

## DIGITAL ANNOTATED CORPORA OF BRAZILIAN INDIGENOUS LANGUAGES WITH AUTOMATIC TRANSLATIONS (DACILAT)

“Corpora anotados digitais de línguas indígenas brasileiras com traduções automáticas (DACILAT)”.

**Fapesp Processo 22/09158-5**

**Relatório 1 Revisado, junho de 2024**

Maria Filomena Sandalo Coordenador UNICAMP Email: sandalo@unicamp.br
Charlotte Galves Pesquisador Principal UNICAMP E-mail: charlotte.mgc@gmail.com
Pablo Feliciano de Farias Colaborador UNICAMPE-mail: fariap.@unicamp.br
Luiz Veronesi Colaborador UNICAMP E-mail: luiz@texugo.com.br
Leonel Figueiredo de Alencar Araripe Colaborador Email: leonel.de.alencar@ufc.br
Michael Becker Colaborador University of Massachusetts, Amherst Amherst USA

Email: becker@phonologist.org

Vanda Pires

Colaboradora

UNICAMP

E-mail: v264435@dac.unicamp.br

## 1. Introdução

É importante ressaltar que este trabalho se baseia em uma concepção de construção de corpus diferente do *mainstream* de abordagens computacionais que partem de uma enorme quantidade de dados (*big data*), como perfeitamente exposto no seguinte trecho de um artigo intitulado “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete”, publicado em 2008 na revista *Wired Magazine*:

“This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology.

The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world.”<sup>1</sup>

(Chris Anderson, *Wired Magazine* 16:7, 2008)

Assumimos uma posição diametralmente oposta que contesta a obrigatoriedade de ‘big data’ para constituir corpora de línguas utilizáveis na pesquisa e na tradução automática de línguas com poucos materiais disponíveis. Para línguas em perigo de extinção, em particular, grandes corpora não podem ser constituídos, e isso por si só as exclui desse tipo de projetos. Defendemos que a anotação acrescida aos textos é que os torna relevantes para pesquisas e aplicações para fins educativos e sociais. Afirmamos que grandes corpora são inúteis se não contiverem informações extras que permitam aos pesquisadores recuperar dados para responder às suas perguntas sobre a linguagem e sua dinâmica. Essas informações extras são adicionadas por meio de anotação e a anotação é baseada em modelos linguísticos. O que precisamos é desenvolver métodos que permitam a anotação de textos de forma

<sup>1</sup> Este é um mundo onde grandes quantidades de dados e matemática aplicada substituem todos os outros instrumentos que podem ser usados. Fora com todas as teorias sobre o comportamento humano, da linguística à sociologia. A nova disponibilidade de grandes quantidades de dados, juntamente com as ferramentas estatísticas para processar esses números, oferece uma nova maneira de entender o mundo (tradução dos autores).

rápida e confiável. Isso implica interdisciplinaridade, em particular, cooperação com cientistas da computação. Portanto, este relatório se concentra em nossa metodologia de anotação e na apresentação das nossas ferramentas para elaboração, edição e anotação de corpora desenvolvidas ou em desenvolvimento.

A *Plataforma Tycho Brahe* é o conjunto destas ferramentas **totalmente online** e é pioneira em sua aplicação para as línguas originárias da América do Sul. A sua interface baseada na web permite a disseminação imediata para as comunidades indígenas, além de permitir a criação de corpora significativamente maiores do que era possível anteriormente.

Além da Plataforma, este projeto entregará dois corpora linguísticos: um Corpus de narrativas do kadiwéu e um Corpus de narrativas do nheengatu, duas línguas de tipologias bem distintas, além de dicionários online das duas línguas. Tudo gramaticalmente anotados para alimentar traduções automáticas e estudos das línguas.

A plataforma Tycho Brahe teve seu início em projetos anteriores, sobre o português. Nossa inovação é adequar a plataforma para línguas originárias e avançar em tradução automática, bem como oferecer os dados online. A ferramenta oferece ferramentas multifuncionais para corpora linguísticos, especialmente adequadas para línguas de alta complexidade morfológica, conhecidas na linguística como aglutinativas ou polissintéticas, em que cada palavra pode codificar os significados de frases inteiras. A Plataforma também inclui diversas funções: recursos multimídia, tradução frase por frase, marcação morfológica, ferramentas de análise sintática, mecanismos de busca, metadados ricos, resultados estatísticos com arquivamento histórico e um fluxo de trabalho completo para edição de árvores sintáticas, bem como ferramentas para importação e exportação de documentos do corpus. A plataforma é baseada em microestruturas sintáticas armazenadas em um banco de dados sintático em nuvem e em regras computacionais paramétricas que permitem uma experiência rica para o usuário. Também incluiremos a integração de uma ferramenta de tradução automática, que será especificamente desenvolvida durante o projeto.

Nosso trabalho tem ainda importante benefícios sociais. As línguas do Brasil, já gravemente ameaçadas antes da pandemia, contaram com a morte de idosos por COVID19, enfrentando, assim, mais uma etapa de enfraquecimento. Este trabalho apresenta nossos esforços na busca de inovação digital inclusiva desenvolvendo uma Plataforma computacional para línguas originárias do Brasil com materiais que são culturalmente e gramaticalmente significativos para as comunidades indígenas. Esses materiais fornecem uma base digital para a educação bilingue aprimorada, conexões intergeracionais, e transmissão de saberes ancestrais. Assim, temos como objetivo poder colaborar com a documentação e preservação nas escolas de línguas nativas do Brasil através de nossa ferramenta para elaboração de corpora linguísticos. As anotações linguísticas dos corpora são feitas de modo a já formar uma gramática das línguas documentadas. Estas gramáticas também servirão como ferramentas de ensino para melhorar o letramento dos falantes locais em suas próprias línguas, bem como a preservação de narrativas culturalmente significativas. Além disso, um melhor estudo das

correspondências, nem sempre óbvias das línguas indígenas e europeias, também melhorará a formação de professores de línguas indígenas para uma educação bilíngue mais eficiente e equilibrada, com resultados positivos para o acesso ao ensino superior. E de um ponto de vista linguístico permitirá novas pesquisas sobre a gramática destas línguas em processo de enfraquecimento e de extinção. O corpus do Kadiwéu na plataforma foi particularmente desenvolvido durante este primeiro ano do projeto DACILAT e sua anotação será discutida neste relatório.

### 1.1.A Plataforma online e links para navegação

Estamos surpresos pela recusa do relatório anterior, cujo principal motivo, entendemos, foi a recusa de abrir links de internet. Verdade que várias afirmações feitas pelo parecerista nos fizeram rever o relatório apresentando mais explicações e figuras ilustrativas. No entanto, nosso projeto lida com ferramentas de web, ou seja, ferramentas dinâmicas que exigem navegação na internet, cujas impressões são impossíveis. Consideramos que links são a maneira mais simples de demonstrar o trabalho feito e de permitir que o parecerista navegue e avalie nosso desenvolvimento no projeto. Reconhecemos que faltamos com algumas informações e esquecemos de abrir alguns links no relatório anterior, mas agora tentamos oferecer mais informações/orientações para a navegação e um guia passo a passo com prints de páginas cruciais. No entanto não podemos evitar links dada a natureza web do trabalho.

## 2. Cronograma

O cronograma apresentado no projeto para esta fase foi (copiado do projeto enviado para a FAPESP) o seguinte:

### 1. Tycho Brahe Platform:

2023: Search in document, apply editions to multiple words, inline edition for faster data input, and edit comments for multiple editors and reviews, hyphen auto detection in transcription. Improve lexicon management. General layout improvements for best user experience. Testing and bug fixes.

### 2. Machine Translation

2023: Incorporate Kadiwéu (ISO 639-3 kbc) into GammKYEP, producing GammKYEP, so that all sentences in the Nheengatu positive test set can be translated into that language and vice versa. GammKYEP will automatically enable automatic translation between Kadiwéu, English and Portuguese. The first step to implement the concrete syntax of Kadiwéu is to translate the positive test

set of Nheengatu into that language and produce a variant of the result with violations of Kadiwéu grammar rules. The implementation of the grammar of Kadiwéu will occur incrementally, starting from the simplest examples to the most complex ones, adopting a spiral development design.

### 3. Morpho-syntactic annotation

2023: conclusion of the parsing system for Kadiweu. Elaboration and testing of a rule-based parser for Kadiwéu. Elaboration of a POS tagging system for Nheengatu.

## 3. Etapas realizadas de acordo com o cronograma

### 3.1. A Plataforma Tycho Brahe

A Plataforma Tycho Brahe tem sido desenvolvida pelo aluno de doutorado no projeto, Luiz Henrique Lima Veronesi, sob supervisão de Charlotte Galves. Neste projeto, as ferramentas da plataforma estão sendo desenvolvida especialmente para corpora de línguas indígenas, como já mencionado. O Corpus Kadiwéu está desenvolvido, embora longe de acabado (supondo que exista um fim para um corpus). O Corpus Nheengatu ainda não foi iniciado dentro da plataforma Tycho. A plataforma Tycho é central ao projeto e esta seção do relatório contém a descrição das atividades técnicas e ferramentas da plataforma desenvolvidas até o momento.

Todas as metas prometidas no cronograma foram cumpridas e fomos além, como será apresentado mais adiante nesta seção.

Segue abaixo uma foto da página da Plataforma (<https://www.tycho.iel.unicamp.br/home>) que apresenta as suas ferramentas desenvolvidas e/ou em desenvolvimento:

## Ferramentas

**Visualizador**

Explore os corpora disponíveis

Utilizado para visualização dos corpora disponíveis.

**eDictor**

Uma ferramenta para edição filológica e anotação linguística automática

Editor de textos (eDictor): inclusão e edição de textos e sentenças com diferentes interfaces dependendo das parametrizações do corpus: transcrição de imagens, áudios, traduções, etc.

**Syntrees**

Revisão de árvores sintáticas

Ferramenta gráfica para revisão de estruturas sintáticas com uma interface intuitiva e integrada ao parser.

**Synviewer**

Visualizador de árvores sintáticas

Ferramenta para converter expressões de colchetes sintáticos em árvores gráficas. É integrado ao analisador e exporta árvores para imagens.

**Parser**

Mecanismos para parser sintático e etiquetagem POS

Módulo para configuração dos mecanismos de etiquetagem e anotação sintática, com seus respectivos pre-processamentos.

**Search**

Ferramenta filológica e morfosintática para corpora anotados

Ferramenta de busca: para buscas de palavras, etiquetas categoriais e morfológicas e estruturas sintáticas.

**IO**

Importe e exporte documentos para seu corpus em vários formatos

Importação e exportação de dados, funcionando como módulo de interoperabilidade entre diferentes formatos de dados utilizados por outras ferramentas.

**Corpora Management**

Acesso e configuração geral de parâmetros

Administração: módulo administrativo com configuração de dados e parâmetros em geral, além do controle de acesso e backups.

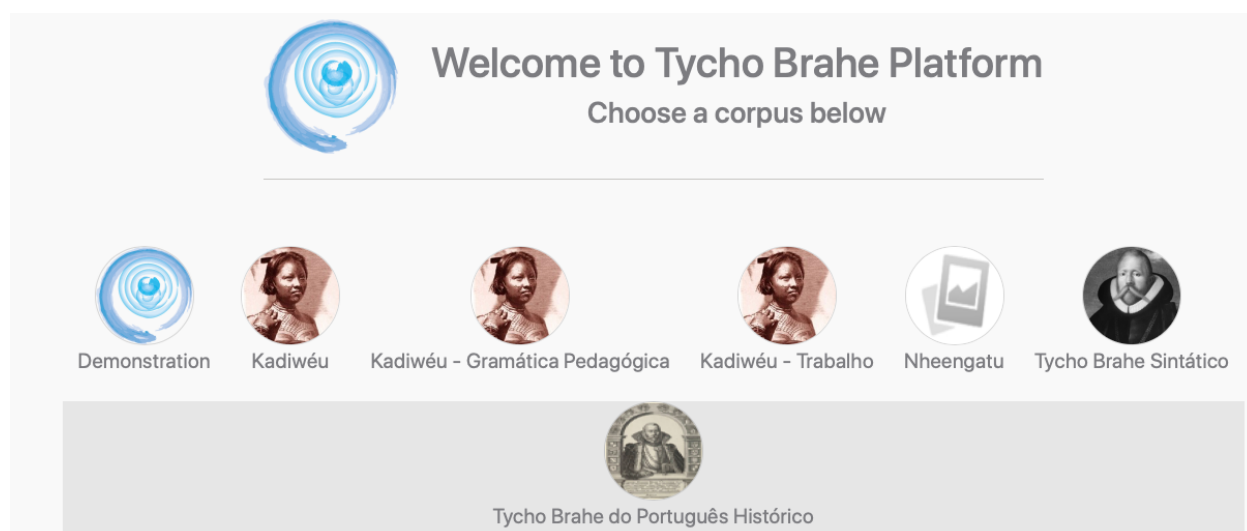
As seguintes ferramentas são totalmente abertas para visitação sem qualquer tipo de senha: Visualizador, Search e Synviewer. Abaixo apresentamos cada ferramenta, aberta para visitação ou não.

### 3.1.1. Visualizador

Nesta ferramenta, os corpora existentes na Plataforma podem ser visitados por todos os interessados, sem senha, no seguinte link:

<https://www.tycho.iel.unicamp.br/viewer>

Abaixo está uma imagem que mostra a entrada do Visualizador com nosso catálogo de corpora:



Basta clicar nas figuras e depois nas sentenças para adentrar nas anotações dos corpora. Os corpora do português não pertencem ao presente projeto, mas ajudarão na tradução automática.

Sugiro visitar Kadiwéu (5369 palavras), uma vez que este corpus de narrativas originárias já está anotado quase em sua totalidade. O corpus Kadiwéu Trabalho (2815 palavras) são as narrativas em processo de anotação. E Kadiwéu Gramática são sentenças para iniciar uma gramática pedagógica da língua (para a comunidade) e dar início a uma *Toy Grammar* para programação na linguagem *Grammatical Framework* para tradução automática. Para Nheengatu, coletamos e transcrevemos dados, mas os dados desta língua entrarão na Plataforma Tycho para o próximo relatório. Seguem imagens do navegação no Corpus Kadiwéu na narrativa *Ejiwajegi dinibolodi* ‘Os tabus de alimentos kadiwéus’.



