Grammar Induction for Brazilian Indigenous Languages

Diego Pedro Gonçalves da Silva
Núcleo Interinstitucional
de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas
e de Computação
diegopedro@usp.br

Thiago Alexandre Salgueiro Pardo
Núcleo Interinstitucional
de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas
e de Computação
taspardo@icmc.usp.br

Abstract

This paper investigates the issue of grammar induction for Brazilian indigenous languages, mainly focusing on unsupervised methods, but also testing a large language model for the task. Grammar induction poses several challenges, particularly when applied to low-resource languages, a characteristic commonly associated with indigenous languages. The primary objective of this paper is to discover syntactically related words in sentences. In addition to the contributions to linguistic studies, as in language description and structural analysis, grammar induction may help in varied Natural Language Processing tasks, as it could help detect parsing errors, enhance parsing results, and reveal pertinent relations for open information extraction purposes. The findings reveal that, even with a limited corpus, it is feasible to identify syntactically related words, especially for some relations. To the best of our knowledge, this represents a pioneering attempt to undertake grammar induction for Brazilian indigenous languages.

1 Introduction

In the year 2001, there were 6,981 languages spoken globally, some of which linguists predict will confront the threat of extinction by the year 2100 (Harrison, 2008). One of the reasons for this decline may be associated with political and social discrimination directed toward its speakers, thereby exerting an influence on subsequent generations. This influence may manifest as parents refraining from transmitting their native languages to their offspring, driven by concerns regarding perceived limitations in future opportunities (Harrison, 2008; Cruz, 2011). The consequences of a language extinction across social, political, and cultural spheres are profound and incalculable. The cumulative wisdom amassed across generations, transmitted exclusively through oral communication, irreversibly dissipates (Harrison, 2008).

In Brazil, according to data provided by *Instituto Brasileiro de Geografia e Estatística* (IBGE), there were 244 indigenous languages documented in the country in 2010 (Morello, 2016). Predominantly, these languages belong to the Tupi family, which comprises more than 40 distinct languages (Ferraz Gerardi et al., 2023). The expansive influence exerted by the Tupi language family constitutes the most extensive diffusion globally. This facilitates mutual comprehension among languages within this linguistic group, many of which share cognates (Ferraz Gerardi et al., 2023). Among the indigenous languages prevalent in Brazil, Ticuna, spoken by 46 thousand individuals, Guarani-Caiuá, with 43 thousand speakers, and Caingangue, with 37 thousand speakers, emerge as the most widely spoken ones according to IBGE (Morello, 2016). A considerable number of Brazilian indigenous languages are spoken by fewer than 100 individuals (Cruz, 2011).

Promoting literacy among indigenous children in their native language and attempting to digitalize their language constitutes strategic initiatives to mitigate language decline (Taylor, 1985; Azevedo, 2016). However, the rise of the internet may have hastened the extinction of indigenous languages, given that the prevalence of dominant languages significantly contributes to the functional loss of indigenous languages (Kornai, 2013). The content deficit of the indigenous languages adversely affects the development of technological tools for these languages, such as translation systems. These tools would be useful for disseminating information and facilitating learning, consequently, contributing to preserving the language.

Artificial Intelligence systems emerge as a significant initiative to contribute to the advance of language technologies (Pinhanez et al., 2023; de Lima et al., 2021). Addressing this challenge involves considering alternatives, such as the use of comparable texts to build parallel corpora[1], and the use of grammar induction for learning syntactical structuring patterns and lexical clustering for detecting semantically-related terms for a (probably low-resource) language of interest. Grammar induction is the focus of this paper.

In Natural Language Processing (NLP) applications, Grammar Induction (GI) proves useful for various tasks, including grammar checking, information extraction, and text simplification, to name a few. Grammar induction can be approached in an Unsupervised way (UGI), in a Semi-Supervised way (SSGI), or in a Supervised way (SGI). SGI methods demonstrated remarkable efficacy in many works, achieving accuracy rates exceeding 95% (Lin et al., 2022) for the English language, while their unsupervised counterparts present a considerable challenge, often falling short of this benchmark.

