NHEENTIQUETADOR: UM ETIQUETADOR MORFOSSINTÁTICO PARA O SINTAGMA NOMINAL DO NHEENGATU

Dominick Maia Alexandre, Juliana Lopes Gurgel, Leonel Figueiredo de Alencar Araripe

RESUMO

Este trabalho tem o objetivo de apresentar os resultados da construção do primeiro etiquetador morfossintático para o sintagma nominal da Língua Geral Amazônica (LGA), ou nheengatu. Em face do decrescente número de falantes e da escassez de recursos de processamento de linguagem natural (PLN) para as línguas indígenas faladas na América Latina, a construção de um etiquetador morfossintático para o nheengatu representa um avanço importante em favor da pesquisa, descrição e preservação dessa língua. A abordagem empregada na construção do etiquetador foi baseada no conhecimento, por meio da implementação de regras, com base nas descrições gramaticais de Navarro (2011) e Cruz (2011). Em sua versão beta, o Nheentiquetador foi testado com relação a uma amostra de 10% das sentenças do corpus compilado, utilizando a métrica F-score. O resultado obtido com esta medida foi 0.83, ou seja, a acurácia da ferramenta na etiquetagem do conjunto de sentenças foi de 83%. Os produtos derivados desta pesquisa envolvem um corpus anotado do nheengatu, um conjunto de etiquetas morfossintáticas, um dicionário em Python e um etiquetador morfossintático. Todos os produtos estão sendo disponibilizados, paulatinamente e sob licença livre, à comunidade acadêmica pela internet.

Palavras-chave: etiquetador morfossintático; Língua Geral Amazônica; processamento de linguagem natural; línguas indígenas.

ABSTRACT

This work aims to present the results of the construction of the first part-of-speech tagger (POS Tagger) for the noun phrase of the Amazonian Lingua Franca (LGA), or nheengatu. Due to the declining number of speakers and the scarcity of natural language processing (PLN) resources for indigenous languages spoken in Latin America, the construction of a POS tagger for nheengatu represents an important advance in favor of the research, description and preservation of this language. The approach used in the construction of the tagger was knowledge-based and involved the implementation of rules based on the grammatical descriptions by Navarro (2011) and Cruz (2011). In its beta version, the Nheentiquetador was tested with a sample of 10% of the sentences of the compiled corpus, using the F-score measure. The result obtained with this measure was 0.83, that is, the tagging accuracy of the software for the set of sentences was 83%. The products derived from this research involve an annotated corpus of nheengatu, a set of morphosyntactic tags, a Python dictionary, and a part-of-speech tagger. All products are gradually being made available open source, to the academic community over the internet.

Keywords: part-of-speech tagger; Amazonian Lingua Franca; natural language processing; indigenous languages.

> "A language is not just words. It's a culture, a tradition, a unification of a community, a whole history that creates what a community is. It's all embodied in a language." (Noam Chomsky)

1. APRESENTAÇÃO

1.1 Introdução

Combinando Linguística e Ciência da Computação, o processamento de linguagem natural (PLN) é tradicionalmente considerado uma subárea da inteligência artificial no Brasil. A princípio, esse campo tecnológico permite o tratamento computacional dos diversos níveis da linguagem humana e o aperfeiçoamento da comunicação entre humanos e computadores (GUINOVART, 2000).

Contudo, o trabalho descritivo empreendido pelo PLN pode, ainda, contribuir para a preservação de línguas em risco de extinção, tal qual o nheengatu, cujos bancos de dados disponíveis e inserção no cenário tecnológico são ínfimos ou inexistentes. Nesse contexto, o presente trabalho apresenta as etapas de construção e testagem do Nheentiquetador 1.0, o primeiro etiquetador morfossintático para as classes de palavras que ocorrem no sintagma nominal da Língua Geral Amazônica.

A etiquetagem morfossintática (*POS tagging*, em inglês) é uma das etapas iniciais do PLN e consiste em atribuir uma etiqueta morfossintática a cada palavra de um dado *corpus* (JURAFSKY; MARTIN, 2019). Apesar da manutenção de corpora eletrônicos anotados exigir, linguística e computacionalmente, tratamento constante, a sua existência é imprescindível em tarefas de processamento de linguagem natural. Além disso, enquanto a existência de etiquetadores morfossintáticos é comum para as línguas majoritárias (ALENCAR, 2013, 2015), às línguas minoritárias é reservado um lugar à margem no que se refere ao tratamento computacional.