Plataforma Tycho Brahe: Viewer		Corpus Kadiwéu	PT-BR
Documento: <b>Ejiwajegi dinibolodi</b>		Lingua: <b>portuguese</b>	anterior próximo Compartilhar
1	ica ejiwajegi inoa icoa ane doitalo me yeligo	O/um kadiwéu tem isso que é um medo de comer (algumas coisas)	▶ 📄
2	ejiwajegi aona yeligo labidaGa	Diz que o kadiwéu não come raspa queimada.	▶ 📄
3	atone yoe lotidi ica iwaalo	pois se diz/acredita (surpreendentemente) que a mulher não dá leite (se comer)	▶ 📄
4	codaa me lioneGa ayeligo	Dai que pessoas jovens não comem isso.	▶ 📄
5	niale ela nowadi eledi ayeligo	Outra coisa que não se come é fruta gêmea de árvore.	▶ 📄
6	doitigi daGa owadi nige dinigaje	O temor é de parir gêmeos.	▶ 📄
7	odaa aginaGa lioneGa idaGee ayeligo	Então os homens jovens também não comem.	▶ 📄
8	doitigi deGetacideteci aca lodawa	Eles tem medo de judiar da esposa.	▶ 📄
9	niGina lati waca nigaanigi lioneGa awikije aGoikateke moyeligo	As crianças, moços e moças jovens, não podiam comer pâncreas de vaca.	▶ 📄
10	atoneo me jomololaGati niGina me iweniti ica oko	Porque não deixa afundar quando você mergulha pessoas.	▶ 📄
11	tibige leeGodi ica me idinaGataGatinigi ica ninyoGodi	Talvez porque escondíamos na água.	▶ 📄
12	eledi aonaGa oyeligo niGina beGee nigaanigipawaanigi	Outra coisa que também não come enquanto é criança:	▶ 📄
13	niGina latikilo libitagi	o tutano do osso.	▶ 📄
14	one doita ica ejiwajegi daGa dilaike	Diz que o kadiwéu também tem medo de ter cabelos grisalhos.	▶ 📄
15	niGinoa waca lotidi liwakapadi eledi odolienaGatidi ica beGee noji	Outra coisa que eles colocavam medo enquanto novo é de tomar a nata do leite da vaca.	▶ 📄
16	leeGodi one iwakapaGadi latobi	Porque diz que enrugava o rosto.	▶ 📄

Plataforma Tycho Brahe: Viewer
Corpus Kadiwéu
PT-BR

Documento: Ejiwajegi dinibolodi
anterior
próxima
Compartilhar

ica ejiwajegi inoa icoa ane doitalo me yeligo

original	ica		ejiwajegi		inoa			icoa			ane	doitalo			me	yeligo	
etiqueta POS	D		N		Q			D			WPRO	VBAPL			C	VB	
gloss-br	-		-		-			-			-	-			-	-	
gloss	o		kadiwéu		um			isso			que	tem medo de			que	comer	
morfemas	i	ca	ejiwa	jegi	i	na	wa	i	ca	wa		d	oi	talo		y	eligo
etiqueta	Gnr	Ncl	n	Der	Gnr	Ncl	Plu	Gnr	Ncl	Plu		lnv	v	Apl		Erg	v
gloss-br	masc	ausente	-	nom	masc	vindo	-	masc	ausente	-		-	-	-		-	-
gloss	-	-	-	-	-	-	-	-	-	-		-	-	-		-	-

Áudio

Traduções

Portuguese


Olum kadiwéu tem isso que é um medo de comer (algumas coisas)

English

Syntactic Rules

### 3.1.2. Search

A ferramenta Search foi desenvolvida em 2023 e está pronta. É exemplificada abaixo a partir do Corpus Kadiwéu. As buscas são abertas, como mencionado acima, e é possível fazer buscas por palavras ou por etiquetas sintáticas, já que o corpus é anotado sintaticamente. Segue uma imagem de uma busca sintática:



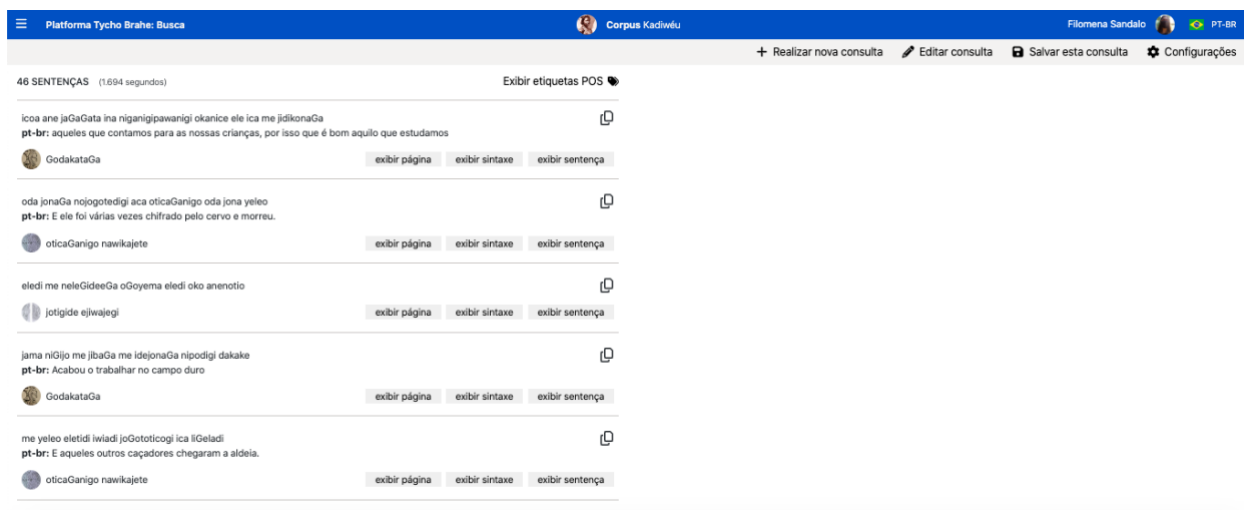
Pesquisa Simples
Pesquisa Sintática

escreva sua pesquisa

TAG
ADJ
iPrecedes
TAG
N
↑
↓
🗑️
+

Executar Pesquisa





O link para a realização de buscas segue abaixo se o parecerista quiser fazer buscas exploratórias. As etiquetas (tags) usadas serão apresentadas em seção mais adiante. As buscas simples são buscas por palavras da língua Kadiwéu.

<https://www.tycho.iel.unicamp.br/search>

### 3.1.3. Ferramenta para criação de parser sintático (Parser)

Neste primeiro ano de projeto, foram elaboradas as funcionalidades para criação, edição e testes de parsers sintáticos na ferramenta de Parser da Plataforma Tycho Brahe. O link desta ferramenta não é aberto, para evitar modificações na programação da sintaxe das línguas.

<https://www.tycho.iel.unicamp.br/parser/> (fechado)

A partir desta ferramenta, será possível criar um parser sintático de regras para as línguas indígenas do projeto, iniciando-se pelo Kadiwéu. Durante o primeiro ano de estudo, analisamos uma narrativa de 50 sentenças e elaboramos regras a serem implementadas para esta língua (veja seção 3.2).

Note que a finalização desta ferramenta não era objetivo deste primeiro ano de projeto. Esta ferramenta em total funcionamento é objeto do próximo relatório. Mas estamos tentando apenas demonstrar que já temos desenvolvimentos neste sentido.

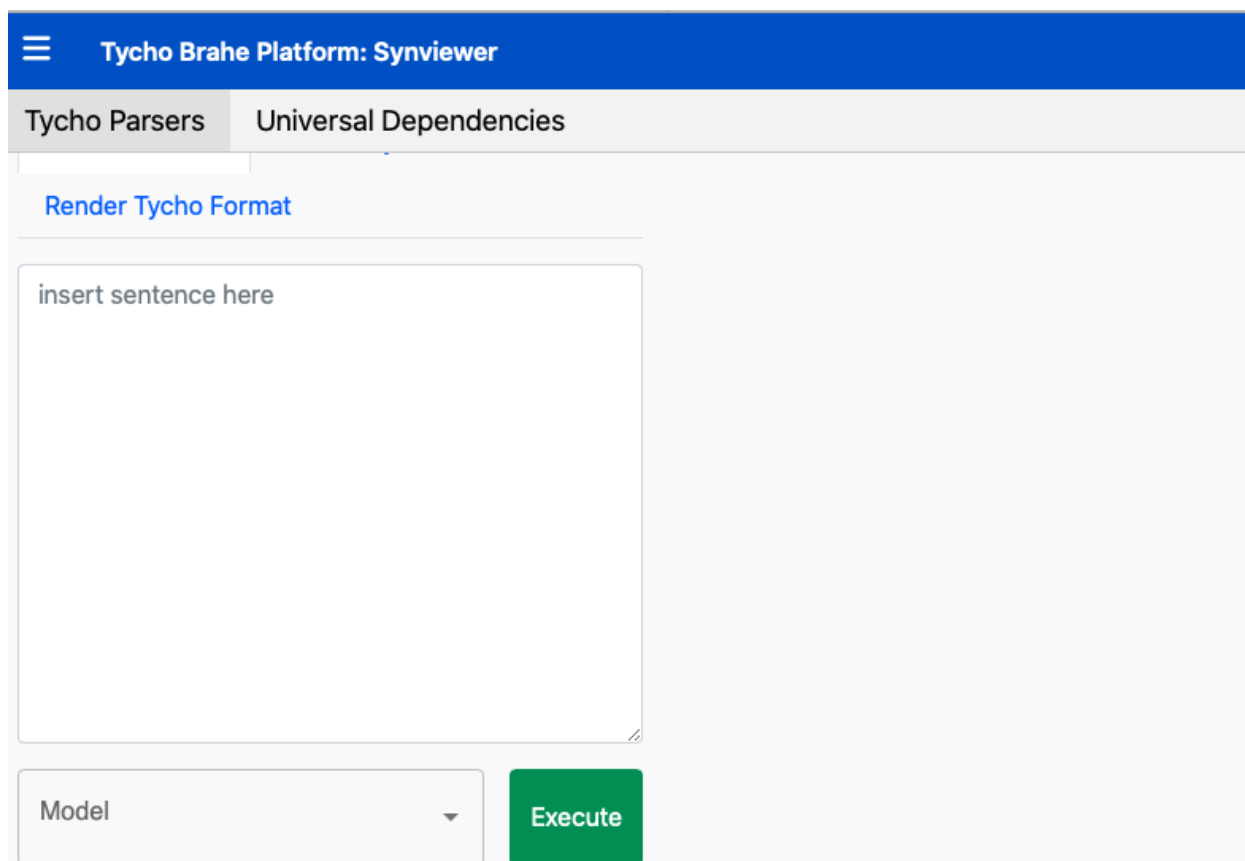
### 3.1.4. Visualizador de Árvore Sintáticas e Dependências Universais (Synviewer)

Foi realizada uma alteração na ferramenta Synviewer da Plataforma Tycho Brahe, que já existia a partir de projetos anteriores com o português, para visualização e parsing de sentenças no modelo de árvores sintáticas e no modelo de Dependências Universais (Universal Dependencies – UD). Além da visualização da estrutura gerada, com possibilidade de download de imagens das sentenças, ela

também possui um conversor de sentenças no formato CoNLL-U. O trabalho com este modelo será importante para as traduções automáticas e programação na linguagem Grammatical Framework.

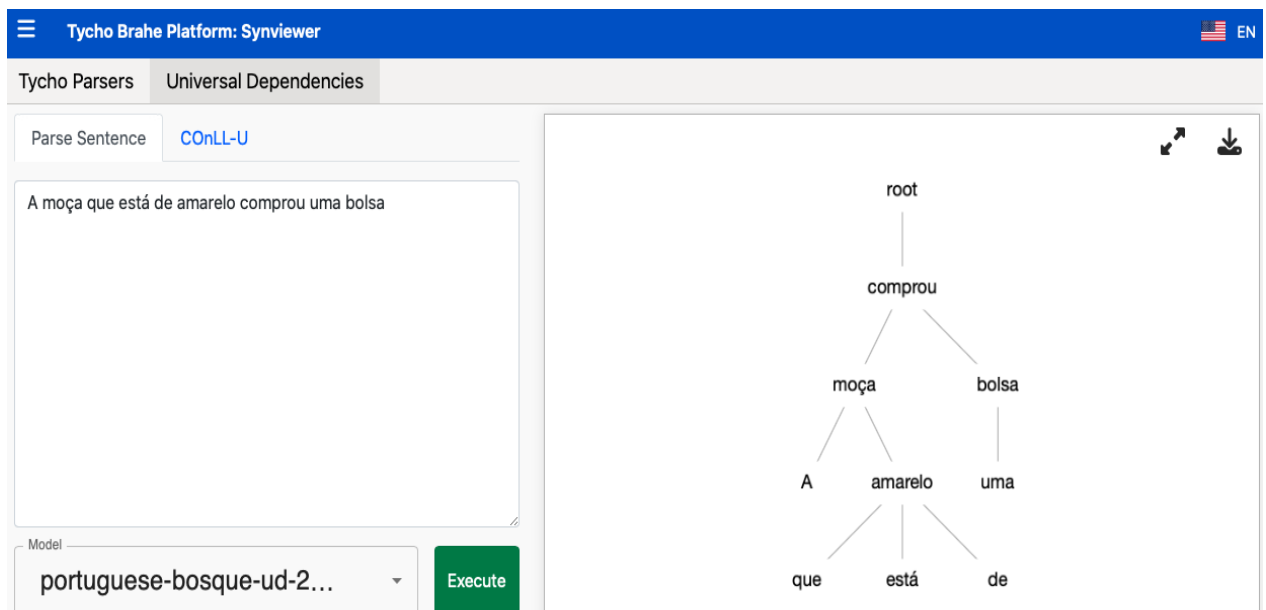
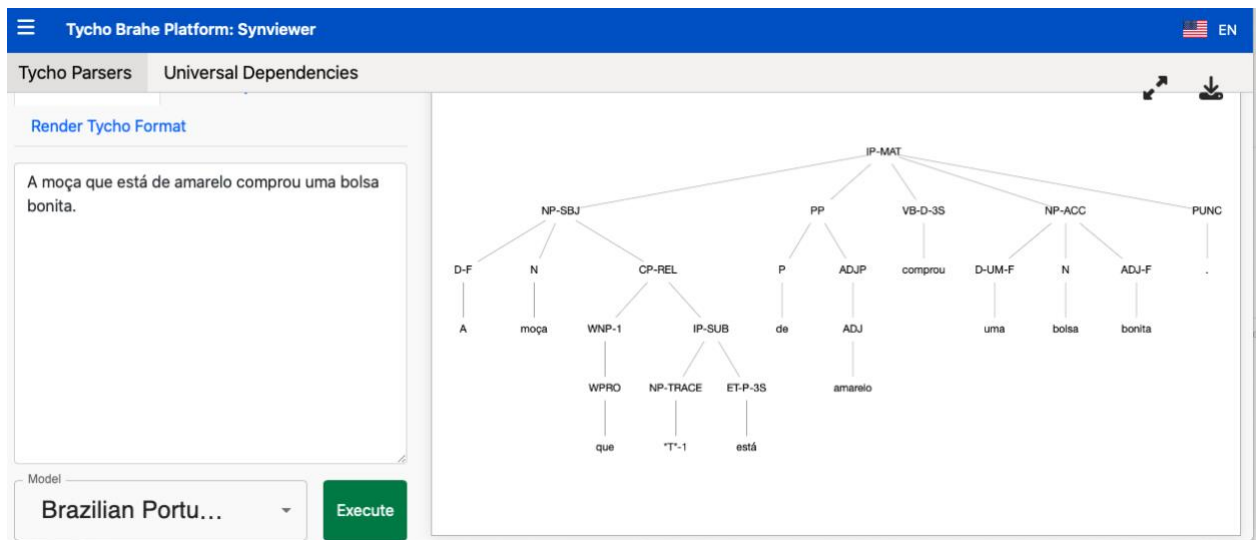
A ferramenta está disponível de forma aberta. E segue abaixo algumas imagens para exemplificar.

