Analysing and Validating Language Complexity Metrics Across South American Indigenous Languages

Felipe Ribas Serras
Miguel de Mello Carpi
Matheus Castello Branco
Marcelo Finger
Institute of Mathematics and Statistics, University of São Paulo
R. do Matão, 1010 - Butantã, São Paulo - SP, Brazil, 05508-090
{frserras, miguel, matheus.castello, mfinger}@ime.usp.br

Abstract

Language complexity is an emerging concept critical for NLP and for quantitative and cognitive approaches to linguistics. In this work, we evaluate the behavior of a set of compression-based language complexity metrics when applied to a large set of native South American languages. Our goal is to validate the desirable properties of such metrics against a more diverse set of languages, guaranteeing the universality of the techniques developed on the basis of this type of theoretical artifact. Our analysis confirmed with statistical confidence most propositions about the metrics studied, affirming their robustness, despite showing less stability than when the same metrics were applied to Indo-European languages. We also observed that the trade-off between morphological and syntactic complexities is strongly related to language phylogeny.

1 Introduction

The development of means for quantifying linguistic properties is essential for cognitive approaches to computational linguistics, becoming simultaneously more challenging and useful as the property of interest is transversal to different languages and, therefore, an important clue for accessing cognitive processes behind human language. This is the case of language complexity.

The concept of language complexity, whether of an utterance or of a language as a whole, is instinctive for us. People know how to recognize when a text is written in a difficult or elaborate way and they usually recognize that certain languages are less or more complicated to learn depending on their linguistic background.

Informally, we can say that: (i) the *complexity of an utterance* encompasses the quantity and sophistication of linguistic constructs necessary to form and understand the utterance and (ii) *the complexity of a language as a whole* refers to the quantity and sophistication of communicative strategies available for the formation of such utterances in that language.

Despite a relative consensus around these intuitions, we lack established formal and quantifiable definitions of language complexity. It is difficult to find a definition that encompasses the heterogeneous range of human language manifestations, both in terms of different languages and of different levels in which meaning can be conveyed within a language.

Even in light of these challenges, it is crucial to establish rigorous, theoretically and experimentally validated definitions of language complexity. Both cognitive and non-cognitive approaches to Linguistics can significantly enhance their expressive capacity and theoretical framework. In NLP, complexity measures can be used in automatic text simplifiers, translators, domain-sensitive correctors and completers (Leal et al., 2023), but can also be integrated into the of training machine learning models, to increase performance (Sarti et al., 2021).

Another challenge for the construction of a robust theory for language complexity is that of inclusion: historically, the construction of tools and theories of human language has included Indo-European languages, to the detriment of other linguistic manifestations, e.g. American native languages. For a concept that aims to be transversal to different languages and provide universal insights into them and their underlying cognitive processes, as is the case with language complexity, it is necessary to include the broadest possible range of languages in its development and validation.

This inclusion is the focus of our work. Here, we examine a set of language complexity metrics derived from Information Theory, proposed in Juola (1998, 2005, 2008), and Ehret and Szmrecsanyi (2016). The authors ran several experiments with the proposed metrics, drawing on data from a sub-

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 152–165

stantial set of languages, with a predominant focus on those belonging to the Indo-European language family.

Here, we repeat these experiments with data from South American indigenous languages, attempting to ascertain whether the desirable properties of these metrics remain solid when incorporating frequently excluded languages. We seek to verify the robustness of the proposed metrics and include a more diverse set of linguistic manifestations in the construction of a quantifiable theory of human language complexity.

Our text is structured as follows: In section 2, we present our theoretical background and related works; section 3 outlines our methodology, complexity metrics and its properties, the data used and the experimental pipeline employed; in section 4 we exhibit our results and, in section 5, we present our conclusions.

2 Related Works

Nichols (1998) was a pioneer in proposing a quantifiable language complexity metric. She defines the morphological complexity of a language as the number of inflection points in its typical sentence. She computed it for more than 200 languages. In this, work we evaluate the consistency of our targeted metrics with hers.

In contrast to the computational challenges of Nichols (1998)'s metric, Juola (1998, 2005, 2008) proposes a set of compression-based complexity metrics based on information theory. The author compares these metrics to alternatives, extend them to different linguistic tiers, and evaluate them on different parallel *corpora*. This family of metrics is the main object of study in this paper, with a focus on their behavior when applied to indigenous South American languages, not explored in the original works. Ehret and Szmrecsanyi (2016) suggests modifications to them, proposing improvements for eliminating potential spurious correlations, and experiment on semi-parallel and non-parallel data.

Several subsequent works draw directly or indirectly from the notion of compression-based complexity metrics (Juvonen, 2008; Sadeniemi et al., 2008; Fenk-Oczlon and Fenk, 2008; Ehret et al., 2021; Szmrecsanyi, 2021; Pellegrino et al., 2011; Ackerman and Malouf, 2013; Kettunen, 2014; Housen et al., 2019), in particular, to quantify the difficulty of acquiring a second language (Bulté and Housen, 2014; Clercq and Housen, 2019).

An alternative approach, which characterizes complexity as a function of linguistic features, was explored in Graesser et al. (2004, 2011); Graesser and McNamara (2011) for English and in Leal et al. (2023) for the Portuguese language. Similar works study language complexity from the perspective of readability, instead of the typological approach adopted here, focusing on text simplification or elaboration (McNamara et al., 2014; Carroll et al., 1998; Max, 2006; Shardlow, 2014; Siddharthan, 2006; DuBay, 2007; Leal and Aluísio, 2024).

Regarding the study of indigenous languages of South America, several classic works studied and documented the languages explored here, e.g. Callow (1962); Derbyshire and Pullum (1986–1991); Dixon and Aikhenvald (2006) *inter alia*. The investigation into the computational complexity of indigenous languages remains much less explored and our work is completely original, to the best of our knowledge. Bentz et al. (2017), Gutierrez-Vasques et al. (2023), Oh and Pellegrino (2023), Bentz et al. (2016), Nichols and Bentz (2018) and Bentz et al. (2023) are the works that closely resemble the work we present here, assessing various complexity metrics or associated measures on language sets that incorporate South American languages. Nevertheless, these studies diverge from ours in terms of goals, methodology, and/or the quantity of included indigenous languages, typically covering a significantly smaller number compared to our assessment.

