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

KEYWORDS: INDIGENOUS LANGUAGES, SOUTH AMERICA, TEXTUAL CORPUS, MACHINE TRANSLATION

Maria Filomena Sandalo Coordinator Full professor UNICAMP Campinas Brazil Indigenous languages documentation and analysis Email: sandalo@unicamp.br
Charlotte Galves Principal Investigator Full Professor UNICAMP Campinas Brazil Corpus Linguistics and Syntactic Theory E-mail: charlotte.mgc@gmail.com
Pablo Feliciano de Farias Principal Investigator Assistant Professor UNICAMP Campinas Brazil Computer Science and Computational Linguistics E-mail: fariap.@unicamp.br
Luiz Veronesi Associate Investigator Computer Scientist UNICAMP Campinas Brazil Computer Science and Corpus Linguistics E-mail: luiz@texugo.com.br

DACILA

Leonel Figueiredo de Alencar Araripe
Associate Investigator
Full Professor
Universidade Federal do Ceará
Fortaleza
Brazil
Machine Translation and Grammar Engineering
Email: leonel.de.alencar@ufc.br

Michael Becker
Associate Investigador
Associate Professor
University of Massachusetts, Amherst
Amherst
USA
Computational Linguistics
Email: becker@phonologist.org

Vanda Pires
PhD Student
UNICAMP
Campinas
Brazil
Native Speaker of Kadiwéu and Indigenous school teacher
E-mail: v264435@dac.unicamp.br

START DATE: March, 1, 2023- 60 MONTHS

Table of Contentes

Abstract

Drawing on the vast experience of the researchers of the three partner institutions in field work, language documentation, corpus linguistics, computational linguistics, and computer science, our main goal in this project is to build a digital platform to accommodate oral corpora for South American endangered languages. Oral corpora serve as an essential empirical foundation for linguistic research, and they also provide a foundation for literacy projects for the local communities, raising their own meta-linguistic and cultural awareness. Our ultimate goal is to implement automatic translation of Kadiwéu and Nheengatu. Kadiwéu is a language with quite complex morphology, and this is a pioneer attempt in machine translation in Brazil.

1 Statement of Significance

Brazil's native languages, already severely endangered before the pandemic, face increased menace from the death of native elders by COVID-19. Our project aims to advance inclusive digital innovation by developing a computational platform for indigenous language materials that are culturally and grammatically significant. These materials will provide a digital foundation for enhanced bilingual education, intergenerational connections, and the transmission of ancestral knowledge within indigenous communities. The ultimate goal of the project is language (re)vitalization, which implies the following secondary objectives:

a. Provide training and digital resources for language fieldworkers using methods from anthropology, historical linguistics, formal linguistics, and computational linguistics.
b. Build oral corpora for endangered languages, with multilevel linguistic analysis and translations to Portuguese. The primary indigenous languages involved are Kadiwéu and Nheengatu.
c. Develop automated tools for scalable digitization, syntactic parsing, and archiving of linguistic resources.
d. Implement computational grammars of the two Indigenous languages to support machine translation, linguistic hypothesis testing, semantic parsing, and language documentation, reconstruction and generation.
e. Elaborate a pedagogical grammar of Kadiwéu to support language reintroduction.

The main contribution of the project will be digital online corpora for Kadiwéu and Nheengatu. These corpora will have automatic syntactic and morphological annotations as well as translations into Portuguese. A second contribution will be native language grammars that are guided by the annotation in the proposed corpora. These grammars will also serve as

teaching tools to improve local speakers literacy in their own languages, as well as the preservation of culturally significant narratives. Moreover, a better study of the intricate correspondences between European and native languages will also improve indigenous language teachers training for a better and more balanced bilingual education, with positive consequences for the teaching of/in Portuguese and the access of native language speakers to higher education. An indigenous school teacher, who is also a PhD student in Linguistics, is part of our team to guarantee a productive interaction with native communities, in order to achieve their needs in education. Therefore, we expect the following social impacts:

a. Transfer of grammatical tools from expert linguists to local communities, focusing on the usefulness of these tools to indigenous teachers and language experts, facilitating the access of local experts to Brazilian higher education
b. Improving the preservation of indigenous languages, cultures, and traditions.
c. Strengthening the linguistic awareness of teachers and their communities, offering native communities an active role in the systematic study of their own languages, cultures, and identities.

One preliminary version of Kadiwéu's digital corpus[1] is already available at the Tycho Brahe Platform (henceforth cited as *the platform*), a web browser-based computational system with tools for search, visualization and edition of linguistic corpora, integrated with a multi-level morphosyntactic tagger and parser mechanisms which provides ways for analyzing narratives from sentence level, down to word's internal structure. This pilot material shows our engagement and strength to achieve larger digital corpora of languages that are rapidly weakening their transmission to the next generation.

2. Scientific and technological challenges

It is important to stress that this project is based on a conception/philosophy of corpus building and use which is different from the mainstream in computational approaches to big data which is well expressed in the following excerpt of a paper entitled "The End of Theory:

1 https://www.tycho.iel.unicamp.br/c/kadiweu

The Data Deluge Makes the Scientific Method Obsolete", published in 2008 in the journal *Wired Magazine*: [2]

> "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."
>
> (Chris Anderson, Wired Magazine 16:7, 2008)

We assume a diametrically opposed position which disputes that size, and good mathematical tools are sufficient to guarantee the relevance of corpora for research and applications to educational and social purposes. We claim that big corpora are useless if they do not contain extra information that allows researchers to retrieve the data they need to address their questions about language and its dynamics. This extra information is added through annotation, and annotation is based on linguistic models. What we need is to develop methods that permit the annotation of great quantities of texts in a rapid and reliable way. This implies interdisciplinary work with computer scientists. This also implies to explore all the opportunities that the web offers, in particular, collaborative work.