<https://www.tycho.iel.unicamp.br/synviewer>



Basta inserir uma sentença do corpus, e executar. Pode-se escolher a forma de árvores gerativas ou de dependências universais. Os testes podem ser feitos pelo parecerista usando português. As regras das línguas indígenas em estudo ainda estão sendo elaboradas ou alimentadas como mencionado acima. Segue abaixo um teste com português:<sup>2</sup>

<sup>2</sup> Na verdade, já é possível a análise de sentenças em nheengatu, pois Leonel Alencar já possui um modelo neural treinado com o UDPipe 1.2 (ver de Alencar 2024). No entanto, os materiais de nheengatu ainda não foram inseridos na Plataforma, como já mencionado.



### 3.1.5. Edictor

O Edictor é nossa ferramenta de edição. Não pode ser visitador sem senha, dado que os dados poderiam ser alterados. Mas seguem algumas imagens mostrando frases do Corpus Kadiwéu anotadas, bem como um vídeo.

Plataforma Tycho Brahe: eDíctor Corpus Kadiwéu Filomena Sandalo PT-BR

Documento: nigedioli Parse Adicionar nova Remover anterior próxima

oda onigotibecewaji yawaligitibiwaji ica miditaGa ica eledi Godoigi .

original	oda	onigotibecewaji						yawaligitibiwaji				ica		miditaGa				ica
etiqueta POS	ADV	VBAPL						VB				D		C+DAPL				D
gloss-br	Então	foram						visitar				alguém		que com ele				um
gloss	Then	go out						visit				someone		with whom				a
morfemas		o	ni	go	tibe	ce	waji	y	awali	tibi	waji	i	ca	me	i	di	taGa	i
etiqueta		Plu	Ant	v	Mot	Dir	Plu	Erg	v	Mot	Plu	Gnr	Ncl	c	Gnr	Ncl	vazio	Gnr
gloss-br		plural	ant	ir	intensivo	para fora	vazio	3	passar	intensivo	plural	masc	ausente	c	masc	deitado	com	masc
gloss		vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio

Áudio

Traduções

Portuguese  
E resolveram visitar uma outra aldeia nossa. ✓

English  
And they decided to visit another village of our own people ✓

Syntactic Rules ✓

Esta foi nossa primeira ferramenta desenvolvida, cujo desenvolvimento iniciou antes do presente projeto. No ano de 2023, corrigimos bugs meramente. Segue um vídeo do youtube de uma apresentação na Associação Brasileira de Linguística (ABRALIN) em 2021, que fala sobre esta ferramenta, mostra o processo de anotação e apresenta traços tipológicos da língua kadiwéu:

<https://youtu.be/Zp0rayUyHvU>

**Grande parte da anotação já pode ser feita automaticamente, mas grande parte é ainda manual aguardando a finalização do banco lexical/dicionários.**

### 3.1.6. Ferramenta de criação e gerenciamento de dicionários

Esta ferramenta é totalmente nova e por isso ainda não figura na página da plataforma. Não estava prevista no cronograma, mas sentimos a necessidade de dicionários para alimentar a tradução automática e para pesquisas nas línguas. Neste período, foi realizado um levantamento das necessidades para construção de uma ferramenta web para criação e gerenciamento de itens lexicais para dicionários. Esta ferramenta será integrada às outras ferramentas da Plataforma Tycho Brahe (<https://www.tycho.iel.unicamp.br>). A proposta desta ferramenta é possibilitar o desenvolvimento de dicionários de línguas indígenas de acordo com a “Proposta de dicionário nheengatu-português”, na qual consiste a tese de doutorado de Marcel Avila<sup>3</sup>, com a proposta da Gramática do kadiwéu da professora Filomena Sandalo<sup>4</sup> e com o *Dicionário da Língua Kadiwéu* de Glyn Griffiths<sup>5</sup>.

3 AVILA M. T., 2021.

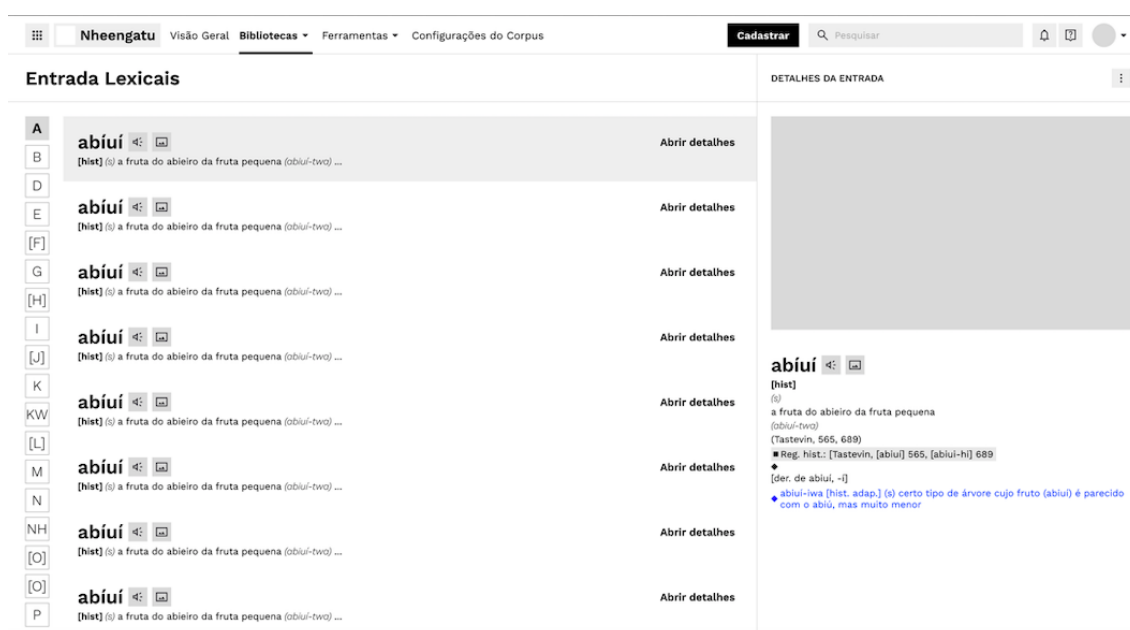
4 SANDALO, M. F., 1995.

5 GRIFFITHS, Glyn, 2002.

No link abaixo, encontram-se as descrições das definições sobre cada item que integrará o sistema e a explicação de cada tela de cadastro dos elementos do dicionário. Estas informações serão utilizadas para criação das telas do sistema. Não há senha, a entrada é livre. Pedimos desculpas pelo erro no relatório anterior que pediu senha de entrada.

<https://muddy-trapezoid-37f.notion.site/Tycho-Dicion-rio-cfe82e89e3224a2ca77fe73ec596dc19>

Abaixo seguem algumas imagens “wareframes” do dicionário Nheengatu em elaboração. O corpus de narrativas do kadiwéu está mais avançado. Para o nheengatu, o dicionário está mais avançado, pois esta língua conta com mais materiais publicados. Os dados do kadiwéu são todos buscados em campo pela coordenadora do projeto e pela aluna de doutorado Vanda Pires.



Nheengatu

Visão Geral

Bibliotecas

Ferramentas

Configurações do Corpus

CADASTRAR

Pesquisar

← VOLTAR

Cadastrar Entrada Lexical

ADICIONAR ATRIBUTO +

PRÉ-VISUALIZAR

Entrada lexical

Digite o verbete

Tipo

☐ ENT. LEXICAL
 ☐ LEMA
 ☐ ADENDO

Prefixos de relação

Digite os prefixos separados por vírgula, se houver

Marcas de uso

SELECIONAR

Entradas relacionadas

Adicionar entradas parentes ou variantes relacionadas, se houver

ADICIONAR RELAÇÃO

Acepção

É necessário cadastrar pelo menos uma acepção

1

CAT. GRAMAT.

Digite a descrição da acepção

ADICIONAR ATRIBUTO +

Exemplo

Descrição do exemplo

Reg. Histórico

Descrição do registro

a)

Digite a descrição da subacepção

+

Exemplo

Descrição do exemplo

Ref. Bibliográfica

Descrição da referencia

2

CAT. GRAMAT.

Digite a descrição da acepção

ADICIONAR ATRIBUTO +

3

CAT. GRAMAT.

Digite a descrição da acepção

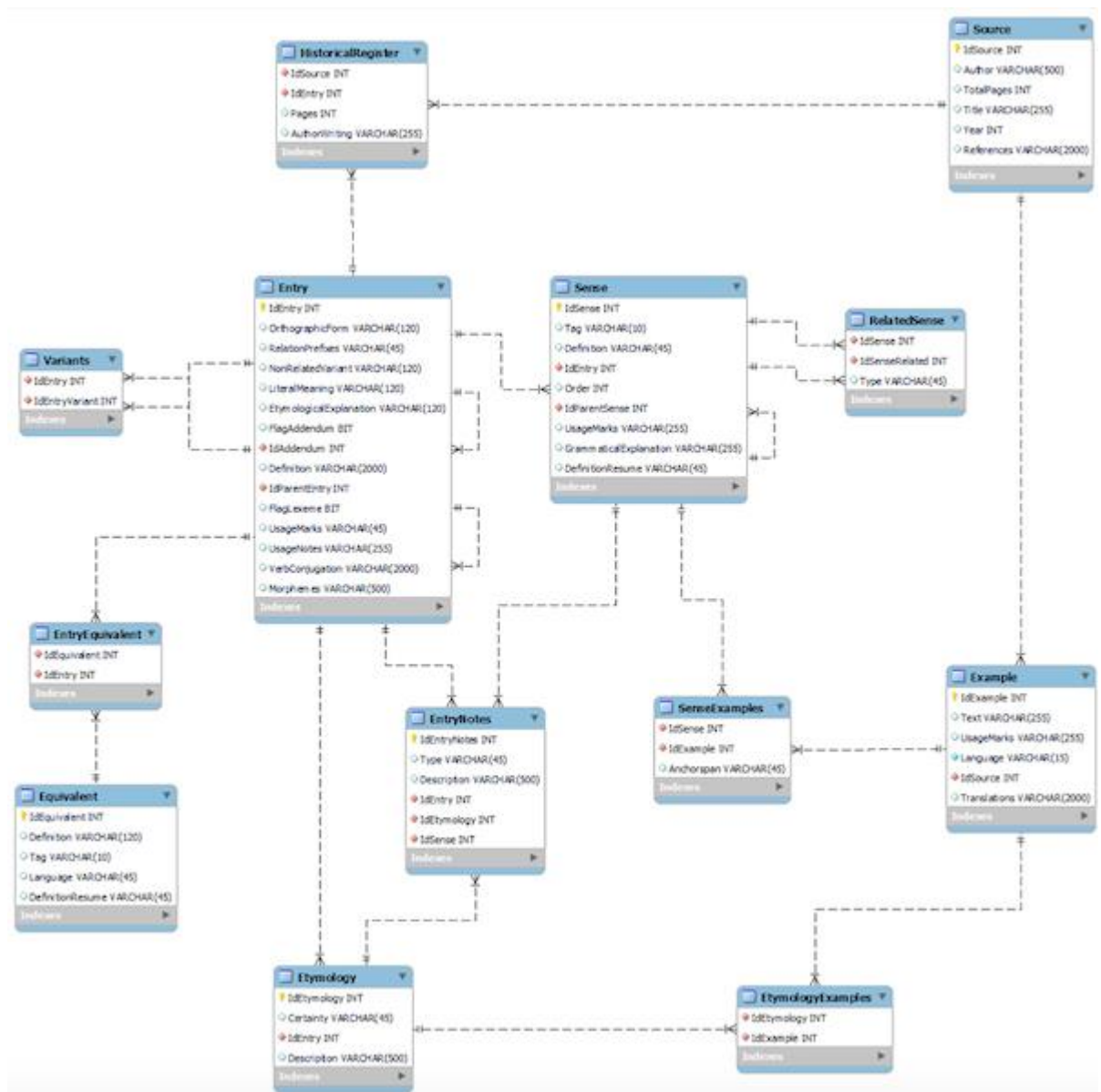
ADICIONAR ATRIBUTO +

NOVA ACEPÇÃO

Morfemas

Original	Digite aqui o verbete original		
POS Tag	Selecione a tag		*
Gloss-BR	Selecione uma opção		*
Gloss	Selecione uma opção		*
Morphemes	Digite o morfema	Digite o morfema	+
Tag	Tag	Tag	
Gloss-BR	Gloss-br	Gloss-br	
Gloss	Gloss	Gloss	

Utilizando as informações levantadas acima e aproveitando parte da modelagem realizada para o *nheengatu*, foi desenvolvida a modelagem de dados necessária para a criação do banco de dados responsável pelo armazenamento dos dados do dicionário que, posteriormente, será necessária para a automatização das traduções entre línguas.



### 3.1.7. Ferramenta para importação de dados (IO)

Para a importação dos dados de dicionário já existentes, foi finalizada e disponibilizada uma ferramenta na Plataforma Tycho Brahe que permite a extração do conteúdo de PDFs através de técnicas de reconhecimento de caracteres (OCR) e interfaces para transcrição e edição. Esta ferramenta pode ser acessada pelos membros do projeto e busca acelerar a importação das informações publicadas em pdf para os dicionários em elaboração.

Plataforma Tycho Brahe: IO			
Corpus Nheengatu			
Filomena Sandalo PT-BR			
+ Criar novo			
input file	type	status	date
nheengatu.lift	LIFT	PROCESSED	2023-06-22 00:45:36
Exibindo 10 resultados Total: 1			
< 1 >			

## 3.2. O Sistema de Anotação de Dados

Esta seção do projeto foi desenvolvida por Filomena Sandalo e Charlotte Galves.

A primeira língua indígena na *Plataforma Tycho Brahe* é o kadiwéu, uma língua da família Guaikurú falada no Mato Grosso do Sul, cujos falantes atualmente não passam de 500 pessoas. O material já anotado do kadiwéu mostra nosso engajamento e força para alcançar corpora digitais maiores de línguas que estão rapidamente enfraquecendo em seus usos e transmissão para as próximas gerações. Temos ainda um número significativo de narrativas nheengatu, mas ainda não foram inseridas na Plataforma.

Para o ano de 2023, propusemos terminar a proposta de **anotação gramatical** do kadiwéu no corpus Tycho Brahe. E cumprimos.