Nesse sentido, o desenvolvimento de um recurso de PLN básico como um etiquetador morfossintático, bem como a disponibilização de um *corpus* etiquetado, estabelecem um novo lugar para o nheengatu no atual contexto científico e tecnológico. Nos últimos tempos, empresas de tecnologia e telecomunicações, como a multinacional estadunidense *Motorola*, por exemplo, têm estabelecido parcerias com centros de pesquisa em linguística computacional e PLN, com o fito de incluírem línguas indígenas da América Latina em extinção dentre as suas opções de

configuração. Em 2021[1], o nheengatu foi um dos idiomas incluídos nos aparelhos dessa marca; e a opção já se encontra disponível em todos os dispositivos que receberam a décima primeira atualização do sistema operacional *Android*.

Atualmente, existem aproximadamente 6.000 falantes de nheengatu no Brasil, na região do Alto Rio Negro na Amazônia, cerca de 8.000 na Colômbia e um número ínfimo na Venezuela (EBERHARD; SIMONS e FENNIG 2021). Não obstante, dentro do curto intervalo de cinco anos, o número total de falantes diminuiu de 19.060 em 2016 (LEWIS; SIMONS e FENNIG, 2016), para 14.000 em 2021 (EBERHARD; SIMONS e FENNIG, 2021). Presentemente, o idioma é cada vez menos falado pelas crianças e pelos jovens da região (NAVARRO, 2012), contudo, em seu período de máxima difusão, em meados do século XVII, a LGA já foi falada do Maranhão à fronteira brasileiro-peruana.

Nesse contexto, o presente trabalho possibilita a testagem de descrições gramaticais do nheengatu já existentes, como os trabalhos de Cruz (2011) e de Navarro (2011), bem como contribui para o reconhecimento e visibilidade da LGA, inserindo-a no atual contexto tecnológico dos estudos linguísticos (ALENCAR, 2021). Uma vez que a existência de corpora anotados morfossintaticamente é condição obrigatória para o desenvolvimento de tecnologias de processamento automático de textos, esta pesquisa, que é resultante de um projeto de Iniciação Científica[2], representa um primeiro passo na construção de um banco de dados do nheengatu voltado para o PLN (ALENCAR, 2020). Com tal iniciativa, pretendemos contribuir para a visibilidade do nheengatu e fornecer, para a comunidade científica, um *corpus* útil ao desenvolvimento de ferramentas de PLN e de pesquisas em diferentes áreas das ciências humanas (ALENCAR, 2021).

Isto posto, apresentamos, nas seções 1.2 e 1.3, respectivamente, a relevância e os objetivos deste estudo. Ao longo da seção 2, apresentamos os materiais e métodos utilizados na pesquisa. Na seção 3, apontamos quais foram os resultados obtidos em cada etapa da pesquisa. Em seguida, analisamos as características e capacidades dos produtos gerados. Finalmente, na seção 4, apresentamos uma

[1] Notícia veiculada no portal de notícias brasileiro G1 em 25 de Março de 2021. Disponível em: https://g1.globo.com/economia/tecnologia/noticia/2021/03/25/linguas-indigenas-brasileiras-kaingang-e-nheengat-podem-ser-utilizadas-em-smartphones-motorola.ghtml. Acesso em: 10/10/2021.

[2] Projeto de pesquisa intitulado “Técnicas em softwares livres para a linguística de corpus (12a Etapa)” do Edital PIBIC No 1/2020.

reflexão sobre os próximos passos e trabalhos futuros, haja vista o potencial dos dados e das ferramentas computacionais produzidas até o momento.

1.2 Justificativa

Este trabalho visa preencher uma lacuna existente na Linguística de Corpus, que carece de recursos computacionais para o processamento da Língua Geral Amazônica. Indiretamente, a pesquisa também contribui para a preservação da LGA. Afinal, um *corpus* eletrônico etiquetado é um recurso útil e valioso para a pesquisa e para a documentação de qualquer língua. Não obstante, a produção de tal recurso computacional amplia até mesmo o potencial de popularização da língua, haja vista que o interesse de empresas na inclusão ou no aperfeiçoamento de tecnologias de reconhecimento e processamento de línguas minoritárias têm crescido nos últimos anos.

