel	discprt	illocprt	n:(Possessor)	n	nprop	
	Word Cat.	n		vi		rel	discprt	illocprt	n		nprop	

Free Seu irmão mais novo era Paulo.

Dependency Annotation

- Canonical UD format: CoNLL-U tab separated text



```
# text = Yma je oiko mboapy avakue .
# text_pt = Antigamente havia três homens.
1       Yma     yma     ADV     adv     _       3       advmod  _       in.the.past
2       je      je      PART    illocprt        _       1       discourse       _       HSY
3       oiko    iko     VERB    vi      _       0       root    _       A3-live,be
4       mboapy  mboapy  NUM     num     _       5       nummod  _       three
5       avakue  ava     NOUN    n       _       3       nsubj   _       man-PL
6       .       .       PUNCT   punct   _       3       punct   _       _

# text = Peteĩ tujave va'e ma je hery Pedro .
# text_pt = Um deles, o mais velho, tinha o nome de Pedro.
1       Peteĩ   peteĩ   NUM     num     _       2       nummod  _       one
2       tujave  tuja    VERB    vi      _       6       nsubj   _       old-more
3       va'e    va'e    REL     rel     _       2       mark    _       REL
4       ma      ma      PART    discprt _       2       discourse       _       BDY
5       je      je      PART    illocprt        _       2       aux     _       HSY
6       hery    ery     NOUN    n       _       0       root    _       B3-name
7       Pedro   Pedro   PROPN   nprop   _       6       obj     _       Pedro
8       .       .       PUNCT   punct   _       6       punct   _       _
```