A anotação gramatical do kadiwéu, língua com pouco tradição gramatical, nos confronta à questão da universalidade das categorias linguísticas, tanto a nível das palavras quanto a nível das orações. Chamamos de problema de Anchieta a questão de saber até que ponto é possível encaixar línguas desse tipo dentro dos esquemas de anotação funcionando para línguas indo-europeias como o português e o inglês. Anchieta, bem como os jesuítas que escreviam gramáticas para as línguas do novo mundo, usavam, por exemplo, o sistema de casos do latim, ou os conceitos de tempos, modos, e aspectos de línguas como o português para organizar os paradigmas nominais e verbais encontrados nelas. Da mesma maneira, procuramos usar noções como orações completivas, relativas, clivadas, relevantes para o português e o inglês, para expressar distinções gramaticais entre tipos de orações do kadiwéu. Em certos casos, essas categorias se mostram eficientes na descrição, em outros, elas se revelam de pouca serventia para estabelecer as distinções necessárias entre diferentes construções. Enquanto gerativistas, acreditamos, contudo, que as línguas são todas o produto da mesma máquina cognitiva, a faculdade de linguagem, e que, no nível sintático, as diferenças não devem impedir o uso de categorias de descrição as mais semelhantes, ou comparáveis, possível.

Escrevemos e publicamos um artigo que apresenta nossa proposta de anotação para o kadiwéu em todos os níveis propostos (morfológico, POS (categorias sintáticas) e sintático), cuja referência aparece abaixo, podendo ser visitado em (link aberto da revista Caderno de Estudos Linguísticos):<sup>6</sup>

<https://periodicos.sbu.unicamp.br/ojs/index.php/cel/article/view/8673592>

Segue sua referência bibliográfica:

Sandalo, Filomena & Galves, Charlotte. 2023. ANOTANDO SINTATICAMENTE UMA LÍNGUA ORIGINÁRIA DO BRASIL: O PROBLEMA DE ANCHIETA. Caderno de Estudos Linguísticos, Campinas, v.65, p. 1-27.

<sup>6</sup> A produção em publicações deste projeto seguirá também anexada.



O artigo está organizado da seguinte maneira. Na seção 2, apresentamos nossa proposta de anotação para o kadiwéu, começando pela anotação de palavras (2.1), e propondo em seguida um nível suplementar de anotação morfêmica devido à natureza polissintética da língua, que a diferencia fortemente de línguas como o português e o inglês (2.2). Este nível é o lugar em que a especificidade do kadiwéu está maximamente representada, acarretando etiquetas totalmente diferentes daquelas usadas até agora em corpora sintaticamente anotados no modelo dos *Penn Parsed Corpora of Historical English* (Kroch et al. 2000; Kroch et al. 2004; Kroch, et al. 2016) como o *Corpus anotado do português histórico Tycho Brahe*. Finalmente, em 2.3, propomos uma primeira versão do sistema de anotação sintática, que se inspira fortemente desse modelo. A seção 3 apresenta casos que desafiam a mera transposição de categorias emprestadas de línguas europeias.

As anotações de POS e morfológicas respectivamente seguem abaixo. O corpus é anotado seguindo essa anotação e as anotações alimentam a anotação sintática, que será apresentada mais adiante.

POS TAGS	Morpheme Tags		Examples	
VB	Plu	plural	o-y-a:lGe Plu-Erg-v	‘They kidnap him.’
	Imp	impersonal	eti-Ga-d:-d:egi Imp-Abs-Inv-v	‘Someone brought you.’
	Erg	ergative agreement	j-awi: Erg-v	‘I hunt it.’
	Abs	absolutive agreement	i-d:-abi-d Abs-Inv-v-Asp	‘I’m standing up.’
	Inv	inverse voice	Go-d:-ili: Abs-Inv-v	‘We grow.’
	Ant	antipassive	n-ema-ta Ant-v-Obl	‘She/he loves him/her in distance.’
	Hit	hither	n-ad:e:gi Hit-v	‘He brings it.’
	v	verbal root		
	Val	valence change morpheme	j-otaGan-Gen:- aGa Erg-v-Val-Plu	‘We talk to him.’
	Asp	aspect	o-y-aqage-di Plu-Erg-v-Asp	‘They cut it.’
	Obl	oblique argument agreement	me-ta v-Obl	‘He says to him.’
	Dir	directional morpheme	ji-l:o-ko-tigi Erg-v-Val-Dir	‘I look up at something.’
	Mot	motion	ji-n-otiqa-tijo Erg-Hit-v-Mot	‘I come wistling.’
	Apl	aplicative	j-ao-tGa-domi Erg-v-Obl-Apl	‘I make it for you.’

POS TAGS	Morpheme Tags		Examples	
N	Gen	genitive agreement	l-okaGe-te-di Gen-n-Cla-Plu	‘his friends’
	Ant	antipassive	n-gato-je Ant-n-Cla	‘a bullet’
	n	root	dom:o:jya n	‘car’
	Cla	classifier	apaqa-co-di n-Cla-Plu	‘rheas’
	Der	derivation	n-dele-Gikajo Ant-v-Der	‘warrior’
	Dim	diminutive	l-atope-nig:i Gen-n-Dim	‘his gun’
	Plu	number	Gonel:egi-wa-tedi n-Cla-Plu	‘groups of man’
D	Anf	anaphoric	nG-i-jo nG-Gnr-Ncl	This/the/ An one_ mentioned before
	Gnr	gender	i-di Gen-Ncl	This/the/An one
	Ncl	numeral classifier	i-di Gen-Ncl	This/the/An one
	Plu	number	i-di-wa Gnr-Ncl-Plu	These/the ones
NUM	Num	numeral	i-ni-wa-ta:le Gnr-Ncl-Plu-Num	two
Q	Qnt	quantifier	oni-ni-te-k-beke Num-Gnr-Ncl-Obl-Apl-Qnt	each
WPRO	Int	interrogative pro-noun	am-i:-na Int-Gnr-Ncl	‘who’
WADV	Whs	Wh-support	ig-ame Whs-Int	‘why’
PRO	Pro	pronoun	aqa:m:-i Pro-Plu	‘you’

Segue novamente uma sentença do corpus com esta anotação para avaliação. O vídeo apresentado acima também é informação relevante neste sentido, pois mostra o processo de anotação.

naigitece jonoGonadi ica apakanigo .

original	naigitece		jonoGonadi			ica		apakanigo		
etiqueta POS	NAPL		T+VB			D		N		
gloss-br	Pelo caminho		viram			uma		ema		
gloss	along the way		they have seen			a		rhea		
morfemas	naigi	tece	janaG	o	nadi	i	ca	apakani	go	
etiqueta	n	Dir	t	Plu	v	Gnr	Ncl	n	Cla	
gloss-br	caminho	para fora	passado	3plu	ver	masc	ausente	ema	cla	
gloss	way	outward	past	3pl	see	masc	absent	rhea	cl	

**Áudio**



**Traduções**

Portuguese

No caminho viram uma ema.

English

Along the way, they have seen a rhea

Syntactic Rules

A anotação sintática do kadiwéu ainda está em fase de construção e por isso não está integralmente implementada na plataforma, como já foi mencionado. O que apresentamos aqui é o sistema que foi elaborado a partir da anotação manual de 50 frases de um dos textos que compõem o corpus do kadiwéu, *nigedioli* “a mulher onça”. Esta anotação, além de constituir a base do manual que será disponibilizado para os usuários do corpus, servirá também de base para o anotador automático (*parser*) baseado em regras, construído no modelo do parser para o português (Magro 2017)<sup>7</sup>, que faz uso da função de revisão da linguagem de busca *Corpus Search*,<sup>8</sup> para construir as árvores sintáticas associadas às orações (cf. Faria et al. 2023).

A publicação do pesquisador colaborador Pablo Faria, da pesquisadora principal Charlotte Galves, e da colaboradora Catarina Magro, citada acima, embora sobre o português, é, portanto, também crucial ao nosso projeto e merece ser apontada como resultado de nossa pesquisa:

FARIA, P., GALVES, C., MAGRO, C. (2023) Syntactic annotation for Portuguese corpora: standards, parsers, and search interfaces. *Language Resources and Evaluation, Especial Issue on Computational Approaches to Portuguese*.

A anotação sintática do kadiwéu toma como ponto de partida o sistema de anotação inicialmente elaborado para anotar o inglês histórico, e adaptado mais tarde ao português. Contudo, modificações envolvendo tanto supressões como acréscimos tiveram que ser efetuadas. Por exemplo, a ausência de Preposições na língua prescinde a categoria PP. A nível das orações, uma vez que a finitude não é expressa, as categorias correspondendo a orações infinitivas, gerundivas e participiais também não são necessárias. Dois fatos aliás sugerem fortemente que a oposição finito/infinitivo não existe em Kadiwéu. O primeiro é que não há distinção morfológica suportando a oposição finito/não finito. O segundo é que todas as construções de subordinação têm complementador, menos no caso

<sup>7</sup> A partir de uma ideia inicialmente implementada por Beatrice Santorini para o francês.

<sup>8</sup> Cf. <https://corpussearch.sourceforge.net/>

dos verbos seriais. A questão que não pode deixar de ser colocada é se o projeto como um todo não deve esbarrar na grande diferença tipológica entre kadiwéu por um lado e inglês e português por outro lado. Mas, observamos que, numa primeira abordagem, pelo contrário, são múltiplas as semelhanças sintáticas, e que os sistemas já elaborados para outras línguas são perfeitamente utilizáveis

Segue abaixo um exemplo de nossas análises de algumas frases do corpus e anotações sintáticas elaboradas especificamente para Kadiwéu. Deste material, criamos as regras que alimentam o Parser.

### **Frase 1**

- IP-MAT com tempo vazio e cópula vazia
- CP-REL com me

IP-MAT > T\* COP\* ADVP NP

NP > D N CP-REL

CP-REL > WPRO IP-SUB

IP-SUB > VBAPL NP-PRD

NP-PRD > CP-REL

CP-REL > C (me) IP-SUB

IP-SUB > NP-PRD

### **Frase 2**

IP-MAT > EV COP\* NP NP NP

NP > Q D N

NP > Q D N

NP > Q D N\$ NP

NP > D N\$

### **Frase 3**

IP-MAT > ADVP T\* VBAPL VB NP-OB2

NP-OB2 > D CP-REL

CP-REL > C (me) +DAPL IP-SUB

IP-SUB > COP\* NP-SBJ

NP-SBJ > D ADJ N\$

### **Frase 4**

IP-MAT > PP T\* VB NP-OB1

NP-LOC > NAPL

NP-OB1 > D N

### **Frase 5**

IP-MAT > EV T\* VB NP-SBJ

NP-SBJ > D N

CP-FRL > C (me) +DAPL IP-SUB

IP-SUB > COP\* NP-SBJ

NP-SBJ > D-N

### **Frase 6**

IP-MAT > EV T\* NP-SBJ (exp) COP\* NP

### **Frase 7**

IP-MAT > ADVP CP-ADV NP-SBJ VBAPL NP-APL CP-QUE-SPE

NP-SBJ > D N

NP-APL > D N

CP-ADV > C@ IP-SUB

IP-SUB > @T EV VB NP-SBJ

CP-QUE-SPE > IP-IND

IP-IND > T\* NP-SBJ \*pro\* VB CP-THT

CP-THT > C (me) IP-SUB

IP-SUB > IP-SUB-1 CONJP

CONJP > IP-SUB-1

IP-SUB-1 > AUX VB NP-OB1

NP-OB1 > N\$

IP-SUB-1 > VB

### **Frase 8**

IP-MAT > EV NP-SBJ EV VB

NP-SBJ > D N

### **Frase 9**

CP-QUE > EV WADVP C IP-IND

IP-IND > ADVP \*T\* NP-SBJ COP \*

NP-SBJ > N\$

### **Frase 10 (9)**

a.

FRAG > ADVP EV NP IP-MAT-SPE

IP-MAT-SPE > INTJ NP-SBJ AUX VB NP-ACC

b.

IP-MAT > NP-SBJ VBAPL NP-APL NP-OB1

NP-OB1 > CP-FRL

CP-FRL > WNP IP-SUB

WNP > WPRO

IP-SUB > NP-ACC \*T\* NP-SBJ

NP-SBJ > N\$

c.

IP-IMP > CONJ CP-ADV<sup>1</sup> NEG VB NP-OB1 CP-ADV<sup>2</sup>

CP-ADV<sup>1</sup> > CT IP-SUB

IP-SUB > T \* NP-SBJ \*pro\* VBAPL NP-APL NP-OB1

CP-ADV<sup>2</sup> > CT IP-SUB

IP-SUB > T\* NP-SBJ VB NP-OB1 \*pro\*

### **Frase 11**

a

P-MAT> ADVP EV@ @T@ @VB NP-SBJ

b

IP-MAT > NP-SBJ \*pro\* VBAPL ADVP

ADVP > CP-FRL

CP-FRL > C D APL IP-SUB

IP-SUB > PP\* COP\* NP-SBJ

### **Frase 12**

IP-MAT > ADVP CP-ADV<sup>1</sup> EV NP-SBJ \*pro\* VBAPL ADJP-SPR CP-ADV<sup>2</sup>