3 Methodology

This paper aims to evaluate a collection of compression-based language complexity metrics $\mathcal{M}$ introduced in previous works (Juola, 1998, 2005, 2008; Ehret and Szmrecsanyi, 2016). The evaluation is conducted on a dataset $\mathcal{D}$ encoded in a broad range of South American indigenous languages $\mathcal{L}$. The objective is to determine the validity of the theoretical and experimental propositions $\mathcal{P}$ regarding $\mathcal{M}$, as observed in the aforementioned studies, when $\mathcal{M}$ is applied to the languages in $\mathcal{L}$.

In this section, we present the methodology adopted to achieve this goal. In subsection 3.1, we define the set $\mathcal{M}$ of language complexity metrics evaluated; in subsection 3.2, we present the set $\mathcal{L}$ of South American languages tested and the data $\mathcal{D}$ used to represent them; in subsection 3.4 we present the propositions $\mathcal{P}$ about $\mathcal{M}$, whose validity we wish to verify when $\mathcal{M}$ are applied

to $\mathcal{L}$ through $\mathcal{D}$; in subsection 3.5, we outline the experimental processing pipeline employed for conducting this verification. Subsection 3.3 presents a brief interlude on the writing systems used to encode the languages in $\mathcal{L}$.

3.1 Complexity Metrics

The language complexity metrics $\mathcal{M}$ evaluated in this work (Juola, 1998, 2005, 2008; Ehret and Szmrecsanyi, 2016) are based on a teleological approach to human language, that can be traced back to (Zipf, 1949) . This view reduces natural language to its primary functionality - the transmission of meaning or information - in line with Shannon's Information Theory (MacKay, 2003).

In this approach, each textual excerpt is seen as a message encoding a certain amount of information. The complexity of the message is the amount of information encoded. For a sufficiently long message, the amount of information can be approximated by the size of the message when compressed by an efficient compression algorithm.

However, experimental results show that natural languages tend to maintain a relatively uniform information density during communication (Manin, 2006; Aylett and Turk, 2004; Jaeger, 2006; Jaeger and Levy, 2006; Jaeger, 2010). Some works model this through the hypothesis that natural languages try to maximize information transmission without overloading the cognitive systems of senders and receivers (Piantadosi et al., 2011), but with limited results (Pimentel et al., 2023). Regardless of the exact mechanisms behind this uniformity, one consequence is that longer texts contain more information, resulting in larger compressed versions.

This correlation between a text's length and its compressed size must be considered by any complexity metric using compressed text size to estimate overall text complexity. To address this issue, Ehret and Szmrecsanyi (2016) proposes that the overall language complexity should be computed as a measure of how much the size of the compressed message deviates from the expected correlation with the size of the uncompressed version. This can be computed from the residuals (res) of the linear regression between the compressed message size and its original size. This definition of overall complexity $\mu^{\mathbb{A}}$ is shown in Equation 1 (For details about the mathematical notation, see InfoBox 1).

$$\mu^{\mathbb{A}}(\mathscr{T}) = res(|C(\mathscr{T})|, |\mathscr{T}|) \quad (1)$$

Mathematical Notation Key

Throughout this text we will use the following notation conventions:

- $\mathscr{T}$ represents a textual excerpt or message encoded in a natural language;
- Degraded texts are represented with subscripts and superscripts. The subscript symbol represents the type of degradation ($\circ$ for replacement, $\times$ for deletion). The superscript represents the target tier of the degradation process.
- Language complexity metrics are functions represented by $\mu_{\_}^{\mathbb{Y}}(\cdot)$, where, the subscript symbol indicates the type of degradation associated with the metric and the superscript symbol indicates the target language tier accessed by the metric;
- $C(\mathscr{T})$ represents the text $\mathscr{T}$ after compression ;
- $|\cdot|$ represents the size of an object in bytes.

InfoBox 1: Mathematical notation adopted throughout this text.

Nevertheless, the complexity of a text cannot be determined solely by the overall information transmitted ($\mathbb{A}$). Natural languages have different mechanisms of encoding information and adopt different strategies to distribute the information transmitted through these mechanisms. Finnish, for example, has a rich morphological case system, in which a noun such as "talo" (house) becomes "talolta" to express the concept "from the house". This same concept is expressed syntactically in English through the association with a preposition, external to the word "house".

In information theory terms, each message encoded in a natural language consists of different tiers through which one can distribute the information conveyed by the message, and therefore its complexity. A text would then have a different level of complexity for each tier.

In an effort to grasp these subtleties, Juola (2008)

introduce a set of metrics designed to capture the relative complexities across three distinct linguistic tiers: morphological ($\mathbb{M}$), syntactic ($\mathbb{S}$), and pragmatic ($\mathbb{P}$). The principle underlying these three metrics is the same: to degenerate[1] the information conveyed only by the targeted linguistic tier and to compute the ratio between the size of the degenerated compressed text to that of the original compressed text. In this way it is possible to access how much of the overall information is being transmitted by the targeted tier.

The more dependent a language is on a particular tier for conveying information, the more the degradation of that tier leads to information loss in the text. This intensified information loss hinders pattern recognition for compression algorithms, resulting in reduced compressibility and higher complexity metric values for that tier.

Juola (2008) achieves degeneration through a deletion process, wherein 10% of the units in the text are randomly erased. The choice of textual unit to be erased depends on the targeted linguistic tier: characters for morphology, words[2] for syntax, and verses for pragmatics[3].

Ehret and Szmrecsanyi (2016) argue that an expected exception to this general template is morphological complexity: languages with rich morphology use systems to convey information within words that other languages express through external elements. As a result, a single word in this languages can have several allowed forms. Thus, in languages with high morphological complexity, deleting a character still often yields a valid word form, minimizing disruption in text compressibility. To address this, a negative sign is incorporated in the definition of morphological complexity. Ehret and Szmrecsanyi (2016) also experimentally confirms the need for this sign correction.