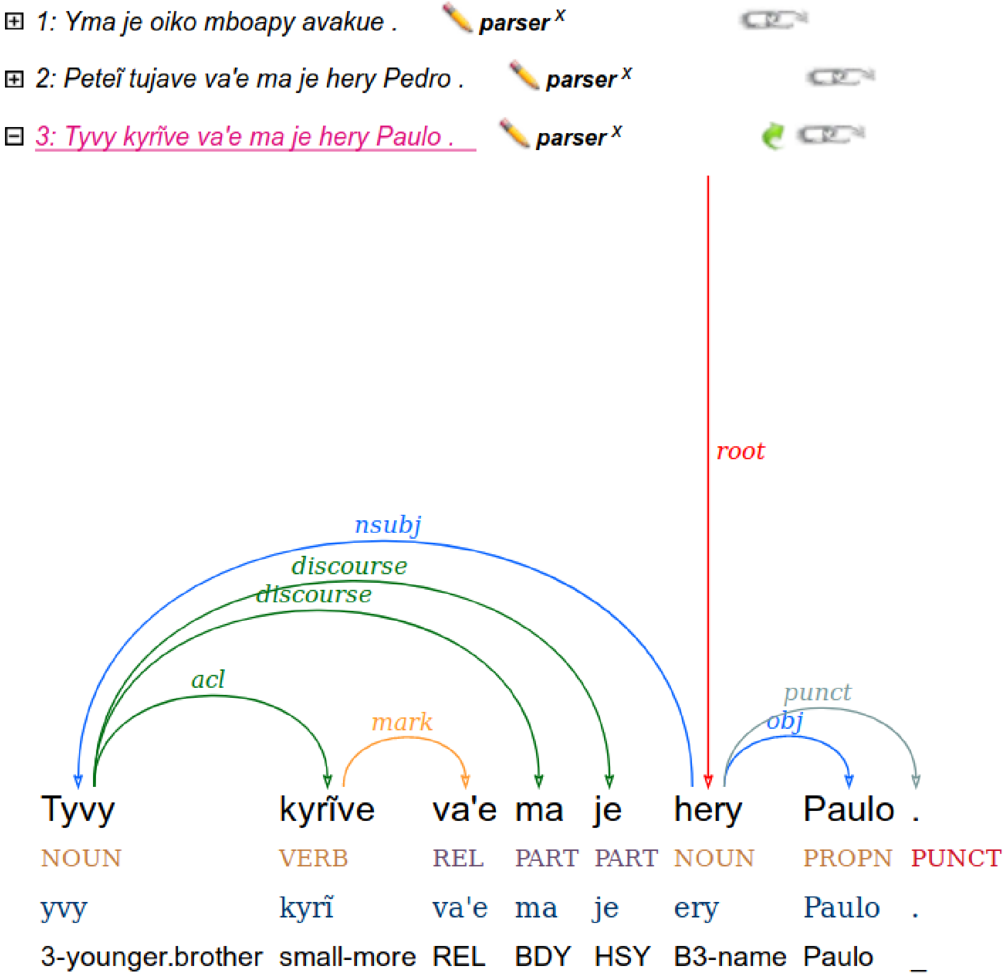
Dependency Annotation

- Interlinearization in FLEX exported to flextext XML format.
- Python script:
  - converts FLEX XML to CoNLL-U
  - perform rudimentary lemmatization
  - maps Mbyá POS tags to UD POS tags
- Must still be implemented:
  - grammatical feature extraction

Dependency Annotation: Arborator

- Dependency annotation can be manual or automatic.
- In both cases, human annotators needed.
- Online interface for collaborative dependency annotation:
    - Arborator
    - Written by Kim Gerdes, Université de Paris III

Dependency Annotation: Arborator



⊞ 1: *Yma je oiko mboapy avakue .* **parser** X

⊞ 2: *Peteĩ tujave va'e ma je hery Pedro .* **parser** X

⊟ 3: *Tyvy kyrĩve va'e ma je hery Paulo .* **parser** X

Tyvy	kyrĩve	va'e	ma	je	hery	Paulo	.
NOUN	VERB	REL	PART	PART	NOUN	PROPN	PUNCT
yvy	kyrĩ	va'e	ma	je	ery	Paulo	.
3-younger.brother	small-more	REL	BDY	HSY	B3-name	Paulo	_

Coreference Annotation

- Tagging of Referential expressions:
  - Overt: Noun Phrases, Pronouns, Proper names, etc
  - Zero subjects and objects on verbs
- Addition of Coreference relations:
  - Referential **identity**
  - **Inferred** relation: bridging anaphora, implicit partitives