CP-ADV<sup>1</sup> > CT@ IP-SUB

IP-SUB > T\* NP-SBJ \*pro\* @VBAPL NP-APL \*pro\*

CP-PUR> C IP-SUB

IP-SUB > NP-SBJ \*pro\* VB

### **Frase 13**

IP-MAT > NP-SBJ \*pro\* CP-ADV T\* ADVP @VB

CP-ADV > CT IP-SUB

IP-SUB> NP-SBJ \*pro\* T\* VB

ADVP > ADV@

#### **Frase 14**

IP-MAT > NP-SBJ \*pro\* CP-ADV T\* ADVP @VB

CP-ADV > CT IP-SUB

IP-SUB> NP-SBJ \*pro\* T\* VB

ADVP > ADV@

#### **Frase 15**

IP-MAT > NP-SBJ-1 \*exp\* CP-ADV T\* NEG@ ADVP COP\* NP-1 CP-CLF

CP-ADV > CT NUMP C IP-SUB

IP-SUB> NP-SBJ \*pro\* T\* VB

ADVP > @ADV

CP-CLF > C IP-SUB

IP-SUB > NP-SBJ \*T\* VB

#### **Frase 16**

IP-MAT > **CP-ADV** EV NP-SBJ \*pro\* VB NP-OB1

**CP-ADV** > CT NP-LFD C **IP-SUB**

**IP-SUB** > NP-SBJ \*ICH\* T\* VBAPL

NP-OB1> NP1 CONJP CONJP

CONJP > NP2

CONJP > NP3

NP1 > N\$

NP2 > D N\$

NP3 > N\$

#### **Frase 17**

IP-MAT > IP-MAT-1 CONJP

CONJP > IP-MAT=1

IP-MAT-1 > **ADVP** NP-SBJ \*exp\* COP\* T@ NP-PRD

NP-PRD > N

ADVP> NUM

IP-MAT=1 > NEG@ ADVP NP-PRD

ADVP > @ADV@

NP-PRED > @PRO



### **Frase 18**

IP-MAT > ADVP NP-SBJ T\* VBAPL NP-APL

NP-SBJ > D N

NP-APL > D N

### **Frase 19.1**

IP-MAT > [ CP-ADV CP-ADV1 CP-ADV2] ADVP NP-SBJ \*pro\* T\* VBAPL NP-APL

CP-ADV1 > CT IP-SUB1

IP-SUB1 > T\* VB-APL NP-SBJ NP-APL

CP-ADV2 > CT IP-SUB2

IP-SUB2 > NP-SBJ \*pro\* VB-APL NP-APL \*pro\*

### **Frase 19.2**

IP-MAT > NP-SBJ \*pro\* T VBAPL NP-APL IP-ADV

IP-ADV > VB NP-OB1

NP-OB1 > N\$

### **Frase 20**

IP-MAT > ADV NP-SBJ T V CP-ADV

CP-ADV > CT IP-SUB

IP-SUB > \*pro\* VB-APL NP-APL NP-OB1

*cupim*

### **Frase 21.1**

IP-MAT > ADV CP-ADV EV NP-SBJ \*pro\* VB-APL NP-APL \*pro\*

CP-ADV: CT NP-SBJ \*pro\* VB

*Quando voltou*

### **Frase 21.2**

IP-MAT > ADV NP-SBJ \*pro\* T+EV VB-APL VB NP-OB2

NP-OB2 > N\$

*alisou seu cabelo*

### **Frase 22.1**

CP-D > ADV ADVP NP-LFD CNEG@ @EV IP-IND

IP-IND > NP-SBJ \*ICH\* VB-APL NP-APL

NP-APL > D N

*a ema*

### **Frase 22.2**

IP-MAT > ADV NP-SBJ \*pro\* T@ @VB-APL NP-APL \*pro\*

*Foi até ela*

### **Frase 22.3**

IP-MAT > CONJ T NP-EXP-1 COP\* NP-1 IP-ADV

NP-1 > NP CONJP

NP > PRO CP\*

CP\*: C (*me*) IP-SUB

IP-SUB > NP

CONJP > CONJ NP

NP > NEG PRO VB

IP-ADV > VB-APL NP-APL \*pro\*

*Era como mulher mas não ela falar sendo diferente*

### **Frase 23**

IP-MAT > T EV COP\* ADVP CP-ADV

CP-ADV > CT@ @EV NP-SBJ \*pro\* VB

*Quando ela falou*

### **Frase 24**

IP-MAT > EV@ @VB-APL NP-APL \*pro\* CP-D-SPE

CP-D-SPE > PRO C IP CP-D

CP-D > PRO C IP

*Ela que é ela nossa carne*

## **Anotação Sintática**

### **1. Noun Phrases:**

*NP-SBJ*

Subject NPs

*NP-OBI* –

direct objects of single verbs

*NP-OB2* –

Applicative objects of serial verbs

OBS: there is no IP-INF projected in this case

Ex. Frase III

*NP-APL* – applicative objects

*NP-LOC* – locative verbs objects

*NP-PRD* – predicative NPs

## **2. Clauses:**

*IP-MAT*

Matrix clauses

Double evidential (cf. double complementizer structure)

*IP-SUB*

Subordinate clauses dominated by CP

*IP-IND*

IPs dominated by root CPs

*IP-ADV*

Adjoined bare IP (cf. jovem Exabigo)

naGa didele icoa liwoneGa ikaqe lodaajo naGa d-idele i-ca-wa liwoneGa i-baqen l-odaajo CT INV-combater CL jovem 3subj-use 3poss-faca

“Foi quando os combateu o jovem rapaz usando sua faca”

Cf. also Mulher onça sentence 20.3

Sentença 40b

*IP-IMP* (cf. 41c)

## **3. CP**

*CP-THT*

$C = Me$

- Double complementizer structure: *naGa* Top *me*

- Top *me* (em orações controladas. Cf. 27, com conseguir)

CP-PUR Purpose clause – só com *me* (corresponde a uma oração infinitiva)

(*me* equivalente do *to*)

*CP-REL*

Relative clauses with antecedent

-*Ane* cf. 41, 42b

-*Me*

*CP-FRL*

Relative clauses without antecedent

CP-CLF

$C = Me$

CP\* com *me*

cf frase 22.3 (*ela era como mulher*)

*CP-QUE*

*Me* can occur after the WH-phrase (cf. 27)

*CP-D*

Top *Me* cf. *Mulher onça* 22h

*CP-ADV*

Cf. *Mulher Onça* sentence 21 (20?)

Sentença 15 e 16: *NaGa* – que parece significar a conjunção “quando”

CP-ADV tem recursão (cf. sentenças 15 e 16). E a recursão é com *Me*

(obs. *Me* pode indicar topicalização sem que haja recursão)

>> CP-TMP?

CP-CON (cf. 44) = conditional CP “ade”

Obs. IP-INF não existe

1. porque não tem distinção morfológica
2. porque todas as construções de subordinação têm complementador (isso exclui a serialização)

#### **4. ADVP**

Adverbial phrases

#### **5. FRAG**

Two or more constituents that do not form a clause

#### **6. Categorias vazias**

COP \* (cópula nula) em clivada, cf. frase 16. Em apresentativa , cf. frase 1.

EXT \* (Existencial nulo)

VB\* (verbo leve nulo)

T\* (quando tem CT em C)

**\* pro \***

**\* exp \***

**\* ICH \***

**\* T \***

#### **6. NEG**

**a(k)**

*Etiquetas de palavras*

NDIR

#### **Categorias funcionais**

CT

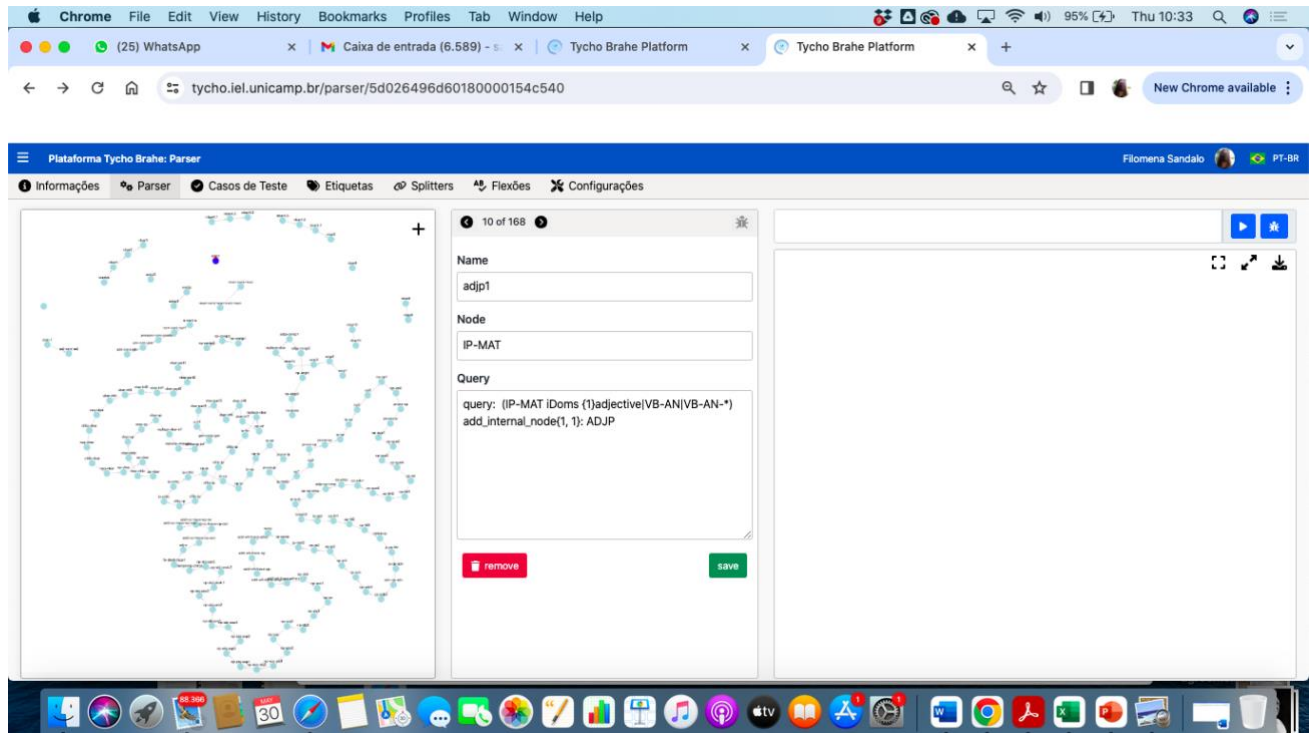
CNEG

CNEG+EV

T

T+EV

Já temos 168 regras implementadas e segue uma imagem para ilustração:



Como já mencionado, o corpus kadiwéu conta atualmente com 5369 palavras anotadas ou parcialmente anotadas com nosso Sistema morfológico e de POS, que podem ser visitadas aqui:

<https://www.tycho.iel.unicamp.br/viewer/C12>

E mais 2815 palavras ainda não anotadas e que podem ser visitadas aqui:

<https://www.tycho.iel.unicamp.br/viewer/6f27da65-4d9a-4823-8d47-1a97c0955db1>

As narrativas ainda não anotadas foram coletadas recentemente pela nossa aluna colaboradora Vanda Pires, falante nativa do kadiwéu. A aluna coletou as narrativas em campo em **Mato Grosso do Sul**.

A figura abaixo apresenta uma árvore da frase traduzida como “Diz que o kadiwéu não come raspa queimada” que foi criada automaticamente pela Plataforma:

A anotação de POS do nheengatu, bem como sua inserção na Plataforma Tycho Brahe está sendo trabalhada pela bolsista TT3 Juliana Lopes Gurgel, cuja bolsa iniciou em Janeiro de 2024, portanto ainda não temos um relatório.

Resta ainda informar que em dezembro de 2023, visitei a Charles University (Institute of Formal and Applied Linguistics), em Praga (república Tcheca) e conversei com os seguintes professores de linguística computacional: Daniel Zeman e Magda Sevcikova. Fizemos um projeto conjunto na intenção de elaborar estruturas de dependências morfológicas no futuro e estes contatos estão sendo valiosos. Seguem as páginas dos pesquisadores mencionados:

<https://ufal.mff.cuni.cz/magda-sevcikova>

<https://ufal.mff.cuni.cz/daniel-zeman>

E a carta convite para o projeto:

I hope this email finds you well.

I am writing to ask you for a favour regarding a research grant proposal that we are about to submit to the Czech Science Foundation. Would you please be so kind as to consider supporting our project with a short letter of intent to collaborate?

The proposal we are working on is a five-year linguistically oriented basic-research project entitled "Metamorphosis - Morpheme-centric Multilingual Modeling of Language Variability", involving five researchers from our Institute (Zdeněk Žabokrtský as Principle Investigator, David Mareček, Anna Nedoluzhko, Tomasz Limisiewicz, and me) and a couple of PhD and postdoc positions.

I am attaching a short summary of the project proposal to this email.

The point is that each accepted project will have to lead to a submission of an European Research Council project later, and cooperation with scientists abroad is thus an important evaluation criterion in this call. Given your top expertise in linguistics, if you could see a collaboration potential between your group and our project team, a short confirmation from you would substantially increase our chances for funding.

Here is some basic information about the project:

- Principal Investigator: Doc. Ing. Zdenek Zabokrtsky, Ph.D. (Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague)
- Project title: Metamorphosis: Morpheme-Centric Multilingual Modeling of Language Variability
- Submitted to: Czech Science Foundation

Please let me know if you find it useful to learn more about the proposed research first, I will send you a more detailed version of the proposal by the start of next week.

Thank you very much for your consideration and I apologize for taking up your time.

Kind regards,

Magda

### **3.3. Tradução Automática**

A tradução automática ainda não é um objetivo inicial do projeto, mas algumas atividades já foram iniciadas e são relatadas aqui.

As atividades aqui desenvolvidas têm como pesquisador principal o colaborador Leonel Figueiredo de Alencar e se agrupam em duas categorias principais.

Primeiramente, destaco as atividades de pesquisa sobre o processamento computacional da Língua Geral Amazônica (nheengatu) no âmbito de projetos de iniciação científica da Universidade Federal do Ceará, contemplados com Bolsa da Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico (FUNCAP). Foram incluídos em 2023 cerca de 1000 novas sentenças, totalizando mais de 10 mil tokens, no UD\_Nheengatu-CompLin, o banco de árvores (*treebank*) do *nheengatu* que desde novembro de 2022 faz parte da coleção Dependências Universais (<https://universaldependencies.org/>) (de Alencar, 2024). No momento, esse banco de árvores é o maior dos cerca de 15 *treebanks* de línguas indígenas das três Américas. Paralelamente a isso, o *Yauti*, o analisador morfossintático automático do *nheengatu*, que teve seu desenvolvimento iniciado em 2022, constrói representações arbóreas dependências de sentenças em *nheengatu*, incluindo lematização e informações morfológicas (de Alencar, 2023), conforme as Figuras 1-3. Observe-se que o *Yauti*, a



partir do exemplo com a ambiguidade de etiqueta de classe de palavra resolvida, produz análise com apenas um erro, que consiste na em tratar como objeto direto o sujeito pós-verbal (a ferramenta ainda não contempla a análise de verbos inacusativos).

```
>>> import Yauti
>>> s='''Ape, paá, usika sesé wirawasú. (p. 66, No. 2) Então o gavião passou perto dele - Ape, paá, usika sesé
-wasú.'''
>>> Yauti.parseExample(s, 'Casasnovas2006', 1, 3, 3, annotator='Juliana Lopes Gurgel')
# sent_id = Casasnovas2006:1:3:3
# text = Ape, paá, usika sesé wirawasú.
# text_eng = Then the hawk passed by him
# text_por = Então o gavião passou perto dele
# text_source = p. 66, No. 2
# text_orig = Ape, paá, usika sesé wirá-wasú.
# text_annotator = Juliana Lopes Gurgel
1      Ape      ape      ADV      ADVDI      AdvType=Loc|Deixis=Remt|PronType=Dem      5      advmod      _      SpaceAfter=No|TokenRange=0:3
No|TokenRange=0:3
1      Ape      ape      ADV      ADVJ      AdvType=Cau      5      advmod      _      SpaceAfter=No|TokenRange=0:3
1      Ape      ape      ADV      ADVT      AdvType=Tim      5      advmod      _      SpaceAfter=No|TokenRange=0:3
2      ,      ,      PUNCT      PUNCT      5      punct      _      TokenRange=3:4
3      paá      paá      PART      RPRT      Evident=Nfh|PartType=Mod      5      advmod      _      SpaceAfter=No|TokenRange=5:8
Range=5:8
4      ,      ,      PUNCT      PUNCT      5      punct      _      TokenRange=8:9
5      usika      sika      VERB      V      Person=3|VerbForm=Fin      0      root      _      TokenRange=10:15
6      sesé      resé      ADP      ADP      AdpType=Post|Number[grnd]=Sing|Person[grnd]=3|Rel=NCont      5      obl
TokenRange=16:20
7      wirawasú      wirawasú      NOUN      N      Number=Sing      5      obj      _      SpaceAfter=No|TokenRange=21:29
Range=21:29
8      .      .      PUNCT      PUNCT      5      punct      _      SpaceAfter=No|TokenRange=29:30
```