These complexity metrics, as described, are represented by equations 2, 3, and 4. In all cases, we follow the mathematical notation conventions outlined in InfoBox 1.

$$\mu_{\times}^{\mathbb{M}}(\mathscr{T}) = -\frac{|C(\mathscr{T}_{\times}^{\mathbb{M}})|}{|C(\mathscr{T})|} \tag{2}$$

$$\mu_{\times}^{\mathbb{S}}(\mathscr{T}) = \frac{|C(\mathscr{T}_{\times}^{\mathbb{S}})|}{|C(\mathscr{T})|} \tag{3}$$

$$\mu_{\times}^{\mathbb{P}}(\mathscr{T}) = \frac{|C(\mathscr{T}_{\times}^{\mathbb{P}})|}{|C(\mathscr{T})|} \tag{4}$$

Juola (2008) proposes an alternative technique to morphological degeneration using substitution instead of deletion. He replaces all tokens of the same type in the original text with an integer, removing information about the internal structure of words without affecting information about their relative positioning within the sentences. This is represented in Equation 5. Here the numerator and denominator are inverted compared to the previous metrics. This inversion is an attempt to address the same problem related to the morphological complexity that led to the proposition of sign inversion in equation 2, but with a different mathematical strategy.

$$\mu_{\circ}^{\mathbb{M}}(\mathscr{T}) = \frac{|C(\mathscr{T})|}{|C(\mathscr{T}_{\circ}^{\mathbb{M}})|} \tag{5}$$

The process of calculating a complexity metric value from a text $\mathcal{T}$, as previously described, is illustrated in Figure 1. This example employs the metric $\mu_{\times}^{\mathbb{M}}$, defined in Equation 2.

Juola (2008) provides a cognitive argument to why compression-based metrics would work for measuring language complexity. Any measure of the complexity of an object is computed as the number of primitive operations necessary for its functioning. We can reduce the compression process to storage and querying operations over a lexicon of frequent textual patterns. These operations, he argues, align with how human mind uses language, storing frequent linguistic patterns and querying them. Compression-based metrics should thus work well, as they are an approximation, albeit simple, of human cognitive linguistic procedures.

3.2 Data

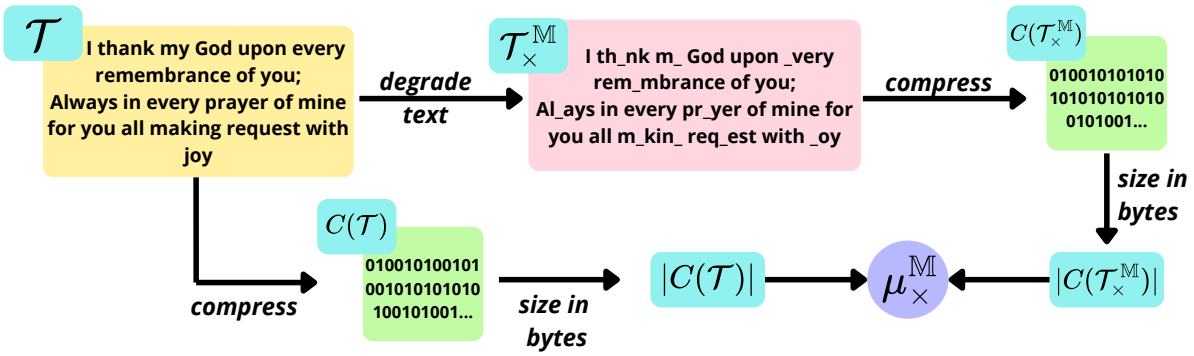
In order to access the properties of complexity metrics across different languages, Juola (2008) opts to eliminate other potential factors of complexity variation, conducting experiments with a parallel corpus comprising the same text translated into different languages. The Christian Bible was his main

[1] In this text, the terms "degeneration" and "degradation" are used interchangeably.

[2] As in Juola (2008), we adopt here the work definition of words as maximal non-blank sequences.

[3] In Juola (2008), as well as here, the main text used in the experiments is a subset of the Christian Bible, given the high availability of translations into different languages. As the Bible is divided into verses and verses correspond roughly to sentences, this is used as the pragmatic unit for computing pragmatic complexity metrics.

Figure 1: Diagram exemplifying the pipeline for computing complexity metrics. This example refers to the metric of morphological complexity through deletion $\mu_{\times}^{\mathbb{M}}$ defined in equation 2.



selected text, chosen for its extensive range of translations and convenient accessibility. In an effort to maintain maximum fidelity to his experiments and isolate potential factors of variation that could undermine the validity of our results, we also have opted to use texts from the Christian Bible.

Another reason for using these texts in our case is the unfortunate scarcity of translations simultaneously available in a wide range of indigenous South American languages. Notably, even the Brazilian constitution lacks versions in the various indigenous languages spoken within its territory. The Bible stands out as one of the rare texts extensively translated into these languages, primarily due to its central role in the colonization process of these communities. A further contributing factor to the limited data availability is the lack of written tradition in the languages studied here. Historically, many of them were primarily oral and only recently adopted a writing system, often developed specifically for the translation and dissemination of christian texts, such as the Bible.

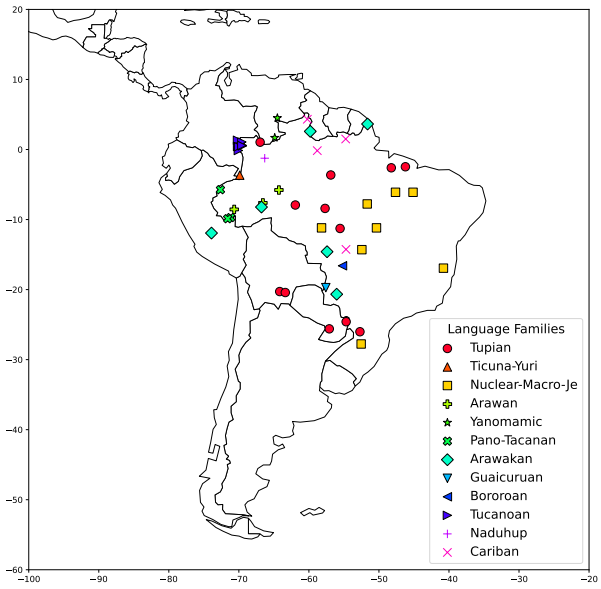
Acknowledging the problematic context in which these translations were produced, we refrain from disclosing the data or deploying any models based on it. Our sole purpose is to leverage these translations to explore aspects of these languages that might otherwise be challenging to investigate. Our aim is to emphasize the importance of considering these languages in the examination of properties that are said to be universal, encompassing human diverse cultural manifestations in our view of natural languages.

We are also aware that these translations were probably produced with very little care for the languages and its cultural meanings and nuances, and that the distribution of language in the Bible may be not representative of these languages as a whole and can be skewed. These may be confounding factors reflected in the obtained values of the aforementioned complexity metrics.

