UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE HUMANIDADES
DEPARTAMENTO DE LETRAS VERNÁCULAS
PROGRAMA DE PÓS-GRADUAÇÃO EM LINGUÍSTICA

JULIANA LOPES GURGEL

DESAMBIGUADOR MORFOSSINTÁTICO BASEADO EM REGRAS PARA O NHEENGATU

FORTALEZA
2023

JULIANA LOPES GURGEL

DESAMBIGUADOR MORFOSSINTÁTICO BASEADO EM REGRAS PARA O NHEENGATU

Dissertação apresentada ao Programa de Pós-Graduação em Linguística, do Departamento de Letras Vernáculas, da Universidade Federal do Ceará, como requisito parcial para obtenção do título de Mestre em Linguística. Área de concentração: Descrição e Análise Linguística.

Orientador: Prof. Dr. Leonel Figueiredo de Alencar Araripe.

FORTALEZA

2023

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

G987d Gurgel, Juliana Lopes.
Desambiguador morfossintático baseado em regras para o nheengatu / Juliana Lopes Gurgel. – 2023.
140 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Humanidades, Programa de Pós-Graduação em Linguística, Fortaleza, 2023.
Orientação: Prof. Dr. Leonel Figueiredo de Alencar Araripe.

1. Nheengatu. 2. Língua Geral Amazônica. 3. Processamento de Linguagem Natural. 4. Etiquetagem Morfossintática. 5. Desambiguação. I. Título.

CDD 410

JULIANA LOPES GURGEL

DESAMBIGUADOR MORFOSSINTÁTICO BASEADO EM REGRAS PARA O NHEENGATU

Dissertação apresentada ao Programa de Pós-Graduação em Linguística, do Departamento de Letras Vernáculas, da Universidade Federal do Ceará, como requisito parcial para obtenção do título de Mestre em Linguística. Área de concentração: Descrição e Análise Linguística.

Aprovada em 29/08/2023

BANCA EXAMINADORA

Prof. Dr. Leonel Figueiredo de Alencar Araripe (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Victor Pereira do Nascimento Santos
Universidade Federal do Ceará (UFC)

Prof. Dr. Eduardo de Almeida Navarro
Universidade de São Paulo (USP)

Franciena supé, Joana supé, Justo supé, Nicole supé yuíri, aité kwá se anama.
Para Franciena, Joana, Justo e Nicole, minha família.

Nheengatú-itá supé, nheengatú yumbuesara-itá supé, sikarisara-itá supé yuíri.
Aos falantes, aprendizes e pesquisadores de nheengatu.

AGRADECIMENTOS

Ao professor Leonel, meu orientador e referência acadêmica, por todos os *feedbacks*, aulas, indicações de leitura e orientações, desde a Iniciação Científica até hoje, que foram fundamentais no meu caminho em direção à minha aspiração de ser cientista. Agradeço também o apoio e a imensa compreensão em todas as etapas do mestrado, que tornaram muito mais tranquilas mesmo as fases mais desafiadoras desse processo.

À Universidade Federal do Ceará (UFC), por ter sido escola e casa para mim nos últimos dez anos, como tenho certeza que foi para muitos alunos que vieram antes de mim e que será para tantos outros que virão.

À Coordenação e aos professores do Programa de Pós-Graduação em Linguística (PPGL) da UFC, pela estrutura necessária para o desenvolvimento desta dissertação.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela bolsa concedida para a realização desta pesquisa.

Aos queridos professores da graduação e da pós-graduação que contribuíram de maneira especial para a minha formação: Tércia Montenegro, Américo Saraiva, Hebe Carvalho, Elisângela Teixeira e Ronaldo Mangueira, pelas aulas inspiradoras que me fizeram me apaixonar por linguística; Paulo Roberto, por ter me ensinado quase tudo do que hoje sei sobre didática e inglês; Robert de Brose, a quem sou extremamente grata pelas excelentes aulas introdutórias de programação; Suene Honorato, pela iniciativa de criar o grupo de estudos de Tupi Antigo da UFC, que, para mim, foi o pontapé inicial no estudo do nheengatu; e Carlos Augusto, pelos projetos de pesquisa que escrevemos em suas disciplinas, fundamentais para o aprimoramento da minha escrita acadêmica.

Ao professor Eduardo Navarro e à professora Aline da Cruz, pelas obras que foram minhas principais companheiras de mesa nos últimos três anos.

Ao Victor Santos, pela amizade, pelo apoio e pela imensa ajuda na fase mais desafiadora do mestrado, fundamental para o desenvolvimento desta dissertação.

À Dominick Maia, pela amizade, pelo apoio em diversas etapas do desenvolvimento desta pesquisa, pela companhia em todos os grupos de estudos que "inventei" de criar durante o mestrado e pela sua insaciável sede de aprender, que tanto me inspira.

À Juliana Barroso, pela amizade, por sempre me apoiar, por tudo o que me ensinou quando trabalhamos juntas na Iniciação Científica e por topar aprender comigo praticamente tudo o que "invento" de estudar.

Aos amigos Katiusha de Moraes, Felipe Franklin e Leandro Vidal, pela disponibilidade e pelos *feedbacks* na etapa de estruturação da dissertação, fundamentais para o amadurecimento do trabalho.

Aos amigos do CompLin: Daniel Brasil, Hélio Leonam, Viviane Anselmo, João Paulo Abreu, Letícia Farias e Fernando Saraiva, pela construção de um ambiente acolhedor e colaborativo para o exercício do pensar.

Aos servidores Valdirene Silva, Eduardo Xavier, Rodrigo Dantas e Antônia dos Santos, pela disponibilidade sempre que precisei e pela dedicação e competência na execução dos trabalhos na secretaria do PPGL.

Aos meus amigos de ingresso na pós-graduação: Gabriela Zaupa, Paulo Ricardo e Ana Cherllany, pela companhia nesses últimos três anos.

À Naryany Morais, pela amizade, companhia e incentivo em tantos momentos da minha trajetória acadêmica.

À Yasmin Lobo, grande amiga, pela sua presença alegre, pela disponibilidade e pelo suporte fundamental no mestrado e na vida.

Ao Pedro Florêncio, pela amizade e pelos passeios de bicicleta, que tornaram a fase final desta pesquisa muito mais feliz e mais leve.

À Dayana Oliveira, minha amiga mais antiga, pelo apoio, por se disponibilizar a ler este trabalho e pela constância em minha vida. Só consigo dizer essas palavras, pois o resto não caberia aqui.

À Franciena, minha amada mãe, exemplo de força e perseverança, por ter enfrentado todas as dificuldades de criar três filhos sozinha. Por sua inteligência, curiosidade e criatividade, que me fizeram querer ser cientista mesmo antes de eu conhecer o significado de ciência.

À Joana, minha amada irmã, por sua inteligência, sua incrível capacidade de ser multitarefas e seu senso de responsabilidade, que me serviram de exemplo a vida inteira, e por assumir tantas funções na minha criação, mesmo sem serem suas.

Ao Justo, meu amado irmão e principal companheiro de infância e adolescência, por sua mente incomparável e sua coragem e curiosidade de conhecer todos os lugares e experiências do mundo, inspiração para mim sempre que me deparei com o desconhecido.

À Nicole, minha amada sobrinha, cuja existência trouxe muito mais propósito para a família Lopes Gurgel, por me ensinar tanto sobre resiliência e sobre a complexidade do crescer.

Ao meu amado avô Chico, que há vinte e seis anos foi estudar a geologia dos campos santos.

Ao Bruno, amor da minha vida e melhor amigo, pela parceria nos últimos oito anos, por dividir comigo as responsabilidades de forma a tornar este momento possível e pela presença e incentivo em todos os momentos desta jornada.

À Mel, amada enteada, por tudo o que me ensinou e me ensina sobre a vida e a maternidade.

Ao Yargo, meu amigo-irmão, por todas as afinidades e diferenças, por sua generosidade, apoio e presença em minha vida.

À Brenda Souza, minha amiga-irmã, pelo apoio durante minha preparação para a seleção do mestrado e por ter estado ao meu lado em tantos outros momentos importantes da minha vida.

Ao Roger, meu amigo, por estar disponível para o que der e vier.

À Nonna, ao Luiz, à Tata, aos Tios Rose, Rosa, César e Ana, querida família, por todo o incentivo e por celebrarem comigo quaisquer das minhas conquistas, grandes ou pequenas.

RESUMO

Neste trabalho, descrevemos a implementação do módulo desambiguador do Nheengatagger (ALENCAR, 2020), um etiquetador morfossintático para o nheengatu. Esta língua indígena, também conhecida como Língua Geral Amazônica, tem aproximadamente 14.000 falantes e está atualmente em risco de extinção (EBERHARD; SIMONS; FENNIG, 2023). Um dos fatores de risco para a extinção de línguas minoritárias é a indisponibilidade de ferramentas voltadas para o seu processamento computacional. Nesse sentido, o Nheengatagger figura como uma das poucas iniciativas voltadas para o processamento automático do nheengatu. Apesar de anotar corretamente a maioria das palavras de textos com a ortografia adotada por Navarro (2016), é necessário, ainda, que o etiquetador tenha a capacidade de resolver ambiguidades, isto é, atribuir a etiqueta correta a palavras que tenham mais de uma classe gramatical. Para preencher essa lacuna, este trabalho tem como objetivo a implementação de um desambiguador morfossintático baseado em regras. Dividimos nossa metodologia em duas etapas principais: a compilação de textos em nheengatu a partir das obras de Navarro (2016), Navarro e Ávila (2017), Casasnovas (2006) e Trevisan (2017) e a implementação da ferramenta. A compilação dos textos resultou em um corpus com 4176 sentenças. As etapas de implementação do desambiguador foram: (i) o levantamento de ambiguidades; (ii) a análise dos contextos das classes de palavras do nheengatu; (iii) a implementação do algoritmo; e (iv) a avaliação. Na etapa (i), identificamos 55 tipos de ambiguidades, com um total de 1047 ocorrências no corpus de desenvolvimento (NAVARRO, 2016). Na etapa (iv), o desambiguador alcançou as acurácias de 52% e 74% nos dois testes preliminares realizados com relação a um conjunto de 50 sentenças, resultado abaixo do estado da arte para esse tipo de ferramenta, que é 95%. A partir dos resultados desses testes, decidimos avaliar a performance da ferramenta a partir de contextos extraídos de sentenças com e sem ambiguidades resolvidas. Nos três testes realizados após os ajustes na ferramenta, obtivemos, respectivamente, de 80.9%, 60% e 57.5% de acurácia. Por outro lado, o desambiguador aumentou significativamente a taxa de acerto do Nheengatagger. Após a integração do módulo, o etiquetador alcançou os índices de 88.9%, 95.4% e 96.2% nos três testes.

Palavras-chave: nheengatu; língua geral amazônica; processamento de linguagem natural; etiquetagem morfossintática; desambiguação.

ABSTRACT

In this study, we describe the implementation of a disambiguation module for Nheengatagger (ALENCAR, 2020), a part-of-speech tagger for Nheengatu. This indigenous language, also known as the Amazonian Lingua Franca, has an estimated number of 14,000 speakers and is currently at risk of extinction (EBERHARD; SIMONS; FENNIG, 2023). One of the risk factors for minority language extinction is the unavailability of low-resource language tools aimed at their computational processing. In the perspective of automatic text processing of Nheengatu, Nheengatagger is one of the few initiatives. Despite correctly labeling most of the words in texts in which the orthography adopted by Navarro (2016) was used, it is still necessary for the tagger to be able to resolve ambiguities, that is, to assign the correct label to words that have more than one part-of-speech. Thus, this work aims at implementing a rule-based disambiguation module. We divided our methodology into two main stages: the compilation of texts in Nheengatu from the works of Navarro (2016), Navarro and Ávila (2017), Casasnovas (2006) and Trevisan (2017) and the implementation of the module. The compilation of the texts resulted in a corpus with 4176 sentences. The stages of implementation were: (i) the identification of ambiguities; (ii) the analysis of the contexts of the parts-of-speech in Nheengatu; (iii) the implementation of the algorithm; and (iv) the evaluation. In stage (i), we identified 55 types of ambiguities, with a total of 1047 occurrences in the development corpus (NAVARRO, 2016). In stage (iv), the disambiguator achieved accuracies of 52% and 74% in the two preliminary tests carried out with a set of 50 sentences, a result below the state-of-the-art for this type of tool, which is 95%. Based on the results of the preliminary tests, we decided to evaluate the performance of the tool using contexts extracted from sentences with and without ambiguities. In the three tests carried out after adjustments to the tool, we obtained accuracies of 80.9%, 60% and 57.5%, respectively, a result still below the state-of-the-art. On the other hand, the disambiguator significantly increased Nheengatagger's hit rate. After the integration of the module, the POS tagger achieved rates of 88.9%, 95.4% and 96.2% in the three tests, respectively.

Keywords: Nheengatu; Amazonian Lingua Franca; natural language processing; part-of-speech tagging; disambiguation.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E CONVENÇÕES

2PL	segunda pessoa do plural
3PL	terceira pessoa do plural
1SG	primeira pessoa do singular
2SG	segunda pessoa do singular
3SG	terceira pessoa do singular
ART	artigo
COND	condicional
DEM	demonstrativo
DIST	distal
FUT	futuro
INDF	indefinido
IPFV	imperfectivo
NEG	negação
PFV	perfectivo
PL	plural (partícula)
PROX	proximal
PST	passado (partícula)
Q	interrogativo (partícula)
REL	relativo
*	dado agramatical

SUMÁRIO

1 INTRODUÇÃO

1.1 Breve panorama da linguística computacional

O processamento automático das línguas naturais é um dos empreendimentos tecnológicos e científicos mais desafiadores do século XXI. O nascimento desta área está vinculado aos primeiros esforços dedicados à tradução automática realizados no início dos anos 1950, como a publicação do *Relatório Weaver*, relevante trabalho de Warren Weaver que destacou a potencial aplicação de técnicas matemáticas e computacionais para a tradução. Na mesma década, em 1957, Noam Chomsky publicou o livro *Estruturas Sintáticas*, no qual apresenta a gramática gerativa, considerado um dos pilares da linguística moderna. Além de ter elaborado uma obra revolucionária no que se refere à teoria sobre a natureza da linguagem, à luz da gramática gerativa Chomsky deu um passo adiante na utilização de formalismos para a representação de fenômenos linguísticos, fato que foi de fundamental importância para o avanço da linguística computacional nas décadas posteriores, pois tornou evidente que modelos gramaticais assim formalizados são adequados à implementação.

A linguística computacional, que combina os domínios da ciência da computação e da linguística, é a disciplina dedicada ao estudo e à construção de modelos computacionais das línguas humanas. Esta área tem se expandido gradativamente à medida que se torna cada vez mais relevante para o desenvolvimento de sistemas que possam entender, interpretar e gerar dados de línguas naturais de maneira semelhante aos seres humanos. Nesse sentido, pesquisas em linguística computacional contribuem para o avanço tecnológico e para a teoria linguística. Do ponto de vista teórico, possibilitam a testagem de hipóteses sobre o funcionamento da linguagem por meio da implementação de modelos gramaticais (KARTTUNEN, 1996; SAG; WASOW; BENDER, 2003; MÜLLER, 2020). Do ponto de vista tecnológico, auxiliam na construção de diversos recursos voltados para o processamento automático das línguas naturais que utilizamos cotidianamente, como tradutores automáticos, *chatbots*, corretores automáticos, entre outros. Em razão de sua complexidade, esse tipo de tecnologia precisa ser capaz de lidar com vários aspectos não triviais da linguagem, o que exige, além de habilidades de programação, conhecimentos em áreas da linguística como a fonologia, a sintaxe, a semântica e a pragmática (MITKOV, 2004; JURAFSKY; MARTIN, 2009).Guinovart (2000) reconhece três vertentes da linguística computacional: (i) a linguística computacional teórica, voltada para “a elaboração de modelos linguísticos em termos formais

e implementáveis", para "a aplicação desses modelos a qualquer nível de descrição linguística" e para "a comprovação automatizada da consistência de uma teoria linguística das suas predições" (GUINOVART, 2000, p. 221, tradução nossa); (ii) a linguística computacional aplicada ou Processamento de Linguagem Natural (PLN), área em que o nosso trabalho se insere, dedicada ao estudo e ao desenvolvimento de ferramentas voltadas para a compreensão e para a geração de dados linguísticos de maneira automática; e (iii) a informática aplicada à linguística, caracterizada pela "aplicação de computadores à pesquisa linguística" (GUINOVART, 2000, p. 222, tradução nossa) na coleta e análise de dados linguísticos reais, os quais são utilizados como base para a formulação de generalizações sobre o funcionamento da linguagem e como recursos para a construção de aplicações computacionais.