1.3 Objetivos

O objetivo geral da pesquisa é implementar o Nheentiquetador, um etiquetador morfossintático para as classes de palavras constitutivas do sintagma nominal no nheengatu (ALENCAR, 2020). Para cumprir tal propósito, determinamos os seguintes objetivos específicos: (i) compilar um *corpus* da LGA a partir dos textos das lições do *Curso de Língua Geral*, de Navarro (2011); e (ii) testar a hipótese de que a acurácia do etiquetador construído atinge um F-Score de pelo menos 0.95 na etiquetagem morfossintática do conjunto de sentenças.

2. DESENVOLVIMENTO

Com base nos objetivos geral e específicos da pesquisa, tomamos como princípio metodológico uma abordagem baseada no conhecimento para a construção do Nheentiquetador, que consistiu em implementar regras com base em descrições gramaticais preexistentes, haja vista a inexistência de dados eletrônicos anotados por especialistas humano para treinar um modelo estatístico (ALENCAR, 2020). A seguir, apresentamos as etapas de compilação do *corpus* e da construção do etiquetador.

2.1 Metodologia

Este trabalho envolve duas áreas da linguística: a linguística computacional e a descritiva. Uma vez que o objetivo principal desta pesquisa é a construção de um etiquetador morfossintático para o sintagma nominal (SN) de sentenças em nheengatu, a primeira parte deste trabalho consistiu na compilação do *corpus* do nheengatu para a etiquetagem. A compilação, por sua vez, foi dividida em duas etapas: (i) a compilação dos textos e exemplos das lições do *Curso de Língua Geral* (NAVARRO, 2011) a partir de um arquivo tokenizado do livro; e (ii) a compilação do glossário de Navarro (2011) em uma tabela passível de conversão para a estrutura de dados *Dictionary*, da linguagem de programação Python. A primeira tarefa, isto é, a compilação dos textos de Navarro (2011) para a constituição do *corpus*, consistiu na extração manual das sentenças e na sua compilação em arquivos de texto à parte, utilizando a codificação UTF-8.

Devido à complexidade da segunda etapa, que diz respeito à compilação do glossário, esta foi subdividida em mais três partes: (i) revisão manual das classes de palavras; (ii) pré-processamento para extração; e (iii) extração e finalização. De início, revisamos as entradas lexicais do glossário e listamos verbetes cujas classificações eram ausentes ou incompatíveis com a descrição gramatical do livro ou com a versão mais recente do livro-texto de Navarro, publicada em 2016.

Na parte (ii), dedicada ao pré-processamento dos textos para extração, realizamos a conversão de caracteres especiais, por meio de um programa, denominado “replace-char.py”. Em seguida, modificamos ou adicionamos as classes de palavras nas entradas lexicais que foram listadas na parte (i).

Na parte (iii), extraímos as entradas lexicais que ocorrem no sintagma nominal do nheengatu. No presente trabalho, consideramos as descrições da estrutura do SN e das classes de palavras conforme Cruz (2011), mas adotamos a terminologia das classes do nheengatu segundo Navarro (2011), devido ao seu aspecto formal e simplificador. À luz das duas descrições gramaticais, inventariamos as seguintes classes do SN do nheengatu: nomes, adjetivos, pronomes pessoais, indefinidos, quantificadores, demonstrativos e numerais.

Uma vez extraídas e inventariadas as classes de palavras constitutivas do SN, geramos tabelas de duas colunas por meio de um programa denominado “tag-words.py”. A primeira coluna contém a entrada lexical, que constitui uma chave, e a segunda, a etiqueta correspondente à sua classe gramatical, que constitui um valor atribuído à chave (ver Figura 1).

Figura 1: Tabela de nomes do nheengatu após o processamento pelo programa tag-words.py

```
16 iakumã N
17 sapu N
18 sesaiukisé N
19 kupixaua N
20 garapá N
21 maniuatiua N
```

Fonte: elaboração própria.

Feito isso, expandimos a lista de nomes do nheengatu com as formas flexionadas no plural, por meio de um programa chamado “nominal-flexionizer-yrl.py” (ver Figura 2).