3. Roles and responsibilities

Our team is composed of Filomena Sandalo, who will coordinate the project, a specialist in language documentation; Charlotte Galves, the initiator and coordinator of the Tycho Brahe Parsed Corpus of Historical Portuguese and the Tycho Brahe Platform. In collaboration with Filomena Sandalo and Leonel F de Alencar Araripe, she will supervise the Part of Speech and syntactic annotation of Kadiweu and Nheengatu. Leonel F. de Alencar Araripe will be the main responsible for the development of the automatic translation mechanism. Dr. Araripe is a full professor from the Universidade Federal do Ceará who has already successfully worked on Nheengatu automatic translation. Our team also includes Michael Becker, who is a specialist in computational

[2] This section is based on Charlotte Galves talk at the TRANS-ATLANTIC PLATFORM (-AP) Workshop: Digital Scholarship in the Social Sciences and Humanities: New Forms of Data for Research, held at the NSH, Washington DC, in January 2015.

linguistics from the University of Massachusetts, crucial for a corpora project. Dr. Becker will collaborate on dictionaries tools and on the linguistic analysis of the languages. Pablo Faria is specialist in language acquisition, has computational experience with editing and parsing of texts and also with automatic methods for inconsistency detection in annotated corpora. Two PhD students are also part of the team: Luiz Veronesi is a computer scientist studying computational linguistics who is and will be responsible for the development of the Tycho Brahe platform. Vanda Pires is a native speaker of one of the languages approached and she is a member of the Kadiwéu community of Mato Grosso do Sul. Vanda Pires will be responsible for the contact with the native speakers needs for language vitalization and education. She will also collect original narrative and translate data according to the needs of the native populations. Moreover, we will have a collaborator, Fabio Kepler, PhD from the São Paulo University in Computer Science, current at Unbabel, Portugal. He will collaborate with the efforts on automatic translation. That is, we have an integrated team essential for the development of the proposal. Fabio Kepler is the author of the automatic tagger currently used in the annotation of the Tycho Brahe corpus.

The members of this project have long collaborations. Sandalo has been part of two team projects (thematic projects) coordinated by Galves funded by FAPESP and they developed in collaboration the pilot corpus of Kadiwéu in the Tycho Brahe platform. Pablo Faria was one of the developers of eDictor in collaboration with Galves' projects. Sandalo, Becker, and Galves have participated together in the project "Putting Fieldwork on Indigenous Language to New Usages" funded by FAPESP (São Paulo Advanced Science School). Becker and Sandalo have worked together in a thematic project funded by FAPESP and coordinated by Sandalo (Edges and Asymmetries in Phonology and Morphology) and have joined publications. Leonel Araripe has joining this team for the first time bringing new ideas and experience in computational linguistics applied to indigenous languages.

This is a proposal for Computer Science and Linguistics, which would be enough already, but our goal is broader: it is a proposal for languages in danger of extinction, without leaving out cultural aspects and the inclusion of a member of one of the communities of this project to work actively in language vitalization.

4. Goals

Our main goal in this project is to build a digital platform to accommodate oral corpora for South American endangered languages. Oral corpora serve as an essential empirical foundation for linguistic research, and they also provide a foundation for literacy projects for the local communities, raising their own meta-linguistic and cultural awareness.

The data will be archived on the Tycho Brahe Electronic Platform (tycho.iel.unicamp.br). 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. It also includes a number of different functions: multimedia resources, sentence-by-sentence translation, morphological tagging, syntactic parsing tools, search mechanisms, rich metadata, statistical results with history archiving, and a complete workflow for syntactic trees edition, as well as tools for importing and exporting corpora documents. The platform is based on syntactic microstructures stored in a cloud-based syntactic database and parametric computational rules that allow a rich user experience. It will be further developed within this project, including mechanisms and the integration of an automatic machine translator tool, which will be specifically developed during this project.

Our product is an important tool for research and linguistic preservation, especially given the acute shortage in language materials and computational materials for South American languages.

Automatic translations to Kadiwéu, Nheengatu, and Portuguese is our final goal. At this point we do not have anything implemented in such direction for Kadiwéu. But we have a specialist is our team.

4.1. The Tycho Brahe Platform: language gathering, annotation, and comparison methodology

The Tycho Brahe Platform, which won the 2nd prize of the Abralin Award for technological innovation in 2022, is a web browser-based computational system with tools for search, visualization and edition of linguistic corpora, integrated with a multi-level morphosyntactic tagger and parser mechanisms which provides ways for analyzing narratives from sentence level, down to word's internal structure.

DACILA

It is still under development (with some tools already available to use), and it will facilitate the collection, annotation, and comparison of languages when it reaches its full operational capabilites. The development team (led by Prof. Galves) has been working since 2012, initially supported by the FAPESP grant # 2012/0678-9. This proposed project will improve and expand the existing functionalities. The platform joins and complements parallel efforts such as ANNIS (corpus-tools.org/annis), developed at the Humboldt University, Berlin, and applied to a variety of high-resource languages such as German, Arabic, and others. Similar to Tycho, ANNIS is a browser-based search and visualization architecture for complex multilayer linguistic corpora with diverse types of annotation. One advantage of the platform is its integrated multi-level tagger and parser tools that analyzes both words and the building blocks of words, making it particularly suitable for the highly complex word structure of the agglutinative languages of South America.