Figura 1: Análise de uma sentença em nheengatu de uma lenda de Casasnovas (2006) por meio do Yauti (de Alencar, 2023) com tradução automática da versão em português para o inglês via Google Translator.

```
>>> s='''Ape/advt, paá, usika sesé wirawasú. (p. 66, No. 2) Então o gavião passou perto dele - Ape, paá, usika
wirá-wasú.'''
>>> Yauti.parseExample(s, 'Casasnovas2006', 1, 3, 3, annotator='Juliana Lopes Gurgel')
# sent_id = Casasnovas2006:1:3:3
# text = Ape, paá, usika sesé wirawasú.
# text_eng = Then the hawk passed by him
# text_por = Então o gavião passou perto dele
# text_source = p. 66, No. 2
# text_orig = Ape, paá, usika sesé wirá-wasú.
# text_annotator = Juliana Lopes Gurgel
1      Ape      ape      ADV      ADVT      AdvType=Tim      5      advmod      _      SpaceAfter=No|TokenRange=0:3
2      ,      ,      PUNCT      PUNCT      5      punct      _      TokenRange=3:4
3      paá      paá      PART      RPRT      Evident=Nfh|PartType=Mod      5      advmod      _      SpaceAfter=No|TokenRange=5:8
Range=5:8
4      ,      ,      PUNCT      PUNCT      5      punct      _      TokenRange=8:9
5      usika      sika      VERB      V      Person=3|VerbForm=Fin      0      root      _      TokenRange=10:15
6      sesé      resé      ADP      ADP      AdpType=Post|Number[grnd]=Sing|Person[grnd]=3|Rel=NCont      5      obl
TokenRange=16:20
7      wirawasú      wirawasú      NOUN      N      Number=Sing      5      obj      _      SpaceAfter=No|TokenRange=21:29
Range=21:29
8      .      .      PUNCT      PUNCT      5      punct      _      SpaceAfter=No|TokenRange=29:30
```

Figura 2: Desambiguação manual da análise da sentença em nheengatu da figura anterior e reaplicação do Yauti (de Alencar, 2023).

Select all: ☐ Basic ☐ Enhanced

```
# sent_id = Casasnovas2006:1:3:3
# text = Ape, paá, usika sesé wirawasú.
# text_eng = Then the hawk passed by him
# text_por = Então o gavião passou perto dele
# text_source = p. 66, No. 3
# text_orig = Ape, paá, usika sesé wirá-wasú.
# text_por_transcriber = Juliana Lopes Gurgel
# text_annotator = Juliana Lopes Gurgel
# acknowledgement = DACILAT Project, FAPESP's Process No. 2022/09158-5
# reviewer1 = Leonel Figueiredo de Alencar
```

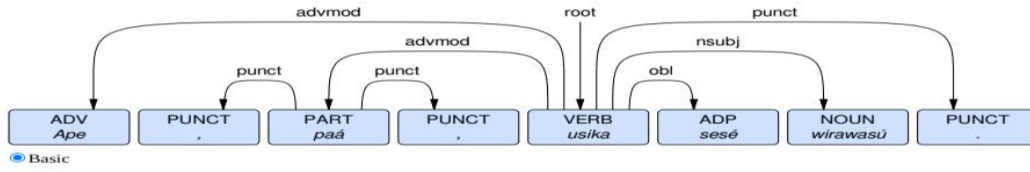


Figura 3: Representação arbórea no formato Universal Dependencies gerada pela ferramenta de visualização <https://urd2.let.ruq.nl/~kleiweq/conllu/> a partir da correção manual da análise automática do Yauti (de Alencar, 2023).

Na

```
$ echo "Ape, paá, usika sesé wirawasú." | udpipe --tokenize --tag --parse model-10.output
Loading UDPipe model: done.
# newdoc
# newpar
# sent_id = 1
# text = Ape, paá, usika sesé wirawasú.
1   Ape    ape    ADV    ADVDI    AdvType=Loc|Deixis=Remt|PronType=Dem    5    advmod    _    SpaceAf
0
2   ,      ,      PUNCT  PUNCT    3      punct
3   paá    paá    PART   RPRT     Evident=Nfh|PartType=Mod    5      advmod    _    SpaceAfter=No
4   ,      ,      PUNCT  PUNCT    3      punct
5   usika  sika   VERB    V        Person=3|VerbForm=Fin    0      root
6   sesé   resé   ADP     ADP      Number[grnd]=Sing|Person[grnd]=3|Rel=NCont    7      nmod:poss
7   wirawasú    wira   NOUN    N        Degree=Aug|Number=Sing    5      obl
8   .      .      PUNCT  PUNCT    _      5      punct    _    SpacesAfter=\n
```

Figura 4, exemplificamos o parsing neural por meio do UDPipe 1.2. Nesse exemplo específico, o Yauti, baseado em regras, produziu uma análise mais correta do que esse parser baseado em aprendizagem de máquina. No entanto, essa última abordagem obteve melhores resultados de uma maneira geral, conforme relatado em de Alencar (2024).

```

$ echo "Ape, paá, usika sesé wirawasú." | udpipe --tokenize --tag --parse model-10.output
Loading UDPipe model: done.
# newdoc
# newpar
# sent_id = 1
# text = Ape, paá, usika sesé wirawasú.
1   Ape    ape    ADV    ADVDI    AdvType=Loc|Deixis=Remt|PronType=Dem    5    advmod    _    SpaceAf
0
2   ,      ,      PUNCT  PUNCT    3      punct
3   paá    paá    PART   RPRT     Evident=Nfh|PartType=Mod    5      advmod    _    SpaceAfter=No
4   ,      ,      PUNCT  PUNCT    3      punct
5   usika   sika   VERB    V        Person=3|VerbForm=Fin    0      root
6   sesé    resé   ADP     ADP      Number[grnd]=Sing|Person[grnd]=3|Rel=NCont    7      nmod:poss
7   wirawasú    wira   NOUN    N        Degree=Aug|Number=Sing    5      obl      _    SpaceAfter=No
8   .      .      PUNCT  PUNCT    5      punct    _    SpacesAfter=\n

```

Figura 4: Análise da sentença *nheengatu* anterior pelo parser neural UDPipe 1.2 com base em modelo treinado em 90% das sentenças do *treebank* (de Alencar, 2024).

Entre as sentenças incluídas no *treebank* encontram-se algumas lendas inteiras e trechos de diversas outras narrativas, do século XIX até o século XXI. Todo esse material poderá ser incorporado no *treebank* do projeto da DACILAT na plataforma Tycho Brahe, que poderá preservar a anotação original ao lado de uma possível anotação adicional no formato Penn Treebank, baseado no modelo de estrutura sintagmática. Graças à quantidade expressiva de sentenças e *tokens* do UD\_Nheengatu-CompLin

(

The screenshot shows the Universal Dependencies website. The main table lists treebanks, with UD\_Nheengatu-CompLin highlighted. A tooltip displays statistics: 14,874 tokens, 15,036 words, 1,470 sentences. The description for UD\_Nheengatu-CompLin states: "The [UD\_Nheengatu-CompLin](https://doi.org/10.5753/stil.2023.234131) is a treebank of [Nheengatu] (https://glottolog.org/resource/languoid/id/nhen1239) (ISO-639: 'yri'), also known, inter alia, as Modern Tupi and \*Língua Geral Amazônica\*. It comprises sentences from diverse published sources, e.g., spontaneous speech, grammatical descriptions, fables, myths, coursebooks, and dictionaries." Below the description, there are links for Contributors, Repository, README, Treebank hub page, and Download.

Figura 5), considerando que se trata de uma língua que antes não possuía qualquer recurso análogo, conseguimos atingir um índice LAS de  $81.17 \pm 1.02$  por meio de treino de um parser neural profundo utilizando o UPPipe 1.2 (de Alencar, 2024). Esse índice é superior ao obtido pelo Yauti (de Alencar, 2023).



The screenshot shows the Universal Dependencies website. At the top, there's a navigation bar with the site name and search icons. Below it, a list of languages is displayed with their respective token counts. Nheengatu is highlighted with a blue box showing its statistics: 14,874 tokens, 15,036 words, and 1,470 sentences. Below this, the 'CompLin' treebank is detailed, including its description, contributors, and download links.

Language	Count	Size	Icons	Family
Naija	1	140K		Creole
Nayini	1	<1K		IE, Iranian
Neapolitan	1	<1K		IE, Romance
Nheengatu	1	15K		Tupian, Maweti-Guarani

**Nheengatu treebanks**

14,874 tokens 15,036 words 1,470 sentences

**CompLin** 15K L F

The [UD\_Nheengatu-CompLin](https://doi.org/10.5753/stil.2023.234131) is a treebank of [Nheengatu] (https://glottolog.org/resource/language/id/nhen1239) (ISO-639: 'yrl'), also known, inter alia, as Modern Tupi and \*Língua Geral Amazônica\*. It comprises sentences from diverse published sources, e.g., spontaneous speech, grammatical descriptions, fables, myths, coursebooks, and dictionaries.

- Contributors: Leonel Figueiredo de Alencar
- Repository [master dev](#)
- [README](#)
- [Treebank hub page](#)
- [Download](#)

Figura 5: Informações sobre a versão atual do treebank do nheengatu da coleção Universal Dependencies (<https://universaldependencies.org/>)

O segundo grupo de atividades abrange iniciativas que contaram com financiamento da Fapesp no âmbito do projeto da DACILAT. A primeira parte desse financiamento consistiu na remuneração da transcrição de todas as lendas da obra “O selvagem” de Couto de Magalhães (1876), alinhando o texto em nheengatu à tradução em português, preservando, quando existente nessa publicação, a correspondência entre de n-gramas da língua fonte e da língua alvo

The screenshot shows a text transcription in Nheengatu. The text is presented in a dark-themed editor with red underlines for specific words. The transcription includes dialogue between characters, with some words in Nheengatu and others in Portuguese.

JAUTI TAPIIRA CAHAUARA  
Jabuti e anta do mato

Iautí mira catú, intimãhã mira puxí.  
Jabuti gente é boa, não gente é má.

Oikô itapereiúá uirpe, oçanhãna i temiú.  
Estava do taperebá embaixo, ajuntando sua comida.

Tapiíra cahaúára oçika ápe, onhehê ixupé: – Retirica iautí, retirica kí (iké) xií.  
Anta do mato chegou ali, disse a ele: – Retire-se, jabuti, retire-se aqui de.

Iautí oçuxára ixupé: – Ixê kí xií (çuí) intí xa tãrica mãhá recê xa ikô cê iúá iua uirpe.  
Jabuti respondeu a ela: – Eu aqui de não me retiro que por (porque) eu estou de minha de fruta árvore embaixo.

– Retirica, iautí, curumú xa pirú indê.  
– Retira-te, jabuti, senão eu piso você.

– Repirú!... rê mahê arãma, inê nhũ será apgáua!  
– Pisa!... tu veres para, se tu só és macho!

Tapiíra, iuruparí, opirú iautí teté.  
Anta, juruparí, pisou jabuti coitado.

Tapiíra oçô âna.  
Anta se foi embora.

Figura 6). Essa transcrição foi realizada por Dominick Maia Alexandre, bolsista de iniciação científica em vários períodos. Para assegurar a fidelidade da transcrição dos textos de Couto Magalhães, que se

valeu da notação fonética de Lepsius, com muitos caracteres inexistentes nos teclados convencionais, foi desenvolvida uma tabela de equivalência entre esses caracteres e codificações equivalentes ou análogas em Unicode. Essa tabela inclui também combinações de teclas que permitem a digitação desses caracteres de forma confortável utilizando o sistema operacional Linux Ubuntu. Também foi implementado um script na linguagem de programação Python capaz de converter os caracteres digitados por meio de combinações de tecla nos correspondentes símbolos em Unicode. Esse programa possui também uma função que verifica o alinhamento dos n-gramas referidos. A segunda parte desse financiamento se refere à bolsa de Treinamento Técnico TT3 concedida a Juliana Lopes Gurgel, que iniciou suas atividades em janeiro. Essa bolsista transcreveu todas as lendas de Casasnovas (2006), das quais anotou algumas.

```

JAUTI TAPIIRA CAHAUÁRA
Jabuti e anta do mato

Iautí mira catú, intimãhã mira puxí.
Jabuti gente é boa, não gente é má.

Oikô itapereiúá uirpe, oçanhãna i temiú.
Estava do taperebá embaixo, ajuntando sua comida.

Tapiíra cahaíúára oçika ápe, onhehê ixupé: – Retirica iautí, retirica kí (iké) xií.
Anta do mato chegou ali, disse a ele: – Retire-se, jabuti, retire-se aqui de.

Iautí oçuxára ixupé: – Ixê ki xií (çuí) intí xa tîrica mǎhá recê xa ikô cê iúá iua uirpe.
Jabuti respondeu a ela: – Eu aqui de não me retiro que por (porque) eu estou de minha de fruta árvore embaixo.

– Retirica, iautí, curumú xa pirú indê.
– Retira-te, jabuti, senão eu piso você.

– Repirú!... rē mahê arāma, inê nhū será apgáua!
– Pisa!... tu veres para, se tu só és macho!

Tapiíra, iurúparí, opirú iautí teté.
Anta, juruparí, pisou jabuti coitado.

Tapiíra oçô āna.
Anta se foi embora.

```

Figura 6: Trecho inicial da transcrição de lenda de Magalhães (1876).

Prometemos ainda incorporar o Kadiwéu no processo de tradução automática. Isso ainda não foi possível porque ainda estamos trabalhando no sistema de Parser para a língua, mas coletamos um material em campo para alimentar a primeira programação do Kadiwéu na linguagem Grammatical Framework que se encontra no Copus Gramática Pedagógica já apresentada e aberta para visitaç o. Trata-se de senten as que cobrem o funcionamento do sintagma nominal do kadiw u.

As publica  es desta se  o foram:

DE ALENCAR, Leonel Figueiredo. Yauti: A Tool for Morphosyntactic Analysis of Nheengatu within the Universal Dependencies Framework. In: SIMP SIO BRASILEIRO DE TECNOLOGIA DA INFORMA  O E DA LINGUAGEM HUMANA (STIL), 14. , 2023, Belo Horizonte/MG. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computa  o, 2023. p. 135-145. DOI: <https://doi.org/10.5753/stil.2023.234131>.