Figura 2: Script do programa flexionador

```
#! /usr/bin/env python3
# -- coding: utf-8 --

# Author: Dominick Maia Alexandre
# <domimaia@alu.ufc.br>
# Date: 11/05/2021

infilename = input('Infile name ')
outfilename = input('Outfile name ')

infile = open(infilename, 'r', encoding='utf-8')
outfile = open(outfilename, 'w', encoding='utf-8')

for i in infile:
    suf_pl = i.replace("-kunhã N-FEM", "-kunhã-itá N-FEM-PL").replace("-apigaua N-MASC", "-apigaua-itá N-MASC-PL")
    if " N-FEM" not in i and " N-MASC" not in i:
        suf_pl = i.replace(" N", "-itá N-PL")
    print(suf_pl, end="", file=outfile)

infile.close()
outfile.close()
```

Fonte: elaboração própria.

Em síntese, o programa adiciona o sufixo *-itá*, morfema que no nheengatu marca o plural, a todos os substantivos definidos por Navarro (2011), isto é, neutros, masculinos e femininos. Na Figura 3, apresentamos uma amostra de como a tabela de nomes se parece depois do processamento pelo programa flexionador.

Figura 3: Tabela de nomes do nheengatu após o processamento pelo programa nominal-flexionizer-yrl.py

```
16 iakumã-itá N-PL
17 sapu-itá N-PL
18 sesaiukisé-itá N-PL
19 kupixaua-itá N-PL
20 garapá-itá N-PL
21 maniuatiua-itá N-PL
```

Fonte: elaboração própria.

Por último, as tabelas de cada classe foram unidas em um só arquivo (ver Figura 4), que foi convertido em uma estrutura de dados do tipo *Python Dictionary* e processado pelo algoritmo do etiquetador morfossintático. Como vemos no exemplo abaixo, na presente versão da tabela, a cada uma das chaves (ou itens lexicais) está atribuída um valor, ou seja, a sua respectiva etiqueta morfossintática. A construção da versão beta do etiquetador foi feita por meio da implementação de uma função capaz de aplicar essa estrutura de dados aos *tokens* de um texto recebido como entrada no etiquetador. Para cada *token*, portanto, é atribuída uma etiqueta, desde que a palavra esteja contida no dicionário; caso contrário, o programa retorna a palavra, sem anotação (ALENCAR, 2020).

Isto posto, realizamos um primeiro teste a fim de verificar a aplicabilidade do dicionário e do *corpus* compilado para utilização na construção do Nheentiquetador, bem como em outras tarefas de PLN. Além disso, buscamos identificar quais aspectos do algoritmo precisam ser aperfeiçoados. No teste, a versão 1.0 do etiquetador recebeu como entrada o conjunto TEST-SET 1, um arquivo contendo as sentenças do texto da primeira lição de Navarro (2011), que perfazem um total de 16 sentenças e 74 palavras, entre itens que ocorrem ou não no sintagma nominal.

Na referida versão do programa (ver Figura 4), implementamos uma função capaz de anotar *tokens* com a etiqueta N-PL, sempre que o sufixo *-itá* estiver presente no item e retornar nomes próprios com letra maiúscula. Assim, além de fazer a atribuição de etiquetas a partir do dicionário, o etiquetador é capaz de diferenciar palavras flexionadas no plural e de identificar nomes próprios em diferentes contextos numa dada sentença.

Figura 4: Fragmento do código da primeira versão do Nheentiquetador 1.0

```
import pandas as pd

yrl_glossary = 'sn-dictAZ3.txt'

infilename = input('Infile name: ')
outfilename="%s_output.txt" % (infilename.split(".",1)[0])

with open(infilename, 'r', encoding='utf-8') as f:
    infile = f.readlines()

# defining function to tag words
def tag_word(line, glossary):
    tagged_word = line.lower()

# tagging words
    for word, tag in glossary.items():
        tagged_word = tagged_word.replace(word.lower(), word.lower() + '\\' + tag)

# tagging nouns inflected in the plural
        if '\\N-itá' in tagged_word:
            tagged_word = tagged_word.replace('\\N-itá', '-itá')
            tagged_word = tagged_word.replace(word.lower(), word.lower() + '\\' + tag)

# printing proper names with capital letters
        if 'N-PRO' in tag:
            tagged_word = tagged_word.replace(word.lower(), word.capitalize())
# printing uppercase tags
    for tag in glossary.values():
            tagged_word = tagged_word.replace(tag, tag.upper())

    return tagged_word[0].upper() + tagged_word[1:]

df = pd.read_csv(yrl_glossary, sep='\t', header=None, index_col=0)
glossary = df.to_dict()[1]