Em geral, ferramentas de PLN têm sido predominantemente desenvolvidas para as línguas humanas majoritárias, como o inglês, o português, etc. (MARCUS *et al.*, 1993; GALVES; ANDRADE; FARIA, 2017). Já as línguas minoritárias, isto é, as línguas faladas por pequenos grupos de habitantes em uma determinada região ou país, ainda dispõem de poucos recursos voltados para o processamento automático de textos. Por outro lado, cresce o número de cientistas interessados em desenvolver ferramentas de PLN para línguas minoritárias, como as línguas indígenas, e já existem projetos relevantes nesse sentido, como o Universal Dependencies[1] (NIVRE *et al.*, 2016), projeto voltado para a construção de bancos de árvores sintáticas (do inglês *treebanks*) de diversas línguas do mundo, e o projeto Naki[2] (MAGER *et al.*, 2018), que objetiva a criação de ferramentas de PLN para línguas indígenas das Américas, entre as quais destacamos Mbya Guarani, Shipibo Konibo e Cusco Quechua, que estão entre as línguas com recursos disponíveis no Universal Dependencies.

1.2 O lugar do nheengatu no processamento automático das línguas naturais

As línguas oficiais do Brasil são o português e a Língua Brasileira de Sinais (Libras). Além da Libras, o Brasil possui dezenas de línguas minoritárias autóctones, faladas pelos povos que, segundo Bueno *et al*. (2003), há mais de 10 mil anos habitam a região que hoje corresponde ao território nacional. Atualmente, estas línguas se encontram ameaçadas de extinção por diversas razões, como o número ínfimo de falantes vivos, a ausência de

[1] Atualmente, as línguas indígenas brasileiras disponíveis são: Akuntsu, Apurina, Bororo, Guajajara, Kaapor, Karo, Madi, Makurap, Munduruku, Nheengatu, Tupinamba, Xavante. Disponível em: https://universaldependencies.org/. Acesso em: 28 nov. 2023.
[2] Disponível em: https://github.com/pywirrarika/naki. Acesso em: 28 nov. 2023.

transmissão intergeracional, a documentação insuficiente ou inexistente, entre outras. O último censo do IBGE (2010)[3] registrou 305 etnias e 275 línguas indígenas no Brasil. Desse total, 190 estão incluídas no *Atlas of the World's Languages in Danger*[4] (MOSELEY, 2010), uma publicação da *United Nations Educational, Scientific and Cultural Organization* (UNESCO). Por outro lado, o *Ethnologue*[5], publicação mais recente, registrou[6] um total de 22 línguas indígenas extintas e 202 línguas existentes no Brasil, a maior parte em perigo de extinção ou praticamente extinta, entre elas o nheengatu (EBERHARD; SIMONS; FENNIG, 2023).

O nheengatu é uma língua indígena falada, sobretudo, na região da bacia do Rio Negro por aproximadamente 14.000 pessoas. O termo *nheengatu* (língua boa) é formado a partir da composição dos vocábulos *nheenga* (língua, palavra) e *katu* (bom). De acordo com Edelweiss (1969, p. 198), "quem lançou[7] o termo *nheengatu* foi o general Couto de Magalhães", na obra *O selvagem*, publicada em 1876. Além de nheengatu, também é conhecida como Língua Geral Amazônica (LGA), Tupi Moderno ou *yeral*, sendo, esta última, a forma utilizada para referência da língua na Venezuela e na Colômbia. A formação do nheengatu teve origem no século XVI, início do período colonial, e foi formada a partir do tupi antigo, que, possivelmente, já era utilizado como língua franca por povos indígenas que habitavam a costa do Brasil antes da chegada dos portugueses, ainda que "em estado embrionário" (FREIRE, 2011, p. 138). No processo de ocupação, especula-se que a língua tenha sido adotada pelos colonizadores como meio de comunicação com os nativos, e, em algumas circunstâncias, entre os nativos entre si, fato que pode ter contribuído para a extinção de diversas línguas que outrora eram faladas na região amazônica (AVILA, 2021; RODRIGUES, 1993).

A língua entrou em declínio no século XIX devido a diversos fatores, sobretudo econômicos e políticos, como as Guerras dos Cabanos e do Paraguai, que, juntas, exterminaram mais de 42.000 falantes (FREIRE, 2011). Segundo o *Ethnologue*, estima-se que existem hoje 6.000 falantes do nheengatu no Brasil, 8.000 na Colômbia e um número ínfimo

---

[3] Ainda não estão disponíveis os resultados do Censo 2023 sobre as etnias e línguas indígenas.
[4] Disponível em: https://unesdoc.unesco.org/ark:/48223/pf0000187026. Acesso em: 01 jun. 2023.
[5] O *Ethnologue* é um banco de dados sobre as línguas conhecidas do mundo, cuja primeira publicação data de 1951. Disponível em: https://www.ethnologue.com/. Acesso em: 01 jun. 2023.
[6] O *Ethnologue* utiliza as fontes de dados oficiais, como Censos e outras pesquisas realizadas por órgãos multilaterais, como a Organização das Nações Unidas. O número de línguas é aferido a partir da separação do total de línguas vivas em relação ao total de línguas extintas. Cf. mais informações em: https://www.ethnologue.com/methodology/#languagePages. Acesso em: 28 nov. 2023.
[7] Aqui, podemos entender a palavra *lançar* com o sentido de *divulgar* ou *popularizar*. Como o próprio autor aponta, há registros do termo *nheengatu* anteriores à publicação de *O selvagem*.

na Venezuela (EBERHARD; SIMONS; FENNIG, 2023). Esta mesma publicação classifica o nheengatu com *status Threatened* (ameaçada)[8], enquanto a UNESCO classifica a língua com *status Severely endangered* (severamente ameaçada)[9]. Embora em comparação a outras línguas indígenas brasileiras com o mesmo *status*[10] a LGA ainda tenha uma quantidade considerável de falantes vivos, o número diminuiu notavelmente em apenas cinco anos, de um total de 19.060 falantes em 2016 para 14.000 em 2020[11] (LEWIS; SIMONS; FENNIG, 2016; EBERHARD; SIMONS; FENNIG, 2023). Apesar da situação crítica em que se encontra, há mais de uma década o nheengatu tem sido consideravelmente difundido por meio de diversas iniciativas dentro e fora do ambiente acadêmico, de um lado com publicações científicas, grupos de estudos e disciplinas em universidades, e de outro com oficinas, minicursos e capacitação de professores promovidos por organizações e lideranças indígenas (CRUZ, 2001; SILVA, VAZ FILHO, 2019). Além disso, foi estabelecida como uma das línguas oficiais do município São Gabriel da Cachoeira, no Amazonas, e do município Monsenhor Tabosa, no Ceará[12].

Do ponto de vista do processamento automático de textos, contudo, ainda existem poucas iniciativas voltadas para a construção de recursos para o nheengatu. Temos conhecimento de três trabalhos: (i) uma pesquisa de Iniciação Científica (ALENCAR, 2020) que objetivou implementar um etiquetador morfossintático para o sintagma nominal desta língua e compilar um conjunto de textos em nheengatu para anotação (NAVARRO, 2011); (ii) a GrammYEP, uma gramática computacional multilíngue composta por fragmentos equivalentes do nheengatu, do inglês e do português, implementada no formalismo GF (do inglês *Grammatical Framework*) (ALENCAR, 2021); e (iii) o UD_Nheengatu-CompLin[13],

---

[8] Descrição do *status* no site do *Ethnologue*: "A língua é usada para comunicação face a face em todas as gerações, mas está perdendo usuários" (EBERHARD; SIMONS; FENNIG, 2023, tradução nossa). Disponível em: https://www.ethnologue.com/methodology/#Status2. Acesso em 01 jun. 2023.

[9] Descrição do *status* no *Atlas*: "A língua é falada apenas pelos avós e pelas gerações mais velhas; embora a geração dos pais ainda possa entender a língua, eles normalmente não a falam com seus filhos ou entre si." (MOSELEY, 2010, p. 12, tradução nossa). Na versão mais recente do site do Atlas, o nheengatu é classificado como *Definitely endangered* (definitivamente ameaçada), entretanto, o site não fornece informações sobre o critério de classificação ou a fonte dos dados que fundamentam a classificação. Site disponível em: https://en.wal.unesco.org/discover/languages. Acesso em: 01 jun. 2023.

[10] Outras línguas também classificadas como *Threatened* (ameaçada) pelo *Ethnologue* contam com menos de mil falantes, como a Pirahã e a Karapanã, que têm, respectivamente, 390 e 63 falantes registrados (dados de 2012) (EBERHARD; SIMONS; FENNIG, 2023). Disponível em: https://www.ethnologue.com/country/BR/. Acesso em 01 jun. 2023.

[11] No Brasil, o número de falantes registrados no *Ethnologue* caiu de 10.300 para 6.000 entre 2016 e 2020.

[12] Em 2002, a Lei Municipal 145/2002, de São Gabriel da Cachoeira, no estado do Amazonas, estabeleceu a LGA como uma das línguas oficiais. Em maio de 2021, a Lei Municipal 13/2021, de Monsenhor Tabosa, no Ceará, reconheceu o nheengatu, referenciado no texto da lei por Tupi-nheengatu, como língua cooficial.

[13] Disponível em: https://github.com/UniversalDependencies/UD_Nheengatu-CompLin/tree/dev. Acesso em: 01 jun. 2023.

um *treebank* (banco de árvores) em desenvolvimento no âmbito de outra pesquisa de iniciação científica (ALENCAR, 2022), que contém sentenças em nheengatu de diversas fontes anotadas no modelo Universal Dependencies.

Pelo que sabemos, a iniciativa de Alencar (2020) é pioneira na construção de ferramentas voltadas para a anotação automática de corpora desta língua. O etiquetador morfossintático que começou a ser desenvolvido para subdomínio do sintagma nominal foi posteriormente refeito, expandido para a anotação de sentenças inteiras em nheengatu e denominado Nheengatagger[14].

A tarefa de etiquetagem consiste em atribuir, a cada palavra de um texto, uma etiqueta indicando sua classe gramatical e pode ser realizada por meio de uma abordagem estatística ou baseada em regras. Os etiquetadores que representam o estado da arte são estatísticos, ou seja, fazem a anotação por meio de algoritmos que calculam a probabilidade de uma determinada palavra ambígua ocorrer como uma ou outra classe gramatical a partir de um corpus previamente etiquetado e selecionam a etiqueta mais adequada para o contexto considerando a frequência (JURAFSKY; MARTIN, 2023a). Uma vez que inexistem corpora do nheengatu anotados, na construção do Nheengatagger foi aplicada a abordagem baseada em regras (ALENCAR, 2020). Um léxico robusto foi construído a partir do glossário de Navarro (2016), do dicionário de Ávila (2021) e de outros trabalhos, e atualmente conta com com mais de 1.000 entradas lexicais. Além disso, com o objetivo de expandir o léxico, em um dos componentes do etiquetador foram implementadas funções que modelam processos morfológicos do nheengatu, como a flexão nominal e verbal.

Por outro lado, apesar de ser capaz de anotar morfossintaticamente a maioria das palavras dos textos com a ortografia utilizada por Navarro (2016), a ferramenta ainda não resolve ambiguidades, isto é, não atribui a etiqueta correta a palavras do nheengatu que, de acordo com a sua entrada no léxico, têm mais de uma classe gramatical, como *casa*, em português, que pode ser um substantivo ou um verbo. Para preencher esta lacuna, este trabalho visa a construção de um módulo desambiguador para o Nheengatagger. Nesta pesquisa, três fatores foram determinantes para a escolha da língua: (i) o risco de extinção; (ii) a disponibilidade de descrições gramaticais e de textos em nheengatu, considerando a complexidade do trabalho e o tempo necessário para a sua realização; e (iii) a quantidade ainda insuficiente de ferramentas voltadas para o processamento computacional da língua. Por

[14] O código está disponível em: https://github.com/CompLin/nheengatu/tree/main/src e o corpus etiquetado pelo Nheengatagger está disponível em: https://github.com/CompLin/nheengatu/tree/main/data/corpus/navarro-2016. Acesso em: 01 jun. 2023.

meio do aprimoramento do Nheengatagger com este módulo desambiguador, pretendemos dar mais um passo no desenvolvimento de recursos computacionais para a LGA e contribuir para a expansão dos corpora do nheengatu anotados morfossintaticamente, para os estudos linguísticos, para a área de PLN e, acima de tudo, para a preservação do nheengatu.

1.3 Objetivos

O objetivo principal desta pesquisa é implementar o desambiguador do Nheengatagger em Python. Para realizar esta tarefa, estabelecemos os seguintes objetivos específicos: (i) compilar um corpus do nheengatu a partir dos trabalhos de Navarro (2016), Navarro e Ávila (2017), Trevisan (2017) e Casasnovas (2006); (ii) implementar o algoritmo do desambiguador; (iii) testar a hipótese de que o desambiguador apresentará uma acurácia de no mínimo 95% na resolução das ambiguidades de um conjunto de sentenças, índice que é atingido pelos sistemas de etiquetagem mais sofisticados; e (iv) testar a hipótese de que, com o módulo desambiguador, o Nheengatagger apresentará uma acurácia de no mínimo 95%.

1.4 Estrutura do trabalho

Este trabalho está dividido em cinco partes, além desta introdução. No capítulo 2, apresentamos os principais aportes teóricos que baseiam nosso estudo. A primeira parte deste capítulo diz respeito à tarefa de etiquetagem morfossintática e a segunda parte trata das abordagens gramaticais do nheengatu utilizadas neste estudo. No capítulo 3, descrevemos os materiais e os métodos utilizados na compilação dos textos em nheengatu e na implementação do desambiguador. No capítulo 4, apresentamos os resultados obtidos nos testes da ferramenta e, por fim, no capítulo 5, apresentamos nossas considerações finais.

2 FUNDAMENTAÇÃO TEÓRICA

O presente trabalho envolve duas áreas do estudo da linguagem: a linguística computacional e a linguística descritiva. Na seção 2.1, descrevemos, de forma não exaustiva, a tarefa de etiquetagem morfossintática. Na seção 2.2, apresentamos uma breve descrição da morfologia do nheengatu, tendo em vista o nosso objetivo principal, que é a implementação do desambiguador do Nheengatagger.

2.1 Etiquetagem morfossintática

A área do Processamento de Linguagem Natural (PLN) dedica-se à construção de sistemas computacionais capazes de lidar com as modalidades oral e escrita das línguas naturais. As tarefas de PLN envolvidas na construção desses sistemas e as técnicas utilizadas estão amplamente descritas em vários livros e manuais, entre os quais destacamos o *Oxford Handbook of Computational Linguistics*, organizado por Mitkov (2004), o *Speech and Language Processing*, de Jurafsky e Martin (2009)[15] e o *Natural language processing with Python*, manual da biblioteca NLTK[16], de Bird, Klein e Loper (2009). Para o propósito desta pesquisa, limitaremos a revisão da literatura à tarefa de etiquetagem morfossintática.

Para um indivíduo fluente em uma determinada língua, a leitura de um texto escrito nessa língua é um processo quase intuitivo. No entanto, para um computador, o processamento de um texto requer uma série de procedimentos complexos que vão além do seu funcionamento padrão. Isto se deve ao fato de que, na sua forma mais básica, os textos digitais são meras cadeias de caracteres, desprovidas de distinções de tipos, como alfabéticos, numéricos, etc. Existem diversas técnicas de PLN para lidar com a língua escrita, que vão desde as de baixo nível, como a tokenização e a segmentação sentencial, empregadas na delimitação de sequências de palavras e sentenças, até as de alto nível, como a sumarização, aplicada na geração de resumos de textos mais longos, e a análise de sentimento, utilizada para determinar se um texto expressa uma opinião positiva, negativa ou neutra. Portanto, para que um computador seja capaz de "ler" e "compreender" um texto de maneira similar a um humano – como o Chat GPT, por exemplo –, antes de empregar técnicas avançadas de PLN é

[15] Os capítulos do livro *Speech and Language Processing*, de Jurafsky e Martin, têm sido constantemente atualizados ao longo dos anos. As versões mais recentes, publicadas em 2023, estão disponíveis sob licença livre em: https://web.stanford.edu/~jurafsky/. Acesso em: 03 jun. 2023.
[16] Disponível em: https://www.nltk.org/book/. Acesso em: 03 jun. 2023.

preciso empregar técnicas basilares, como a etiquetagem morfossintática (JURAFSKY; MARTIN, 2023a; OpenAI, 2023).