DE ALENCAR, Leonel Figueiredo. A Universal Dependencies Treebank for Nheengatu. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF PORTUGUESE (PROPOR 2024), 16, March 12-15, 2024, Santiago de Compostela, Galicia, Spain. **Proceedings** [...]. Stroudsburg, PA, USA: Association for Computational Linguistics, 2024. v. 2, p. 37-54. Disponível em: <https://aclanthology.org/2024.propor-2.8>. DOI: <https://doi.org/10.5281/zenodo.11372209> ISBN: 979-8-89176-062-2

Este último trabalho foi apresentado no dia 12 de março na Universidade de Santiago de Compostela no âmbito do First Workshop on NLP for Indigenous Languages of Lusophone Countries (<https://propor2024.citius.gal/index.php/first-workshop-on-nlp-for-indigenous-languages-of-lusophone-countries/>), oficina satélite da *16th International Conference on Computational Processing of Portuguese* (PROPOR 2024) (<https://propor2024.citius.gal/>). Destaco que o colaborador Leonel Figueiredo de Alencar também integrou a comissão científica desse evento, no qual figurou como *keynote speaker*, proferindo a palestra de abertura intitulada “NLP for Brazilian Indigenous Languages: A (computational) linguist's perspective”, atendendo a convite dos organizadores.

Além desses dois artigos, Leonel Figueiredo de Alencar, pela sua contribuição com o treebank do dheengatu, consta como coautor da coleção Universal Dependencies, que, na sua versão mais recente, tem esta referência:

Zeman, Daniel; et al., 2024, Universal Dependencies 2.14, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, <http://hdl.handle.net/11234/1-5502>.

Um terceiro eixo das pesquisas desta seção, que ainda não resultou em publicação, e já foi apresentado anteriormente, consistiu na modelagem computacional da microestrutura do dicionário dheengatu-português de Marcel Twardowsky Ávila. Leonel Figueiredo de Alencar implementou, na linguagem de programação Python, um módulo intitulado Nheengariru com uma série de classes representando os diferentes níveis de informação dos verbetes, contemplando os aspectos morfológicos, sintáticos, semânticos, etimológicos etc. (Figuras 7-9). Essa modelagem, cujo código se encontra no repositório <https://github.com/leoalenc/nheengariru>, foi utilizada por Luiz Henrique Lima Veronesi para transformação do dicionário de Ávila em um banco de dados, conforme discutido acima (seção 3.1.6), visando à construção de uma interface amigável na Web e uma API a ser utilizada pelas diferentes ferramentas de processamento computacional do dheengatu, em desenvolvimento ou a serem desenvolvidas, como o próprio sistema de tradução automática.

```

>>> import Nheengariru
>>> import Entries
>>> entry=Entries.ExampleEntry()
>>> entry.lemma.form
'teité'
>>> entry.senses
[<Nheengariru.Sense object at 0x7fcec27af970>, <Nheengariru.Sense object at 0x7fcec0b880d0>]
>>> entry.senses[0].definition
'(exprime compadecimento:) coitado (a, os, as)!, pobrezinho (a, os, as)!, TEITÉ! (PA)'
>>> entry.senses[0].cat.lexcat[0].label
'interj'
>>> for sense in entry.senses:
    for cat in sense.cat.lexcat:
        print(cat.label,sense.definition)

```

```

interj (exprime compadecimento:) coitado (a, os, as)!, pobrezinho (a, os, as)!, TEITÉ! (PA)
s coitado (de), coitadinho (de), pobre, miserável
adj coitado (de), coitadinho (de), pobre, miserável

```

Figura 7: Explorando a modelagem computacional no Nheengariru dos diferentes sentidos do verbete de **teité** de Avila (2021, p. 747)

```

>>> help(sense)
Help on Sense in module Nheengariru object:

class Sense(builtins.object)
| Sense(cat, definition, num=None, subsenses=[], examples=[], usage=None, sourcelist=[], usage_note='', equivalent=None)
|
| A class to represent a sense of a lemma.
|
| Args:
|     cat (Category): Categorical information.
|     definition (str): Sense definition.
|     num (int): Sequence number of the sense.
|     examples (list): List of SenseExamples instances.
|     usage (Usage): Sense usage.
|     sourcelist (Sourcelist): List of bibliographical sources.
|     usage_note (str): Explanation about the use of the sense.
|     equivalent (Equivalent): Synonymous lemma.
|
| Methods defined here:
|
| __init__(self, cat, definition, num=None, subsenses=[], examples=[], usage=None, sourcelist=[], usage_note=equivalent=None)
|     Initialize self. See help(type(self)) for accurate signature.

```

Figura 8: Documentação da classe Sense no Nheengariru.

```

>>> entry.senses[1].examples[1].example.por
'A máscara da Mãe do Barro voltou e aderiu em nós. Por isso, nós, coitados, não saímos muito bonitos.'
>>> entry.senses[1].examples[1].example.yrl
'Nhaã Tuyuka Manha pírera kwera uyuíri uyari yané resé. Yawé arā paá yandé, taité, ti yasemu purapuranga.'
>>> entry.senses[1].examples[1].example.source.author
'Leetra Indígena. n. 17'
>>> entry.senses[1].examples[1].example.source.page
37

```

Figura 9: Acessando no Nheengariru as diferentes abonações de um dos sentidos de **teité**.

Para concluir, *nheengatu* permite iniciar os trabalhos em tradução automática porque esta língua já conta com uma significativa quantidade de materiais textuais providos de anotação morfológica e sintática, disponíveis no UD\_Nheengatu-CompLin (de Alencar, 2024), que atualmente



totaliza 1470 sentenças e 15036 palavras, conforme a

universaldependencies.org

Language	Sentences	Tokens	Words
Naija	1	140K	
Nayini	1	<1K	
Neapolitan	1	<1K	
Nheengatu	1	15K	

**Nheengatu treebanks**

14,874 tokens 15,036 words 1,470 sentences

**CompLin** 15K

The [UD\_Nheengatu-CompLin](https://doi.org/10.5753/stil.2023.234131) is a treebank of [Nheengatu] (https://glottolog.org/resource/languoid/id/nhen1239) (ISO-639: `yri`), also known, inter alia, as Modern Tupi and \*Língua Geral Amazônica\*. It comprises sentences from diverse published sources, e.g., spontaneous speech, grammatical descriptions, fables, myths, coursebooks, and dictionaries.

- Contributors: Leonel Figueiredo de Alencar
- Repository [master dev](#)
- [README](#)
- [Treebank hub page](#)
- [Download](#)

Figura 5. O kadiwéu é uma língua que necessita totalmente de materiais e dedicamos o tempo do primeiro ano coletando, anotando e analisando dados desta língua para nosso Parser de regras. Não foi um trabalho braçal de pouca importância, pois a anotação do Kadiwéu gerou convites para dois projetos internacionais que serão apresentados abaixo.

#### 4. Outros projetos internacionais a partir do projeto DACILAT/FAPESP

Dois projetos internacionais foram submetidos em cooperação com o projeto DACILAT, um já mencionado, na República Tcheca. E outro para a União Europeia. Seguem os resumos. A coordenadora do projeto DACILAT será membro dos projetos internacionais, se aprovados.

##### 1. União Europeia

**LoI 1317766**

**To:** Human Frontier Science Program

**Program:** Research Grants - Program

**Title:** Glassy landscape of syntax spin evolutionary dynamics (Treves, Alessandro)

**Deadline:** 3/28/2024 1:00:00 PM (U.S. Eastern Time)

##### Abstract

Has language arisen only once? Analyses of cognitive functions often assume that their neural mechanisms have evolved towards some “optimal” form. Yet syntax, at the core of the human



language faculty, presents a striking variety among the ca 5,000 extant natural languages, challenging both optimality ideas and language learners. Can all those divergent forms of syntax be nearly optimal? E.g. Latin, verb-last and rich in case morphology, transitioned to object-last, largely case-free Romance languages: was it a sub-optimal communication tool, for hundreds of years?

If not convergent, is syntax evolution a process of divergence from a single original form, like words within language families, or rather a set of quasi-random trajectories along a manifold of similar efficiency, with many long-lasting metastable states? Longobardi's syntactic phylogenetics yield a dataset of 58 languages, described by 94 binary parameters [1], to analyze with tools from the physics of spin glasses (Treves, De Giuli). To assess whether the apparent tree structure stems from a focus on Eurasia, Sandalo will bring in American languages, enabling inferences from measures of treeness and ultrametricity.

1197/1200

### **Subproject 1**

The limbo group of A Treves will develop network models of syntax dynamics and in parallel explore their implementation in the brain, ie, the self-organization and operation of syntactic parameters in sentence understanding and production. Crucial to the first effort will be the integration of widely used models of phylogenetic evolution – such as Early Burst, Brownian Motion and Ornstein-Uhlenbeck models – with the notion of syntactic parameters as a network of interacting binary units subject to external fields. Effective interactions that include strong asymmetric 'implications' as well as weaker random couplings have been shown in our current work to lead to a multiplicity of metastable states, in some conditions. Importantly, the discovery of the speed inversion effect [2] indicates that non-binary hidden variables – ubiquitous in language – may slow down considerably the glassy dynamics of binary parameters. Previously developed measures of metric and ultrametric content will be applied to assess the divergence vs Liouville contrast, based on n-tuples of remote languages.

1094/1200

### **Subproject 2**

In the past 10 years Longobardi has suggested that through syntactic comparison traces of treeness and language systematics can be found across Eurasian families, supporting divergence models of word evolution within established families. Syntactic differences are modeled as binary parameters with crucial emphasis on the analysis of implications. Syntax variability is limited by the implications among parameters: for example, a language that does not represent gender at all (value 0 of parameter FGG) cannot logically represent it on cardinal numbers (value 1 of parameter GPC).

Particularly insidious are empirical/statistical rather than logical implications, eg it appears that no natural language represents gender if it does not represent number (parameter FGN).

772/1200

### **Subproject 3**

Critical to address the challenge of extending the database to native American languages, Sandalo is coordinating a project in Brazil (FAPESP # 22/09158-5) whose goal is to build digital grammatically annotated corpora for South American endangered languages. The data from this project is archived on the Tycho Brahe Electronic Platform. The platform provides multi-purpose tools for linguistic corpora especially suited for highly complex languages, known in linguistics as agglutinative or polysynthetic languages, where each word can encode the meanings of entire sentences. The first language to integrate our corpus is Kadiwéu; thus, a preliminary version of Kadiwéu's digital corpus is already available at the Tycho Brahe Platform (<https://www.tycho.iel.unicamp.br/viewer/C12>). Kadiwéu is a polysynthetic language spoken in Midwest Brazil from the Guaikuruan linguistic family, with intriguing properties [3]. Toba, also from the same family, will be added to the platform, with two languages from the Tupian family: Nheengatu and Guarani.

Sandalo will now see how to extract from the extended-POS tagged corpora 'syntactic spins' to be used by the HFSP team, eg De Giuli and Treves.

1192/1200

### **Subproject 4**

DeGiuli proposed in [4] an ensemble of stochastic context-free grammars and showed that the phase space is divided into grammatical and ungrammatical regions, whose boundaries are understood. It was also shown that the grammatical region hosts metastable states, ripe for comparison with real data. Context-free grammars for each language in the Longobardi database will be constructed and placed in the phase diagram of [4]. This data will be scrutinized to see if plausible histories of these languages are compatible with a series of symmetry-breaking transitions in the framework of the model. Once cast in the appropriate format, also grammars for the more complex agglutinative/polysynthetic languages investigated by Sandalo will be considered in relation to the phase diagram. Iteration with LIMBO approaches and possible extensions of [4] will aim for compatibility with neurodynamics and phylogenetics in the simplest model. Simulations of [4] coupled to an external field can then validate or refute models of learning, quantitatively extending the Principles & Parameters approach within a continuous framework.

1123/1200

## **Frontier Science**

Linguists, and particularly syntacticians, have often been considered aloof from the life sciences, in part due to their formalism based on individual examples. Large language models, conversely, approach syntax statistically, but remain largely non-transparent to the mechanisms implementing it in any particular language, and how they could be instantiated in the brain. The gaps are even wider in regard to syntax evolution, with its temporal dimension.

Recently, a breakthrough has been afforded by the Longobardi database [1] and the subsequent demonstration, in an ERC-funded collaboration with geneticists, that it allows for the reconstruction of a putative ‘language tree’ extending  $O(10^4\text{yr})$  into the past, consistent with the phylogenetic tree inferred from the DNA of the speaker populations. Unlike genetic variability, though, syntax variability does not appear to be constrained by the ‘fitness’ of the language phenotype. It is limited, instead, by the so called ‘implications’, detailing which requires extensive expertise.

The stochastic character of grammaticalization expressed eg by the DeGiuli approach [4] is compatible with glassy metastable states during syntax evolution, but to be incorporated in a concrete model of language change it must be based on linguistic analyses. Then it can help vis-à-vis both actual language history and specific syntax-related neural mechanisms. Understanding which, beneath the surface of a phenomenological description, would be a major breakthrough in answering the question of how the brain expresses higher cognition.

1584/1600

## **Interdisciplinarity**

Longobardi and Treves, a theoretical linguist and a statistical physicist turned theoretical neuroscientist, have been talking for years, but only lately, after breakthroughs in their research, have seen the potential for combining their different expertise to better understand the implementation and self-organization of language parameters in the brain and ultimately language evolution; leading to a joint paper. It [5] shows major violations of ultrametricity, but from distant language triplets, raising the issue: does an ultrametric tree description applies only within an Eurasian context?

598/600

## **Collaboration**

To wade outside it, and beyond classical Indo-European linguistics, they have contacted Sandalo, leading expert on unrelated languages from South America. Her corpora also provide evidence for syntax change, which has to be cast in ‘spin’ format and then inform with real data the ground-breaking but so far abstract statistical-physics-style syntax model by De Giuli, an expert in disordered systems. With the ideas and tools from genetics already adopted in part by Longobardi,

Treves will model parameter diachronics and link them to transitions in the self-organization of cortical networks.

595/600

## 5 References

- [1] Ceolin, A., Guardiano, C., Longobardi, G., Irimia, M. A., Bortolussi, L., & Sgarro, A. (2021). At the boundaries of syntactic prehistory. *Phil Trans Roy Soc B*, 376(1824), 20200197.
- [2] Ryom, K. I., & Treves, A. (2023). Speed Inversion in a Potts Glass Model of Cortical Dynamics. *Phys Rev X Life*, 1, 013005.
- [3] Sandalo, F. (2023). On the Guaikuruan inverse system: interpreting Kadiwéu and Mocoví person hierarchies. *International Journal of American Linguistics*, 89(1), 105-135
- [4] DeGiuli, E. (2019). Random language model. *Phys Rev Lett*, 122(12), 128301.
- [5] Longobardi, G., & Treves, A. (2024). Grammatical Parameters from a Gene-like Code to Self-Organizing Attractors: a research program. *arXiv preprint arXiv:2307.03152*. and in press in *A Cartesian Dream*, Greco, M. & Mocci, D., eds. Vol.17 of RGG monographs, LingBuzz Press.