The platform is a pioneer in its application to indigenous languages of South America, and its web-based interface allows immediate dissemination to indigenous communities, allowing the creation of corpora significantly larger than previously possible. The platform was created and is being entirely developed in South America by South American researchers for the benefit of linguistic work here.

The current proposal, also built on the FAPESP grant # 2012/17869-7, produced a pilot corpus in the Kadiwéu language (see Galves, Sandalo, Sena & Veronesi 2017), that will serve as a model for a larger corpus built in the current project (tycho.iel.unicamp.br/c/kadiweu).

Kadiwéu is a language from the Guaikurúan linguistic family, spoken in the Chaco area of South America. Of the 1,500 members of the community, distributed over an area of 538,000 hectares, only about 50 families, or about 200 people of all ages, still speak the language, which is in great danger of extinction. Only three other languages of the Guaikurúan family are still spoken: Toba, Pilagá, and Mocoví, all in Argentina. The existing Kadiwéu corpus, built along the model of the Tycho Brahe Corpus, currently contains 15 narratives. The tagging system of the *Penn Parsed Corpora of Historical English* (cf. Santorini 2022) had been already expanded to accommodate the rich inflectional morphology of Portuguese, each word being tagged with its primary category and optionally with one or more secondary tags that encode morphological properties (see Britto et al. 2002). This two-level system was not yet rich enough for the morphology of Kadiwéu, requiring further layers of labeling. In Kadiwéu, words can be decomposed into multiple smaller units, each with its own meaning, requiring its own label. In effect, in Kadiwéu and other agglutinative languages, the components of words are treated similarly to independent words, requiring parallel layers of labels (cf. Galves et al. 2017) .

The platform was expanded to allow a wider range of word tags, in order to accommodate the informationally rich inflectional morphology of Portuguese by using the part-

of-speech tagging system of the *Penn Parsed Corpora of Historical English*. In this expanded system, each word is tagged with its primary category and optionally with one or more secondary tags that encode morphological properties (see Britto et al. 2002). This two-level system was not rich enough for the morphology of Kadiwéu, requiring further layers of labeling. In Kadiwéu, words can be decomposed into multiple smaller units, each with its own meaning, requiring its own label. In effect, in Kadiwéu and other agglutinative languages, the components of words are treated similarly to independent words, requiring parallel layers of labels.

The automatic tagging process that is currently being piloted consists of two levels: first, the tagger runs at the sentence level, assigning a part-of-speech tag for each word. Second, the process runs inside of each relevant word, assigning tags to the internal constituents of words, based on the lexicon extracted from the narratives. This system forms the basis for the syntactic annotation of texts without loss of the detailed morphological information typical of the richly structured languages of the Americas. The automatic tagger under development uses machine learning technology ("artificial intelligence"), learning from manually tagged sentences.

Moreover, the platform intends to provide two different approaches for searching over documents and corpora. The first one is a word/POS tag linear mechanism that allows the user to search for any word either its specific value or its POS tag, i.e., the user may search for a specific word at a specific position in the sentence. The second approach will consist in a purely syntactic search with a simple graphical interface for building blocks to generate queries for searching over the structures stored into a NOSQL database, in order to retrieve all sentences that match the given criteria. Both mechanisms may also match the results based on the document's metadata information like geographical and historical information, literary genre and others. Metadata are defined and configured per corpus and may be shared among corpora. Another important feature for those mechanisms is that they allow the users to continue searching over the results of a query as many times as possible and also choose which mechanism to use. For example, the user may first realize a linear search and then a syntactic one over the results, and vice-versa, allowing many kinds of comparison between the languages in the project.

In the future, we aim to use our tools in other languages of Brazil. To sum up, our goal is to produce flexible tools that are customizable to accommodate data from many different language families and typological configurations, based on the insights of modern parametric grammatical theory. Such tools will also facilitate the analysis of data from new languages (especially morphological and syntactic analysis), allowing the researcher to uncover significant patterns. This kind of work will also contribute to language preservation and

education. Finally, our design of the language corpora will allow comparative studies among Guaikurúan languages as well as other indigenous or poorly documented languages around the world.

The development status for each platform tool is described below.

Tool	eDictor
Description	It contains functionalities for text and audio transcription, word and morpheme level edition, different ways of document organization using categories and metadata, integration with tagger and parser mechanisms, translations, syntactic tree visualization, and more.
Done	Major edition tools are available for use. Document organization, metadata edition, corpus access control (for different roles and corpora), tagger and parser integration, audio and image upload. Importing of documents generated in other tools (eDictor, Flex, PSD and text files). Transcription tool integrated with browser language for spell checking. OCR for text recognition. Basic lexicon management with sentence extraction.
To do	Search in document, apply editions to multiple words, inline edition for faster data input, and 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.

Tool	Browser
Description	Tool for corpora visualization.
Done	Visualization available for all types of supported corpora, shareable short links to corpora, documents, pages and sentences. Only public corpora may be visualized in browser, this is set on the corpus configuration and the management tool.
To do	Document setting for public/private visualization. General layout improvements for best user experience. Testing and bug fixes.

Tool	Search
Description	This tool contains mechanism for searching by a combination of words, POS tags and syntactic structure with tree visualization and links to check

	the sentence in the document (browser and eDictor). Search also allows to query sentences based on document metadata or specific corpus document. It does not include a search over multiple corpora.
Done	Word and POS tag search. Syntactic search basic mechanism with a very restricted set of functions. It already allows to choose for specific documents to search.
To do	Development of selected functions for syntactic search, based on Corpus Search. Usage of document metadata for criteria selection. Improvement of mechanism to allow morpheme search. Search over results of another search.