Our dataset, kindly provided by IBM Research Brazil, was originally assembled to explore language classification and machine translation between indigenous languages. It consists of the texts of the Catholic Bible's New Testament, translated into a diverse set of South American indigenous languages and is separated by books, chapters and verses.

The dataset includes 51 South American indigenous languages: Apalaí (*apl*), Apinayé (*api*), Apurinã (*apu*), Asheninka (*cax*), Bakairi (*bki*), Borôro (*brr*), Canela (*cnl*), Culina (*cul*), Desano (*des*), Guajajara (*gjj*), Guarani Eastern Bolivian (*crg*), Guarani Mbya (*[gun]*), Guarani Paraguay (*gua*), Guarani Western Bolivian (*[gnw]*), Hixkaryána (*hix*), Jamamadi (*jmm*), Kaapor (*urk*), Kadiwéu (*kdw*), Kaigang (*kng*), Kaiwá (*kaw*), Karajá (*jva*), Kashinawa (*csh*), Kayabí (*kyz*), Kayapó (*kyp*), Kubeo (*cub*), Macushi (*mac*), Makuna (*mcn*), Matsés (*myr*), Maxakali (*max*), Mundurukú (*muu*), Nadeb (*nad*), Nambikuára (*nmb*), Nheengatu (*[yrl]*), Palikúr (*plk*), Parecís (*pex*), Paumarí (*pau*), Piratapúya (*prt*), Rikbaktsa (*rik*), Sanumá (*snm*), Sateré-Mawé (*[mav]*), Siriano (*sri*), Tenharim (*[pah]*), Terêna (*trn*), Ticuna (*tic*), Tucano (*tuc*), Tuyúca (*tuy*), Wanana (*gno*), Wapishana (*wps*), Xavante (*xav*), Yaminawa (*yam*), and Yanomami (*[guu]*). The geographical distribution of these languages is represented in Figure 2. Additional information about them can be found in Appendix B.

Figure 2: Geographical distribution by family for the languages explored in our experiments. Latitude, Longitude and Phylogenetic data were obtained from the Glottolog Database (Hammarström et al., 2024).



Furthermore it also includes 5 Indo-European languages: English (*eng*), French (*fre*), German (*ger*), Portuguese (*por*), and Spanish (*spa*), which we use for comparison purposes.

We also collected the New Testament in Ancient Greek (*[grc]*) [4] for verifying the proposition that overall complexity of a text is always smaller in its original language (see Section 3.4).

3.3 Writing Systems

The metrics defined in previous sections assess language complexity through the degradation of orthographic elements and sequences such characters and words, thus linking these metrics to the writing systems employed by the targeted languages under evaluation. Both Juola (2008)'s and our experiments focus on languages with alphabetic and low-logographic writing systems derived from the Latin alphabet (Sproat, 2000). Consequently, our conclusions are constrained to this region of the orthographic space. Further research is needed to validate these complexity metrics across diverse regions of the orthographic space that are beyond the scope of this paper. Figure 5 in Appendix A provides a visual representation of the types of writing systems not addressed in our experiments.

[4] https://www.greekbible.com/

3.4 Propositions

We used the data described above to assess whether the desirable properties of the proposed complexity metrics remain consistent when evaluated over our broad set of native South American languages.

These expected properties can be formulated as propositions falling into two broad groups: *prior hypotheses* about how a language complexity metric should behave, and *a posteriori observations*, found in Juola (2008)'s experiments.

The prior hypotheses evaluated in this work are:

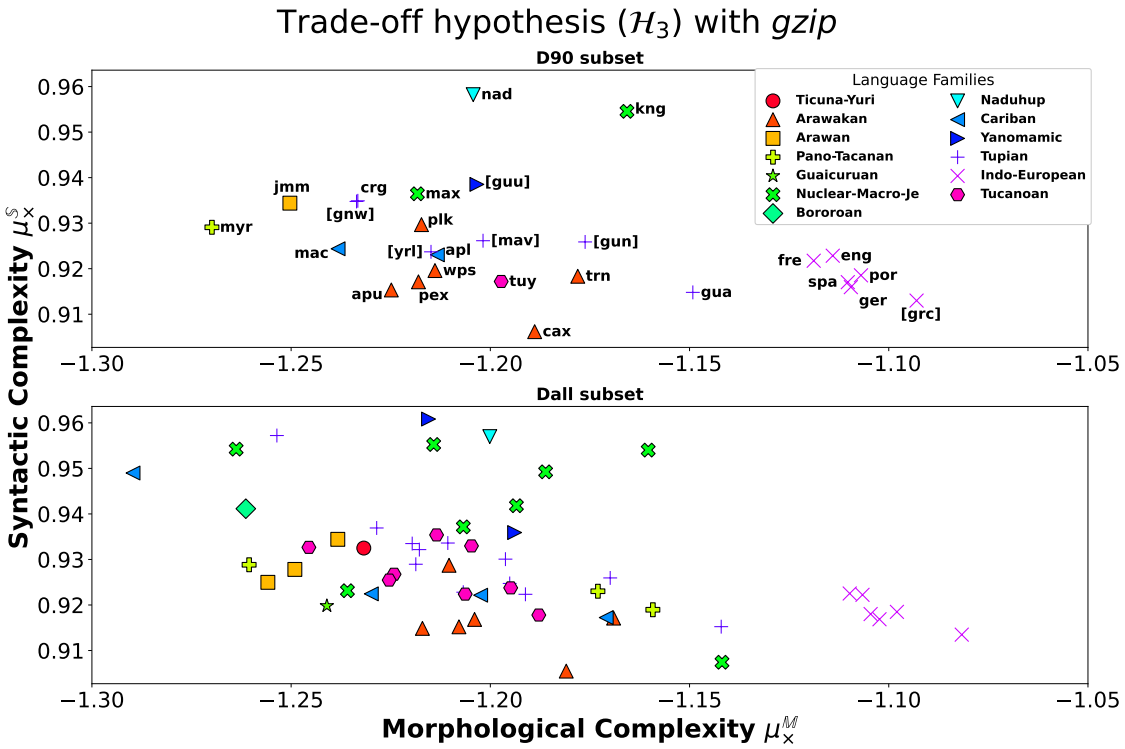
- $\mathcal{H}_1$: the overall complexity ($\mu^{\mathbb{A}}$) of a text in its original language is lower than in other languages, as a result of the introduction of cultural clarifications in the translation process;
- $\mathcal{H}_2$ (equi-complexity hypothesis): all languages have (approximately) the same overall complexity ($\mu^{\mathbb{A}}$)[5];
- $\mathcal{H}_3$ (trade-off hypothesis): there is a trade-off between the syntactic ($\mu^{\mathbb{S}}_{\times}$) and morphological ($\mu^{\mathbb{M}}_{\times}$) complexities of a language.