Etiquetagem morfossintática ou *part-of-speech tagging* é a tarefa do PLN que consiste em atribuir a cada *token* de um texto dado como entrada num etiquetador uma etiqueta indicando a sua classe gramatical, considerando suas propriedades morfológicas e distribucionais (JURAFSKY; MARTIN, 2023a). O termo *part-of-speech* (*POS*) pode ser traduzido para o português como "classe de palavra" ou como "partes do discurso". Na área de PLN, a tradução para "classe de palavra" tem sido a mais utilizada na literatura quando o termo se refere às classes de palavras em si (ALENCAR, 2013; KEPLER, 2005). Contudo, algumas diferenças entre os termos *word classes* (classes de palavra) e *parts-of-speech* (partes do discurso) apontadas por Jurafsky e Martin (2019) valem ser destacadas:

> Até agora, utilizamos termos de parte do discurso como **substantivo** e **verbo** livremente. Nesta seção, damos uma definição mais completa dessas e de outras classes. Embora as classes de palavras tenham tendências semânticas — adjetivos, por exemplo, geralmente descrevem *propriedades* e substantivos *pessoas* — as partes do discurso são tradicionalmente definidas com base na função sintática e morfológica, agrupando palavras que têm palavras vizinhas semelhantes (suas propriedades **distribucionais**) ou recebem afixos semelhantes (suas propriedades morfológicas). (JURAFSKY; MARTIN, 2019, p. 1-2, grifo dos autores, tradução nossa)

Neste trecho, Jurafsky e Martin (2019) apresentam dois critérios para a definição de categorias gramaticais, um semântico, outro sintático e morfológico. Além dos autores supracitados, Schachter (1985) também explora essa questão, afirmando que "os critérios primários para a classificação de partes do discurso são gramaticais, não semânticos", pois existem muitos casos em que as definições de natureza semântica "falham em fornecer base adequada para a classificação das partes do discurso, uma vez que existem muitos casos em que sua aplicabilidade ou inaplicabilidade não é clara" (SCHACHTER, 1985, p. 01, tradução nossa). Consideremos os seguintes exemplos:

(1) Maria está no rio
(2) *Maria no está rio
(3) *está rio no Maria

Do ponto de vista morfológico, sabemos que a palavra *rio* pode ser submetida à flexão de número (*rios*), mas não de tempo, enquanto a palavra *está* sofre flexão de tempo e número. No exemplo (1), os itens lexicais que formam a sentença têm propriedades

distribucionais diferentes e não podem ser organizados em qualquer ordem, tal como demonstramos em (2)-(3). Por esta razão, a etiquetagem morfossintática não é uma tarefa trivial e exige uma perspectiva gramatical das classes de palavras, que utilize critérios precisos para a classificação.

No processo de anotação de um texto dado como entrada num etiquetador, cada palavra e cada sinal de pontuação corresponde a um *token* que será anotado, como podemos observar nos exemplos abaixo:

(4) Maria está no rio.

(5) [“Maria”, “está”, “no”, “rio”, “.”]

(6) Maria N
rio N
está V
no P
. PUNCT

(7) Maria/N está/V no/P rio/N ./PUNCT

O exemplo (4) corresponde ao texto dado como entrada no etiquetador. Os exemplos (5-6) apresentam, respectivamente, a lista dos cinco *tokens* que compõem a sentença após a segmentação e o léxico dos cinco *tokens* atribuídos a uma etiqueta indicando sua classe. Na tarefa de anotação, o etiquetador recebe o texto de entrada, faz a atribuição das etiquetas correspondentes e fornece como saída a sentença anotada, conforme o exemplo (7). Neste item, vemos que a etiqueta N (substantivo) foi atribuída ao *token Maria*, V (verbo) foi atribuída ao *token está*, e assim por diante. Sobre esse exemplo, vale ressaltar dois pontos: primeiro, a etiqueta N foi atribuída a *rio* e *Maria* porque a distinção entre nome comum e nome próprio, que pode ser feita por meio da criação de outra etiqueta, não foi considerada nesse modelo; segundo, o ponto final foi anotado como PUNCT porque o etiquetador deve ser capaz de diferenciar, por exemplo, casos em que o ponto final faz parte de uma abreviatura ou se marca o limite de uma sentença (MIKHEEV, 2004; JURAFSKY; MARTIN, 2023a, 2023b). Além disso, grande parte da tarefa de etiquetagem consiste na desambiguação de palavras que tem mais de uma classe gramatical, de forma a atribuir a essas palavras as etiquetas apropriadas ao contexto em que se encontram na sentença, por exemplo:

(8) a **casa** de Maria é bonita

(9) Maria **casa** com Pedro hoje

A palavra *casa* pode ser classificada como substantivo feminino singular no exemplo (8) e como verbo flexionado na terceira pessoa do singular do presente do indicativo em (9). Para resolver ambiguidades como a exemplificada, existe uma linha de base comum em modelos estatísticos conhecida como *Most Frequent Class Baseline*: "dada uma palavra ambígua, escolha a etiqueta mais frequente do corpus de treino" (JURAFSKY; MARTIN, 2023a, p. 5, tradução nossa). Com base na frequência, um etiquetador pode atribuir corretamente as etiquetas não apenas a palavras conhecidas, ou seja, aquelas que já estão contidas no corpus de desenvolvimento, mas também a palavras novas. Como a maioria das palavras desconhecidas pertencem às chamadas classes abertas (substantivos, verbos, adjetivos, etc.), em alguns casos esse problema pode ser solucionado por meio da modelagem de processos de formação de palavras (ver seção 2.1.1). Na construção do Nheengatagger, por exemplo, foram implementadas funções que modelam a flexão verbal e nominal tendo em vista a expansão do léxico, mas que potencialmente também podem ser utilizadas na resolução de ambiguidades (ALENCAR, 2020)[17].

Existe uma ampla variedade de métodos e modelos de *POS tagging*, cada um com suas vantagens e desvantagens, portanto, compreender as diferenças entre os modelos é fundamental para selecionar a abordagem mais adequada para cada situação. Em geral, os modelos utilizados para a tarefa de etiquetagem morfossintática são baseados em regras ou estatísticos. O modelo baseado em regras, já brevemente introduzido, utiliza um conjunto de regras gramaticais escritas à mão e codificadas para identificar a classe das palavras com base nas suas propriedades morfológicas e/ou sintáticas. O modelo estatístico, que também já mencionamos, é treinado a partir de um conjunto de sentenças previamente anotadas e utiliza a probabilidade para atribuir as etiquetas corretas a cada *token* de um texto não anotado que recebe como entrada (VOUTILAINEN, 2004; JURAFSKY; MARTIN, 2023a). Para o propósito deste trabalho, nosso foco será no modelo baseado em regras.

[17] Cf. o componente BuildDictionary do Nheengatagger, disponível em: https://github.com/CompLin/nheengatu/tree/main/src. Acesso em: 03 jun. 2023.

2.1.1 Modelos baseados em regras

Um dos primeiros modelos de etiquetagem automática desenvolvidos foi o modelo baseado em regras, que podem ser morfológicas e/ou sintáticas. Regras morfológicas são elaboradas com base na estrutura interna das palavras, por exemplo, palavras em português que terminam com a desinência *-amos* podem ser verbos na primeira pessoa do plural no presente do indicativo, enquanto palavras que terminam com o sufixo *-mente* são advérbios de modo. Por sua vez, as regras sintáticas ou contextuais levam em consideração a distribuição das palavras na sentença, como *casa* em português, que, por exemplo, pode ser um substantivo se ocorrer depois de um artigo e antes de um verbo, ou um verbo, se ocorrer depois de um substantivo e antes de um advérbio. As regras morfológicas geralmente incorporam listas de exceções para formas irregulares e são codificadas por meio de técnicas de correspondência de padrões, como as expressões regulares, que são uma ferramenta comum para a implementação desses padrões, enquanto as regras sintáticas consideram a palavra e a etiqueta em diferentes contextos (KARTTUNEN, 1996; JURAFSKY; MARTIN, 2023a). Etiquetadores baseados em regra são, portanto, ferramentas não triviais, na medida em que a sua implementação envolve a complexa tarefa de codificar o conhecimento linguístico.

Um dos primeiros etiquetadores desse tipo foi provavelmente um dos componentes do *parser* desenvolvido no âmbito do projeto *Transformations and Discourse Analysis*, coordenado pelo renomado linguista Zellig Harris na Universidade da Pensilvânia (HARRIS, 1962; JURAFSKY; MARTIN, 2023). O módulo desambiguador do *parser* contava com 14 regras escritas a mão e utilizou também a frequência relativa das etiquetas para cada palavra, fato que, segundo Jurafsky e Martin, "prefigura algoritmos modernos" (JURAFSKY; MARTIN, 2023, p. 23, tradução nossa). Além do desambiguador de Harris (1962), foram desenvolvidos outros sistemas que utilizam regras e estatística para lidar com ambiguidades. Esses modelos, conhecidos como modelos híbridos, são soluções construídas com base na "hipótese de que pelo menos algumas fontes de erro típicas de sistemas baseados em dados podem ser evitados com uma quantidade razoável de regras linguísticas bem elaboradas" (VOUTILAINEN, 2004, p. 230, tradução nossa).

A seleção do modelo apropriado para desambiguação depende das necessidades e restrições específicas de cada projeto. Embora os modelos baseados em regras possam ser precisos, eles podem apresentar dificuldades com textos que violam ou desviam dessas regras.

Por outro lado, enquanto os modelos estatísticos geralmente oferecem um desempenho superior em comparação com os modelos baseados em regras, eles podem requerer uma grande quantidade de dados para o treinamento. Uma vez que não dispomos de dados para aplicar um modelo estatístico, na construção do desambiguador do Nheengatagger utilizaremos regras de contexto e a frequência das etiquetas em determinados contextos para a resolução das ambiguidades.

2.1.2 Métrica de avaliação

No que diz respeito à performance, um etiquetador geralmente é avaliado em função da sua acurácia, métrica amplamente utilizada na avaliação de etiquetadores (JURAFSKY; MARTIN, 2023a). No que diz respeito à etiquetagem morfossintática, esta métrica indica a quantidade de etiquetas que um etiquetador consegue atribuir corretamente aos *tokens* de um dado corpus. Essa medida é calculada com os seguintes termos: verdadeiros positivos, que são os *tokens* etiquetados corretamente; verdadeiros negativos, que são os *tokens* que não devem ser etiquetados; falsos positivos, que são os *tokens* etiquetados incorretamente; e falsos negativos, que são aqueles que deveriam ter sido etiquetados, mas não foram. O cálculo[18] da acurácia pode ser simplificado da seguinte forma (KEPLER, 2010):

$$Acurácia = \frac{Total\ de\ palavras\ etiquetadas\ corretamente}{Total\ de\ palavras\ etiquetadas}$$

Existem ferramentas para diversas línguas, como o português e o inglês, cuja performance chega a 95% e 97% (ALENCAR, 2013; MARCUS *et al.*, 1993; JURAFSKY; MARTIN, 2023a). O Ship-LemmaTagger, etiquetador desenvolvido para a língua indígena peruana Shipibo-konibo, atingiu uma acurácia de 81.4% na etiquetagem de 274 sentenças (PEREIRA NORIEGA *et al.*, 2017). A depender do modelo, diversos fatores podem concorrer para o resultado da performance do etiquetador, como a complexidade do conjunto de etiquetas, a precisão das regras elaboradas, a porcentagem de dados separados para o corpus de desenvolvimento e de teste, a quantidade de sentenças do corpus, entre outros, os quais são analisados nos resultados dos testes de forma que a ferramenta seja aprimorada (ALENCAR, 2015; JURAFSKY; MARTIN, 2023a).

[18] A acurácia é calculada pela fórmula: (VP + VN) / (VP + VN + FP + FN), em que VP=Verdadeiros Positivos; VN=Verdadeiros Negativos; FP=Falsos Positivos e FN=Falsos Negativos.

2.2 Nheengatu

Nesta pesquisa, fundamentamos nossas descrições da morfologia e da sintaxe do nheengatu nos trabalhos de Navarro (2016) e Cruz (2011). Por se tratarem de tipos diferentes de textos, com finalidades distintas, as duas abordagens apresentam algumas diferenças relevantes para o nosso trabalho. Em razão de seu propósito didático e não exaustivo, a descrição das classes de palavras segundo Navarro (2016) aproxima-se da perspectiva gramatical tradicional, enquanto o trabalho de Cruz (2011), devido ao seu caráter descritivo e documental, descreve detalhadamente os aspectos morfológicos do nheengatu, dividindo as categorias de palavras em classes lexicais e classes gramaticais.

Segundo Navarro (2016), as classes do nheengatu são: substantivos; artigos definidos; verbos de primeira e de segunda classe; pronomes pessoais de primeira e de segunda classe; pronomes (demonstrativo, indefinido, quantificador, interrogativo); adjetivos de primeira e de segunda classe; posposições; interjeições; e partículas. Por outro lado, Cruz (2011) faz a distinção entre classes lexicais e classes gramaticais e considera certos aspectos gramaticais da morfologia da língua na terminologia, que adiante discutiremos. Segundo a autora, as classes lexicais são: nomes (dêiticos e substantivos); índices de pessoa (de série dinâmica e série estativa); verbos transitivos e intransitivos (dinâmicos e estativos); e expressões adverbiais (advérbios e posposições). Já as classes gramaticais são: partículas intra-oracionais (posição inicial, segunda posição, existenciais e flutuantes) e extrassentenciais (fáticas e interjeições); conjunções (nativas e emprestadas); subordinadores; e clíticos.

2.2.1 O sintagma nominal do nheengatu

Na descrição da morfossintaxe do nheengatu, adotamos uma perspectiva topológica da estrutura da sentença, análoga à abordagem de Alencar (2020) para a construção do Nheengatagger. Por esta razão, nos debruçamos, primeiramente, sobre as classes que ocorrem no sintagma nominal (SN) do nheengatu. No quadro a seguir, apresentamos a estrutura do sintagma nominal do nheengatu.

Quadro 1 – A estrutura do sintagma nominal do nheengatu

Quantificação	**Determinação**	**Complemento nominal**	**Núcleo**
panhe mui(ri)	Demonstrativos Indefinidos Numerais	Nome substantivo IPE (índice pessoal da série estativa)	NOME

Fonte: Fonologia e Gramática do Nheengatú (CRUZ, 2011, p. 282).

Quadro 2 – Classes de palavras que ocorrem no sintagma nominal do nheengatu segundo Cruz (2011) e Navarro (2016)

	Zona Prefixal						**Modificador Pré-nominal**	**Zona Nuclear**		**Modificador Pós-nominal**
	Quantificação	**Determinação**			**Complemento nominal**			**Núcleo**		
Navarro	Quantificadores	Demonstrativos	Indefinidos	Numerais	Substantivos	Pron. de 2ª classe	Adjetivos de 1ª classe	Pron. de 1ª classe	Substantivos e pronomes[19]	Adjetivos de 1ª e 2ª classe
Cruz	Quantificação contínua	Referenciação	Indefinitude e alteridade	Quantificação discreta	Nome substantivo	Índice pessoal da série estativa	Verbos intransitivos estativos	Nomes dêiticos (D) e pronomes anafóricos (A)	Nomes substantivos	Verbos intransitivos estativos
Exemplos[20]	*panhẽ* (todo/a/os/as)	*nhaã* (aquele/a) *kwá* (este/a)	*yepé* (um/a)	*mukũi* (dois), *mukũisawa* (segundo/a)	*igara* (canoa), *Maria*	*se, ne, i, yané, pe, aintá*[21]	*pisasu* (novo/a), *puranga* (bonito/a)	D: *ixé, indé, yandé, penhẽ* A: *aé, aintá*[22]	*mimbira* (filho), *Maria, awá* (quem)	*puranga* (bonito/a), *pusé* (pesado)
Etiquetas do Nheengatagger[23]	QUANT	DEMS, DEMX	IND	CARD, ORD	N, PROPN	PRON2	A	PRON	N, PROPN, IND	A e A2

Fonte: Cruz (2011) e Navarro (2016). Elaboração própria.

[19] Cf. Apêndice A.
[20] Adotamos a convenção ortográfica e a tradução utilizada por Navarro (2016) em todos os exemplos deste trabalho.
[21] *Se* (eu, meu/s, minha/s), *ne* (tu, você, teu/s, tua/s, seu/s, sua/s), *i* (ele/a, seu, sua, dele/a), *yané* (nós, nosso/a/os/as), *pe* (vós, vocês, vosso/a/os/as), de vocês, seu/s, sua/s), *aintá* (eles/as), deles/as, índice de indeterminação do sujeito).
[22] *Ixé* (eu, pronome objeto: me), *indé* (tu, pronome objeto: te), *yandé* (nós, pronome objeto: nos), *penhẽ* (vós, vocês, pronome objeto: vos), *aé* (ele/a, pronomes objetos: o, a), *aintá* ou *ta* (eles/as).
[23] Conjunto de etiquetas do Nheengatagger. Disponível em: https://github.com/CompLin/nheengatu/blob/main/docs/tags.md. Acesso em: 04 jun. 2023.

No Quadro 1, apresentamos a estrutura do SN segundo Cruz (2011), e no Quadro 2, sistematizamos as classes do SN considerando as diferenças de ordem terminológica entre Cruz (2011) e Navarro (2016), a fim de analisarmos a abordagem adequada para o nosso trabalho. Em relação à estrutura do SN proposta por Cruz (2011), é possível observar que os adjetivos não estão incluídos no Quadro 1. Isso decorre do fato de que, para a autora, os adjetivos inexistem em nheengatu. Uma vez que adotaremos a terminologia de Navarro (2016), incluímos essa classe no Quadro 2.

Conforme indicamos no Quadro 2, a classe que Navarro (2016, p. 12) trata como adjetivos "qualificativos e predicativos", Cruz (2011, p. 180) denomina como "verbos intransitivos estativos". De acordo com a autora, os itens lexicais dessa classe não podem ser combinados com os índices pessoais da série dinâmica, ou seja, prefixos que indicam a concordância número-pessoal entre sujeito e verbo segundo Navarro (2016). No que se refere à posição, os adjetivos qualificativos podem ocorrer antes ou depois do núcleo N do sintagma nominal, ver (10)-(11)[24], enquanto os adjetivos predicativos ocorrem em posição pós-nominal, ver (12)-(14) (NAVARRO, 2016, p.12-13).