Tool	Parser
Description	This tool contains the mechanisms for POS tagging and syntactic parsing and its configurations.
Done	Creation of a rule-based parser for any platform corpus using a graphical editor with testing and debugging. Online tagger training and use of dictionary-based tagging. Configurations for tags (syntactic, morpheme and POS), inflections and splitter rules. APIs for integration with other systems.
To do	Integration with available probabilistic parsers with online training. Interface bug fixes.

Tool	IO
Description	Import and export tool that provides interoperability between the platform and third-party tools.
Done	Importing for the following: CHAT Transcription, eDictor, FieldWorks FLEx (lexicon), PDF files (with multiple pages), POS (Part-Of-Speech sentences), PSD (parsed syntactic sentences), text files. Exporting is available only for the platform format.
To do	Full export with pictures and sounds. Inclusion of more formats for both importing and exporting. General layout improvements for best user experience. Testing and bug fixes.

DACILA

Tool	Corpora Management
Description	Administrative tool for corpora creation and configuration.
Done	New corpus creation with a single click with parameter configuration allowing different visualization and usage. Configuration to set which tagger and parser the corpus may use from the available ones. Upload of PDF files for citations. Catalog information with number of sentences, words by document and full document management (edit main information, remove a document, send the document to another corpus). Full access control by associating an existing user to the corpus or creating an invitation link for single or multiple users. Access levels are for administrators, coordinators and editors, which one with restrictions applying to all other tools. Configuration for tiers (word, morphemes and sentence), translations, hierarchical categories and general system parameters. Metadata management allowing creation of different types (date, geographical, list of values, number, text, year) used for specific tools (search, catalog, import). Backup tool processing isolated for each corpus with external storage.
To do	General layout improvements for best user experience. Testing and bug fixes.

Tool	Syntrees
Description	Tool with graphical interface for syntactic tree revision.
Done	Revision workflow controlled by status. Multiple POS tag review. Complete graphical interface for syntactic tree revision integrated with the parser tool allowing realtime parsing with step-by-step revision.
To do	Integration with other parsers. Configuration sharing for rule-based parser revision. General layout improvements for best user experience.

Tool	Portal
Description	Main Tycho Brahe Platform site containing user profile page, access to the tools, project information and request to corpus creation. Used for communication with community.
Done	User profile page with easy access to the corpora and tools. Access link for new users and invitations for corpora collaboration.

To do	Implementation of new design, improvement for profile page with detailed instructions on how to use the platform and access to tutorials. Form to request creation of new corpus and step by step configuration for a new corpus.

4.2. Pedagogical Grammars

This project aims to support language documentation efforts of several indigenous communities in Brazil by creating a corpus with multi-level parallel analyses of auditory and written texts. Words and phrases will be analyzed at multiple levels of granularity in terms of their sounds and meanings, which may overlap and mismatch. This type of multi-analysis requires us to uncover the building blocks of the sounds and meanings of the project's language.

We will work on uncovering the elements that build words of Kadiwéu and Nheengatu: consonants, vowels, accents, stress, and intonation. We start by recording native speakers as they conjugate verbs and decline nouns, and we analyze the recordings acoustically, structurally, and computationally. The work builds on the results of an ongoing collaboration between Prof. Becker and Prof. Sandalo using computational tools that Dr. Becker developed at UMass Amherst in R and Python. Further computations at UMass will allow the analysis to be integrated into the platform and apply the word-level grammatical analysis to the narratives hosted by it.

As mentioned before, the platform will contain a tool for syntactic search. In order to feed this system, the program is trained with a toy grammar of the language. This toy grammar is a syntactic analysis of the sentences in the narratives of the corpora. The syntactic annotation in this system will feed the elaboration of the pedagogical grammars and dictionaries of the languages.

4.3. Machine Translation and Grammar Engineering

The first attempts at machine translation date back to the early 1950s (Booth, 1955/2003; Zarechnak, 1959/2003). Given the limitations of electronic equipment at the time, the first results were only experimental. The following years saw progress in the area driven by the evolution of digital technology and computational linguistics, a discipline focused on the mathematical modeling of linguistic knowledge. To meet the demand of government agencies and companies, several approaches were proposed and several systems were implemented for translation between major languages such as Russian, German, English, French, and Japanese.

Among these systems, Systran stands out (Loffler-Laurian, 1996). Created in 1968 and still in use today, it has had different versions and implementations, for example at Xerox and the Nuclear Energy Research Center Karlsruhe. Using a rule-based approach, it was used for online translation on the internet by Alta Vista in 1997 and initially by Google in the next decade, before being replaced by this company's system in 2007 (Oliveira & Anastasiou 2011). It currently adopts a neural architecture. Its repertoire of 50 languages features Brazilian Portuguese as the only language of South America.

Since the 1990s, and especially in the past decade, great advances in the area of machine learning have had a strong impact on machine translation. As a result, symbolic approaches, based on the modeling of linguistic structures by means of hand-written rules, became restricted to the translation of texts in very specific genres and domains, typically using a controlled language.

In machine learning approaches, rules for analysis and generation of the source and target language are not manually formulated. Instead, the algorithms are fed a massive volume of texts in the languages to be translated. This allows them to determine the highest probability that an expression *x* in the source language is equivalent to an expression *y* in the target language. This type of approach handles the problem of ambiguity better than rule-based approaches. However, the choice of the translation equivalent of an ambiguous term is not always felicitous, but rather reflects a bias in the training data. As the quality of translations fundamentally depends on the number of texts available for training the models, only a small fraction of the world's languages have been able to take advantage of this technology. Besides, the performance of the systems varies across the languages covered by the technology in proportion to the volume of texts used in the training.