The *a posteriori* observations of Juola (2008), accessed in this work are:

- $\mathcal{O}_1$: there is a positive correlation between morphological complexity by replacement ($\mu^{\mathbb{M}}_{\circ}$) and the number of types in the sample and a negative correlation with the number of tokens;
- $\mathcal{O}_2$: all languages are approximately equal in terms of their pragmatic complexity ($\mu^{\mathbb{P}}_{\times}$); in other words, the variance of pragmatic complexity is significantly lower than that of morphological and syntactic equivalents;
- $\mathcal{O}_3$: the morphological complexity metric $\mu^{\mathbb{M}}_{\circ}$ is consistent with Nichols (1998) morphological complexity metric (see Section 2);
- $\mathcal{O}_4$: the results were equivalent when varying the compression algorithm between *gzip* and *bz2*.

[5] Contact languages are possible exceptions to this, but without a representative dataset of contact languages, we cannot verify this hypothesis.

Figure 3: Trade-off between syntactic and morphological complexities by deletion, computed with *gzip* for both $D90$ and $DAll$ sets. The legend is the same for both plots. Phylogenetic data was obtained from Glottolog (Hammarström et al., 2024).



3.5 Experimental Pipeline

Our experimental pipeline consists of five steps:

1. Data normalization: this step ensures that characters that appear identical are indeed encoded identically in UTF-8 representation;

2. Data processing: here, we create two datasets Dall and D90. Dall contains only verses that appear in all languages (2585 verses across 27 languages), while D90 contains verses from languages where the intersection of verses makes up at least 90% of the total (7159 verses across 27 languages).

3. Outlier detection and removal: we analyzed the dataset, detecting the Nambikuára data as a potential anomaly. Nambikuára is a language family spoke in *Mato Grosso*, Brazil. These are tonal languages, i.e. languages in which the pitches produced are grammatically or lexically distinctive, with tones marked orthographically by special characters "¹," "²," and "³" in all syllables of our sample (Lowe, 1999). Orthographic tone marking varies widely across languages, but even where pervasive, it typically evolves organically with compensatory mechanisms to ensure easy written communication. Nambikuára, like many of the languages studied here, does not have a long written tradition, and the development of its writing system is connected the contact between native speakers and peoples of European descent. It is likely that our sample's ubiquitous tonal marking reflects the needs of people unfamiliar with tonality rather than those of it's native speakers. Consequently, this marking likely increases information redundancy without appropriate compensation, affecting the comparability of complexity metrics [6]. We thus removed Nambikuára from our analysis.

4. Encoding choice: since UTF-8 is a variable-size encoding, we encoded our data in UTF-16, to ensure all characters use exactly the same amount of storage;

[6] A similar argument can be found in Sproat (2000, pp. 21–23), using as an example the differences between the standard Hebrew writing system and the Masoretic Hebrew system. The later includes annotations designed to help people who don't speak Hebrew to read the Bible with the correct pronunciation.

5. Compression: we employed Gzip (*gzip*) and Bzip2 (*bz2*) from Python's 3.11.8 standard library. In both cases, we used the maximum compression level available (level 9).

With this pipeline, we have a total of four experimental settings for each metric *(gzip, D90)*, *(bz2, D90)*, *(gzip, Dall)*, *(bz2, Dall)*.

The programs developed for this work are available in an online repository[7].

4 Results and Discussion

Using the complexity metrics computed from the described experimental pipeline, we conducted analyses to empirically validate each proposition outlined in section 3.4, obtaining the following results:

- $\mathcal{H}_1$: we ordered the languages of each dataset by overall complexity $\mu^{\mathbb{A}}$ in ascending order. For instance, the ranking obtained for the setting *(gzip, D90)* was *Nheengatu, Jamamadi, Eastern Bolivian Guarani, Western Bolivian Guarani, Matsés, Maxakali, Guarani Mbya, Parecís, Nadeb, Asheninka, Paraguay Guarani, Tuyúca, Apurinã, Apalaí, English, Kaigang, Macushi, Portuguese, French, Palikúr, Wapishana, Yanomami, German, Spanish, Terêna, Ancient Greek*. It's clear that this ranking does not follow the expectation posed by $\mathcal{H}_1$ that Ancient Greek would be the least complex language, however we have no information on the translation history that connects the different versions of the Bible and which could influence our ranking. Another confounding factor is that Juola (2008) uses the Old Testament as experimental data, originally written in Hebrew. Hebrew, being a Semitic language, lacks orthographic representation of vowels, thus reducing character count. Ancient Greek, our approximate basis for the New Testament original language, features a highly intricate orthographic system with numerous diacritics, significantly increasing character count. Analyzing the correlation between overall complexity and number of distinct characters per language reveals a non-negligible correlation ($\rho = 0.45$, p-value$= 0.019$), suggesting orthographic complexity as a confounding variable that warrants consideration for a more precise assessment of this hypothesis. The fact that *Nheengatu* is in all settings one of the languages of least complexity may be related to its role as a *lingua franca*, or to the possibility that it was used as a basis for the other translations. The evaluation of $\mathcal{H}_1$ is therefore inconclusive, and is subject to a more in-depth study of the translation history of the different versions of the text and orthographic complexity;

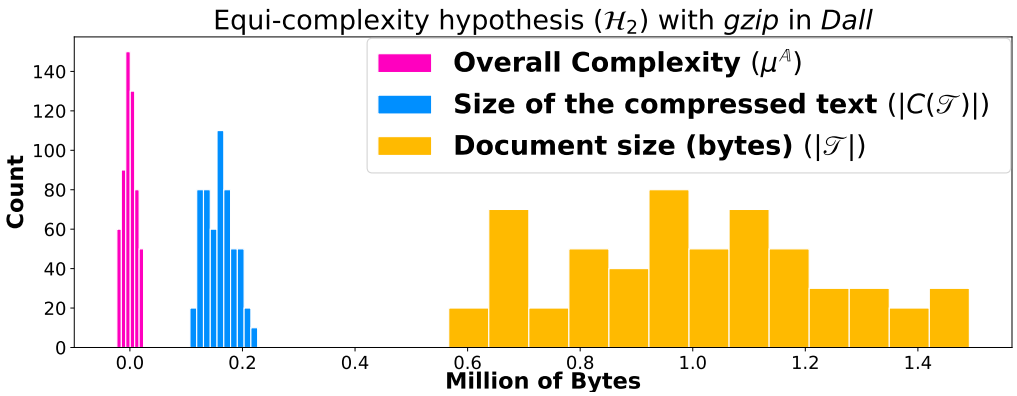
- $\mathcal{H}_2$: we observed that the variance of the compressed text sizes is two orders of magnitude smaller than the variance of the original text sizes while the variance of the overall complexity metric $\mu^{\mathbb{A}}$ is three orders of magnitude smaller, in all scenarios, confirming the hypothesis within our experimental limitations. This is illustrated in Figure 4.