```

Fonte: elaboração própria

Embora o Nheentiquetador 1.0 não apresente erros de execução, o modelo mostrou-se limitado para lidar com grandes volumes de dados. Por esta razão, implementamos o Nheentiquetador 1.5 (ver Figura 5), cuja arquitetura é constituída por comandos mais simples, envolvendo, basicamente, a concatenação das etiquetas às palavras das sentenças recebidas como entrada.

Figura 5: Código do Nheentiquetador 1.5

```
import pandas as pd

yrl_glossary = 'sn-dictAZ3.txt'

with open('licoes1a6.txt', 'r', encoding='utf-8') as f:
    infile = f.readlines()

df = pd.read_csv(yrl_glossary, sep='\t', header=None, index_col=0)
glossary = df.to_dict()[1]

outlines = []

for line in infile:
    list_of_words = line.lower().split()

    new_line = ''

    for word in list_of_words:

        if word in glossary:
            new_line += word + '\\' + glossary[word] + ' '
        else:
            new_line += word + ' '

    outlines.append(new_line.strip() + '\n')

with open('licoes1a6etiquetadas.txt', 'w', encoding='utf-8') as f:
    f.writelines(outlines)
```

Fonte: elaboração própria.

Para avaliar a performance das ferramentas desenvolvidas, realizamos três testes distintos. A seguir, apresentaremos os resultados dos testes e os próximos passos da pesquisa.

3. RESULTADOS DA PESQUISA

Como resultado do primeiro teste, o Nheentiquetador 1.0 obteve uma acurácia de 100% (ver Figura 6) na etiquetagem do conjunto TEST-SET 1. Vale ressaltar, contudo, que todos os itens pertencentes ao sintagma nominal presentes no arquivo de teste constavam no dicionário utilizado e que a parcela do *corpus* testada representa um escopo bastante limitado do banco de dados atual. Portanto, a acurácia alcançada neste primeiro teste é meramente ilustrativa, servindo apenas ao propósito de verificar a aplicabilidade do *corpus* compilado para a construção de ferramentas voltadas para o PLN e de identificar os erros do algoritmo que precisam ser corrigidos a seguir. À esquerda, encontra-se o arquivo de teste dado como entrada no teste da primeira versão do etiquetador; à direita, o arquivo de saída, em que todas as palavras passíveis de ocorrência no SN foram devidamente etiquetadas em todas as sentenças.

Figura 6: Conjunto TEST-SET 2 antes e após etiquetagem

```
Maria anama.

Puranga ara!

Auá taá indé?

Ixé Maria.

Indé puranga, Maria!

Auá taá uiku iké?

Pedro, Maria mena, uiku iké.

Puranga pituna, Pedro!

Maié taá indé resasá?

Puranga tẽ asasá.

Auá taá aé?

Aé Antônio, Maria mimbira.

Pedro, Antônio, aintá Maria anama-itá.

Puranga karuka, Antônio!

Indé puranga!

Kuekatu reté!
```

```
Maria\N-PRO anama\N.\PUNCT

puranga\A1 ara\N!\PUNCT

auá\INT taá indé\PRON1?\PUNCT

ixé\PRON1 Maria\N-PRO.\PUNCT

indé\PRON1 puranga\A1,\PUNCT Maria\N-PRO!\PUNCT

auá\INT taá uiku iké?\PUNCT

Pedro\N-PRO,\PUNCT Maria\N-PRO mena\N,\PUNCT uiku iké.\PUNCT

puranga\A1 pituna\N,\PUNCT Pedro\N-PRO!\PUNCT

maié taá indé\PRON1 resasá?\PUNCT

puranga\A1 tẽ asasá.\PUNCT

auá\INT taá aé\PRON1?\PUNCT

aé\PRON1 Antônio\N-PRO,\PUNCT Maria\N-PRO mimbira\N .\PUNCT

Pedro\N-PRO,\PUNCT Antônio\N-PRO,\PUNCT aintá\PRON1 Maria\N-PRO anama-itá\N-PL.\PUNCT

puranga\A1 karuka\N,\PUNCT Antônio\N-PRO!\PUNCT

indé\PRON1 puranga\A1!\PUNCT

kuekatu\A1 reté!\PUNCT
```