In the beginning, the data-driven approach had the main advantage of robustness, enabling, in many cases, only a general understanding of the original text and demanding a large number of post-editing to achieve the quality of a human expert’s translation. In recent years, however, there has been a significant increase in the quality of results, not only due to the greater volume of training data but above all due to the development of new approaches in machine learning applied to translation. A milestone was Google's 2016 neural deep learning approach (dubbed GNMT), often capable of producing results ready to be published or requiring minimal post-editing (see Wu et. Al. 2016).

Technological and scientific advances have made machine translation more and more beneficial not only for ordinary people (Vieira; O’Sullivan; Zhang; O’Hagan, 2022) but also for professional human translators, as it reduces the translation effort to post-editing (Jia; Carl; Wang, 2019). This language technology has become part of the daily life of both government

and business organizations, shortening the product launch cycle and making documents accessible to a wider audience (Loffler-Laurian, 1996).

With the cheapening and expansion of the reach of internet access, people in the most remote places, through mobile devices with modest processing capacity, can make use of free instant translation services for both written and spoken language. Machine translation is also an ally in language teaching and learning. With all this, this technology becomes a factor for the survival of smaller languages, since it favors the use of the language by facilitating the translation to and from a majority language.

Despite enormous recent advances, the technology, however, still suffers from limitations in translation quality. First, it does not work equally well for all language combinations (Aiken, 2019) and performs poorly in certain communication contexts, e.g., healthcare (Taira; Kreger; Orue; Diamond, 2021). Some of these problems stem from the GNMT architecture itself, which does not take into account the situational context necessary to understand the language, as observed four years ago by Hofstadter (2018), who exemplified this deficiency with the translation into French of the following text:

> In their house, everything comes in pairs. There's his car and her car, his towels and her towels, and his library and hers.

Google's translation of this text into French hasn't improved since then. It still fails to adequately convey the idea of a male-female couple with separate belongings: "Il y a sa voiture et sa voiture, ses serviettes et ses serviettes, et sa bibliothèque et la sienne."

Covering just a few languages at first, Google Translate has expanded its coverage considerably in recent years, now translating between 133 languages. Even so, it covers only a tiny portion of the world's linguistic diversity, as the company itself acknowledges (Caswell; Bapna, 2022). Only in June 2022 did the system incorporate three Amerindian languages into its repertoire: Aymara, Quechua, and Guarani (Caswell, 2022). All three are languages with more than 1 million speakers, the last two having more than six million speakers each (Eberhard; Simons; Fennig, 2022). In comparison to most Amerindian languages, they are relatively high-resource languages with a considerable presence on the Internet, for example, on Wikipedia. Of the other many Amerindian languages, only Yucatec Maya, with around 800,000 speakers, is in the beta development stage, which means that it will soon be supported by Google Translate. The company, however, has no schedule to begin developing machine translation for other Amerindian languages with fewer speakers such as Cree, reported being spoken by 96,000 people in Canada (Beattie, 2021; Chidley-Hill, 2021; Hilleary, 2021).

Therefore, it is uncertain whether the tool will ever enable the translation of South American minority languages with few or no digital text resources (Hilleary, 2021), such as the Língua Geral Amazônica, also known as Modern Tupi and Nheengatu, which has only 14,000 speakers, 6,000 of which in Brazil (Eberhard; Simons; Fennig, 2022). For languages with far fewer speakers, such as Kadiwéu, the situation is even more serious. Indeed these two languages were not included in a recent experiment involving 1500 languages targeted at extending coverage of Google Translate to additional 1000 languages (Bapna et al, 2022).

Fields inaugurated in the USA and European countries in the late 1950s, computational linguistics and natural language processing, which include machine translation, arrived in Brazil with about three decades of delay. In this country, research in these areas has focused on the development of tools and resources for the computational processing of Portuguese (Pardo et al., 2010). Research on machine translation has focused on majority language pairs, with Portuguese as the source or target language (Aziz; Pardo; Paraboni, 2008; Soares, 2019).

In pioneering work in Brazil, Alencar (2021) developed GrammYEP, the first automatic translator of a Brazilian indigenous language, in this case, Nheengatu. The tool adopts the pivot approach and was implemented in the *Grammatical Framework* (GF) formalism. It translates between Nheengatu, Portuguese, and English. Compared to the transfer approach, the pivot approach substantially reduces the complexity of implementing a machine translation system for three or more languages (Boitet, 1988/2003; Lima; Nunes; Vieira, 2007).

GF is a typed functional programming language based on the Haskell language (Ranta, 2011). The fundamental operations of this language boil down to function application, which allows building objects of certain types from objects of other types. GF was designed specifically for the elaboration of multilingual grammars, having been successfully tested in the implementation of computer grammars of languages of different families and types of languages, including most European languages (from Slavic languages to Romance and Germanic through Basque, Finnish, Hungarian, and Maltese), Turkish, Chinese, Japanese, Persian, Hindi, Swahili, Somali, etc.

The central component of GrammYEP is a so-called abstract syntax, which models the conceptual organization of the three languages involved. In this syntax, primitive types are defined such as *SimpleKind*, *Quality,* etc. These types correspond to concepts expressed by lexical roots in at least one of the three languages, as exemplified in Table 1. The last three types are functions that apply to an Item (an entity or a set of entities) to produce a Location (expressed in English, for example, by the PP *in that house*).