This study focuses on unsupervised approaches to induce grammar within the context of dependency paradigm, which seeks to model the dependency relations among syntactic elements. Illustrative instances are provided in the form of a Nheengatu sentence presented in Figure 1, along with its Portuguese translation portrayed in Figure 2. These sentences were extracted from the Nheengatu CompLin treebank (Avila, 2021) identified with ID *Avila2021:0:0:647*. The arrows delineate the relationships between two tokens, wherein the arrow originates from the head term and is directed toward the dependent term.

Good methods for grammar induction include Large Language Models (LLM) (Shen et al., 2021) and neural networks (He et al., 2018) and both methods need a huge amount of data for training. Due to the limited amount of available digital data in indigenous languages, we test two different approaches to discover related words in an unsupervised way: Dependency Model with Valence (DMV) (Klein and Manning., 2004), the most influential model in grammar induction tasks; and Mutual Information (MI), a measure that has demonstrated efficacy to retrieve syntactic structures (Futrell et al., 2019; Hoover et al., 2021). Furthermore, we also evaluate an LLM for the taks.

The investigation specifically centers on twelve indigenous languages spoken in Brazil, most of which were annotated as a part of the TuLaR (Tupían Language Resources) project within the "Universal Dependencies" (UD) framework (Nivre et al., 2020). Notably, seven of these languages are affiliated with the Tupi family. To the best of our knowledge, this is the first unsupervised grammar induction study within the domain of Brazilian indigenous languages. We provide the code from this project at Github[2].

The next section brings a brief literature review on the topic of grammar induction. Section 3 presents the methods that we test, while Section 4 shows and discusses the achieved results. Discussion and final remarks are presented in Sections 5 and 6.

2 Related Work

In recent decades, Grammar Induction has been applied in different contexts and diverse applications. Varied methodologies have been employed, with the DMV (Klein and Manning., 2004) emerging as the most prevalent and widely recognized approach. This approach was the first to surpass the right-branching baseline, wherein the rightmost word functions as the head of the immediately adjacent left word, for grammatical structure induction.

Contemporary methods involve the utilization of neural networks (He et al., 2018) and LLM (Shen et al., 2021). Nevertheless, these innovative models may exhibit limitations when applied to languages with limited resources, particularly indigenous languages, and notably in the context of dependency grammar.

A noteworthy approach is the application of the MI measure, which has been harnessed to induce constituent grammar (Solan et al., 2005), and dependency relations for languages like Japanese (de Paiva Alves, 1996) and Portuguese (da Silva and Pardo, 2023).

Several initiatives have advanced in the domain of grammatical induction for languages with limited linguistic resources. Dahl et al. (2023) introduced a method employing Womb Grammars, a technique designed for the translational mapping of languages, in which grammar has been described to languages with no grammar description, to facilitate the induction of the Ch'ol language[3].

[1] It is not rare to use the Bible for such end, as it is published in many languages.

[2] https://github.com/diegodpgs/PROPORInd

[3] Ch'ol is an indigenous language of Mexico that lacks a formally documented grammar. However, it is noteworthy that the grammatical induction methodology articulated in this study relies on the use of syntactic relations, by definition, using supervised training.

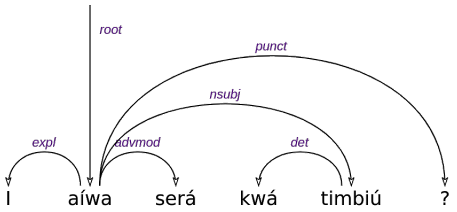

Figure 1: An example sentence for Nheengatu Language (Avila, 2021)

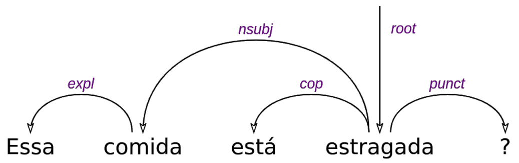

Figure 2: The translation to Portuguese of the sentence presented in Figure 1 (Avila, 2021)

In what follows we present the data and the methods that we explore in this paper.

3 Methodology

We use data from UD version 2.13[4]. This contains 245 treebanks (i.e., corpora with sentences and their corresponding syntactical dependency analyses) for 141 languages. Almost 50 languages are indigenous or ethnic representative. Of these, twelve are spoken in Brazil and nine in Russia. The twelve languages used in this work are: Akuntsu, Guajajara, Kaapor, Karo, Makurap, Munduruku, Tupinamba, Nheengatu, Apurina, Bororo, Xavante, and Madi.