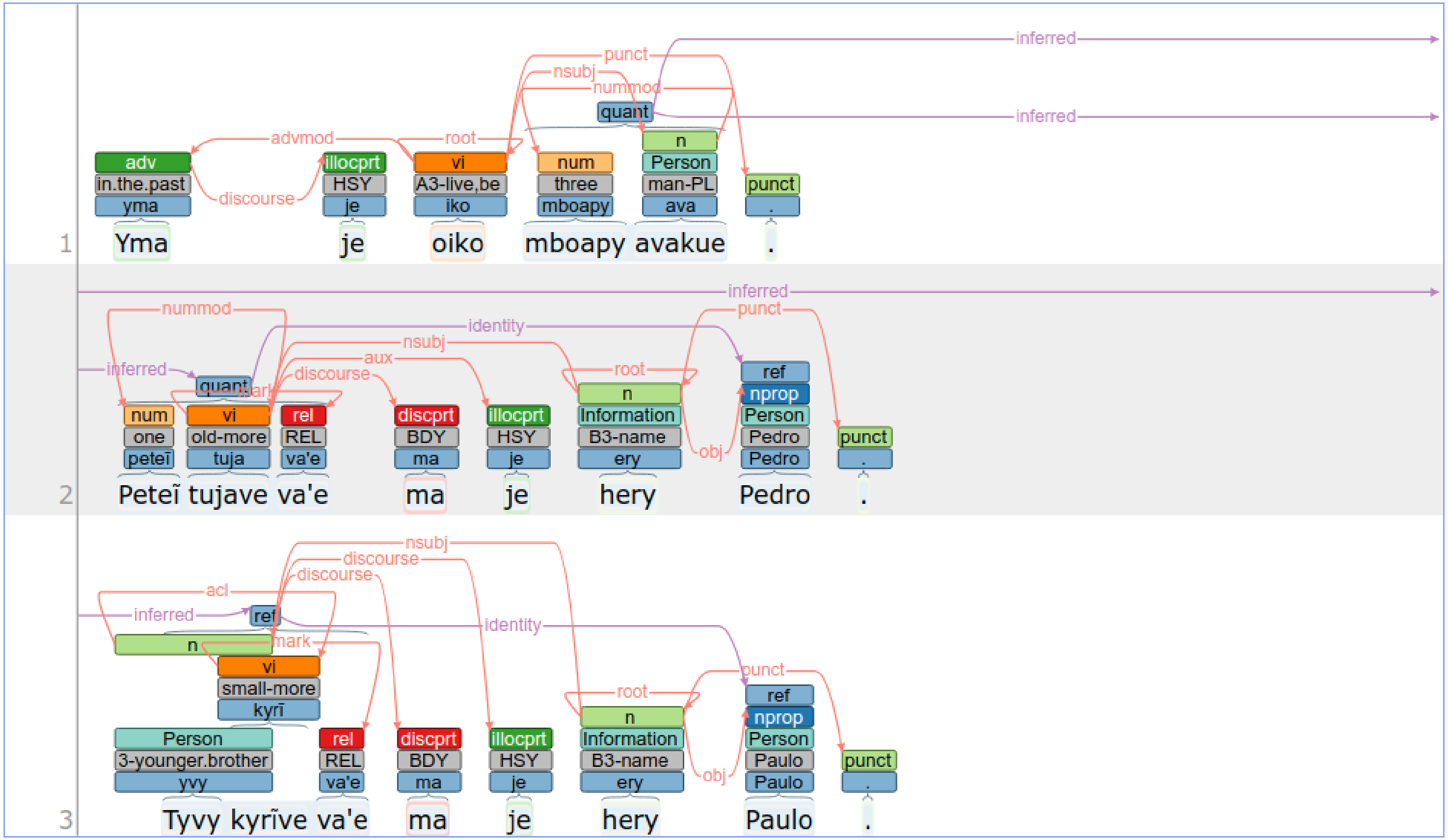
Coreference Annotation: WebAnno 3

- More flexible annotation platform:
  - Multiple annotation layers
  - Tagging of spans of words
  - Relations across sentences
- Webanno 3:
  - Project Lead: Richard Eckart de Castilho
  - Technische Universität Darmstadt

Coreference Annotation: WebAnno 3

- CoNLL-U files were imported into WebAnno 3 and annotated for coreference.
- A layer of Ontological Category for noun phrases was added using a Python script and the FLEX lexicon:
    - Physical Objects > Person; Animal; Artefact; …
    - Eventuality > State; Event
    - Abstract Objects > Property; Information; …
    - …