Type	**Concept**	**Nheengatu**	**Portuguese**	**English**

SimpleKind	House	*uka*	*casa*	*house*
SimpleKind	Daughter_Of_Man	*taiera*	*filha*	*daughter*
SimpleKind	Daughter_Of_Woman	*mimbira*		
SimpleKind	Son_Of_Woman		*filho*	*sound*
SimpleKind	Son_Of_Man	*taira*		
Quality	Red	*piranga*	*vermelho*	*red*
Item → Location	in	*upé*	*em*	*in*
Item → Location	on	*upé*	*em*	*on*
Item → Location	near	*ruaki*	*perto*	*near*

Table 1: Some concepts and their linearization in Nheengatu, Portuguese, and English.

As Table 1 shows, primitive concepts are defined relative to one or more specific languages. The term *casa* and its equivalents in English and Nheengatu express the primitive concept of *House.* The three languages coincide in the designation of the *Red* concept, expressing it in simple lexemes. They differ, however, in the lexicalization of concepts denoting immediate and collateral descendants of an ego. In Nheengatu, these terms lexicalize the gender of the ego, unlike English and Portuguese, which only lexicalize the gender of the alter. In the expression of the spatial location of a Figure relative to a Ground, Portuguese (Brazilian) and Nheengatu coincide in neutralizing the distinction between *In* and *On*, expressed in English by the corresponding prepositions *in* and *on*.

From primitive concepts, complex concepts are built up in abstract syntax. For example, a State is formed from a Quality or a Location. A Quality, in turn, can be constructed from another Quality through a function expressed by an intensifier. Entities or groups of entities can be formed by combining *Kind* with the types *Deitic*, *PossPro*, and *Num*, linearized in English by demonstratives, possessive pronouns, and grammatical number. Sentences (1)-(4) in Nheengatu exemplify these processes of conceptual construction. The examples are accompanied by some of the Portuguese and English translations generated by GrammYEP:

(1) *igara pisasu retana*
a canoa é muito nova
the canoe is very new
(2) *i mimbira-itá ruka-itá uiku paranã ruaki*

as casas dos filhos dela estão perto do rio
the houses of her sons are near the river
(3) *pe rendaua suaki*
a comunidade de vocês é perto dela
your community is near her
(4) *tapiira-kunhã kambi-yukisé saku uiku*
o leite da vaca está quente
the milk of the cow is hot

Following the GF architecture, automatic translation between the three languages in GrammYEP is made possible by the respective concrete syntaxes. These specifications mathematically model the syntactic, morphological, and lexical structures of each language, expressing the mapping onto the conceptual structures modeled in the abstract syntax. Note that GF allows one to easily handle large structural discrepancies between languages. For example, in (3), while the third person possessive is realized as a prenominal possessive determiner in English and a postnominal prepositional phrase in Portuguese, in Nheengatu it is realized as the relational prefix *s*.

An important advantage of the GF formalism and programming language is that the developer of a machine translation system does not need to elaborate the parsing and generation algorithms, their only task is to formulate the abstract syntax, on the one hand, and the concrete syntaxes of the languages involved, on the other. The system automatically produces the abstract syntax representations for each grammatical source language sentence. From these representations, it generates the corresponding target language sentences.

GrammYEP is still a limited-range prototype. It only handles qualifying predication, translating sentences with a limited vocabulary that describe states and locations of people and things, as in (1)-(4). In the field of referential expressions, it models modification by adjectives, demonstratives and genitive complements, not handling yet, e.g., adjectival relative sentences. Another limitation of GrammYEP is that it does not disambiguate ambiguous sentences, producing all target language linearizations corresponding to the conceptual representations generated for the source language sentences.

Despite small coverage, the concrete syntax of Nheengatu successfully deals with several grammatical phenomena of reasonable computational complexity, which give it the identity of a legitimate descendant of Tupi, e.g., postpositions, the genitive construction, the distinction between active and inactive verbs, relational prefixes, and transcategorial morphology (Rodrigues; Cabral, 2011; Moore, 2014; Cruz; Magalhães; Praça, 2019; Magalhães; Praça; Cruz, 2019).

DACILA

In grammar engineering, one works with two test sets to evaluate a grammar implementation: a positive test set and a negative test set, consisting, respectively, of grammatical sentences and ungrammatical sentences (Butt et al., 1999). The first set makes it possible to measure the coverage of the grammar, while the second makes it possible to assess whether it is sufficiently constrained. The negative set derives from the first by injecting, into the grammatical examples, violations of the formulated grammatical rules. GrammYEP's two Nheengatu test sets contain 142 and 171 examples, respectively. All sentences from the first set are parsed, while only two from the second are parsed.

In this project, we first intend to incorporate Kadiwéu (ISO 639-3 code kbc) into GammYEP, 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 implementing 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 (Zelle, 2009).

In the second phase of the project, the abstract syntax of GammKYEP will be successively expanded to encompass all Nheengatu grammatical phenomena and vocabulary discussed in Navarro (2020)’s 13 lessons. At the same time, the concrete syntaxes of Nheengatu, Portuguese, and Kadiwéu will be expanded. The basis of these implementations will be positive and negative test sets for the three languages. To this end, the examples of Nheengatu that were built based on Navarro (2020) will be translated into Kadiwéu. Negative versions of these test sets will also be produced. GrammYEP's concrete English syntax will not be developed in the project, as the English translation can be done from the Portuguese translation through available automatic translators, such as Google.

In the third and final phase of the project, grammatical phenomena of Kadiwéu and Nheengatu that were not yet included in the test sets of the previous phases will be chosen, repeating the previously adopted development methodology.

Besides the machine translation system itself, the project will provide the following additional contributions:

1. Computational grammars of Nheengatu, Kadiweu, and Portuguese.
2. A multilingual treebank derived from the aligned positive test sets and their translations.