(10) yepé pisasu ara u-sika
INDF novo dia 3SG-chegar
'um novo dia chega'

(11) igara pisasu
canoa novo
'a canoa é nova'

(12) igara i pusé
canoa 1SG[25] pesado
'a canoa é pesada'

[24] Nas glosas interlineares, assim como Alencar (2021), optamos por seguir as convenções das Leipzig Glossing Rules (LGR). Disponível em: https://www.eva.mpg.de/lingua/resources/glossing-rules.php. Acesso em: 13 ago. 2021.

[25] Alencar (2021) utiliza símbolos propostos por Gynan (2017) para distinguir, por exemplo, os índices de pessoa da série dinâmica (ACT) e os índices de pessoa da série estativa (INACT). Em razão da falta de familiaridade com este trabalho, optamos por não incluir essas informações nas nossas glosas interlineares neste momento.

(13) kunhã s-uri u-iku
mulher 3SG-feliz 3SG-estar
‘a mulher está feliz’

(14) ixé se r-uri
eu 1SG REL-feliz
‘eu sou feliz’

Outro aspecto relevante em relação às duas abordagens descritivas diz respeito à zona nuclear do sintagma nominal, especificamente no que concerne às classes que ali ocorrem. Tomemos como exemplo as sentenças abaixo, extraídas de Navarro (2016):

(15) awá kunhã taá u-sika ana
INDF mulher Q 3SG-chegar PST
‘qual mulher chegou?’

(16) awá u-riku turusu iuí u-meẽ-kwau awá
INDF 3SG-ter grande terra 3SG-dar-poder INDF
nti u-riku supé
NEG 3SG-ter para
‘quem tem muita terra pode dar para quem não tem’

(17) manungara a-maã waá a-su a-mbeú penhẽ arama
INDF 1SG-ver REL 1SG-ir 1SG-contar 2PL para
‘algo que vi vou contar para vocês’

O lexema *awá* é classificado como pronome interrogativo e como pronome indefinido segundo Navarro (2016), enquanto Cruz (2011) o classifica como nome genérico de humano. No exemplo (15), *awá* desempenha um papel de determinação, enquanto em (16) ocupa a posição de núcleo do sintagma nominal. No exemplo (17), *manungara*, classificado por Navarro (2016) como pronome indefinido e por Cruz (2011) como nome autônomo, constitui o núcleo do sintagma nominal e não ocorre na zona prefixal. Outro aspecto que vale ser destacado diz respeito à descrição de Cruz (2011) sobre o que Navarro (2016) classifica como pronomes pessoais de primeira classe e sua ocorrência no núcleo do SN. Para a autora,

essa classe corresponde a duas categorias diferentes, os nomes dêiticos e os pronomes anafóricos (ver Quadro 2). Os nomes dêiticos "(...) são nomes - portanto, expressões capazes de funcionar como núcleo de sintagma nominal - cuja referência depende diretamente da situação enunciativa" (CRUZ, 2011, p.140). Os pronomes anafóricos, por sua vez, caracterizam-se por recuperarem um nome enunciado previamente na situação comunicativa e "(...) podem ocupar a posição de núcleo do sintagma nominal que ocupa as principais posições de argumento: sujeito e objeto. Porém, não ocorrem como complemento de posposição" (CRUZ, 2011, p. 145-146). De acordo com Navarro (2016, p. 23), "com as posposições devem-se usar os pronomes pessoais da 2ª classe". Observemos os seguintes exemplos:

(18) aé u-su se irumũ
ele 3SG-ir 1SG com
'ele vai comigo'

(19) *aé u-su ixé irumu

(20) a-nheẽ puranga ara aintá supé
1SG-dizer bonito dia 3PL para
'digo "bom dia" para eles'

(21) ixé a-piripana xirura i xupé
eu 1SG-comprar calça 3SG 3SG-para
'eu compro calça para ele'

(22) re-rúri timbiú ixé arama
2SG-trazer comida 1SG para
'traga comida para mim'

Nos exemplos (18-19), observamos a ocorrência dos pronomes pessoais como complemento de posposição, sendo (18) gramatical, por ser de segunda classe, e (19) agramatical, por ser de primeira classe. Nos exemplos (20-21), apresentamos duas sentenças que evidenciam a ocorrência das posposições com pronomes de terceira pessoa da segunda classe, denominado por Cruz (2011) como índices pessoais da série estativa. Por outro lado,

vale destacar que em alguns casos os pronomes de primeira classe podem ocorrer com posposições, conforme (22).

Embora os pronomes pessoais ocorram como núcleos do SN, não ocorrem em relação genitiva com outros substantivos (23-24)[26], isto é, uma estrutura "em que um NOME realizado como pronome pessoal rege um CN [complemento nominal]" é agramatical, como aponta Alencar (2021, p. 1735).

(23)	panhẽ	nhaã	kurumĩ	igara-itá
	todo	DEM.DIST	menino	canoa-PL
	'todas aquelas canoas do menino'			

(24)	* kurumĩ	penhẽ	
	menino	vós	

Tendo em vista a elaboração das regras de contexto, observamos que é preciso identificar as classes que ocorrem em uma ou outra zona do sintagma nominal. Por esta razão, elaboramos uma lista de itens pertencentes à classe dos pronomes segundo Navarro (2016), indicando com exemplos a sua ocorrência em uma ou outra zona do sintagma nominal (ver APÊNDICE A).

2.2.1 Conjunto de etiquetas do Nheengatagger

Considerando as particularidades distribucionais e semânticas das classes de palavras do nheengatu, Alencar (2020) elaborou para o Nheengatagger um conjunto de 86 etiquetas que exprimem essas diferenças (ver APÊNDICE B). No *treebank* UD_Nheengatu-CompLin[27], Alencar (2022) utilizou as etiquetas do modelo Universal Dependencies (NIVRE *et al.*, 2016) e as etiquetas do Nheengatagger na anotação das sentenças. Para o propósito deste trabalho, utilizaremos o conjunto de etiquetas de Alencar (2020).

A fim de avaliarmos a performance do desambiguador na resolução de ambiguidades em sentenças anotadas sem distinções de subtipos para algumas classes,

[26] Exemplos extraídos de Alencar (2021).
[27] Disponível em: https://github.com/UniversalDependencies/UD_Nheengatu-CompLin/tree/dev. Acesso em: 03 jun. 2023.

elaboramos um conjunto de etiquetas simplificado (ver APÊNDICE C), já que existem casos que as distinções semânticas podem ser desconsideradas, uma vez que não afetam a distribuição das palavras na sentença.

Com base nos resultados obtidos nos Testes 1 e 2, apresentados na seção 4.3 desta dissertação, selecionamos, para a composição do *tagset* Simplificado, as classes: substantivo, pronome demonstrativo, advérbio e adposição. Com base na análise dos dados dos dois testes, hipotetizamos que essas classes poderiam impactar a performance do desambiguador. Na seção de resultados, discutimos de maneira mais aprofundada em que medida os subtipos podem influenciar no desempenho da ferramenta.

3 METODOLOGIA

A metodologia deste trabalho divide-se em duas partes: a compilação do corpus e a construção do desambiguador. Na seção 3.1, apresentamos os materiais e métodos utilizados na compilação dos textos que compõem o nosso corpus, são eles: (i) *Curso de Língua Geral* (NAVARRO, 2016); (ii) *Noções de Língua Geral ou Nheengatu* (CASASNOVAS, 2006); (iii) *Histórias em língua geral da Amazônia* (NAVARRO; ÁVILA, 2017); e (iv) tradução integral da obra *Le Petit Prince* para o nheengatu (TREVISAN, 2017). Na seção 3.2, apresentamos as etapas da implementação do desambiguador.

3.1 Compilação do corpus

Na primeira etapa deste trabalho, compilamos os textos selecionados para o corpus[28] em um formato adequado para serem submetidos ao algoritmo do Nheengatagger e do seu desambiguador. Nas subseções a seguir, apresentamos a metodologia adotada para a compilação de cada uma das fontes.

3.1.1 Curso de Língua Geral

O *Curso de Língua Geral (nheengatu ou tupi moderno): a língua das origens da civilização amazônica* (NAVARRO, 2016)[29] é um manual utilizado em disciplinas de tupi ministradas na Universidade de São Paulo e está disponível em duas edições, uma publicada em 2011[30] e outra em 2016[31]. Neste trabalho, utilizamos a versão mais recente da obra.

O *Curso* é composto por prefácio, introdução, informações sobre a fonologia e a grafia da língua, treze lições e um vocabulário. Tendo em vista o desenvolvimento de um corpus paralelo nheengatu-português-inglês alinhado, extraímos as sentenças do livro, compilamos as sentenças nheengatu-português e incluímos em cada arquivo que compõe o

[28] Disponível em: https://github.com/CompLin/nheengatu/tree/main/data/corpus. Acesso em: 25 maio 2023.
[29] Esta obra compilada está disponível em: https://github.com/juliana-gurgel/nheengatu/tree/main/corpus. Acesso em: 14 ago. 2023.
[30] Disponível em: http://tupi.fflch.usp.br/sites/tupi.fflch.usp.br/files/CURSO%20DE%20L%C3%8DNGUA%20GERAL%20(NHEENGATU).pdf. Acesso em: 15 maio 2023.
[31] Disponível em: https://mega.nz/file/A4xllaIa#m2sh3zN4WzgCIOdF7HwfA-FdXswG7v-lycZfeEKNX_o. Acesso em: 15 maio 2023.

corpus informações sobre a obra e a extração. Por fim, traduzimos as sentenças para o inglês e as incluímos no corpus.

A extração e o alinhamento das sentenças de Navarro (2016) foram feitos de forma manual[32] e os dados foram armazenados em arquivos com a codificação UTF-8[33]. Os títulos dos arquivos compilados foram sistematizados seguindo a estrutura abaixo:

(25) t-yrl-por-eng-01.txt

(26) e-yrl-por-eng-01.txt

(27) p-yrl-por-eng-01.txt

(28) y-yrl-por-eng-01.txt

Cada lição de Navarro (2016) resultou em três arquivos compilados. Nos exemplos acima, a primeira casa corresponde à parte da lição de onde foram extraídas as sentenças, isto é, texto (t), exemplos (e), exercícios (p, do nheengatu *purakisawa-itá*) e *yasú yanheengari* (*y*, "vamos cantar" em português), que é uma seção presente nas lições 3, 4, 6, 7 e 8. A segunda, a terceira e a quarta casa fazem referência aos identificadores das línguas no padrão ISO 639-3:2007[34] (*yrl* para o nheengatu, *por* para o português e *eng* para o inglês) e a última casa indica o número da lição. Em seguida, incluímos em cada arquivo um cabeçalho contendo as informações sobre a obra e a extração.

Após a extração das sentenças, traduzimos e alinhamos as sentenças do português para o inglês, por meio de um programa em Python[35] e, para evitar inconsistências, revisamos cada sentença em inglês com relação à sentença correspondente em nheengatu, fazendo ajustes quando necessário. Por fim, incluímos as informações sobre a tradução no cabeçalho de cada arquivo. O corpus com as sentenças *yrl-por-eng* será disponibilizado na plataforma Github[36] quando a revisão de todos os componentes estiver concluída.

---

[32] À princípio, a tarefa de extração e alinhamento das sentenças poderia ter sido feita automaticamente por meio das técnicas de *sentence splitting* (KOEHN, 2005), utilizando Python ou ferramentas de processamento de texto no Unix, como Grep, Sed, entre outras. Contudo, no primeiro momento da pesquisa, a execução desta tarefa de forma manual foi preferível até nos familiarizarmos com os métodos automatizados.

[33] Unicode Transformation Format 8-bit (UTF-8) é um tipo de codificação que reconhece qualquer caractere Unicode, incluindo caracteres com sinais diacríticos, o que, visto a ortografia adotada em Navarro (2016), é imprescindível para o processamento computacional. Disponível em: https://www.utf8.com/; https://www.unicode.org/standard/standard.html. Acesso em 23 maio 2023.

[34] O padrão internacional ISO 639-3:2007 define os códigos para a representação dos nomes de línguas. Disponível em: https://www.iso.org/standard/39534.html. Acesso em: 16 maio de 2023.

[35] Todos os programas implementados para pré-processamento de textos no âmbito desta pesquisa estão sendo disponibilizados em: https://github.com/juliana-gurgel/nheengatu/tree/main/src/tools. Acesso em: 14 ago. 2023.

[36] Disponível em: https://github.com/CompLin/nheengatu/tree/main/data/corpus/. Acesso em: 18 maio 2023.

3.1.2 Histórias em língua geral da Amazônia

Histórias em língua geral da Amazônia (NAVARRO; ÁVILA, 2017)[37] é uma obra composta por traduções de trinta e sete textos feitas por alunos matriculados na disciplina Tupi IV (Nheengatu), ministrada entre 2008 e 2016 na Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo. A extração e o alinhamento das sentenças foram feitos manualmente a partir de um arquivo editável, disponibilizado pelos autores, e a revisão foi feita com relação à versão final do livro, publicada em 2017.

3.1.3 Noções de Língua Geral ou Nheengatu

O livro *Noções de Língua Geral ou Nheengatu* (CASASNOVAS, 2006)[38] é composto por uma parte introdutória sobre a língua, o alfabeto, trinta e quatro notas gramaticais, doze lendas em nheengatu traduzidas para o português e um vocabulário bidirecional nheengatu-português. A compilação desta obra foi dividida em duas partes: notas gramaticais e lendas.

Devido ao formato do arquivo da obra ao qual tivemos acesso, a extração e alinhamento das sentenças foram feitas manualmente. Os exemplos contidos nas notas gramaticais foram compilados em um único arquivo, as lendas, por sua vez, foram compiladas a partir da seguinte estrutura:

(29) x_y.txt

Em (29), *x* corresponde ao número da lenda na ordem em que aparece no índice do livro e *y* corresponde ao título, com letra minúscula e sem sinais diacríticos, por exemplo, ‘1_urubu_wira_wasu.txt’. No que se refere ao vocabulário, resta conferir se todos os verbetes de Casasnovas (2006) já estão contidos no glossário de Navarro (2016) e quais precisam ser incorporados ao léxico utilizado no Nheengatagger.

---

[37] Esta obra compilada será disponibilizada em: https://github.com/juliana-gurgel/nheengatu/tree/main/corpus. Acesso em: 14 ago. 2023.

[38] Esta obra compilada está disponível em: https://github.com/juliana-gurgel/nheengatu/tree/main/corpus. Acesso em: 14 ago. 2023.

3.1.4 Tradução integral da obra Le Petit Prince

A tradução integral da obra *Le Petit Prince*[39] (*Muruxawamirĩ*, em nheengatu) do francês para o nheengatu é produto da dissertação de mestrado de Trevisan (2017). Uma vez que a tradução é composta pelas sentenças alinhadas em francês e em nheengatu, a compilação foi feita da seguinte maneira: primeiro, extraímos as sentenças da dissertação, em seguida, alinhamos as sentenças automaticamente[40].

3.2 Desenvolvimento do desambiguador

O desenvolvimento do desambiguador foi dividido em quatro etapas: (i) levantamento das ambiguidades presentes no léxico e no corpus etiquetado pelo Nheengatagger[41]; (ii) extração e análise dos contextos das etiquetas ambíguas e não ambíguas; (iii) elaboração das condições para resolução das ambiguidades; e (iv) implementação do desambiguador. Nas seções a seguir, descrevemos os materiais e procedimentos utilizados em cada etapa.

3.2.1 Levantamento das ambiguidades

O levantamento das ambiguidades foi feito por meio da extração manual de todas as entradas lexicais com etiquetas ambíguas contidas no léxico[42] do Nheengatagger e revisadas com relação ao corpus de desenvolvimento (NAVARRO, 2016). Vale ressaltar que as ambiguidades são compostas por etiquetas do *tagset*[43] elaborado no âmbito do desenvolvimento do Nheengatagger e do UD_Nheengatu-CompLin[44]. Após o levantamento das ambiguidades, que somam 55 na versão do léxico utilizada até o momento, determinamos

---

[39] Esta obra compilada será disponibilizada em: https://github.com/juliana-gurgel/nheengatu/tree/main/corpus. Acesso em: 14 ago. 2023.

[40] Todos os programas implementados para pré-processamento de textos no âmbito desta pesquisa estão sendo disponibilizados em: https://github.com/juliana-gurgel/nheengatu/tree/main/src/tools. Acesso em: 14 ago. 2023.

[41] Nesta parte da pesquisa, utilizamos a versão etiquetada em 16 de novembro de 2022, disponível em: https://github.com/CompLin/nheengatu/tree/main/data/corpus/navarro-2016. Acesso em: 25 maio 2023.