## 2. *República Tcheca*

Metamorphosis: Morpheme-Centric Multilingual Modeling of Language Variability

### Our Motivation and Overview of the Research Proposal

Natural languages are complex systems composed of subsystems that interact with each other in various ways. Many of the interactions are conventionally interpreted using higher levels of linguistic abstraction; various multi-layered terminological systems have been designed for that, ranging, e.g., from phonetics to pragmatics. For many language phenomena, data resources annotated by human experts exist. Given that multilinguality is one of the strongest trends in Computational Linguistics research in the last decade or two, substantial efforts have been invested into creating multilingual data collections annotated under common harmonized schemes, be they of corpus-like nature such as OntoNotes (Hovy et al., 2006) or Universal Dependencies (de Marneffe et al., 2021), dictionary-like such as UniMorph (McCarthy et al., 2020) or CLICS (List et al., 2018), or typological such as Grambank (Skirgārd et al., 2023).

Elaborating the higher levels of linguistic abstraction (going “beyond syntax” etc.) sounds like the natural way to go, and seems also somewhat more intellectually appealing at first sight. However, the central claim of our proposal is that there is a surprisingly huge unexploited potential when we go right in the opposite direction. To a lower level of abstraction. To the smallest meaningful units of language. To morphemes.

We argue that focusing on morpheme-centric computational data models would allow to develop radically new empirical methods for language research, which could not only bring new insights into how languages work, but could establish new points of contact with recent advances in Natural Language Processing, whose linguistic interpretability is currently limited.

The morpheme-centric account we propose and advocate for in the present proposal is expected to fill the gaps we see in the mainstream research in morphology. There are several very basic linguistic notions such as prefix that are inherently based on the notion of morpheme and that are not only omnipresent in traditional grammars, but are crucial also in less classical approaches to language. For example, competitive exclusion (“struggle for existence”) is not only considered a general principle in evolutionary biology, but can be used also for understanding complementary distributions of various language means that share similar niches, such as allomorphic variants, rivaling affixes, competing synonyms (e.g. native words vs. loanwords), or alternative morphological means of expressing the same function (e.g. possessive forms vs. genitive forms); naturally, linguists explain competition in terms of (different types of) morphemes too. In the light of these facts, we find it extremely surprising that morphemes are not properly substantiated in modern language data resources. Even for many resource-rich languages, data in which segmentation to morphemes is available are scarce (surprisingly scarcer than, for instance, syntactically annotated corpora). Thus, linguists are doomed to focus only on narrow tasks and work only with very limited language material and/or to rely on *ad-hoc* approximations of the morphological structure of words (resorting, e.g., to regular expressions when searching for prefixed verbs, with all the negative consequences of such simplifications).

Further challenges can be found in the cross-lingual perspective. Morphology-based typology of languages constituted one of the earliest approaches to linguistic typology in the 19th century. More recently, typological databases came into existence, in which admirable expertise of generations of comparative linguists is accumulated. Current typological databases are focused on a small amount of relatively abstract features, but do not contain information about large amounts of individual morphemes in individual languages, and they certainly do not offer information about the flow of such morphemes across language boundaries (lexicostatistical approaches exist too, e.g. those based on the famous Swadesh 1952 list and its derivatives, but they are limited either to very small core vocabulary, or to typologically not very diverse languages (Nelson-Sathi et al., 2011)). Even though the field of Quantitative Typology is fascinating, we claim that one could go much further with morpheme-centric data about the world’s languages. First, even if the current typological databases are sufficient, e.g., for automated induction of phylogenetic

trees, and even if such tree-shaped structures seem useful as a highly simplified model of language evolution, one should not disregard substantial horizontal transfers of morphemes along lines that cross such phylogenetic trees, resulting in well-known phenomena such

as Sprachbunds, dialect continua, or neoclassical compounding. Second, tasks such as identification of borrowings would become more clearly specified and resolved at the level of morphemes too as it is often the case that a single word is composed of morphemes of very different origin. Third, cross-lingual patterns in inflection, word formation, or even syntax could be studied more in depth and in a much wider scale if morphological segmentations are available.

Research to be carried out in the proposed project is structured around three pillars following upon the given motivation.

1. **Morphemes in an Intra-Lingual Perspective.** Intra-lingual topics will be elaborated ('intra-lingual' only in the sense that languages under study will be processed separately, but still, there will be a large number of them; to be exact, the first research direction is multi-lingual too). We will survey existing approaches and data resources relevant for morphological segmentation, design a formal model that would make it possible to accommodate various corner cases and heterogeneities of massively multilingual data (including usage of various writing systems, code switching, substandard language forms etc.), and create multilingual collections of morpheme-segmented wordlists. We will focus on written evidence and synchronic data. We will design empirical methods for modeling paradigmatics (classification) as well as syntagmatics (sequential combinatorics) of morphs. We will use morphological segmentation for designing novel methods for modeling competition processes in languages. Purposely, we will make only minimal use of more abstract linguistic concepts such as the conventional distinction between inflection and word formation (or replace such dichotomies with gradual scales) so that our models can be applied across as many languages as possible without too much risk of introducing biases from Indo-European linguistics.
2. **Morphemes in a Cross-Lingual Perspective.** We want to perform a number of cross-lingual studies whose common denominator is profiting from availability of morphological segmentation in individual languages. Using multilingual resources such as parallel and comparable corpora, and possibly also modern Machine Translation methods, we will identify morpheme-level correspondences across languages, and distinguish homologous correspondences (morphemes having the same origins, cognates) from the other ones. Cross-lingual patterns valid for larger groups of languages will be extracted from the data. Usage of different means (on the morpheme level) for expressing comparable semantic functions across different languages will be studied contrastively across languages and language groups; we will combine morphological segmentations with Universal Dependencies in order to be

able to study phenomena that belong to morphology in one language but to syntax in some other language (such as compounding or incorporation). Findings in all these subdirections will be contrasted with information in existing typological databases.

3. **Morphemes in a Language Change Perspective.** Natural languages are dynamic systems, and morphology serves as a vital component in tracing linguistic transformations over time. Some morphological structures erode, while, conversely, languages may also innovate, introducing new morphological patterns, for example through borrowing. Using the data and experience collected in the previous two pillars and the ‘natural selection’ metaphor, we will explore various mechanisms leading to gradual changes in languages, such as rivalry among affixes or between native and borrowed morphemes. Channels of borrowing of lexical and grammatical material (on the morpheme level, again) between and among languages will be modeled on an empirical basis.

The proposal is based on two key decisions. First, we are going to include a growing number of languages in all three research directions; three phases will be explicitly distinguished in individual subtasks: 3L – developing proof-of-concept approaches using data from 3 languages, 30L – extending the approaches to 30 languages, and 300L – further extension to 300 languages. This pace is ambitious, but we find it realistic; we will profit from our past experience with highly multilingual data, and we also assume various synergies and ‘economy of scale’ savings to apply. However, it is clear at the same time that accuracy reachable for individual subtasks may vary greatly across languages, depending on many factors ranging, e.g., from sizes of available corpora to existence of stable orthographic norms.

The second essential methodological decision is that we will gradually accumulate all pieces of information gathered in the three research directions in a shared database that will serve as a “meeting



point” for the three research subteams. Emphasis will be put on scalability of the database, so that it can accommodate virtually all languages for which written forms exist. The resulting database will be released publicly and will be capable of serving as a universal international standard reference for many specialized subfields of linguistics (computational typology, etymology, translation studies, etc.). We claim that this approach, which is very broad both in terms of the lexical coverage of individual languages AND of the number and typological diversity of included languages, is unexplored to a large extent; this is partly because of clearly limited competence of any human individual when it comes to large-scale multilinguality (in light of the richness of the world’s languages), and partly because of the previous lack of data and computational methods (and perhaps also because of the possible illusion that morphological topics have been already more or less resolved in linguistics and it is time to aim ‘higher’). Thus, we assume that we will be able to unravel truly novel insights into language patterns and self-organization mechanisms that were hidden to linguistic research up to now. These insights will be empirically grounded in replicable experiments, and can possibly lead to enrichment or even redefinition of some traditional linguistic notions towards their more general applicability. Finding such new insights is the main objective of the project.

## References

- de Marneffe, M.-C., Manning, C., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2):255–308.
- Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60.
- List, J.-M., Greenhill, S. J., Anderson, C., Mayer, T., Tresoldi, T., and Forkel, R. (2018). CLICS2: An improved database of cross-linguistic colexifications assembling lexical data with the help of cross-linguistic data formats. *Linguistic Typology*, 22(2):277–306.
- McCarthy, A. D., Kirov, C., Grella, M., Nidhi, A., Xia, P., Gorman, K., Vylomova, E., Mielke, S. J., Nicolai, G., Silfverberg, M., et al. (2020). Unimorph 3.0: Universal morphology. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association.

- Nelson-Sathi, S., List, J.-M., Geisler, H., Fangerau, H., Gray, R. D., Martin, W., and Dagan, T. (2011). Networks uncover hidden lexical borrowing in Indo-European language evolution. *Proceedings of the Royal Society B: Biological Sciences*, 278(1713):1794–1803.
- Skirg ard, H. et al. (2023). Grambank reveals global patterns in the structural diversity of the world’s languages. *Science Advances*, 9.
- Swadesh, M. (1952). Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. *Proc. of the American philosophical society*, 96(4):452–463.

## 5. Gest o de Dados

Como previsto, e apresentado, todos nossos dados (com transcri  es e  udios) e anota  es est o sendo depositados na Plataforma Tycho Brahe. Nada at  o momento est  fechado e pode ser visitado online e usado para outras pesquisas ou para educa  o ind gena livremente. A aluna de doutorado Vanda Pires, que   kadiw u, foi treinada para entrar dados e explicar em sua comunidade/escola sobre o uso da Plataforma para fins educacionais na pr pria aldeia. Segue novamente o link de entrada aberto para visita  o e navega  o dos dados (ver se  o 3.1):

<https://www.tycho.iel.unicamp.br/viewer>

## 6. Uso da Verba

A verba neste primeiro ano foi usada para viagens de coleta de dados, viagem de acordos de pesquisa e consulta sobre anota  o de corpora (Praga, Rep blica Tcheca), participa  o em congresso (ABRALIN), servi  os de terceiros em transcri  o de dados (kadiw u e nheengatu), compra de equipamento para uso na aldeia (1 computador laptop e seguro). Os microfones foram doados pelo pesquisador colaborador Michael Becker.

Material Permanente	5.000,00	0,00	0,00	5.000,00
Material de Consumo	0,00	0,00	0,00	0,00
Despesas de Transporte	2.676,44	0,00	0,00	2.676,44

Serviços de Terceiros	4.217,85	0,00	0,00	4.217,85
Diárias / Manutenção	14.692,37	0,00	0,00	14.692,37
<b>Total Despesas (A)</b>	26.586,66	0,00	0,00	26.586,66

## REFERÊNCIAS/RELATORIO

- BRITTO, H., Finger, M., GALVES, C. (2002) Computational and linguistic aspects of the construction of the Tycho Brahe Parsed Corpus of Historical Portuguese. Romance Corpus Linguistics - Corpora and Spoken language. Tubingen: Narr.
- FARIA, P., GALVES, C., MAGRO, C. (2023) Syntactic annotation for Portuguese corpora: standards, parsers, and search interfaces. *Language Resources and Evaluation, Especial Issue on Computational Approaches to Portuguese*.
- FINGER, M. (2000) Técnicas de Otimização da Precisão Empregadas no Etiquetador Tycho Brahe. Anais do V Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR2000).
- CHIERCHIA, G. (1998). ‘Plurality of mass nouns and the notion of “semantic parameter”’. In ‘Events and Grammar’, 53–103. Springer Netherlands.
- GALVES, C., SANDALO, F., SENA, T.A., VERONESI, L. (2017) Annotating a polysynthetic language: From Portuguese to Kadiwéu. *Cadernos de Estudos da Linguagem*, (59.3), pp. 631-648.
- GRIFFITHS, G. (1987). *Relative Clause Formation and other Word Parameters in Kadiwéu*. Reading University master thesis.
- GRIFFITHS, G. (2002). *Dicionário da Língua Kadiwéu*. SIL ms. <https://www.sil.org/system/files/reapdata/74/06/08/74060839706011162756896570533590209458/KDDict.pdf>.
- KROCH, A., TAYLOR, A., SANTORINI, B. 2000-. The Penn-Helsinki Parsed Corpus of Middle English (PPCME2). Department of Linguistics, University of Pennsylvania. CD-ROM, second edition, release 4 (<http://www.ling.upenn.edu/ppche/ppche-release-2016/PPCME2-RELEASE-4>).
- KROCH, A., SANTORINI, B., DELFS, L. 2004. The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME). Department of Linguistics, University of Pennsylvania. CD-ROM, first edition, release 3 (<http://www.ling.upenn.edu/ppche/ppche-release-2016/PPCEME-RELEASE-3>).

- KROCH, A., SANTORINI, B., DIERTANI, A. 2016. The Penn Parsed Corpus of Modern British English (PPCMBE2). Department of Linguistics, University of Pennsylvania. CD-ROM, second edition, release 1 (<http://www.ling.upenn.edu/ppche/ppche-release-2016/PPCMBE2-RELEASE-1>).
- MAGRO, C., GALVES, C. (2019) Portuguese Syntactic Annotation Manual. <http://alfclul.clul.ul.pt/portuguesesyntacticannotation/>
- NEVINS, A., and SANDALO, F. (2011). Markedness and morphotactics in Kadiwéu. [+participant] agreement. *Morphology* 21(2): 351-378.
- SANDALO, F. (1997). *A Grammar of Kadiwéu with Special Reference to the Polysynthesis Parameter*. MIT Occasional Papers in Linguistics 11.
- SANDALO, F. (2009). Person hierarchy and inverse voice in Kadiwéu. *LIAMES* 9: 27-40.
- SANDALO, F. (2020). Individuation, counting, and measuring in the grammar of Kadiwéu. *Linguistic Variation* 20(2): 239-254.
- SANDALO, F., and Michelioudakis, D. (2016). Classifiers and Plurality: evidence from a deictic classifier language. *Baltic International Yearbook of Cognition, Logic and Communication*, 11: 1-40.
- SANDALO, F. (2023). Evidencialidade Reportativa, Tempo e Negação em kadiwéu. *Liames* 23(1) <https://doi.org/10.20396/liames.v23i00.8671197>
- SANDALO, F. (2023b). On the Guaikuruan inverse system: interpreting Kadiwéu and Mocoví person hierarchies. *International Journal of American Linguistics* 89(1). <https://doi.org/10.1086/722239>
- SANTORINI, B. (2022) Annotation manual for the Penn Historical Corpora and the York-Helsinki Corpus of Early English Correspondence <https://www.ling.upenn.edu/hist-corpora/annotation/index.html>