Fonte: elaboração própria

Após a compilação dos textos das treze lições do livro de Navarro (2011), fizemos um segundo teste da versão 1.0 do algoritmo. Desta vez, utilizamos um conjunto de teste maior e mais complexo do ponto de vista lexical, contendo 30 sentenças e 158 palavras, denominado TEST-SET 2. Diante da considerável quantidade de erros apresentada neste teste, entretanto, optamos pela construção de um novo modelo, denominado Nheentiquetador 1.5.

Comentado [1]: . Diante da considerável quantidade de erros apresentada neste teste, entretanto, optamos pela construção de um novo modelo, denominado Nheentiquetador 1.5

No teste desta versão, realizado também com o conjunto de sentenças TEST-SET 2, o Nheentiquetador 1.5 atingiu um F-score de 0.83 na etiquetagem morfossintática automática. Do total de 158 palavras, 72 pertenciam ao SN do nheengatu, das quais 60 palavras foram etiquetadas e 12 não receberam etiqueta, embora estas também pertencessem ao SN do nheengatu e constassem no dicionário. Certamente, a acurácia obtida nesse teste ainda distancia-se do estado da arte, que é 95%. Vale ressaltar, porém, que a presente versão do programa ainda está sendo aperfeiçoada, e que as novas regras ainda não foram implementadas. Apesar disso, os resultados e produtos derivados desta pesquisa abrem horizontes para muitas pesquisas envolvendo o processamento automático da Língua Geral Amazônica.

De antemão, os resultados do trabalho empreendido até o momento apresentam produtos com grande potencialidade. Além de entregar o primeiro *corpus* eletrônico composto por sentenças do nheengatu, esta pesquisa resultou em mais dois produtos úteis à aplicação em tarefas de PLN, conforme a Tabela 1: (i) um dicionário do tipo *Python Dictionary*, contendo as entradas lexicais do nheengatu e suas respectivas etiquetas morfossintáticas; e (ii) um conjunto de etiquetas morfossintáticas das classes que ocorrem no sintagma nominal da LGA (ver Anexo).

Tabela 1: Produtos da pesquisa

	DESCRIÇÃO	**TOTAL**
Tagset	Conjunto de etiquetas	18
Dicionário	Itens lexicais e suas *POS-tags*	522
Corpus	Sentenças em Nheengatu	726

Fonte: elaboração própria.

Atualmente, todos esses itens estão disponíveis à comunidade científica através do GitHub[3], uma plataforma na qual os usuários podem hospedar seus projetos privados ou em código aberto, e que vem sendo muito utilizada por cientistas no mundo todo.

[3] Disponível em: https://github.com/juliana-gurgel/yrl/tree/main/pibic-2020-2021. Acesso em: 12 de out. de 2021.

Em contrapartida, como dito anteriormente, a versão 1.5 do Nheentiquetador ainda está passando por aperfeiçoamentos e correções manuais. Para tanto, os trabalhos de Navarro (2011) e de Cruz (2011) oferecem diversos subsídios para lidarmos com os desafios que emergiram da classificação de palavras do nheengatu.

4. CONCLUSÕES

Para trabalhos futuros, urge aperfeiçoar o algoritmo do etiquetador e testá-lo com relação ao restante do *corpus* compilado. Paralelamente, uma pesquisa de mestrado, ainda em andamento, objetiva a construção de uma ferramenta capaz de etiquetar sentenças inteiras do nheengatu (GURGEL, 2021). No que diz respeito ao dicionário e ao *corpus*, por fim, cumpre ampliá-los através da compilação de textos de outras obras, como Casasnovas (2006). Ademais, esta pesquisa originou três publicações científicas aceitas em eventos nacionais e internacionais. Assim como os outros produtos dessa pesquisa, a versão final do etiquetador para o sintagma nominal da LGA estará disponível no GitHub, sob licença livre, até o final de 2021.

REFERÊNCIAS BIBLIOGRÁFICAS

ALENCAR, L. F. de. Projeto de pesquisa "**Técnicas em softwares livres para linguística de corpus (12ª Etapa)**". Fortaleza: Universidade Federal do Ceará, 2020. Não publicado.