All these languages include 36,322 tokens, 8,632 types, and 5,000 sentences. A detailed description of these languages is presented in Table 1. The first column describes the language used in the experiment, the second shows the linguistic family, and the third column describes the number of different Syntactic Relations (SR) used in the annotations. Subsequent columns detail the number of tokens, vocabulary size, and complexity (computed as the type-token ratio). Higher complexity indicates greater sparsity. The final three columns present the number of sentences, the average number of tokens per sentence, and the standard deviation for token counting.

Nheengatu may stand out as the most extensively documented Brazilian indigenous language, dating back to its description in the first Brazilian indigenous language dictionary in 1756 (Avila, 2021). Moreover, numerous texts in Nheengatu were authored during the eighteenth century, further contributing to its rich documentation. The Nheengatu treebank is the largest one: 12,743 tokens (35% of all treebanks) and 1,913 types (22.1% of all treebanks). About 99.8% of all sentences have a length of up to 40 tokens (including punctuation), which is compared to almost all European languages available in UD initiative. For instance, in German, Czech and Russian, which are the biggest treebanks in UD, about 93% of sentences have a length of fewer than 40 tokens.

Since the UD repository only provides test sets, we perform cross-validation such that the test set is split into five folds: one for test and four for training. Three different grammar induction methods are used: MI, DMV, and LLM. In the present study, it is pertinent to emphasize that our approach is entirely unsupervised. Therefore, our training data solely comprises raw text, with the exception of the DMV method which incorporates gold Part of Speech (POS) tags.

The first works on grammar induction applied a dynamic programming algorithm on $O(n^3)$ for constituency grammar (Sankaran, 2010; Cohen et al., 2008), which is computationally expensive for longer sentences. For this reason, most works on grammar induction were trained on sentences up to 40 tokens (Kim et al., 2019) . In this paper, we tested the models on sentences of lengths up to 10 and up to 40 tokens, to evaluate the impact of different sentence size. The tree models used in this work are described in subsections 3.1, 3.2, and 3.3. These models are unsupervised, except for the LLM that, besides the zero-shot approach, we also used one and two-shot learning.

3.1 DMV Model

The DMV stands as a prevalent model for grammar induction, serving as a baseline in several works on unsupervised grammar induction (Shen et al., 2021; Yang et al., 2020). This model operates by generating syntactic trees in a top-down fashion using generative unsupervised training. The idea behind the DMV model is to estimate the syntactic tree by using the Expectation-Maximization (EM) algorithm. For each branch to be generated, it uses probability distributions to make decisions on when and which branch to generate.

We experimented with DMV using the same set-

[4] http://hdl.handle.net/11234/1-5287

Table 1: Indigenous languages in Brazil used in this study

Language	*Family*	*SR*	*Tokens*	*Types*	*Complexity*	*Sentences*	μ	σ
Xavante	Macro-Je	22	1,597	385	0.241	148	10.791	6.423
Tupinambá	Tupian	26	4,508	1,970	0.437	581	7.759	5.946
Nheengatu	Tupian	32	12,743	1,913	0.150	1,239	10.285	6.736
Munduruku	Tupian	26	1,022	399	0.390	158	6.468	5.977
Makurap	Tupian	15	178	95	0.533	37	4.811	1.998
Madi	Arawan	17	115	68	0.591	20	5.750	3.048
Karo	Tupian	25	2,319	773	0.333	674	3.441	1.523
Kaapor	Tupian	22	366	221	0.603	83	4.410	2.024
Guajajara	Tupian	27	9,160	1,515	0.165	1,182	7.750	4.041
Bororo	Bororoan	29	1,905	762	0.400	371	5.135	5.512
Apurina	Arawakan	26	941	373	0.396	152	6.191	3.258
Akuntsu	Tupian	21	1,468	506	0.344	343	4.280	2.556
All	-	35	36,322	8,632	4.208	5,000	7.264	5.450

```
Na sentença "Aiwana, paá, aintá uyaxiú",  as relações de dependência sintática
são mostradas abaixo no formato (token dependente -> token cabeça)
(Aiwana -> uyaxiú)
(, -> paá)
(, -> paá)
(aintá -> uyaxiú)
(. ->  -> uyaxiú)
Liste as relações de dependência sintática na sentença "Yané tuixawa umanú ana
mira amusuaxarawara usikié tenhẽ waá.", usando o formato (token dependente ->
token cabeça).
```

Figure 3: An example of prompt for the Nheengatu language in one shot learning

ting provided by He et al. (2018). It is pertinent to note that this model exhibits limitations in training with longer sentences, attributed to the $O(n^3)$ time complexity of the EM algorithm (Cohen et al., 2008; Spitkovsky et al., 2010). However, given the relatively small treebanks employed in this investigation, the DMV is executed with 10 epochs on each fold using cross-validation assessments.