[42] Na atual versão do desambiguador, utilizamos a versão do léxico atualizada em 13 de novembro de 2022. Posteriormente, atualizaremos os dados com base na versão atual do léxico, que está disponível em: https://github.com/CompLin/nheengatu/blob/main/data/lexicon.json. Acesso em: 25 maio 2023.

[43] Nesta etapa da pesquisa, utilizamos a versão do *tagset* disponível no APÊNDICE B. A última versão do *tagset* está disponível em: https://github.com/CompLin/nheengatu/blob/main/docs/tags.md. Acesso em: 25 maio 2023.

[44] Disponível em: https://github.com/UniversalDependencies/UD_Nheengatu-CompLin/tree/dev. Acesso em: 25 maio 2023.

a frequência de cada ambiguidade no corpus e aferimos a proporção com relação ao total de ambiguidades do corpus (ver APÊNDICE D).

Para a implementação do protótipo do desambiguador, decidimos, a princípio, testar a ferramenta com relação às ambiguidades que correspondem a pelo menos 50% do total presente no corpus de desenvolvimento. Da lista apresentada no APÊNDICE D, selecionamos, para os testes preliminares[45], um conjunto de ambiguidades que equivalem a 52.38% do total a serem resolvidas pelo desambiguador, apresentado na tabela abaixo:

Tabela 1 – Conjunto de ambiguidades analisadas nos testes preliminares

Ambiguidade	**Classes**	**Frequência**	**(%)**
A+N	Adjetivo de 1ª classe + substantivo	64	9.24%
PRON+PRON2	Pronome de 1ª classe + pronome de 2ª classe	65	9.38%
ADP+FUT+SCONJ	Posposição + partícula de futuro + conjunção subordinativa	67	10.67%
A+ADV	Adjetivo de 1ª classe + advérbio	79	11.40%
ADP+SCONJ	Posposição + conjunção subordinativa	81	11.69%
	Total	**356**	**52.38%**

Fonte: Elaboração própria.

Com o objetivo de simular o funcionamento do desambiguador para identificarmos ajustes necessários no algoritmo antes de avaliarmos a performance da ferramenta na resolução de todas as ambiguidades, extraímos os contextos dessas apresentadas na Tabela 1, além dos contextos de cada etiqueta que as compõem. Em seguida, analisamos todos os contextos (das ambiguidades e das etiquetas individualmente) e buscamos descrever como a ferramenta deveria funcionar no processo de desambiguação.

Na próxima seção, descrevemos o procedimento adotado na extração dos contextos.

3.2.2 Extração e análise dos contextos

Após a seleção do primeiro conjunto de ambiguidades, iniciamos a extração automática[46] dos contextos das palavras ambíguas e das etiquetas que compõem as ambiguidades em ocorrências não ambíguas presentes no corpus de desenvolvimento. Por

[45] Apresentamos os resultados dos testes preliminares na seção 4.2 Testes preliminares.
[46] Todos os programas implementados para pré-processamento de textos no âmbito desta pesquisa estão sendo disponibilizados em: https://github.com/juliana-gurgel/nheengatu/tree/main/src/tools. Acesso em: 14 ago. 2023.

exemplo, primeiro extraímos do corpus as etiquetas que ocorrem antes e depois da ambiguidade ADP+SCONJ (posposição + conjunção subordinativa), em seguida, extraímos as etiquetas que ocorrem antes e depois da etiqueta ADP e aquelas que ocorrem antes e depois de SCONJ. Adotamos o mesmo procedimento para as demais ambiguidades presentes na Tabela 1 e para as etiquetas que as compõem. Nos quadros a seguir, apresentamos um fragmento[47] dos quadros dos contextos da ambiguidade ADP+SCONJ e da etiqueta ADP, e todos os contextos de SCONJ.

Quadro 3 – Fragmento do quadro dos contextos da ambiguidade ADP+SCONJ extraídos de Navarro (2016)

Etiqueta anterior	Classe da etiqueta anterior	Etiqueta posterior	Classe da etiqueta posterior
PROPN	Nome próprio	PUNCT	Pontuação
IND	Pronome indefinido	ADV	Advérbio
N	Substantivo	ADVJ	Advérbio conjuncional
PROPN	Nome próprio	N	Substantivo
A+N	Adjetivo de 1ª classe + substantivo	PUNCT	Pontuação

Fonte: Elaboração própria.

Quadro 4 – Fragmento do quadro dos contextos da etiqueta ADP extraídos de Navarro (2016)

Etiqueta anterior	Classe da etiqueta anterior	Etiqueta posterior	Classe da etiqueta posterior
???	Palavra desconhecida	???	Palavra desconhecida
N	Substantivo	ADP	Posposição
None	Etiqueta atual no início da sentença	ADP	Posposição
PRON2	Pronome de 2ª classe	ADP	Posposição
N	Substantivo	ADV	Advérbio

Fonte: Elaboração própria.

Quadro 5 – Contextos da etiqueta SCONJ extraídos de Navarro (2016)

Etiqueta anterior	Classe da etiqueta anterior	Etiqueta posterior	Classe da etiqueta posterior
None	Etiqueta atual no início da sentença	PRON	Pronome de 1ª classe
None	Etiqueta atual no início da sentença	ADV	Advérbio
None	Etiqueta atual no início da sentença	IND	Pronome indefinido
PQ	Partícula de pergunta polar	ADP+SCONJ	Posposição + conjunção subordinativa

Fonte: Elaboração própria.

[47] Cf. APÊNDICES F e G para ver os quadros completos.

Nos Quadros 3, 4 e 5, cada linha corresponde, respectivamente, a um contexto da ambiguidade ADP+SCONJ, da etiqueta ADP e da etiqueta SCONJ. Nos exemplos a seguir[48], retirados do corpus de desenvolvimento, apresentamos, na ordem, as cinco sentenças de onde extraímos os contextos do Quadro 5.

(30) **Mayawé/SCONJ ixé/PRON** maraari/A aikú/V yepé/ART+CARD+FRUST+SCONJ ,/PUNCT ayenũ/V se/PRON2 mimbira/N ruakí/ADP ,/PUNCT akiri/V ana/PFV ./PUNCT

mayawé	ixé	maraari	a-ikú	yepé	a-yenũ
como	1SG	cansado	1SG-estar	IPFV	1SG-deitar
se	mimbira	r-uakí	a-kiri	ana	
1SG	filho.de.mulher	REL-perto.de	1SG-dormir	PFV	

'como eu estava cansado, deitei-me perto de meu filho e dormi'

(31) **Mayawé/SCONJ kuíri/ADV** pekwáu/V ana/PFV maã/IND+INT+N+V kurí/FUT pemunhã/V ,/PUNCT asú/V ana/PFV ./PUNCT

mayawé	kuíri	pe-kwáu	ana	maã	kurí	pe-munhã	a-sú
como	já	2PL-saber	PFV	REL	FUT	2PL-fazer	1SG-ir
ana							
PFV							

'como agora vocês já sabem o que farão, eu já vou'

(32) **Mayawé/SCONJ nti yepé/IND** apigawa/A+N usaisú/V nhaã/DEMS kunhãmukú/A+N ,/PUNCT aé/PRON nti/NEG upurasí/V ./PUNCT

mayawé	nti-yepé	apigawa	u-saisú	nhaã
como	NEG.INDF	homem3SG-amar	DEM.DIST	
kunhãmukú	aé	nti	u-purasí	
moça	3SG	NEG	3SG-dançar	

'como nenhum homem amava aquela moça, ela não dançou'

[48] A etiqueta atual (SCONJ) está em negrito vermelho e os contextos estão em negrito azul. Nos contextos cuja etiqueta anterior é None, a conjunção subordinativa ocorre no início da sentença.

(33) **Mayawé/SCONJ aé/PRON** nti/NEG rẽ/ADV uyupiri-kwáu/V ,/PUNCT upitá/V iwá[49]/??? wirpe/ADP ./PUNCT

mayawé	aé	nti	rẽ	u-yupiri-kwáu	u-pitá	iwá
como	3SG	NEG	ainda	3SG-subir.saber	3SG-ficar	árvore
wirpe						
embaixo.de						

‘como ele/a ainda não sabia subir, ficou embaixo da árvore’

(34) Nti/NEG ana/PFV remandwari/V **será/PQ mayawé/SCONJ resé/ADP+SCONJ** repuká/V sé/A+ADV nhaã/DEMS kunhã-itá/N renundé/ADP ?/PUNCT

nti	ana	re-mandwari	será	mayawé	r-esé
NEG	PFV	2SG-lembrar	Q	como	REL-de
re-puká		sé	nhaã	kunhã-itá	r-enundé
2SG-rir		gostoso	DEM.DIST	mulher-PL	REL-diante

‘você não lembra de como você riu gostoso diante daquelas mulheres’

Nos exemplos (30-32), que correspondem, respectivamente, aos contextos apresentados na segunda, terceira e quarta linhas do Quadro 5, as etiquetas anteriores são None porque a conjunção subordinativa ocorre no início da sentença, e as posteriores são, na ordem, PRON (pronome de 1ª classe), ADV (advérbio) e IND (pronome indefinido). No exemplo (34), que corresponde à quinta linha do referido quadro, a etiqueta anterior é PQ (partícula de pergunta polar) e a posterior ADP+SCONJ (posposição + conjunção subordinativa).

Uma vez finalizada a extração, iniciamos a análise dos contextos das ambiguidades, do ponto de vista das suas propriedades distribucionais, para a elaboração das regras. Dada uma ambiguidade, o desambiguador deve ler o contexto da ambiguidade, isto é, as etiquetas vizinhas, e identificar, dentre as etiquetas que compõem a ambiguidade, qual a etiqueta adequada para o contexto. Por exemplo, a partir de um conjunto de etiquetas A, B, C, D, E e F, para resolver uma ambiguidade qualquer, que chamaremos de X+Y, poderíamos definir as seguintes regras: (i) a etiqueta será X se a etiqueta anterior à ambiguidade for A ou

[49] Correção: *íwa*. Este vocábulo apresenta um erro de revisão, conforme apontado pelo próprio autor na ocasião da defesa desta dissertação. “Quando um ditongo for antecedido por uma vogal, acentua-se graficamente a vogal tônica I ou U do ditongo (...)” (NAVARRO, 2016, p. 8).

B, e se a etiqueta posterior à ambiguidade for B ou C; (ii) a etiqueta será Y se a etiqueta anterior à ambiguidade for D ou E e se a etiqueta posterior à ambiguidade for F. Como podemos observar, nenhuma das etiquetas vizinhas a X e Y coincidem, portanto, as duas regras não se contradizem. Desse modo, a ambiguidade seria resolvida como X se, por exemplo, tivesse como etiqueta anterior A e posterior C ou seria resolvida como Y se tivesse como etiqueta anterior D e posterior F. Por outro lado, algumas ambiguidades presentes no nosso corpus possuem etiquetas em comum nos seus contextos, como é o caso de ART+CARD+FRUST+SCONJ (artigo indefinido + numeral cardinal + partícula de frustrativo + conjunção subordinativa), conforme os exemplos[50] a seguir:

(35) Aé/PRON upitá/V yepé/**ART** igara/**N** mirĩ/A+ADV upé/ADP Maria/PROPN irũmu/ADP+SCONJ ./PUNCT

Aé	u-pitá	yepé	igara	mirĩ	upé	Maria	irũmu
3SG	3SG-ficar	ART	canoa	pequena	em	Maria	com

‘ele fica em uma canoa pequena com Maria’

(36) Rute/PROPN usasá/V kurí/FUT yepé/**CARD** yasí/**N** Maria/PROPN rendawa/N upé/ADP ./PUNCT

Rute	u-sasá	kurí	yepé	yasí	Maria	rendawa	upé
Rute	3SG-passar	FUT	um	mês	Maria	comunidade	em

‘Rute passará um mês na comunidade de Maria’

(37) Apurakí-putari/**V** yepé/**FRUST** ixé/PRON se/PRON2 maraari/ADJ ./PUNCT

A-purakí-putari	yepé	ixé	se	maraari
1SG-trabalhar-querer	IPFV	1SG	1SG	cansado

‘eu queria trabalhar, porém estou cansado’

No nheengatu, o substantivo (N) pode ocorrer imediatamente depois do artigo indefinido (35) e do numeral cardinal (36), e o verbo de 1ª classe (V) pode ocorrer imediatamente antes da partícula de frustrativo (37)[51] e da conjunção subordinativa[52]. Em

[50] As ambiguidades dos exemplos (35, 36 e 37) foram resolvidas manualmente.
[51] Exemplo extraído de Ávila (2021, p. 866) e etiquetado manualmente.
[52] Por enquanto, não temos dados disponíveis para incluirmos um exemplo de *yepé* como conjunção subordinativa.

razão de casos dessa natureza, decidimos utilizar uma abordagem híbrida para a resolução das ambiguidades, considerando não apenas o contexto, mas também a frequência das etiquetas que compõem o contexto. Além disso, ao analisarmos a frequência das etiquetas nos seus respectivos contextos nos testes preliminares (ver seção 4.2), decidimos restringir as etiquetas consideradas válidas na resolução das ambiguidades na presente versão do desambiguador.

A seguir, apresentamos a justificativa para as restrições que fizemos na primeira versão da ferramenta.

3.2.2.1 Restrições da primeira versão do desambiguador

Após a realização de testes preliminares (ver Tabelas 5 e 6) e com base na técnica denominada Desenvolvimento em Espiral (ZELLE, 2009), na versão inicial da ferramenta decidimos restringir os tipos de etiquetas consideradas válidas na resolução das ambiguidades a apenas etiquetas não ambíguas que de fato correspondem a classes de palavras. Deste modo, desconsideramos, neste primeiro momento, ambiguidades ou etiquetas classificadas como inválidas, isto é, sinais de pontuação (PUNCT), palavras desconhecidas (???) ou None. Justificamos nossa decisão de não incluir ambiguidades que ocorram nos contextos daquelas que devem ser resolvidas como forma de evitar considerar na desambiguação alguma etiqueta cuja classe de palavra não pertence, de fato, ao contexto morfossintático da etiqueta em questão. Por exemplo, na versão atual do corpus etiquetado, a ambiguidade ADVD+DEMX+V (advérbio demonstrativo + pronome demonstrativo proximal + verbo de 1ª classe) ocorre antes de ADP (posposição) e não ocorre antes de SCONJ (conjunção subordinativa). No nheengatu, o pronome demonstrativo proximal (DEMX) *kwá* (*este* ou *esta*, em português) não ocorre antes de posposições nem de conjunções subordinativas, enquanto verbos de 1ª classe (V) podem ocorrer antes de posposições. Considerando o funcionamento da versão atual do algoritmo, se a ambiguidade ADVD+DEMX+V fosse incluída no contexto de ADP, as três etiquetas que compõem a ambiguidade seriam incluídas como etiquetas possíveis de ocorrer antes de ADP, mesmo sem necessariamente fazerem parte das propriedades distribucionais das posposições no nheengatu, o que poderia resultar na resolução incorreta da ambiguidade. Por exemplo:

(38) Kwá/**ADVD+DEMX+V** rupí/**ADP** aikwé/**EXST** siía/INDQ mirá/N ./PUNCT

kwá	rupí	aikwé	siía	mirá
DEM.PROX	por	haver	muitas	árvore

‘por aqui há muitas/os árvores’

No exemplo (38), temos a ambiguidade ADVD+DEMX+V seguida da etiqueta ADP. No nheengatu, o *kwá* como advérbio demonstrativo (*aqui*, em português) ocorre seguido das posposições *suí*, *kití* e *rupí* (respectivamente *de*, *para* e *por*, em português). Além do pronome demonstrativo proximal não ocorrer antes de posposições no nheengatu, existe outro problema que poderia resultar na resolução incorreta da ambiguidade em (38). Se as três etiquetas que compõem a ambiguidade (ADVD, DEMX e V) fossem incluídas na tabela de contexto como etiquetas que podem ocorrer antes de ADP e, consequentemente, ADP fosse incluída como etiqueta que pode ocorrer depois das três etiquetas, teríamos, na tabela de contexto, uma ocorrência de ADVD antes de ADP, uma ocorrência de DEMX antes de ADP e duas ocorrências de V antes de ADP. Uma vez que decidimos utilizar a frequência para a desambiguação nos casos em que as etiquetas vizinhas da ambiguidade ocorram no contexto de duas ou mais etiquetas que compõem a ambiguidade, a inclusão de DEMX e de V entre as etiquetas que ocorrem antes de ADP poderia resultar na resolução incorreta da ambiguidade do exemplo (38) como V. Por esta razão, decidimos desconsiderar etiquetas que consistam em ambiguidades até que possamos, a partir de um conjunto de dados mais robusto, calcular a frequência das etiquetas vizinhas considerando todas as propriedades distribucionais das classes de palavras do nheengatu.

Na próxima seção, apresentamos o procedimento adotado para a criação da Tabela de Contexto, que contém, para cada etiqueta do nosso *tagset*, os contextos extraídos do corpus de desenvolvimento e suas respectivas frequências.