- $\mathcal{H}_3$: we computed the correlation between syntactic and morphological complexity, obtaining negative values in all scenarios and confirming the trade-off hypothesis. In particular, for the set containing all languages, we obtained $\rho = -0.45$, p-value$= 0.0004$ with *gzip* and $\rho = -0.47$, p-value$= 0.0002$ with *bz2*. Analysing the relationship between these complexities, as illustrated in Figure 3, we noted (i) a significant cohesion in complexity space between languages that belong to the same family. This is clearly observable, for example, for the Indo-European, Tupian, Nuclear-Macro-Je, and Arawakan families; (ii) a significant separation between the cluster of Indo-European languages and the clusters of South American languages, indicating that the distance in complexity space can be a meaningful metric of language dissimilarity; (iii) that South American languages have a much greater dispersion in complexities between them than Indo-European languages, reinforcing the need to validate the desired properties of this metrics in a more diverse set of languages, instead of generalizing the results obtained for Indo-European languages. We consider these results as evidence that the trade-off between syntactic and morphological complexities may be dependent on the phylogeny of languages, and usable as feature or tool in language differentiation.

- $\mathcal{O}_1$: as expected, we observed significant positive correlations between morphological com-

[7] Our source code is available in this repository

Figure 4: Compared distributions of original text size $|\mathcal{T}|$, compressed text size $|C(\mathcal{T})|$ and overall complexity $\mu^{\mathbb{A}}$ for the *Dall* subset. The differences in the dispersion of the distributions corroborate $\mathcal{H}_2$.



plexity and the number of types and negative correlations with the number of tokens for all settings. In particular, for *(gzip, D90)* we obtained $\rho_{types} = 0.92$ and $\rho_{tokens} = -0.77$, both with p-value $< 10^{-6}$. This hypothesis was therefore validated;

- $\mathcal{O}_2$: in all scenarios, we observed that the variance of the pragmatic complexity metric is one to three orders of magnitude smaller than the variance of the morphological and syntactic complexities, confirming this hypothesis within our experimental limitations. This corroborates Juola (2008)'s hypothesis that the amount of information transmitted at the intersentential level is language universal, perhaps related to the general cognitive processes of sequential reasoning.

- $\mathcal{O}_3$: we collected the available values of Nichols (1998) morphological complexity metric for the languages in our dataset. Unfortunately, this came down to a small set of six languages. This number of points was too small to obtain a statistically reliable measure of correlation. The evaluation of this hypothesis is therefore inconclusive;

- $\mathcal{O}_4$: the assessment of all previously validated propositions yielded equivalent results for both *gzip* and *bz2*. The hypothesis of their equivalence as base for language complexity measurements is therefore validated within our experimental limitations. Despite this, it's evident that *bz2* typically achieves superior compression compared to *gzip*. However, this isn't always advantageous, as *bz2*'s compression capacity may flatten complexities distributions, complicating the assessment of the trade-off hypothesis $\mathcal{H}_3$.

5 Conclusions and Future Steps

The majority of propositions about the studied complexity metrics ($\mathcal{H}_2$, $\mathcal{H}_3$, $\mathcal{O}_1$, $\mathcal{O}_2$, and $\mathcal{O}_4$) were successfully validated in our vast dataset of South American indigenous languages. These results confirm the robustness of such metrics and indicate the universality of the techniques proposed by (Juola, 2008) to compute the different forms of linguistic complexity. As we used a greater variety of languages, we were also able to document that the trade-off between morphological and syntactic complexities strongly relates with language phylogeny.

Although we confirmed most of our propositions, we obtained inconclusive results for $\mathcal{H}_1$ and $\mathcal{O}_3$, and even for the confirmed hypothesis, we found them to be weaker in South American languages compared to the sets of predominantly Indo-European languages used in the original experiments. This highlights the need to validate and adjust these metrics for a wider range of human languages, a task we have initiated here.

In future research, we aim to investigate the inconclusive propositions, particularly focusing on the impact of orthographic complexity on overall linguistic complexity, extending our results to a greater set of writing systems.

Our findings add to those of (Juola, 2008) and (Ehret and Szmrecsanyi, 2016), expanding the set of languages on which these family of language complexity metrics have been validated.

Limitations

Authors

We, the authors, speak Portuguese, English, and Spanish, with Brazilian Portuguese as our native language. Consequently, we cannot provide insights requiring in-depth knowledge of other languages studied in this work.

Nomenclature of Complexity Metrics

We adhered here to Juola (2008)'s classification of complexity metrics as morphological, syntactic, and pragmatic. However, we believe these names might be misleading.

Regarding syntactic and morphological complexity metrics, it is known that polysynthetic languages like Central Siberian Yupik (not studied here) embed almost all sentence information within words. Many researchers view these process as syntactic rather than morphological, constituting an internal syntax within words (de Reuse, 2006). The metrics studied here would categorize this as morphological complexity instead of syntactic, therefore a more appropriate terminology might be "word complexity" and "sentential complexity."

Regarding pragmatic complexity, the metric used here measures relationships between text parts rather than between the text and external context, typically studied by pragmatics. Thus, a term like "intersentential complexity" might be more suitable.

Data

We used data from the New Testament of the Christian Bible for our experiments. The language in these texts has its own bias, not reflecting the cultural reality of the studied languages. Many translations of this text were made to facilitate colonization, with little regard for cultural and linguistic nuances of each language and people. This could affect our results. We also lacked access to a clear history of translation relationships between versions in different languages, which could have provided a more comprehensive interpretation of $\mathcal{H}_1$. We aim to obtain this data in future work.