3.2 MI-based Model

Generally defining, the MI measure indicates the dependency among elements of interest. In our case, it is used to determine words that are more probable to be syntactically related. Equation (1) shows how it is computed for head (h) words and their dependents (d).

$$MI(D,H) = \sum_{d \in D} \sum_{h \in H} P(d,h) log_2 \frac{P(d,h)}{P(d)P(h)} \quad (1)$$

To compute it, we performed word pair permutations within each sentence, considering every possible configuration. The total number of permuted pairs is described by $\sum_{d=1}^{DW} n - d$, where *n* is the number of tokens in the sentence, including punctuation, and *DW* is the distance between the words in the sentence. For instance, for the sentence *"I love the sun"*, the word pairs for DW=1 is <I, love>,<love, the>,<the sun>. Using DW=n, the number of pairs is described by binomial coefficients $\binom{n}{k}$, with *k* representing two (tokens per pair). This setting produces the pairs <I,love>,<I,the>,<I,sun>,<love,the>,<love,sun> and <the,sun>. We train all models using different DW values and choose DW=2 as the best performance.

That permutation process resulted in the creation of the final set of Sentence Permutations (SP), comprising pairs of tokens where the first token precedes the second in the sentence sequence. Following this, MI was computed for each word pair within the SP. Finally, we take the *n* pairs with the highest MI and compare them to manually annotated sentences.

Since corpora used in this work are very small, we perform an edit distance smoothing. For each token in the test that was not in the training set, we searched for the most similar morphological token in the training set using edit distance. For instance, if the token *"uyapí"* does not appear in the training set, the edit distance is applied to find the most lexically related word in the training set, such as *"uyari"*. Then the frequency of the token *"uyari"* is assigned to the token *"uyapí"*. Since there will always be a lexically related token, all tokens in the test set will have a frequency. For bigrams found in the test and not in the training set, we apply a derived simple Laplace smoothing by attributing frequency equal to 1/size of the vocabulary.

3.3 Large Language Model

LLM are models that are trained with a massive amount of data and require a huge computational structure. They can be used in a wide number of tasks such as information extraction, summarization, and question answering, to name a few (Wei et al., 2022). We did not build the LLM using native languages, instead, since we do not have enough data, we used LLM trained in Portuguese. Since the native languages used in this work are spoken in Brazil, and their vocabularies eventually incorporate some Portuguese words, we believe that is possible to find some syntactic relations using LLM even if that language has never been used for training.

We aim to demonstrate the limits and potentialities of LLMs to learn syntactic information in languages with lower resources. We use the chatGPT 3.5 API provided by OpenAI. Differently from the experiments on MI and DMV, we select only three languages to conduct experiments with the LLM. As we wanted to analyze the influence of a larger treebank, we tested with Nheengatu. Average sentence length can also play a role in dependency grammar induction and, therefore, we chose the Karo language, whose sentences are shorter. Finally, we wanted to study the influence of the language family, and language Bororo was chosen for having the largest treebank among those languages not belonging to the Tupian family.

We performed zero, one shot, and two shots learning. In +1 shot learning, we use two different prompts: using a fixed sentence and a random sentence for composing the prompt. For the fixed sentence, we chose a sentence of length seven, which is approximately the average of all languages used in this study. The chosen sentence is the one with the most frequent tokens in the treebank. For the prompt that applies a random sentence, we have random sentences with lengths up to 40 tokens in the training set to be included in the prompt. Since the answers provided by the model are not always the same, we tested the prompts on 30 sentences for each of the five folds of cross-validation. This experiment resulted in 2,250 requests to OpenAI API. We also tested different prompts in Portuguese language and chose the best one. An example of a prompt for one shot is shown in Figure 3.