Coreference Annotation: WebAnno 3



1 Yma je oiko mboapy avakue .

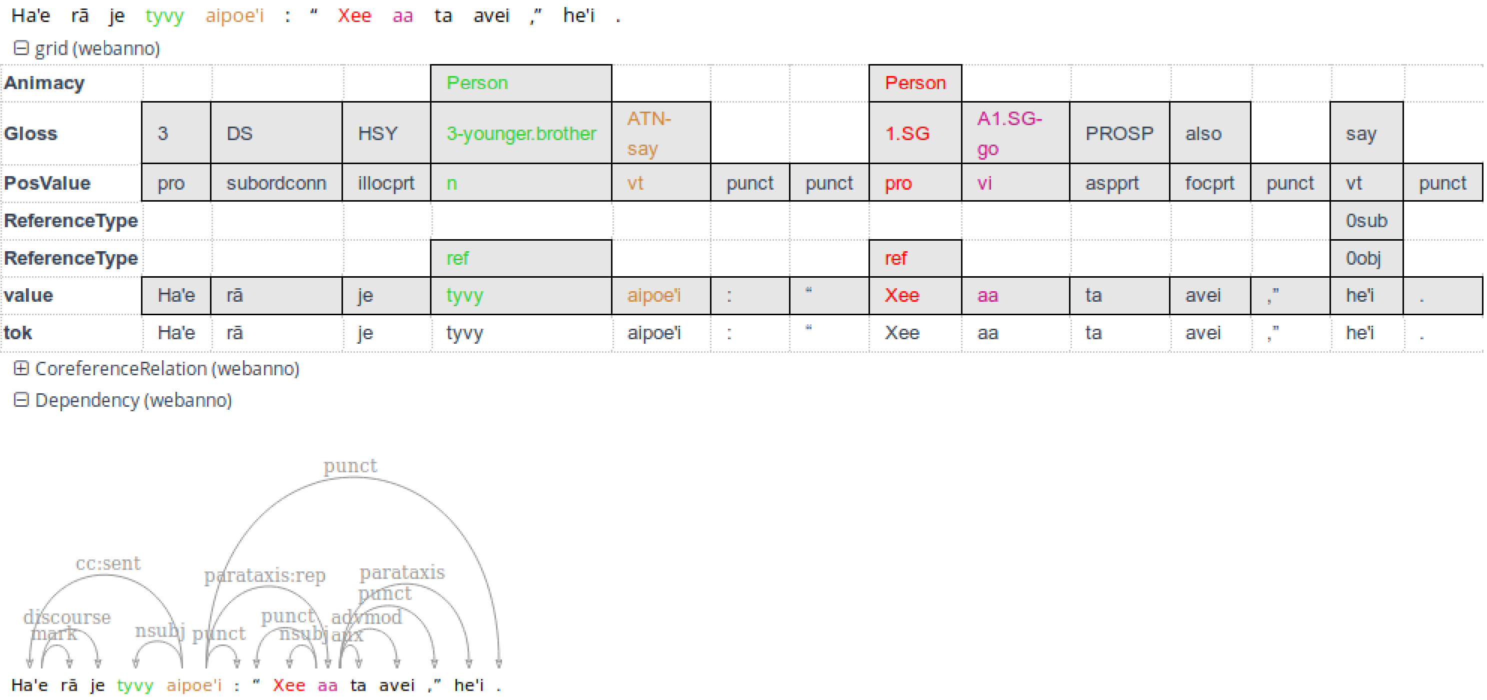
2 Peteĩ tujave va'e ma je hery Pedro .

3 Tyvy kyrĩve va'e ma je hery Paulo .

Visualization in ANNIS

- Plain text files of the treebank are being archived.
- A search and visualization interface is still needed.
- ANNIS (ANNotation of Information Structure):
  - web-browser based search and visualization platform
  - support multi-layer annotation
  - advanced query language
  - Pepper: tool for conversion of different annotation format
- http://corpus-tools.org/annis/
- Thomas Krause & Amir Zeldes (2016)

Visualization in ANNIS



Ha'e rā je tyvy aipoe'i : “ Xee aa ta avei ,” he'i .

⊟ grid (webanno)

Animacy				Person				Person						
Gloss	3	DS	HSY	3-younger.brother	ATN-say			1.SG	A1.SG-go	PROSP	also		say	
PosValue	pro	subordconn	illocprt	n	vt	punct	punct	pro	vi	aspprt	focprt	punct	vt	punct
ReferenceType													0sub	
ReferenceType				ref				ref					0obj	
value	Ha'e	rā	je	tyvy	aipoe'i	:	“	Xee	aa	ta	avei	,”	he'i	.
tok	Ha'e	rā	je	tyvy	aipoe'i	:	“	Xee	aa	ta	avei	,”	he'i	.