Writing Systems

As noted in Section 3.3, the metrics studied here are strongly dependent on the writing systems used to represent target languages. Their applicability is therefore currently limited to alphabetic and low-logographic writing systems. Extensions are needed to apply them to other writing systems.

Acknowledgments

This work was partly supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001; and partly 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. This work was carried out at the Center for Artificial Intelligence (C4AI-USP), supported by FAPESP grant 2019/07665-4 and by the IBM Corporation. Marcelo Finger was partly supported by CNPq grant PQ 302963/2022-7 and Fapesp grant 2023/00488-5. Felipe R. Serras was supported by the IBM Corporation in a grant managed by FUSP under number 3541 and in a PPI-SOFTEX grant managed by FUSP under number 3970. We would like to thank IBM Research Brazil for providing the data that made our experiments possible. We thank Sandro Preto for helping us better understand the Bible's internal structure.

References

Farrell Ackerman and Robert Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language*, 89:429–464.

Matthew Aylett and Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and speech*, 47(1):31–56.

Christian Bentz, Dimitrios Alikaniotis, Michael Cysouw, and Ramon Ferrer i Cancho. 2017. The entropy of words—learnability and expressivity across more than 1000 languages. *Entropy*, 19:275.

Christian Bentz, Ximena Gutierrez-Vasques, Olga Sozinova, and Tanja Samardžić. 2023. Complexity trade-offs and equi-complexity in natural languages: a meta-analysis. *Linguistics Vanguard*, 9(s1):9–25.

Christian Bentz, Tatyana Ruzsics, Alexander Koplenig, and Tanja Samardžić. 2016. A comparison between morphological complexity measures: Typological data vs. language corpora. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 142–153, Osaka, Japan. The COLING 2016 Organizing Committee.

Bram Bulté and Alex Housen. 2014. Conceptualizing and measuring short-term changes in l2 writing

complexity. *Journal of Second Language Writing*, 26:42–65.

John Campbell Callow. 1962. The apinaye language: Phonology and grammar.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 workshop on integrating artificial intelligence and assistive technology*, pages 7–10. Association for the Advancement of Artificial Intelligence.

Paulo Cavalin, Pedro Domingues, Julio Nogima, and Claudio Pinhanez. 2023. Understanding native language identification for brazilian indigenous languages. pages 12–18. Association for Computational Linguistics.

Bastien De Clercq and Alex Housen. 2019. The development of morphological complexity: A cross-linguistic study of l2 french and english. *Second Language Research*, 35:71–97.

W.J. de Reuse. 2006. Polysynthetic language: Central siberian yupik. In Keith Brown, editor, *Encyclopedia of language & linguistics (second edition)*, second edition edition, pages 745–748. Elsevier, Oxford.

Desmond C. Derbyshire and Geoffrey K. Pullum. 1986–1991. *Handbook of Amazonian Languages: Volumes 1–4*. De Gruyter Mounton, Berlin, New York. 4 volumes.

R.M.W. Dixon and A.Y. Aikhenvald. 2006. *The Amazonian Languages*. Cambridge Language Surveys. Cambridge University Press.

William H DuBay. 2007. *Smart Language: Readers, Readability, and the Grading of Text.* ERIC.

Katharina Ehret, Alice Blumenthal-Dramé, Christian Bentz, and Aleksandrs Berdicevskis. 2021. Meaning and measures: Interpreting and evaluating complexity metrics. *Frontiers in Communication*, 6.

Katharina Ehret and Benedikt Szmrecsanyi. 2016. *An information-theoretic approach to assess linguistic complexity*, pages 71–94.

Gertraud Fenk-Oczlon and August Fenk. 2008. *Complexity trade-offs between the subsystems of language*, pages 43–65.

Arthur C. Graesser and Danielle S. McNamara. 2011. Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science*, 3:371–398.

Arthur C. Graesser, Danielle S. McNamara, and Jonna M. Kulikowich. 2011. Coh-metrix. *Educational Researcher*, 40:223–234.

Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36:193–202.

Ximena Gutierrez-Vasques, Christian Bentz, and Tanja Samardžić. 2023. Languages Through the Looking Glass of BPE Compression. *Computational Linguistics*, 49(4):943–1001.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. Glottolog 5.0. Available online at `http://glottolog.org`, Accessed on 2024-07-03.

Alex Housen, Bastien De Clercq, Folkert Kuiken, and Ineke Vedder. 2019. Multiple approaches to complexity in second language research. *Second Language Research*, 35:3–21.

T Jaeger and Roger Levy. 2006. Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems*, 19.

T Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, 61(1):23–62.

Tim Florian Jaeger. 2006. *Redundancy and syntactic reduction in spontaneous speech*. Ph.D. thesis, Stanford University Stanford, CA.

Patrick Juola. 1998. Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Lin-Druckguistics*, 5(3):206–213.

Patrick Juola. 2005. Compression-based analysis of language complexity. *Approaches to Complexity in Language*.

Patrick Juola. 2008. Assessing linguistic complexity. *Language Complexity: Typology, Contact, Change. John Benjamins Press, Amsterdam, Netherlands*.

Päivi Juvonen. 2008. *Complexity and simplicity in minimal lexica: The lexicon of Chinook Jargon*, pages 321–340.

Kimmo Kettunen. 2014. Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics*, 21:223–245.

Sidney Evaldo Leal and Sandra Maria Aluísio. 2024. Complexidade textual e suas tarefas relacionadas. In H. M. Caseli and M. G. V. Nunes, editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, 2 edition, book chapter 23. BPLN.

Sidney Evaldo Leal, Magali Sanches Duran, Carolina Evaristo Scarton, Nathan Siegle Hartmann, and Sandra Maria Aluísio. 2023. Nilc-metrix: assessing the complexity of written and spoken language in brazilian portuguese. *Language Resources and Evaluation*.

Ivan Lowe. 1999. Nambiquara. In R. M. W. Dixon and Alexandra Y. Aikhenvald, editors, *The Amazonian Languages*, pages 269–91. Cambridge University Press, Cambridge.

D.J.C. MacKay. 2003. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.

Dmitrii Manin. 2006. Experiments on predictability of word in context and information rate in natural language. *Journal of Information Processes*, 6(3):229–236.

Aurélien Max. 2006. Writing for language-impaired readers. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 567–570. Springer.

Danielle S McNamara, Arthur C Graesser, Philip M McCarthy, and Zhiqiang Cai. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

Johanna Nichols. 1998. *Linguistic diversity in space and time*. University of Chicago Press.