4 Results

In this study, we adopt the 37 syntactic relations of the UD initiative[5], yet not all languages that we examined utilize all of these relations. As demonstrated in Table 1, Makurap employs only 15 syntactic relations, while Nheengatu utilizes 32. It is noteworthy that Guajajara does not include any occurrence of the subject relation. This study concentrates exclusively on syntactic relations that constitute a minimum of 10% of the respective treebank annotations. Due to limited data, we did not consider the subtypes of some syntactic relations.

We present results for the standard evaluation metrics: Undirected Dependency Accuracy (UDA) and Directed Dependency Accuracy (DDA). Comparing with the reference annotations, these metrics compute how many relations (for word pairs) were correctly predicted, considering or not the relation direction, respectively.

Overall, it is interesting that, despite the limited size of the treebanks, the induction methods for these languages achieved good results, even better than some reported results for non-indigenous languages, such as German, English, and Chinese, using DMV (Klein and Manning., 2004).

In general, Akuntsu and Karo emerged as languages exhibiting the best outcomes, whereas Guajajara and Xavante posed notable challenges. These results are not related to the family origin or annotation. Akuntsu, Karo, and Guajajara were annotated using the same annotation protocol within the same project (Gerardi et al., 2021). However, Akuntsu and Karo are two languages spoken in the state of Rondônia, but Guajajara and Kaapor, which are also spoken in the same state (Maranhão) and come from the same family, Tupian, present different outcomes.

No discernible correlation is observed between vocabulary size and treebank size; however, a subtle correlation is discerned between sentence length

[5] https://universaldependencies.org/u/dep/index.html

and associated scores. Across all settings, the "object" dependency relation was the most correctly detected one, yet substantial variation exists among languages.

MI presented the best results on UDA; on the other hand, DMV was better on DDA. As may be expected, LLM presented the worst results.

The syntactic relations that were more correctly induced (with the highest scores) with DMV are *punct* (punctuation) with 20.8%, *obj* (object) with 18.7%, and *nsubj* (subject) with 16.7%. However, MI presents the highest incidence of *obj* with 26% and *nsubj* with 18%, followed by *advmod* (adverbial modifier) with 8%. The selection of these syntactic relations is based on their prevalence within the treebank. Nonetheless, our code is accessible for retrieving data related to other syntactic relations as well.

The detailed results are presented in Subsections 4.1, 4.2, and 4.3. The summarized results are presented in Table 2. The last three lines present the most correctly induced syntactic relations (1 SR), the second most correctly induced syntactic relations (2 SR), and the third most correctly induced syntactic relations (3 SR), respectively. Due to space limitation, we presented only the results for DMV using the DDA metric[6].

Table 2: Summarized results

	DMV	MI	LLM
UDA 10	0.5135	**0.5692**	0.4165
UDA 40	0,.4654	**0.5089**	0.4212
DDA 10	**0.3201**	0.3122	0.2779
DDA 40	**0.2808**	0.1687	0.2720
1 SR	obj	obj	obj
2 SR	nsubj	nsubj	case
3 SR	punct	advmod	advmod

4.1 DMV

The results for DDA are presented in Table 3. DMV can induce correctly 89% of all object relations on Akuntsu, but only 11% on Kaapor. Despite presenting good results on small corpora such as those of Makurap and Madi, DMV struggles to induce some important syntactic relations. This pattern is similar when evaluated using UDA metrics.

4.2 MI

The use of edit distance yielded notable improvements, showcasing a 29.5% enhancement in MI for UDA and a 13.6% boost for DDA. While the results based on MI lag behind DMV in terms of DDA metrics, it is crucial to highlight the superiority of MI in UDA metrics. Moreover, it manifests superior outcomes in the context of induced object and subject relations. Notably, in the Makurap language, all object relations were accurately induced, and, in the Madi language, every subject was correctly induced.

4.3 LLM

Differently from experiments with DMV and MI, we did not use weighted average for LLM because the Nheengatu language presents 75% of the available corpora. The results presented in Table 2 refer to the average of all settings. As we expected, the zero-shot for all languages and all settings yielded the least favorable results on average, with 0.290 for UDA and 0.142 for DDA; transitioning to one-shot learning, UDA improved to 0.413, and DDA to 0.264; in two-shot learning, the model achieved 0.427 for UDA and 0.285 for DDA. When sentences were not fixed, the model exhibited competence with scores of 0.431 for UDA and 0.286 for DDA. However, when fixed sentences were employed in the prompt for one and two-shot learning, the overall performance deteriorated, resulting in an average of 0.406 for UDA and 0.263 for DDA. This result may be due to the distribution of the sentences, since that, with no fixed sentence, almost 150 different sentences were tested in the prompt. However, to induce object relations, using a fixed sentence in the prompt presented better results.