The grammars will allow automatic testing of theoretical and descriptive claims about the two languages as well as to carry out formal comparisons between them. These grammars

can also be used to generate examples in the two indigenous languages with their translations into Portuguese. This data, together with the test sets and other available annotated corpora, can be used to train or evaluate machine learning models, which, in turn, can perform a wide range of natural language processing tasks, e.g., part-of-speech tagging, parsing, and machine translation.

5. Working Plan

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.

2024: Development of selected functions for syntactic search, based on Corpus Search. Usage of document metadata for criteria selection. Improvement of mechanism to allow morpheme search. Search over results of another search.

2025: Integration with available probabilistic parsers with online training. Interface bug fixes. Full export with pictures and sounds. Inclusion of more formats for both importing and exporting. General layout improvements for best user experience. Testing and bug fixes.

2026: General layout improvements for best user experience. Testing and bug fixes. Integration with other parsers. Configuration sharing for rule-based parser revision. General layout improvements for best user experience.

2027: Implementation of new design, improvement for profile page with detailed instructions on how to use the platform and access to tutorials. Form to request creation of new corpus and step by step configuration for a new corpus.

2. Machine Translation

2023: Incorporate Kadiwéu (ISO 639-3 kbc) into GammYEP, 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.

2024-2025: In the second phase of the project, the abstract syntax of GammKYEP will be successively expanded to encompass all Nheengatu grammatical phenomena

and vocabulary discussed in Navarro (2020)'s 13 lessons . At the same time, the concrete syntaxes of Nheengatu, Portuguese and Kadiwéu will be expanded. The basis of these implementations will be positive and negative test sets for the three languages. To this end, the examples of Nheengatu that were built on the basis of Navarro (2020) will be translated into Kadiwéu. Negative versions of these test sets will also be produced. GrammYEP's concrete English syntax will not be developed in the project, as the English translation can be done from the Portuguese translation through available automatic translators, such as Google.

(2025-2027): In the third and final phase of the project, grammatical phenomena of Kadiwéu and Nheengatu that were not yet included in the test sets of the previous phases will be chosen, repeating the previously adopted development methodology.

Besides the machine translation system itself, the project will provide the following additional contributions since the beginning of its implementation:

1. Computational grammars of Nheengatu, Kadiweu and Portuguese: This will allow one to automate the testing of theoretical and descriptive claims about the two languages as well as to carry out formal comparisons between them.
2. A multilingual treebank derived from the aligned positive test sets and their translations.

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.

2024: Elaboration and testing of a rule-based parser for Nheengatu. Parsing and revision of the Kadiwéu documents.

2025-2027: amplification of the parsed corpora of Kadiwéu and Nheengatu.

6. **Dissemination and Training**

The language data collected during the project will be shared with the community on external hard drives bought with the research stipend. One hard drive per village will be left under the care of the indigenous school teachers. If the community considers that the schools need printed versions of the narratives, they will be provided. The platform will support easy download of sentence by sentence including annotations. Moreover the platform's corpora are **online and easily accessible**. In addition, all researchers or educators will have access to our work via open-source published papers with our results on historical comparison of languages of the Chaco area of South America.

The platform is user friendly, but online documentation will be available. To further increase the dissemination of results from this project, the applicant commits the group to teach a course in the communities so that native speakers can participate in their language's documentation. It will help spread this technology to researchers, indigenous community members and activists in South America.

The applicant has already promoted a FAPESP Advanced Science School in the University of Campinas, already mentioned above (Putting fieldwork on indigenous language on new uses) under the FAPESP grant # 2015/18995-4:

http://espca.fapesp.br/school/sao_paulo_school_of_advanced_science_on_putting_fieldwork_on_indigenous_languages_to_new_uses/59/

This event had young professionals (post-doctoral fellows, PhD students and early career professionals) coming from all over the continents. Other similar events will be promoted. We plan one additional for 2025. Moreover, the project will welcome students from all levels to enroll the project in any of the three universities involved.

7. **References**

Aiken, M. (2019). An updated evaluation of Google Translate accuracy. *Studies in Linguistics and Literature*, 3 (3), 253-260.

Aziz, W. F., Pardo, T. A. S. & Paraboni, I. (2008). An experiment in Spanish-Portuguese statistical machine translation. In G. Zaverucha & A. Loureiro da Costa (Eds.): *Advances in Artificial Intelligence: Proceedings of the 19th Brazilian Symposium on Artificial Intelligence (SBIA 2008)* (pp. 248–257). Springer.

Bapna, A. et al. (2022, July 6). *Building machine translation systems for the next thousand languages*. arXiv. https://doi.org/10.48550/arXiv.2205.03983

Beattie, S. (2021, February 23). *Google Translate's exclusion of indigenous languages a 'squandered' opportunity*. Huffpost. https://www.huffpost.com/archive/ca/entry/google-translate-cree-indigenous-language_ca_6035242ac5b67c329620c3e3

Britto, H., Finger, M., Galves, C. (2002) Computational and linguistic aspects of the construction of theTycho Brahe Parsed Corpus of Historical Portuguese. Romance Corpus Linguistics - Corpora and Spoken language. Tubingen: Narr.

Boitet, C. (2003). Pros and cons of the pivot and transfer approaches in multilingual machine translation. In S. Nirenburg, H. Somers & Y. Wilks (Eds.), *Readings in machine translation* (pp. 273-279). MIT Press. (Original work published 1988)

Booth, A. D. (2003). Mechanical translation. In S. Nirenburg, H. Somers & Y. Wilks (Eds.), *Readings in machine translation* (pp. 19-20). MIT Press. (Original work published 1955)

Butt, M. et al. (1999). *A Grammar writer's cookbook*. CSLI.