3.2.2.2 Frequência

Considerando que as etiquetas que pertencem aos contextos de algumas classes de palavras do nheengatu apresentam interseção, como demonstramos nos exemplos (35) e (36), na construção da primeira versão do desambiguador, além das regras de contexto, a resolução das ambiguidades é feita com base na frequência em que as etiquetas ocorrem em determinados contextos. Na versão atual da ferramenta, utilizamos a frequência absoluta das

etiquetas para a desambiguação, uma vez que não temos ambiguidades resolvidas no nosso corpus de desenvolvimento para calcularmos a frequência relativa, ou seja, a frequência em que determinada palavra ocorre atribuída a uma ou outra etiqueta em diferentes contextos. Assim, nesta versão decidimos que a resolução das ambiguidades será feita por meio da seleção da etiqueta que apresenta a maior média aritmética da frequência das etiquetas que compõem o contexto da ambiguidade ou apenas a média da etiqueta anterior ou da posterior. Por exemplo, na resolução da ambiguidade ART+CARD+FRUST+SCONJ em (36), o desambiguador deve identificar o contexto em que a ambiguidade se encontra (V à esquerda e N à direita), em seguida, buscar a frequência de V e N nos contextos de ART, CARD, FRUST e SCONJ, calcular a média e, por fim, selecionar a etiqueta que apresenta o maior valor para V e N.

Por conta de exemplos como o (36), à medida que ampliamos nosso conhecimento sobre o nheengatu, avaliamos que, eventualmente, haveria necessidade de alterar o *tagset* e os contextos possíveis de cada etiqueta e recalcular as frequências das etiquetas em seus respectivos contextos. Por esta razão, no lugar de implementarmos cada regra de contexto como estrutura condicional em Python, como planejamos inicialmente, decidimos utilizar uma estrutura de dados tabular (*dataframe*), que chamamos de Tabela de Contexto, contendo as etiquetas, seus contextos possíveis e a frequência de cada etiqueta com relação ao seu respectivo contexto, isto é, a frequência em que uma determinada etiqueta ocorre antes ou depois de outra. Desta forma, o desambiguador recebe como entrada um conjunto de etiquetas e um corpus etiquetado pelo Nheengatagger para gerar diferentes Tabelas de Contexto e realiza a desambiguação de uma sentença por meio da tabela gerada, o que possibilita posteriores modificações nas etiquetas, nos contextos e na frequência sem necessidade de alterações substanciais no código. Adiante, na subseção 3.2.4, demonstramos como funciona a geração da tabela de contexto e a desambiguação de uma sentença na linha de comando do Linux.

Para explicitar como a frequência funciona como critério para a resolução de ambiguidades como aquelas apresentadas nos exemplos (35) e (36), apresentamos abaixo o exemplo de uma sentença contendo a ambiguidade A+ADV.

(39) Aé/**PRON** puranga/**A+ADV** ./**PUNCT**

aé puranga

3SG bonito

‘ele/a é bonito/a’

A seguir, apresentamos um fragmento[53] da Tabela de Contexto gerada a partir de Navarro (2016)[54] contendo as etiquetas possíveis de ocorrerem antes e depois de A (adjetivo de 1ª classe) e de ADV (advérbio), e suas respectivas frequências.

Tabela 2 – Fragmento da Tabela de Contexto das etiquetas A e ADV

1	Etiqueta	Etiqueta do contexto	Tipo de contexto	Frequência
2	A	ADVR	Anterior	1
3	A	FUT	Anterior	1
4	A	N	Anterior	15
5	A	None	Anterior	5
17	ADV	A	Anterior	1
18	ADV	ADP	Anterior	4
19	ADV	ADV	Anterior	8
20	ADV	ADVD	Anterior	2
21	ADV	ADVR	Anterior	1

Fonte: Elaboração própria.

Na Tabela de Contexto de A e ADV, a primeira coluna contém o número de linhas, a segunda contém as etiquetas A e ADV, que compõem a ambiguidade do exemplo (39), a terceira coluna contém as etiquetas que fazem parte do contexto de A e de ADV, a quarta contém o tipo de contexto, isto é, se a etiqueta ocorre antes ou depois de A e de ADV, e, por fim, a quinta coluna contém a frequência de cada etiqueta da segunda coluna com relação ao contexto em que ocorre. Como podemos observar, no exemplo (39) a etiqueta PRON ocorre antes da ambiguidade e a etiqueta PUNCT ocorre depois. Uma vez que PUNCT, por enquanto, não é considerada na desambiguação, a ferramenta resolve a ambiguidade com base na média da etiqueta anterior nos contextos de A e ADV. Conforme a tabela acima, a frequência de PRON antes de A é 3 e antes de ADV é 1, portanto, nesse caso, a ambiguidade será resolvida corretamente como A (adjetivo de 1ª classe). O mesmo critério será aplicado para todos os casos de ambiguidades cujos contextos das etiquetas que as compõem tenham interseção.

[53] Cf. a Tabela de Contexto completa no APÊNDICE I.

[54] Nesta parte da pesquisa, utilizamos a versão etiquetada em 16 de novembro de 2022, disponível em: https://github.com/CompLin/nheengatu/tree/main/data/corpus/navarro-2016. Acesso em: 25 maio 2023.

A seguir, descrevemos o processo de elaboração das condições para que a ferramenta resolva as ambiguidades.

3.2.3 Elaboração das condições para desambiguação

A elaboração das condições para resolução das ambiguidades foi feita com base na análise dos contextos. Antes da implementação do algoritmo, decidimos formular todas as condições como proposições lógicas condicionais. Para tanto, primeiro, elaboramos as seguintes proposições lógicas simples:

Quadro 6 – Lista de proposições

a: A etiqueta anterior é ambígua
b: A etiqueta anterior é válida (não é PUNCT, None ou ???)
c: A frequência da etiqueta anterior na tabela de contexto é verdadeira
d: A etiqueta posterior é ambígua
e: A etiqueta posterior é válida (não é PUNCT, None e ???)
f: A frequência da etiqueta posterior na tabela de contexto é verdadeira
g: O algoritmo resolve a ambiguidade

Fonte: Elaboração própria.

Para as ambiguidades que não devem ser resolvidas com base nas restrições da primeira versão da ferramenta (ver subseção 3.2.1.1), elaboramos as nove proposições condicionais[55] apresentadas no quadro a seguir:

Quadro 7 – Condições para não resolver ambiguidades

Condição	**Proposição**
1	$(a \vee \neg b) \wedge (d \vee \neg e) \rightarrow \neg g$
	Se a etiqueta anterior é ambígua **ou não** é válida **e** a etiqueta posterior é ambígua **ou não** é válida, **então** o algoritmo **não** resolve a ambiguidade
2	$(a \vee (b \wedge (\neg c\_X \wedge \neg c\_Y))) \wedge (d \vee \neg e) \rightarrow \neg g$

[55] Utilizamos os seguintes operadores lógicos: *se... então* (→), *não* (¬), *e* (∧), *ou* (∨), *igual* (=), *diferente* (**!**=), *maior que* (>) e *menor que* (<).

	Dada uma ambiguidade X+Y: **se** a etiqueta anterior é ambígua **ou** é válida **e** a frequência da etiqueta anterior na tabela de contexto de X **e** Y **não** é verdadeira **e** a etiqueta posterior é ambígua **ou não** é válida, **então** o algoritmo **não** resolve a ambiguidade
3	$(a \vee \neg b) \wedge (d \vee (e \wedge (\neg f\_X \wedge \neg f\_Y))) \rightarrow \neg g$
	Dada uma ambiguidade X+Y: **se** a etiqueta anterior é ambígua **ou não** é válida **e** a etiqueta posterior é ambígua **ou** é válida **e** a frequência da etiqueta posterior na tabela de contexto de X **e** Y **não** é verdadeira, **então** o algoritmo **não** resolve a ambiguidade
4	$(a \vee (b \wedge (\neg c\_X \wedge \neg c\_Y))) \wedge (d \vee (e \wedge (\neg f\_X \wedge \neg f\_Y))) \rightarrow \neg g$
	Dada uma ambiguidade X+Y: **se** a etiqueta anterior é ambígua **ou** é válida **e** a frequência da etiqueta anterior na tabela de contexto de X **e** Y **não** é verdadeira **e** a etiqueta posterior é ambígua **ou** é válida **e** a frequência da etiqueta posterior na tabela de contexto de X **e** Y **não** é verdadeira, **então** o algoritmo **não** resolve a ambiguidade
5	$(a \vee (b \wedge (c\_X = c\_Y))) \wedge (d \vee \neg e) \rightarrow \neg g$
	Dada uma ambiguidade X+Y: **se** a etiqueta anterior é ambígua **ou** é válida **e** a frequência da etiqueta anterior na tabela de contexto de X é **igual** à frequência da etiqueta anterior na tabela de contexto de Y **e** a etiqueta posterior é ambígua **ou não** é válida, **então** o algoritmo **não** resolve a ambiguidade
6	$(a \vee \neg b) \wedge (d \vee (e \wedge (f\_X = f\_Y))) \rightarrow \neg g$
	Dada uma ambiguidade X+Y: **se** a etiqueta anterior é ambígua **ou não** é válida **e** a etiqueta posterior é ambígua **ou** é válida **e** a frequência da etiqueta posterior na tabela de contexto de X é **igual** à frequência da etiqueta posterior na tabela de contexto de Y, **então** o algoritmo **não** resolve a ambiguidade
7	$(a \vee (b \wedge (\neg c\_X \wedge \neg c\_Y))) \wedge (d \vee (e \wedge (f\_X = f\_Y))) \rightarrow \neg g$
	Dada uma ambiguidade X+Y: **se** a etiqueta anterior é ambígua **ou** é válida **e** a frequência da etiqueta anterior na tabela de contexto de X **e** Y **não** é verdadeira **e** a etiqueta posterior é ambígua **ou** é válida **e** a frequência da etiqueta posterior na tabela de contexto de X é **igual** à frequência da etiqueta posterior na tabela de contexto de Y, **então** o algoritmo **não** resolve a ambiguidade
8	$(a \vee (b \wedge (c\_X = c\_Y))) \wedge (d \vee (e \wedge (\neg f\_X \wedge \neg f\_Y))) \rightarrow \neg g$
	Dada uma ambiguidade X+Y: **se** a etiqueta anterior é ambígua **ou** é válida **e** a frequência da etiqueta anterior na tabela de contexto de X é **igual** à frequência da etiqueta anterior na tabela de contexto de Y **e** a etiqueta posterior é ambígua **ou** é válida **e** a frequência da etiqueta posterior na tabela de contexto de X **e** Y **não** é verdadeira, **então** o algoritmo **não** resolve a ambiguidade
9	$(a \vee (b \wedge (c\_X = c\_Y))) \wedge (d \vee (e \wedge (f\_X = f\_Y))) \rightarrow \neg g$
	Dada uma ambiguidade X+Y: **se** a etiqueta anterior é ambígua **ou** é válida **e** a frequência da etiqueta anterior na tabela de contexto de X é **igual** à frequência da etiqueta anterior na tabela de contexto de Y **e** a etiqueta posterior é ambígua **ou** é válida **e** a frequência da etiqueta posterior na tabela de contexto de X é **igual** à frequência da etiqueta posterior na tabela de contexto de Y, **então** o algoritmo **não** resolve a ambiguidade

Fonte: Elaboração própria.

Para demonstrar como as condições se aplicam na prática, apresentamos a seguir a tabela-verdade da Condição 1 com todos os valores possíveis de serem atribuídos às

proposições simples e compostas que compõem esta condição. Em seguida, apresentamos quatro exemplos de sentenças que contêm ambiguidades cujos contextos satisfazem a Condição 1.

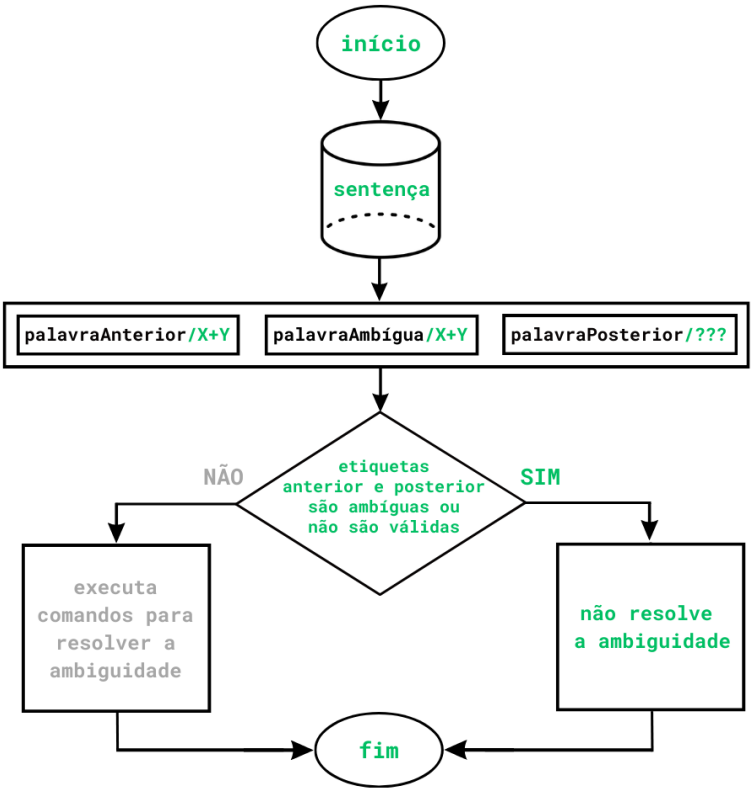
Quadro 8 – Tabela-verdade da Condição 1

1	**a**	**b**	**¬ b**	**a ∨ ¬ b**	**d**	**e**	**¬ e**	**d ∨ ¬ e**	**(a ∨ ¬ b) ∧ (d ∨ ¬ e) → ¬ g**
2	V	V	F	V	V	V	F	V	V
3	V	F	F	V	F	F	V	V	V
4	F	F	V	V	V	V	F	V	V
5	F	F	V	V	F	F	V	V	V
6	V	V	F	V	F	V	F	F	F
7	F	F	V	V	F	V	F	F	F
8	F	V	F	F	V	V	F	F	F
9	F	V	F	F	F	F	V	V	F
10	F	V	F	F	F	V	F	F	F

Fonte: Elaboração própria.

No Quadro 8, a primeira linha contém todas as proposições da Condição 1 e a primeira coluna contém o número das linhas da tabela. A segunda, terceira, quarta, sexta, sétima e oitava colunas da primeira linha contêm as proposições simples que apresentamos no Quadro 6, a quinta e a nona colunas da primeira linha contém as proposições compostas que formam a Condição 1, e a décima coluna da primeira linha contém a própria condição. As linhas 2-5 contêm todos os possíveis valores que os termos podem receber para a Condição 1 ser verdadeira, ou seja, casos de ambiguidades que não serão resolvidas pela versão atual do desambiguador, enquanto as linhas 6-10 contêm todos os possíveis valores que os termos podem receber para a condição ser falsa, isto é, casos em que o desambiguador deve resolver a ambiguidade. No fluxograma a seguir, demonstramos a sequência de operações realizadas pelo algoritmo se, ao identificar uma ambiguidade X+Y, a Condição 1 é satisfeita:

Figura 1 – Fluxograma da execução do algoritmo para a Condição 1 dada uma ambiguidade X+Y



Fonte: Elaboração própria.

Vejamos a seguir alguns exemplos de sentenças que apresentam ambiguidades que satisfazem a Condição 1 (linhas 2-5, coluna 10).

(40) Penhẽ/PRON kirimbawa/**A+N** ramé/**ADP+SCONJ** maã/**IND+INT+N+V** ,/PUNCT pesú/V maã/IND+INT+N+V i/PRON2 irũmu/ADP+SCONJ ./PUNCT

Penhẽ	kirimbawa	ramé-maã	pesú	maã	i	irũmu
2PL	valente	COND	valente	COND	3SG	com

‘se vocês fossem valentes, iriam com ele’

(41) Maria/PROPN upurakari/V tipití/N maniaka/N kitika/**A+V** irũmu/**ADP+SCONJ** ./**PUNCT**

Maria	u-purakari	tipití	maniaka	kitika	irũmu
Maria	3SG-encher	tipití	mandioca	ralada	com

‘Maria enche o tipiti com a mandioca ralada’

(42) Kwá/**ADVD+DEMX+V** apigawa/**A+N** pirasua/A+N piri/ADP+ADVG nhaã/DEMS suí/ADP ./PUNCT

kwá apigawa pirasua piri nhaã suí

DEM.PROX homempobre mais DEM.DIST que

‘este homem é mais pobre que aquele’

(43) Mukũi/**CARD+INDQ** real/**???** ./PUNCT

mukũi real

dois real

‘dois reais’

Cada exemplo acima corresponde, na sequência, às quatro proposições verdadeiras do Quadro 8. Na ambiguidade do exemplo (40) (ADP+SCONJ), a etiqueta anterior é ambígua (A+N) e a etiqueta posterior é ambígua (IND+INT+N+V), no exemplo (41) (ADP+SCONJ), a etiqueta anterior é ambígua (A+V) e a etiqueta posterior não é válida (PUNCT), no exemplo (42) (ADVD+DEMX+V), a etiqueta anterior não é válida (None) e a etiqueta posterior é ambígua (A+N), e, por fim, no exemplo (43) (CARD+INDQ), a etiqueta anterior não é válida (None) e a etiqueta posterior não é válida (etiqueta desconhecida).