Different from MI and DMV, LLM may be influenced by the size of the treebank. When using one and two-shot learning, Nheengatu presents 0.440 DDA, against 0.406 in Karo and 0,410 in Bororo. This result is different from the DMV and MI approaches, in which Nheengatu presents the poorest scores. Nonetheless, the induction of particular dependency relations may not necessarily exhibit a correlation with treebank size. In the cases of Karo and Bororo languages, accurate induction of object relations is achieved with notable proficiency. In contrast, the Nheengatu language demonstrates a lower level of accuracy in this regard. These outcomes align with the findings obtained through both DMV and MI approaches.

[6] Detailed results may be found at https://github.com/diegodpgs/PROPORInd

Table 3: Results for DMV with DDA metric

DDA for sentences $\leq$ 10 tokens							
Language		1 SR		2 SR		3 SR	
Akuntsu	0.5661	0.8957	obj	0.5783	nsubj	0.5551	punct
Apurina	0.4248	0.7460	obj	0.7227	nsubj	0.2321	punct
Bororo	0.3832	0.8696	case	0.6992	obl	0.5489	nsubj
Guajajara	0.1669	0.4690	obl	0.2142	discourse	0.0730	punct
Kaapor	0.2500	0.7843	obj	0.4921	nsubj	0.1765	advmod
Karo	0.3803	0.5882	nsubj	0.4595	advmod		
Madi	0.4186	0.4545	punct	0.2500	obj		
Makurap	0.4696	0.6667	advmod	0.3750	discourse		
Munduruku	0.4074	0.8077	case	0.6846	obl	0.5000	punct
Nheengatu	0.3671	0.5756	advmod	0.5579	nsubj	0.2271	punct
Tupinamba	0.3138	0.5111	punct	0.4100	obl		
Xavante	0.3264	0.7500	dep	0.3099	punct	0.1176	nsubj
μ	**0.3729**	**0.6765**		**0.4795**		**0.3038**	
μ **weighted**	**0.3201**	**0.5907**		**0.4492**		**0.2193**	
DDA for sentences $\leq$ 40 tokens							
Akuntsu	0.5641	0.8800	obj	0.6077	nsubj	0.5879	punct
Apurina	0.3907	0.8488	obj	0.7211	nsubj	0.2153	punct
Bororo	0.3579	0.6647	punct	0.6497	obl	0.5020	nsubj
Guajajara	0.1704	0.4223	obl	0.2135	discourse	0.0900	punct
Kaapor	0.2287	0.8302	obj	0.4242	nsubj	0.2432	advmod
Karo	0.3301	0.5882	nsubj	0.4757	advmod		
Madi	0.3585	0.4167	punct				
Makurap	0.4348	0.6250	advmod	0.4375	discourse		
Munduruku	0.3784	0.9029	case	0.6506	nsubj	0.5909	obl
Nheengatu	0.2943	0.5376	nsubj	0.4918	advmod	0.1613	punct
Tupinamba	0.2572	0.4835	punct	0.3921	obl		
Xavante	0.3110	0.6348	dep	0.2800	punct		
μ	**0.3397**	**0.6529**		**0.4858**		**0.3415**	
μ **weighted**	**0.2808**	**0.5512**		**0.4198**		**0.1916**	

5 Discussion

Despite the effectiveness of modern approaches such as neural networks and LLM, simple methods such as MI can perform better when applied to low language resources. For some sentences, we identified that the LLM likely employed the straightforward right-branching algorithm. It is necessary to note that an explicit evaluation of the comparative efficacy of these methodologies against the right-branching baseline, established at 0.38 for the English language (Klein and Manning., 2004), was not conducted and remains for future work.

The MI models present good results, but the induced syntactic tree could have missing elements, as presented in Appendix A. It can be solved by optimization, which could also be a matter of future work.

It is essential to highlight that the indigenous languages utilized in this study exhibit distinct syntactic characteristics, including the absence of certain crucial syntactic relations (such as nsubj in Guajajara, for example), as well as unique sentence structures. These nuances may influence the obtained outcomes. In-depth linguistic inquiries or even anthropological investigations may be necessary to elucidate the variations in results across different languages.