Caswell, I. (2022, May 11). *Google Translate learns 24 new languages*. https://blog.google/products/translate/24-new-languages/

Caswell, I. & Bapna, A. (2022, May 11). *Unlocking zero-resource machine translation to support new languages in Google Translate*. Google AI Blog. https://ai.googleblog.com/2022/05/24-new-languages-google-translate.html

Chidley-Hill, J. (2021, February 2021). *Online petition asks for Cree language to be added to Google Translate*. CTV News. https://www.ctvnews.ca/sci-tech/online-petition-asks-for-cree-language-to-be-added-to-google-translate-1.5317866

Cruz, A., Magalhães, M. M. S. & Praça, W. N. (2019). A morfologia transcategorial e sua relação com o padrão omnipredicativo em línguas da família Tupi-Guarani. *ReVEL, 17* (32), 69-94.

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).

Galves, C.; Andrade, A.; Faria, P. (2017). Tycho Brahe Parsed Corpus of Historical Portuguese, phase III, University of Campinas, Brazil. URL: http://www.tycho.iel.unicamp.br/corpus/index.html

Galves, C. Sandalo, F. Sena, T. & Veronesi, L. (2017). Annotating a polysynthetic language: from Portuguese to Kadiwéu. Cadernos de Estudos Linguisticos 59.3.

Hilleary, C.. (2021, April 09). *Google explains why app can't translate most native American languages*. Voice of America News. https://www.voanews.com/a/usa_google-explains-why-app-cant-translate-most-native-american-languages/6204275.html

Hofstadter, D. (2018, January 30). *The shallowness of Google Translate*. The Atlantic. https://www.theatlantic.com/technology/archive/2018/01/the-shallowness-of-google-translate/551570/

Jia, Y., Carl, M., & Wang, X. (2019). How does the post-editing of neural machine translation compare with from-scratch translation? A product and process study. *Journal of Specialised Translation*, (31), 60-86. https://www.jostrans.org/issue31/art_jia.pdf

Lima, V. L. S.; Nunes, M. G. V.; Vieira, R. (2007). Desafios do processamento de línguas naturais. In *SEMISH-Seminário Integrado de Software e Hardware*. https://www.inf.pucrs.br/linatural/Recursos/Desafios.pdf

Loffler-Laurian, A-M. (1996). *La traduction automatique*. Presses universitaires du Septentrion.

Magro, C. ; Galves, C. (2019) Portuguese Syntactic Annotation Manual. Lisboa/Campinas: Centro de Linguística da Universidade de Lisboa/Instituto de Estudos da Linguagem. URL: http://alfclul.clul.ul.pt/portuguesesyntacticannotation/home.html

Magalhães, M. M. S., Praça, W. N. & Cruz, A. (2019). Gradação da omnipredicatividade na família Tupi-Guarani. *Forma y Función, 32*(2), 151-189. doi: http://dx.doi.org/10.15446/fyf.v32n2.80818

Moore, D. Historical development of Nheengatu (Língua Geral Amazônica). (2014). In S. S. Mufwene (Ed.), *Iberian Imperialism and Language Evolution in Latin America* (pp. 108-142). University of Chicago Press.

Navarro, E. A. (2020). *Curso de Língua Geral (nheengatu ou tupi moderno): a língua das origens da civilização amazônica*. http://www.amazon.com.br

Oliveira, R. G. & Anastasiou, D. (2011). Comparison of SYSTRAN and Google Translate for English→ Portuguese. *Revista Tradumàtica: tecnologies de la traducció,* 9, 118-136.

Pardo, T. A. S. et al. (2010). Computational linguistics in Brazil: An overview. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas* (pp. 1–7). Association for Computational Linguistics.

Ranta, A. *Grammatical Framework: Programming with multilingual grammars*. Stanford, California: CSLI, 2011.

Rodrigues, A. D. & Cabral, A. S. A. C. (2011). A contribution to the linguistic history of the Língua Geral Amazônica. *Alfa, 55*(2), 613-639.

Santorini, B. (2022) Annotation manual for the Penn Parsed Historical Corpora of English and the York-Helsinki Parsed Corpus of Early English Correspondence. URL: https://www.ling.upenn.edu/hist-corpora/annotation/index.html

Soares, F. (2019). *Machine Translation for the biomedical domain, corpora acquisition and translation experiments* [Doctoral dissertation, UFRGS]. UFRGS Digital Repository. http://hdl.handle.net/10183/199624

Taira, B. R., Kreger, V., Orue, A., & Diamond, L. C. (2021). A pragmatic assessment of Google Translate for emergency department instructions. *Journal of general internal medicine, 36*(11), 3361–3365. https://doi.org/10.1007/s11606-021-06666-z

Vieira, L.N., O'Sullivan, C., Zhang, X., O'Hagan, M. (2022). Machine translation in society: insights from UK users. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-022-09589-1

Zarechnak, M. (2003). Three levels of linguistic analysis in machine translation. In S. Nirenburg, H. Somers & Y. Wilks (Eds.), *Readings in machine translation* (pp. 325-331). MIT Press. (Original work published 1959)

Zelle, J. M. (2009). *Python programming: an introduction to computer science* (2nd ed.). Franklin, Beedle & Associates.

Yang, J. & Lange, E. D. (1998). SYSTRAN on AltaVista: A user study on real-time machine translation on the internet. In D. Farwell, L. Gerber & E. Hovy (Eds.), *Machine Translation and the Information Soup: AMTA 1998* (pp. 275-285). Springer. https://doi.org/10.1007/3-540-49478-2_25

Wu, Y. et al. (2016). Google's Neural Machine Translation System: Bridging the gap between human and machine translation. arXiv. https://doi.org/10.48550/arXiv.1609.08144