No que se refere à resolução das ambiguidades, os casos em que a ferramenta deve fazer a desambiguação são todos aqueles que não satisfazem nenhuma das condições apresentadas no Quadro 7.

Na próxima seção, descrevemos o processo de implementação da ferramenta e o seu funcionamento.

3.2.4 Implementação e estrutura do desambiguador

Após a elaboração das condições, iniciamos a implementação do desambiguador[56]. Uma vez que, ao receber sentenças etiquetadas como entrada, o desambiguador executa uma sequência de comandos para retornar as mesmas sentenças com todas as ambiguidades resolvidas, primeiro descrevemos as instruções dessa sequência de comandos. Dividimos o processo de desambiguação nas seguintes etapas: (i) decisão de resolver ou não a ambiguidade identificada nos dados de entrada do desambiguador (com base nas etiquetas consideradas válidas); (ii) identificação do contexto da ambiguidade com relação

[56] Disponível em: https://github.com/juliana-gurgel/nheengatu/tree/main/src. Acesso em: 14 ago. 2023.

a todos os contextos possíveis de ocorrer com cada etiqueta que compõe a ambiguidade e sua respectiva frequência; e (iii) resolução da ambiguidade. Na primeira etapa, a ferramenta recebe como entrada uma sentença ou arquivos contendo sentenças anotadas pelo Nheengatagger, lê todas as ambiguidades e, para cada ambiguidade, identifica, com base nos seus contextos, se deve seguir para a segunda e terceira etapas, isto é, se a ambiguidade deve ser resolvida, ou se, com base nas condições do Quadro 8, a ambiguidade não deve ser resolvida.

A partir da descrição do processo de desambiguação, implementamos o módulo, que é composto por seis componentes:

Figura 2 – Descrição dos componentes do desambiguador

1. **Cli**

   Permite executar o desambiguador na linha de comando. Ele suporta dois modos de operação: (i) a geração de tabela de contexto a partir de um corpus e um *tagset* em estrutura de dados tabular; (ii) desambiguação de uma sentença a partir da Tabela de Contexto. Este último pode ser utilizado com outros comandos para a desambiguação de arquivos em um diretório.

2. **Corpus**

   Realiza a leitura do corpus para que seja utilizado pelos outros componentes.

3. **Word**

   Identifica as etiquetas de uma palavra e verifica se as palavras são válidas ou ambíguas de acordo com o tipo e a quantidade de etiquetas que apresentam. Além disso, verifica, para cada etiqueta do *tagset* (ver APÊNDICE B), se existem contextos extraídos do corpus na Tabela de Contexto. No caso das ambiguidades, este programa também verifica se todas as etiquetas que as compõem apresentam pelo menos uma ocorrência sem ambiguidade no corpus. Se todas as etiquetas estiverem

presentes na Tabela de Contexto, o desambiguador resolve a ambiguidade, caso contrário, retorna a ambiguidade.

4. **Sentence**

   Realiza o pré-processamento de sentenças individuais. Após a tokenização das sentenças, este componente opera com duas finalidades: (i) verificar se a sentença contém palavras com uma etiqueta específica, tendo em vista a construção da Tabela de Contexto; e (ii) verifica se a sentença contém ambiguidades a serem resolvidas.

5. **Context**

   Gera a Tabela de Contexto a partir do *tagset* e do corpus. A Tabela contém as etiquetas anteriores e posteriores de cada etiqueta do *tagset* as suas frequências nos contextos em que ocorrem.

6. **Desambiguador**

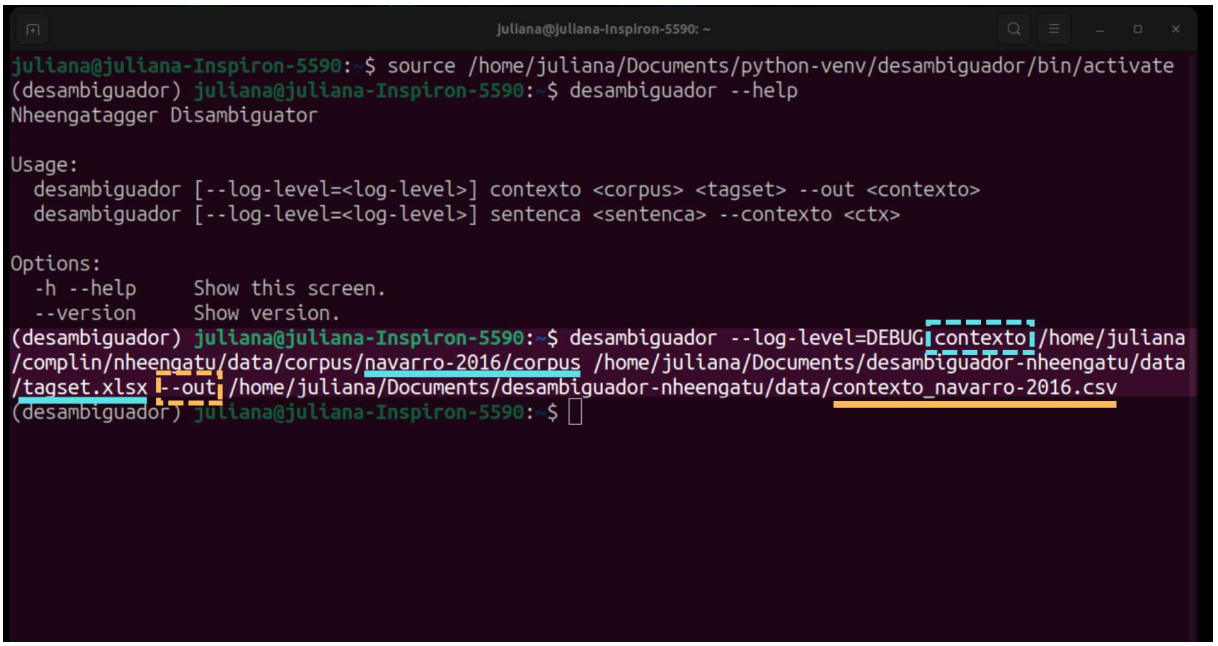
   Faz a desambiguação por meio das funções implementadas nos cinco componentes apresentados acima. Este componente verifica se a sentença possui ambiguidades que se enquadram nas condições descritas no Quadro 8 ou se possui apenas uma palavra. Em caso afirmativo, a função retorna a ambiguidade, ou seja, a sentença original. Caso contrário, faz a seleção das etiquetas adequadas para cada contexto com base nas frequências contidas na Tabela de Contexto por meio da análise das palavras vizinhas.

Fonte: Elaboração própria.

Uma vez que implementamos o desambiguador de forma a ser possível testar a sua performance com diferentes conjuntos de etiquetas e dados de treinamento simplesmente

fornecendo *inputs* diferentes, a ferramenta é capaz de gerar em milissegundos Tabelas de Contexto a partir de diferentes conjuntos de dados, conforme a figura a seguir:

Figura 3 – Geração de uma Tabela de Contexto a partir de Navarro (2016)

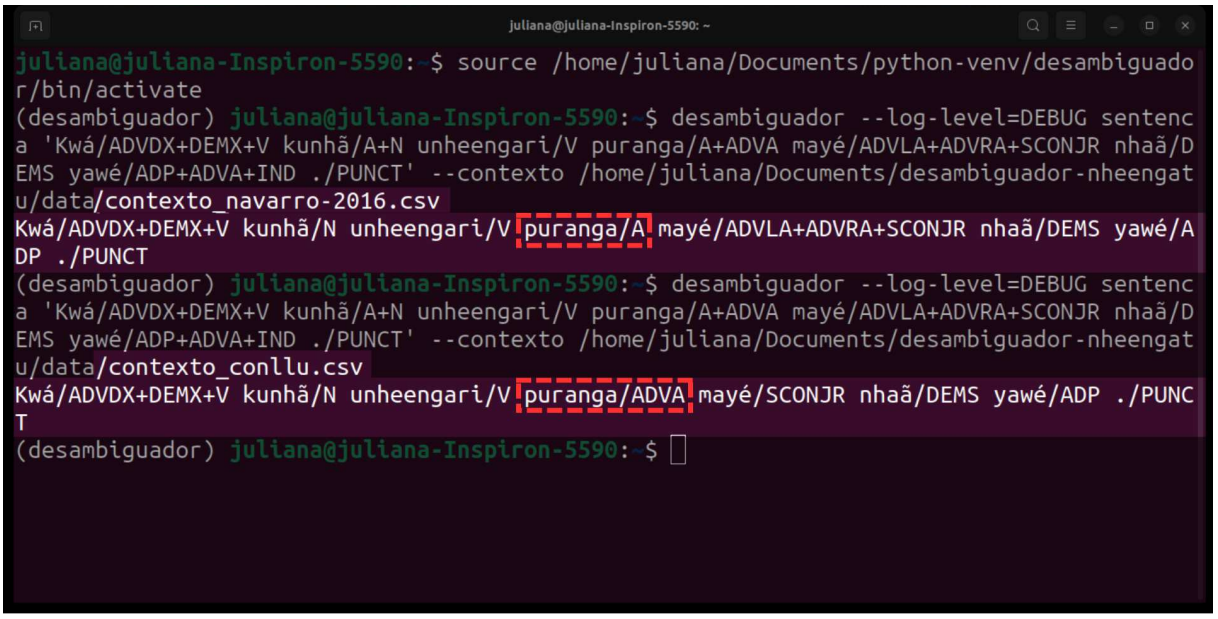


Fonte: Elaboração própria.

De acordo com a Figura 3, o comando para geração de uma Tabela de Contexto é composto por quatro elementos: (i) *desambiguador*, o comando principal; (ii) *--log-level*, uma opção de linha de comando que permite definir o nível de registro de informações sobre a execução do programa; (iii) *contexto*, um subcomando que recebe dois argumentos, o primeiro, o diretório do corpus, e o segundo, o conjunto de etiquetas; e (iv) *--out*, subcomando que recebe um argumento, o arquivo de saída onde a Tabela de Contexto deve ser armazenada. Já para a desambiguação de uma sentença, o argumento do subcomando *sentenca* é uma sentença etiquetada pelo Nheengatagger e do subcomando *--contexto* é uma Tabela de Contexto.

A partir da utilização de diferentes Tabelas de Contextos, podemos observar se existe diferença na performance da ferramenta na resolução de ambiguidades, como demonstramos no exemplo a seguir:

Figura 4 – Resultado da desambiguação da mesma sentença utilizando duas Tabelas de Contexto diferentes



Fonte: Elaboração própria.

Na Figura 4, apresentamos duas possibilidades de resolução da ambiguidade A+ADVA na seguinte sentença[57]:

(44) Kwá/ADVDX+DEMX+V kunhã/A+N unheengari/V puranga/A+ADVA mayé/ADVLA+ADVRA+SCONJR nhaã/DEMS yawé/ADP+ADVA+IND ./PUNCT

kwá	kunhã	u-nheengari	puranga	mayé	nhaã	yawé
DEM.PROX	mulher	3SG-cantar	bem	como	DEM.DIST	assim

‘esta mulher canta tão bem como aquela’

No exemplo (44), temos a palavra *puranga*, que, a depender do contexto, pode ser classificada como adjetivo de 1ª classe (A) ou advérbio de maneira (ADVA). Por modificar o verbo *cantar* na sentença acima, a ambiguidade A+ADVA deve ser resolvida como ADVA. No primeiro resultado da Figura 2, utilizamos a Tabela de Contexto gerada a partir de Navarro (2016)[58], enquanto o segundo resultado foi obtido com a utilização da tabela gerada a partir

[57] Esta sentença foi extraída da versão do corpus de Navarro (2016) etiquetado pelo Nheengatagger em 26 de julho de 2023. Disponível em: https://github.com/juliana-gurgel/nheengatu/tree/main/corpus. Acesso em: 15 ago. 2023.

[58] Esta tabela de contexto foi gerada a partir da versão do corpus de Navarro (2016) etiquetado pelo Nheengatagger em 26 de julho de 2023.
Disponível em: https://github.com/juliana-gurgel/nheengatu/tree/main/corpus. Acesso em: 15 ago. 2023.

das sentenças do UD_Nheengatu-CompLin[59]. Uma vez que V (verbo de 1ª classe) é a etiqueta anterior à ambiguidade A+ADVA e a etiqueta seguinte é outra ambiguidade, a resolução de A+ADVA será feita apenas com base na frequência da etiqueta V presente no contexto de A e ADVA. A seguir, apresentamos a frequência do verbo de 1ª classe nas duas tabelas de contexto.

Tabela 3 – Fragmento da Tabela de Contexto das etiquetas A e ADVA

1	Etiqueta	Etiqueta do contexto	Tipo de contexto	Navarro (2016)	UD_Nheengatu-CompLin
2	A	V	Anterior	7	11
3	ADVA	V	Anterior	6	26

Fonte: Elaboração própria.

Com base na frequência apresentada na tabela acima, o desambiguador resolve incorretamente a ambiguidade como A (adjetivo de 1ª classe) quando recebe como entrada a Tabela de Contexto gerada a partir de Navarro (2016), cujas ambiguidades não estão resolvidas, e resolve corretamente quando recebe a Tabela gerada a partir do UD_Nheengatu-CompLin, cujas ambiguidades estão resolvidas.

Por outro lado, assim como a utilização de um corpus sem ambiguidades pode aumentar a acurácia da ferramenta na resolução de algumas ambiguidades, como A+ADVA, também obtemos resultados superiores na resolução de outras ambiguidades utilizando a Tabela gerada a partir de Navarro (2016).

Na seção a seguir, apresentamos os resultados obtidos para todas as ambiguidades com as Tabelas de Contexto geradas a partir dos dois conjuntos de dados[60].

[59] Disponível em: https://github.com/UniversalDependencies/UD_Nheengatu-CompLin/tree/dev. Acesso em: 01 jun. 2023.

[60] As duas Tabelas de Contexto utilizadas estão disponíveis em: https://github.com/juliana-gurgel/nheengatu/tree/main/src/NheengataggerDisambiguator/tests. Acesso em: 15 ago. 2023.

4 RESULTADOS

4.1 Cobertura do corpus

A compilação dos textos de Navarro (2016), Navarro e Ávila (2017), Casasnovas (2006) e Trevisan (2017) está em andamento[61]. Na tabela abaixo, apresentamos a cobertura atual do corpus e a situação em que cada obra se encontra nessa etapa.

Tabela 4 – Cobertura atual do corpus compilado

Obra	Componente	Cobertura	Situação	Sentenças
Navarro (2016)	Textos	100%	Concluído	317
Navarro (2016)	Exemplos	100%	Concluído	477
Navarro (2016)	Exercícios	30.7%	Compilação	719
Navarro e Ávila (2017)	Lendas	100%	Revisão	1217
Casasnovas (2006)	Exemplos	100%	Revisão	247
Casasnovas (2006)	Lendas	100%	Revisão	238
Trevisan (2017)	Capítulos	100%	Pré-processamento	961
			Total	4176

Fonte: Elaboração própria.

A seguir, apresentamos os resultados dos dois primeiros testes realizados.

4.2 Testes preliminares

A fim de testarmos o funcionamento do desambiguador considerando todas as etiquetas do *tagset* do Nheengatagger (ver APÊNDICE B) incluindo as ambiguidades e as etiquetas válidas e não válidas, realizamos dois testes preliminares a partir de cinco conjuntos de dez sentenças (ver APÊNDICE I), um para cada ambiguidade apresentada na Tabela 1. Nesses testes, avaliamos o desambiguador em função da sua acurácia[62], que foi calculada considerando a performance da ferramenta apenas com relação à resolução das ambiguidades, sem contar as palavras que já estavam etiquetadas corretamente.

[61] À medida que são concluídas e revisadas, as obras compiladas estão sendo disponibilizadas em: https://github.com/juliana-gurgel/nheengatu/tree/main/corpus. Acesso em: 14 ago. 2023.

[62] Cálculo utilizado: (VP + VN) / (VP + VN + FP + FN), em que VP=Verdadeiros Positivos; VN=Verdadeiros Negativos; FP=Falsos Positivos e FN=Falsos Negativos.

Nas tabelas a seguir, os Verdadeiros Positivos (VP) correspondem às palavras cujas ambiguidades foram resolvidas corretamente pelo desambiguador, os Verdadeiros Negativos (VN) são as palavras cujas ambiguidades não foram resolvidas por satisfazerem alguma das condições do Quadro 7, os Falsos Positivos (FP) são as palavras cujas ambiguidades foram resolvidas incorretamente e os Falsos Negativos (FN) correspondem àquelas que não satisfazem as condições do Quadro 7, mas ainda assim não foram resolvidas pelo desambiguador. Na tabela abaixo, apresentamos os resultados do primeiro teste preliminar.