6 Final Remarks

We presented a study on grammar induction for different Brazilian indigenous languages. We demonstrate the efficacy of inducing syntactically related words for low-resource languages using some well-known approaches and a current LLM-based strategy, mainly in inducing specific relations, such as object and subject relations. Such methods may be very useful to uncover syntactic structures for languages for which the grammar was not yet described or to refine NLP parsing methods.

Future work includes the investigation of other induction methods and the exploitation of language-specific features that may improve the results.

The interested reader may find other details about this and other related work at the web portal of the POeTiSA project (*POrtuguese processing - Towards Syntactic Analysis and parsing*)[7].

Acknowledgments

This work was carried out at the Center for Artificial Intelligence of the University of São Paulo (C4AI - `http://c4ai.inova.usp.br/`), with support by the São Paulo Research Foundation (FAPESP grant #2019/07665-4) and by the IBM Corporation. The project was also supported by the Ministry of Science, Technology and Innovation, with resources of Law N. 8.248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC 13, DOU 01245.010222/2022-44.

References

Marcel Twardowsky Avila. 2021. *Proposta de dicionário nheengatu-português*. Ph.D. thesis, Universidade de São Paulo.

Marta Maria Azevedo. 2016. Urbanização e migração na cidade de são gabriel da cachoeira, amazonas. In *Anais do XV Encontro Nacional de Estudos Populacionais*, pages 1–14.

Shay B. Cohen, Kevin Gimpel, and Noah A. Smith. 2008. Logistic normal priors for unsupervised probabilistic grammar induction. In *Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems*, pages 321–328.

Aline Da Cruz. 2011. Fonologia e gramática do nheengatú: A língua geral falada pelos povos baré, warekena e baniwa. *Ph.D. Thesis,Vrije Universiteit Amsterdam*.

Diego Pedro Gonçalves da Silva and Thiago Alexandre Salgueiro Pardo. 2023. Induçao gramatical para o português: a contribuiçao da informaçao mutua para descoberta de relaçoes de dependência. In *Proceedings of the 14th Brazilian Symposium on Information Technology and Human Language*, pages 298–307.

Veronica Dahl, Gemma Bel-Enguix, Velina Tirado, and Emilio Miralles. 2023. Grammar induction for underresourced languages: the case of ch'ol. In *Proceedings of the Analysis, Verification and Transformation for Declarative Programming and Intelligent Systems: Essays Dedicated to Manuel Hermenegildo on the Occasion of His 60th Birthday*, pages 113–132.

Tiago Barbosa de Lima, André C. A. Nascimento, Pericles Miranda, and Rafael Ferreira Mello. 2021. Analysis of a brazilian indigenous corpus using machine learning methods. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 118–129.

Eduardo de Paiva Alves. 1996. The selection of the most probable dependency structure in japanese using mutual information. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 372–374.

Fabrício Ferraz Gerardi, Tiago Tresoldi, Carolina Coelho Aragon, Stanislav Reichert, Jonas Gregorio de Souza, and Francisco Silva Noelli. 2023. Lexical phylogenetics of the tupí-guaraní family: Language, archaeology, and the problem of chronology. *Plos one*, 18(6):1–25.

Richard Futrell, Peng Qian, Edward Gibson, Evelina Fedorenko, and Idan Blank. 2019. Syntactic dependencies correspond to word pairs with high mutual information. In *Proceedings of the 5th international conference on dependency linguistics*, pages 3–13.

Fabrício Ferraz Gerardi, Stanislav Reichert, and Carolina Coelho Aragon. 2021. Tuled (tupían lexical database): introducing a database of a south american language family. *Language Resources and Evaluation*, 55(4):997–1015.

K. David Harrison. 2008. When languages die: The extinction of the world's languages and the erosion of human knowledge. *Oxford University Press*.

Junxian He, Graham Neubig, and Taylor Berg-Kirkpatrick. 2018. Unsupervised learning of syntactic structure with invertible neural projections. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1292–1302.

Jacob Louis Hoover, Wenyu Du, Alessandro Sordoni, and Timothy J. O'Donnell. 2021. Linguistic dependencies and statistical dependence. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2941–2963.