Johanna Nichols and Christian Bentz. 2018. Morphological complexity of languages reflects the settlement history of the americas. *New Perspectives on the Peopling of the Americas*, pages 13–26.

Yoon Mi Oh and François Pellegrino. 2023. Towards robust complexity indices in linguistic typology: A corpus-based assessment. *Studies in Language*, 47(4):789–829.

François Pellegrino, Christophe Coupé, and Egidio Marsico. 2011. Across-language perspective on speech information rate. *Language*, 87:539–558.

Steven T Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Tiago Pimentel, Clara Meister, Ethan Wilcox, Kyle Mahowald, and Ryan Cotterell. 2023. Revisiting the optimality of word lengths. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2240–2255, Singapore. Association for Computational Linguistics.

Markus Sadeniemi, Kimmo Kettunen, Tiina Lindh-Knuutila, and Timo Honkela. 2008. Complexity of european union languages: A comparative approach. *Journal of Quantitative Linguistics*, 15:185–211.

Gabriele Sarti, Dominique Brunato, and Felice Dell'Orletta. 2021. That looks hard: Characterizing linguistic complexity in humans and language models. pages 48–60. Association for Computational Linguistics.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70.

Advaith Siddharthan. 2006. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4:77–109.

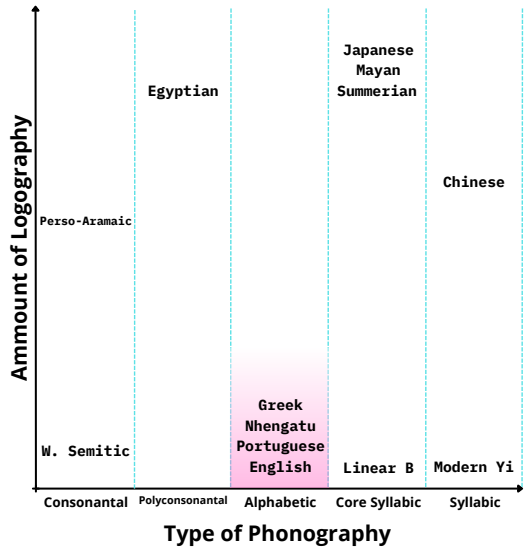
Richard William Sproat. 2000. *A computational theory of writing systems*. Cambridge University Press.

Benedikt Szmrecsanyi. 2021. *Uncovering the Big Picture: Measuring the Typological Relatedness of Varieties of English*, pages 184–208. Cambridge University Press.

G K Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press.

A A contextualization within a planar taxonomy of writing systems

Figure 5: Sproat (2000)[p. 142]'s planar taxonomy of writing systems, organizing them by the *Ammount of Logography*, i.e. the degree to which a system uses single symbols to represent entire words, and the *Type of Phonography*, i.e. which sound units are represented by the symbols in the system. The region of the plane colored in pink (alphabetic and low-logographic systems) corresponds to the types of writing systems where the explored metrics were validated.



B Reference Information for the South American languages studied in this work

Table 1: Reference information about the native South American languages used in this work (Apalaí - Kayapó), partially based on Cavalin et al. (2023)

Language	Code	Family	Countries
Apalaí	apl	Cariban	Brazil
Apinayé	api	Nuclear-Macro-Je	Brazil
Apurinã	apu	Arawakan	Brazil
Asheninka	cax	Arawakan	Peru
Bakairí	bki	Cariban	Brazil
Bororo	brr	Bororoan	Brazil
Canela	cnl	Nuclear-Macro-Je	Brazil
Culina	cul	Arawan	Brazil Peru
Desano	des	Tucanoan	Colombia Brazil
Guajajara	gjj	Tupian	Brazil
Guarani Eastern Bolivia	crg	Tupian	Argentina Bolivia Paraguay
Guarani Mbya	[gun]	Tupian	Argentina Brazil Paraguay
Guarani Paraguay	gua	Tupian	Paraguay
Guarani Western Bolivia	[gnw]	Tupian	Argentina Bolivia Paraguay
Hixkaryána	hix	Cariban	Brazil
Jamamadi	jmm	Arawan	Brazil
Kaapor	urk	Tupian	Brazil
Kadiwéu	kdw	Guaicuruan	Brazil
Kaigang	kng	Nuclear-Macro-Je	Brazil
Kaiwá	kaw	Tupian	Brazil Paraguay
Karajá	jva	Nuclear-Macro-Je	Brazil
Kashinawa	csh	Pano-Tacanan	Brazil Peru
Kayabí	kyz	Tupian	Brazil
Kayapó	kyp	Nuclear-Macro-Je	Brazil

Table 2: Reference information about the native South American languages used in this work (Kubeo - Yanomami).

Language	**Code**	**Family**	**Countries**
Kubeo	cub	Tucanoan	Colombia
Macushi	mac	Cariban	Brazil Guyana Venezuela
Makuna	mcn	Tucanoan	Brazil Colombia
Matsés	myr	Pano-Tacanan	Brazil Peru
Maxakali	max	Nuclear-Macro-Je	Brazil
Mundurukú	muu	Tupian	Brazil
Nadeb	nad	Naduhup	Brazil
Nambikuára	nmb	Nambikwára	Brazil
Nheengatu	[yrl]	Tupian	Brazil Colombia Venezuela
Palikúr	plk	Arawakan	Brazil
Parecís	pex	Arawakan	Brazil
Paumarí	pau	Arawan	Brazil
Piratapúya	prt	Tucanoan	Brazil Colombia
Rikbaktsa	rik	Nuclear-Macro-Je	Brazil
Sanumá	snm	Yanomamic	Brazil Venezuela
Sateré-Mawé	[mav]	Tupian	Brazil
Siriano	sri	Tucanoan	Brazil Colombia
Tenharim	[pah]	Tupian	Brazil
Terêna	trn	Arawakan	Brazil
Ticuna	tic	Ticuna-Yuri	Brazil Peru
Tucano	tuc	Tucanoan	Brazil Colombia
Tuyúca	tuy	Tucanoan	Brazil Colombia
Wanana	gno	Tucanoan	Brazil Colombia
Wapishana	wps	Arawakan	Brazil Guyana
Xavante	xav	Nuclear-Macro-Je	Brazil
Yamináwa	yam	Pano-Tacanan	Brazil Peru
Yanomami	[guu]	Yanomamic	Brazil