ALENCAR, L. F. de. **Uma gramática computacional de um fragmento do nheengatu**. Revista Estudos da Linguagem, Belo Horizonte, v. 29, n. 3, p. 1717-1777, 2021.

CASASNOVAS, A. **Noções de língua geral ou nheengatú**: gramática, lendas e vocabulário. 2. ed. Manaus: Editora da Universidade Federal do Amazonas; Faculdade Salesiana Dom Bosco, 2006.

CRUZ, A. **Fonologia e Gramática do Nheengatú**: A língua falada pelos povos Baré, Warekena e Baniwa. 2011. 626f. Tese (Doutorado em Linguística) - Faculteit der Letteren, Vrije Universiteit Amsterdam, Utrecht, 2011.

EBERHARD, D. M.; SIMONS, G. F.; FENNIG, C. D. (org.). **Ethnologue**: Languages of the World. 24. ed. Dallas: SIL International, 2021. Disponível em: http://www.ethnologue.com. Acesso em: 10 out. 2021. el em: <http://www.tycho.iel.unicamp.br/~tycho/corpus/> Acesso em 05. jul. 2021.

GUINOVART, X. G. Lingüística computacional. In: RAMALLO, F.; Rei-Doval, G.; Yáñez, X. P. R. (org.). **Manual de Ciencias da Linguaxe**. Edicións Xerais de Galicia. 2000.

GURGEL, J. L. **Nheenga-Tagger**: um etiquetador morfossintático para o nheengatu (working title). Projeto de dissertação (Mestrado em Linguística) - Universidade Federal do Ceará, Fortaleza. Não publicado. 2021.

JURAFSKY, D.; Martin, J. H. **Speech and Language Processing**: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2. ed. Upper Saddle River: Prentice Hall. 2009.

LEWIS, M. P.; SIMONS, G. F.; FENNIG, C. D. (org.). **Ethnologue**: Languages of the World. 19. ed. Dallas: SIL International, 2016. Disponível em: http://www.ethnologue.com. Acesso em: 04 jul. 2021.

MIKHEEV, A. Text segmentation. In: MITKOV, R. (org.). **The Oxford handbook of computational linguistics**. Oxford: Oxford University Press, 2004, p.201-218.

MITKOV, R. (org.). **The Oxford handbook of computational linguistics**. Oxford: Oxford University Press, 2004.

NAVARRO, E. de A. **Curso de Língua Geral (Nheengatu ou Tupi Moderno)**: A Língua das Origens da Civilização Amazônica. São Paulo: Paym Gráfica e Editora, 2011a.

NAVARRO, E. de A. O último refúgio da língua geral no Brasil. **Estudos Avançados**, v. 26, p. 245-254, 2012.

VOUTILAINEN, A. Part-of-speech tagging. In: MITKOV, R. (org.). **The Oxford handbook of computational linguistics**. Oxford: Oxford University Press, 2004, p. 219-232.

ANEXO

QUADRO 1: Conjunto de etiquetas implementado

	ETIQUETA	DESCRIÇÃO	EXEMPLO		ETIQUETA	DESCRIÇÃO	EXEMPLO
1	N	substantivo	*mimbira*	**10**	INDF	artigo indefinido	*iepé*
2	N-PL	plural do substantivo	*anama-itá*	**11**	QUANT	quantificador	*muíri*
3	N-PR	nome próprio	*Maria*	**12**	A1	adjetivo de primeira classe	*puranga*
4	N-FEM	substantivo exclusivamente feminino	*iauara-kunhã*	**13**	A2	adjetivo de segunda classe	*pusé*
5	N-FEM-PL	plural do substantivo feminino	*sumuara-kunhã-itá*	**14**	PRON1	pronome de primeira classe	*ixé*
6	N-MASC	substantivo exclusivamente masculino	*sapukaia-apigaua*	**15**	PRON2	pronome de segunda classe	*se*
7	N-MASC-PL	plural do substantivo masculino	*sapukaia-apigaua-itá*	**16**	POS	pronome possessivo	*se*
8	DEM	pronome demonstrativo	*nhaã*	**17**	NUM	numeral	*mukuĩ*
9	DEM-PL	plural do pronome demonstrativo	*kuá-itá*	**18**	PUNCT	pontuação	*.*

Fonte: Elaboração própria.