Yoon Kim, Chris Dyer, and Alexander M. Rush. 2019. Compound probabilistic context-free grammars for grammar induction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 2369–2385. Association for Computational Linguistics.

Dan Klein and Christopher D. Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the*

[7] `https://sites.google.com/icmc.usp.br/poetisa`

42nd Annual Meeting of the Association for Computational Linguistics, page 478–485.

András Kornai. 2013. Digital language death. *PloS one*, 8(10):1–11.

Boda Lin, Zijun Yao, Jiaxin Shi, Shulin Cao, Binghao Tang, Si Li, Yong Luo, Juanzi Li, and Lei Hou. 2022. Dependency parsing via sequence generation. In *Findings of the Association for Computational Linguistics*, pages 7339–7353.

Sidney Facundes Moore, Denny and Nádia Pires. 1994. Nheengatu (língua geral amazônica), its history, and the effects of language contact. In *Proceedings of the Meeting of SSILA and the Hokan-Penutian Workshop*, pages 93–118.

Rosângela Morello. 2016. Censos nacionais e perspectivas políticas para as línguas brasileiras. *Revista Brasileira de Estudos de População*, 33(2):431–439.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Jan Hajič, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4033.

Claudio S Pinhanez, Paulo Cavalin, Marisa Vasconcelos, and Julio Nogima. 2023. Balancing social impact, opportunities, and ethical constraints of using ai in the documentation and vitalization of indigenous languages. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6174–6182.

Baskaran Sankaran. 2010. A survey of unsupervised grammar induction. *Manuscript, Simon Fraser University 47*, pages 1–63.

Yikang Shen, Yi Tay, Che Zheng, Dara Bahri, Donald Metzler, and Aaron C. Courville. 2021. Structformer: Joint unsupervised induction of dependency and constituency structure from masked language modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 7196–7209.

Zach Solan, David Horn, Eytan Ruppin, and Shimon Edelman. 2005. Unsupervised learning of natural languages. In *Proceedings of the National Academy of Sciences*, pages 11629–11634.

Valentin I. Spitkovsky, Hiyan Alshawi, and Daniel Jurafsky. 2010. From baby steps to leapfrog: How "less is more" in unsupervised dependency parsing. In *Proceedings of the Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*, pages 751–759.

Gerald Taylor. 1985. Apontamentos sobre o nheengatu falado no rio negro, brasil. *Amérindia: revue d'ethnolinguistique amérindienne*, pages 5–23.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*, pages 1–30.

Songlin Yang, Yong Jiang, Wenjuan Han, and Kewei Tu. 2020. Second-order unsupervised neural dependency parsing. In *Proceedings of the 28th International Conference on Computational Linguistic*, pages 3911–3924.

A Illustration of grammar induction for Nheengatu

We present a sample of the induced relations for the sentence *Aikwé awá ururi indé u reyuri putari tẽ ne rupí?*, which corresponds to *Was there anybody to bring you or did you yourself want to come?* in English, using DMV, MI, and LLM methods. The cited sentence represents a transcription of speech delivered by an indigenous Nheengatu speaker (Moore and Pires, 1994). It is important to note that the orthography utilized is not the original form, but has been adjusted to adhere to the UD framework.

In Figures 4, 5, and 6, the color orange means that the model correctly predicted the relation according the UDA measure (which does not evaluate the direction of the arrow), and green means that the model correctly predicted the direction too, as informed by the reference annotation (in Figure 7).

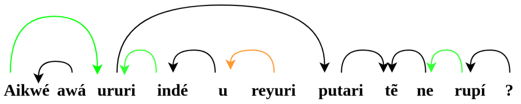

Figure 4: Induced relations using DMV

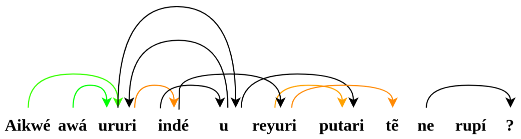

Figure 5: Induced relations using MI

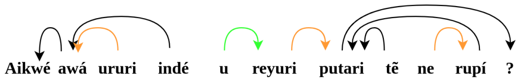

Figure 6: Induced relations using LLM

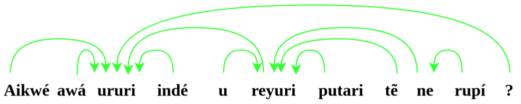

Figure 7: Reference annotation in the treebank