⊞ CoreferenceRelation (webanno)

⊟ Dependency (webanno)

```
node & node & node & node & #2 ->Dependency[DependencyType=/nsubj/] #1

& #4 ->Dependency[DependencyType=/nsubj/] #3

& #3 ->CoreferenceRelation[CoreferenceRelation=/identity/] #1
```

Visualization in ANNIS

- ANNIS's query language also makes it a great tool for corpus correction.

- Look for all adjectival modifiers of PP:

```
cpos=/NOUN/ & tok & cpos=/PART/ & #1 ->dep[func=/case/] #2

& #1 ->dep[func=/amod/] #3 & #2 .* #3
```

- Look for all occurrences of "ma je" not glossed as 'BDY':

```
/ma/ & /je/ & gls!=/BDY/ & #1 . #2 & #1_=_#3
```

Workflow

NLP tools

- POS tagger and dependency parser trained in UDPipe.
- UDPipe:
    - Institute of Formal and Applied Linguistics, Charles University
    - http://ufal.mff.cuni.cz/udpipe
- Lay-person friendly:
    - No advanced knowledge of parsing required,
    - Well documented,
    - Built-in evaluation tools.

NLP tools

- Evaluation on 60 sentences retained from Dooley corpus:
    - POS tagger: upostag: 92.17%, xpostag: 89.81%
    - Parser: UAS: 82.72%, LAS: 74.74%
- Inflated scores:
    - few test sentences
    - normalized spelling
    - limited vocabulary
- Parser was essential in speeding up the annotation process.

Annotation team

- INRC corpus, 2013-2015
  - transcription:
    - 24 Mbyá researchers from RJ and ES
  - translation:
    - Alberto Alvarez, Mbyá & Nhandeva speaker from RJ
  - Funding:
    - National Institute of Historic and Artistic Heritage of Brazil

Annotation team

- Mbyá Treebank, 2017 - present
    - 6 UofT undergraduate students:
        - Gregory Antono, Laurestine Bradford, Vidhya Elango, Jean-François Juneau, Barbara Peixoto, Darragh Winkelman
    - 1 UofT Linguistics graduate student:
        - Angelika Kiss
    - Funding: Connaught Fund, UofT
    - Interlinearization and translation were essential to support annotation work by students.

Dooley Corpus Workflow

1. 2015-2016: Interlinearize corpus
2. 2017: Manual dependency annotation of 300 sentences
3. Spring & summer 2018
    - Train parser on 300 sentences, parse remaining 700
    - Correction in Arborator with student RAs
4. Winter 2018
    - Coreference annotation with student RAs
    - Upload to ANNIS

Archiving

- Dooley's AILLA collection:
    - Parallel collection that will host the treebank already created on AILLA.
    - Treebank will be uploaded soon.
- INRC corpus:
    - Original corpus already archived at the Museu do Indio
    - Treebank will be archived there in the near future.

Looking Forward

Looking forward

- Immediate plans for INRC corpus:
    - Adding full Interlinearization to INRC corpus
    - Correcting POS tagging & Dependency Annotation
    - Integrating ELAN time-aligned audio in corpus
    - Archiving by the end of 2019
- Immediate plans for divulgation:
    - Discuss possible application of corpus with Mbyá educators

Looking forward

- Not so immediate plans:
    - Spelling/tokenization normalization
    - Morphological Analyzer
    - Add Morphological Dependency Annotation: dependencies between morphemes
    - Discuss annotations of published texts with copyright holders
- Definitely nor immediate:
    - Cross-lingual annotation with other Tupi Guarani languages.

Conclusion

Conclusion

- Excellent situation for treebank development:
  - Advanced scholarship on Mbyá/Guarani languages
  - Legacy materials
  - Plethora of linguist friendly NLP and annotation tools
- Rich database for quantitative linguistic research
- Foundation for developing NLP resources
- How to make these resources useful to speakers?