Tabela 5 – Acurácia do desambiguador no Teste Preliminar 1 por ambiguidade

Ambiguidade	VP	VN	FP	FN	Acurácia
A+ADV	4	0	5	1	40%
ADP+FUT+SCONJ	4	0	5	1	40%
ADP+SCONJ	6	0	4	0	60%
A+N	2	0	1	7	20%
PRON+PRON2	10	0	0	0	100%
				Total	**52%**

Fonte: Elaboração própria.

No Teste Preliminar 1, todas as etiquetas foram incluídas na construção da Tabela de Contexto. Desse modo, o cálculo da frequência foi feito considerando as ambiguidades e as etiquetas classificadas como inválidas, isto é, os sinais de pontuação (PUNCT), as palavras desconhecidas (???) e a etiqueta None. No Teste Preliminar 2, cujos resultados apresentamos na Tabela 6, excluímos do cálculo da frequência as ambiguidades e as etiquetas não válidas a fim de avaliar se havia diferença no desempenho da ferramenta.

Tabela 6 –Acurácia do desambiguador no Teste Preliminar 2 por ambiguidade

Ambiguidade	VP	VN	FP	FN	Acurácia
A+ADV	9	1	0	0	90%
ADP+FUT+SCONJ	9	1	0	0	90%
ADP+SCONJ	6	0	4	0	60%
A+N	3	6	1	0	30%
PRON+PRON2	10	0	0	0	100%
				Total	**74%**

Fonte: Elaboração própria.

Como podemos observar nas Tabelas 5 e 6, a utilização da frequência de todas as etiquetas no Teste Preliminar 1 aumentou o número de Falsos Positivos com relação ao Teste Preliminar 2, em que as ambiguidades e as etiquetas não válidas foram desconsideradas. Partindo desse resultado, decidimos desconsiderar essas etiquetas na avaliação da ferramenta.

4.3 Avaliação do desambiguador

4.3.1 Preparação dos conjuntos de dados

Antes de avaliarmos a performance do desambiguador na resolução de todas as ambiguidades do corpus, realizamos três testes da ferramenta utilizando dois conjuntos de etiquetas e três Tabelas de Contexto geradas a partir de dois corpora de desenvolvimento. Para tanto, construímos dois conjuntos-teste: o Testset 1, etiquetado com o conjunto de etiquetas apresentado no APÊNDICE B, que contém todas as etiquetas do Nheengatagger (*tagset* Original), e o Testset 2, etiquetado com o conjunto de etiquetas parcial, mencionado na seção 2.2.1 (*tagset* Simplificado), do qual excluímos as etiquetas que fazem distinções de subtipos para as classes advérbio, substantivo, pronome demonstrativo e adposição (ver APÊNDICE C). Além disso, utilizamos dois corpora de desenvolvimento para a geração das Tabelas de Contexto: o primeiro, Navarro (2016), que não contém ambiguidades resolvidas, e o segundo, o UD_Nheengatu-CompLin[63], doravante UD Nheengatu, cujas ambiguidades estão resolvidas. Assim, foram geradas três Tabelas de Contexto, cada uma utilizada em um dos três testes, conforme o quadro a seguir:

Tabela 7 – Tabelas de Contexto utilizadas nos Testes 1, 2 e 3

Tabela de Contexto	Conjunto de etiquetas	Conjuntos-teste	Testes	Etiquetas	Ambiguidades
Navarro	*Tagset* Original	Testset 1	1	86	55
UD Nheengatu	*Tagset* Original	Testset 1	2	86	55
UD Nheengatu Simplificada	*Tagset* Simplificado	Testset 2	3	56	43

Fonte: Elaboração própria.

[63] Disponível em: https://github.com/UniversalDependencies/UD_Nheengatu-CompLin/tree/dev. Acesso em: 01 jun. 2023.

Por sua vez, os conjuntos-teste (Testsets 1 e 2) apresentados no APÊNDICE J[64] foram elaborados a partir de sentenças extraídas de Navarro (2016) que apresentam palavras ambíguas. No que se refere às listas de ambiguidades, identificamos 55 tipos a partir do conjunto de etiquetas original do Nheengatagger (ver APÊNDICE D), e 43 tipos a partir do conjunto de etiquetas simplificado (ver APÊNDICE E).

Antes da elaboração dos Testsets 1 e 2, identificamos a necessidade de obtermos uma distribuição homogênea dos dados do corpus de teste considerando o total de 1047 sentenças ambíguas presentes em Navarro (2016) em função das 55 ambiguidades. Para tanto, decidimos extrair aleatoriamente um número mínimo de 5 sentenças para cada ambiguidade quando o número *n* de sentenças do corpus com ocorrências da ambiguidade era $n > 5$ e incluímos todas as sentenças das ambiguidades quando o número de ocorrências era $n < 5$. Vejamos o seguinte quadro:

Tabela 8 – Descrição dos conjuntos de sentenças utilizados nos Testes 1, 2 e 3

Testsets	**Testes**	**Sentenças**	***Token*s**	**Etiquetadas**	**Ambíguas**	**Desconhecidas**	**Sinais de pontuação**
Testset 1	1 e 2	224	1887	1261	230	17	385
Testset 2	3	206	1744	1164	212	16	352

Fonte: Elaboração própria.

Conforme o quadro acima, o Testset 1, etiquetado com o *tagset* Original, é composto por 224 sentenças, que correspondem a 23% do total de sentenças ambíguas de Navarro (2016), enquanto o Testset 2, etiquetado com o *tagset* Simplificado, é composto por 206 sentenças, pois utilizamos o mesmo conjunto de sentenças extraído para o Testset 1 subtraído daquelas cujas ambiguidades não foram identificadas no corpus etiquetado com o conjunto simplificado. Por exemplo, ADVJ+ADVT (advérbio conjuncional + advérbio temporal) está presente na lista de ambiguidades original, mas não está contida na lista de ambiguidades simplificada, uma vez que é composta por subtipos de advérbios, os quais não foram considerados na versão simplificada do conjunto de etiquetas. Assim, na elaboração do Testset 2, excluímos as sentenças com essas e outras ambiguidades que não estão presentes nesse conjunto de sentenças.

[64] No APÊNDICE J, a primeira coluna indica se os subconjuntos de sentenças estavam contidos nos dois *testsets* e quais estavam contidos apenas no Testset 1.

4.3.3 Testes

No Teste 1, cujos resultados detalhados apresentamos no APÊNDICE L, utilizamos as sentenças do Testset 1 anotadas pelo conjunto de *tagset* Original do Nheengatagger e construímos a Tabela de Contexto a partir de Navarro (2016)[65], cujas sentenças também estão etiquetadas com o conjunto de etiquetas original. No Teste 2 (ver resultados no APÊNDICE M), utilizamos as sentenças do Testset 1, mas, dessa vez, utilizamos a Tabela de Contexto construída a partir do UD Nheengatu[66] anotado com o *tagset* Original, com o objetivo de avaliarmos a performance da ferramenta com uma Tabela de Contexto gerada a partir de sentenças cujas ambiguidades estão resolvidas. Por fim, no Teste 3 (ver APÊNDICE N), utilizamos as sentenças do Testset 2 e o desambiguador recebeu como entrada a Tabela de Contexto do UD Nheengatu Simplificada[67]. Na tabela a seguir, apresentamos a proporção de acertos e erros obtidos nos três testes:

Tabela 9 – Proporção de erros e acertos do desambiguador nos Testes 1, 2 e 3

Teste	Tipos de ambiguidades	Sentenças	Ambiguidades	Resolvidas	(%)	Acertos	(%)	Erros	(%)
Teste 1	55	224	230	85	36.9	41	48.2	44	51.8
Teste 2	55	224	230	181	78.7	90	49.2	92	50.8
Teste 3	44	206	212	179	84.4	89	49.7	90	50.2

Fonte: Elaboração própria.

No Teste 1, em que utilizamos a Tabela de Contexto gerada a partir do corpus com ambiguidades (Navarro, 2016), 36.9% do total de 230 ambiguidades foi resolvido. Do total de 85 ambiguidades resolvidas, 48.2% foram resolvidas corretamente enquanto 51.8% foram resolvidas incorretamente. No Teste 2, em que utilizamos a Tabela gerada a partir dos dados sem ambiguidades (UD Nheengatu), foram resolvidas 78.7% do total de ocorrências. Das 181 palavras desambiguadas, 49.7% foram resolvidas corretamente e 50.8% foram resolvidas incorretamente. No Teste 3, em que utilizamos a Tabela gerada a partir do UD Nheengatu etiquetado com o conjunto de etiquetas simplificado, 84.4% das 212 ambiguidades foram

[65] Construímos a Tabela utilizando a versão do corpus etiquetado em 26 de julho de 2023, disponível em: https://github.com/juliana-gurgel/nheengatu/tree/main/corpus. Acesso em: 10 ago. 2023.
[66] Disponível em: https://github.com/juliana-gurgel/nheengatu/tree/main/src/NheengataggerDisambiguator/tests. Acesso em: 15 ago. 2023.
[67] Disponível em: https://github.com/juliana-gurgel/nheengatu/tree/main/src/NheengataggerDisambiguator/tests. Acesso em: 15 ago. 2023.

resolvidas. Das 179, 49.7% foram resolvidas corretamente e 50.2% foram resolvidas incorretamente.

A fim de avaliarmos o desambiguador separadamente, calculamos a sua acurácia considerando apenas as ambiguidades a serem resolvidas, sem levar em conta os demais dados presentes no corpus (palavras etiquetadas corretamente e palavras desconhecidas). A seguir, apresentamos os resultados obtidos:

Tabela 10 – Acurácia do desambiguador nos Testes 1, 2 e 3

Testes	Ambiguidades	Resolvidas	VP	VN	FP	FN	Acurácia (%)
Teste 1	230	85	41	145	44	0	80.9
Teste 2	230	181	90	48	92	0	60
Teste 3	212	179	89	33	90	0	57.5

Fonte: Elaboração própria.

Conforme a tabela acima, no Teste 1, o desambiguador alcançou 80.9% de acurácia. Ainda que inferior ao estado da arte, o resultado mostra que a ferramenta teve uma performance alta tanto na seleção da etiqueta correta quanto na identificação das ambiguidades que de fato deveriam ser resolvidas com base nas condições do Quadro 7. Além disso, a ausência de falsos negativos, que correspondem às ambiguidades que deveriam ser resolvidas, mas não foram, demonstra que o desambiguador está aplicando corretamente as referidas condições. Com relação aos falsos positivos, isto é, as ambiguidades resolvidas incorretamente, no Teste 1 a ferramenta retornou um valor significativamente inferior em comparação aos Testes 2 e 3. Por outro lado, vale destacar que o fato da Tabela de Contexto Navarro ter sido gerada a partir do mesmo conjunto de dados de onde foram extraídas as sentenças do corpus de teste pode indicar um sobreajuste (*overfitting*), que é o caso em que o modelo utilizado se adapta bem aos dados de treino, mas não consegue fazer generalizações para dados novos. Para confirmar esta hipótese, será necessário testar a Tabela de Contexto gerada a partir de Navarro (2016) com relação a um conjunto de sentenças de outras obras, algo que não foi possível realizar no âmbito desta pesquisa.

No que diz respeito aos Testes 2 e 3, os valores de acurácia obtidos foram 60% e 57.5%, inferiores ao estado da arte. Nesses testes, buscamos verificar como a performance da ferramenta na resolução de determinadas ambiguidades pode ter sido influenciada por fatores como subtipos de classes gramaticais e escolhas de etiquetas.

Nos dois testes, o desempenho foi semelhante na resolução da ambiguidade ADP+ADVA+IND (posposição + advérbio de maneira + pronome indefinido)[68], indicando que o subtipo de advérbio não teve impacto significativo na resolução dessa ambiguidade no segundo teste com relação ao primeiro. Já para as ambiguidades A+ADVC (adjetivo de 1ª classe + advérbio locativo) e A+ADVS (adjetivo de 1ª classe + advérbio de intensidade)[69], os testes apresentaram um resultado diferente. No Teste 2, apenas uma pequena proporção das ambiguidades foi resolvida, mas no Teste 3, o número de acertos foi substancialmente superior, possivelmente devido à influência do subtipo de advérbio no resultado do Teste 1.

No caso da ambiguidade ADP+N (adposição + substantivo), a escolha da etiqueta teve um efeito na desambiguação. Usar a etiqueta N para nome próprio em vez de PROPN levou a uma melhoria de 50% na resolução da ambiguidade, demonstrando a importância de considerarmos, nas decisões de etiquetagem, as propriedades distribucionais de certas classes gramaticais representadas por mais de uma etiqueta.

A análise da ambiguidade CLADP+PRON2 (posposição enclítica + pronome de 2ª classe)[70] revelou uma diferença positiva entre os dois testes, possivelmente atribuída à influência do subtipo de adposição, o que indica que essa categorização específica desempenha um papel importante na desambiguação a depender da frequência na tabela de contexto, aspecto a ser considerado nas próximas versões da ferramenta. Já na resolução da ambiguidade ADP+N+SCONJ (posposição + substantivo + conjunção subordinativa pós-verbal), o subtipo de adposição não parece ter impactado o resultado, pois os valores obtidos nos dois testes foram iguais.

No caso de ADVDX+DEMX+V (advérbio demonstrativo proximal + pronome demonstrativo proximal + verbo de 1ª classe)[71], o subtipo de demonstrativo não influenciou na resolução, já que o desempenho permaneceu inalterado entre os testes.

Por fim, a ambiguidade ART+CARD+FRUST+SCONJ (artigo indefinido + numeral cardinal + partícula de frustrativo + conjunção subordinativa pós-verbal) apresentou variações entre os testes, possivelmente relacionadas a diferenças na frequência de ocorrência

[68] No Teste 3, essa ambiguidade passa a ser ADP+ADV+IND (adposição + advérbio + pronome indefinido), conforme o APÊNDICE E.
[69] No Teste 3, as duas ambiguidades correspondem a A+ADV (adjetivo de 1ª classe + advérbio).
[70] No Teste 3, ADP+PRON2 (adposição + pronome de 2ª classe).
[71] No Teste 3, ADV+DEM+V (advérbio + pronome demonstrativo + verbo de 1ª classe).

de conjunções subordinativas após verbos, comparado à frequência de artigos indefinidos antes de substantivos[72].

4.3.4 Avaliação do Nheengatagger

Para avaliarmos o desempenho do Nheengatagger após a utilização do desambiguador, calculamos a acurácia do etiquetador antes e depois do módulo[73]. Na tabela a seguir, apresentamos os resultados:

Tabela 11 – Acurácia do Nheengatagger antes e depois do desambiguador

			Antes			Depois		
Teste	**Testset**	**Tabela de Contexto**	**Acertos**	**Erros**	**Acurácia**	**Acertos**	**Erros**	**Acurácia**
1	Testset 1	Navarro (2016)	1261	247	83.6%	1.302	162	88.9%
2	Testset 1	UD Nheengatu	1261	247	83.6%	1.351	65	95.4%
3	Testset 2	UD Nheengatu[74]	1164	228	83.6%	1.253	49	96.2%

Fonte: Elaboração própria.

Com o desambiguador, no Teste 1, no qual utilizamos a tabela de contexto de Navarro (2016), a acurácia do Nheengatagger foi de 88.9%, ou seja, o módulo proporcionou um aumento de 5.3% na taxa de acerto do etiquetador. Com relação aos demais testes, nos quais utilizamos a Tabela de Contexto gerada a partir do UD Nheengatu, a acurácia do Nheengatagger chegou a 95.4% no Teste 2 e 96.2% no Teste 3, valores que representam um aumento de 11.8% e 12.6%, respectivamente, índices atingidos pelos etiquetadores morfossintáticos mais avançados.

Uma vez que as acurácias do Nheengatagger nos Testes 2 e 3 apresentaram uma diferença de apenas 0.8% entre si, concluímos que o *tagset* Simplificado, utilizado no Teste 3, não resultou em uma melhoria significativa na performance do desambiguador que justifique a necessidade da utilização deste conjunto de etiquetas no futuro.

---

[72] Cf. as Tabelas de Contextos utilizadas nos testes em: https://github.com/juliana-gurgel/nheengatu/tree/main/src/NheengataggerDisambiguator/tests. Acesso em: 15 ago. 2023.

[73] Neste cálculo, utilizamos a equação apresentada na seção 2.1.2, considerando que os *Acertos* são todas as palavras etiquetadas corretamente (não incluímos os sinais de pontuação), e os *Erros* são as palavras com etiquetas ambíguas ou palavras desconhecidas (etiqueta ???).

[74] Esta tabela foi gerada a partir das sentenças do UD_Nheengatu-CompLin anotadas com o conjunto de etiquetas simplificado, apresentado no APÊNDICE C.

Por uma limitação de tempo, não tivemos a oportunidade de analisar, no âmbito desta pesquisa, os dados obtidos na resolução de todas as ambiguidades.

5 CONSIDERAÇÕES FINAIS

Neste trabalho, descrevemos a implementação de um desambiguador baseado em regras para o nheengatu que consiste em um dos componentes do etiquetador Nheengatagger (ALENCAR, 2020). No que diz respeito ao processamento de linguagem natural de línguas indígenas brasileiras, pelo que sabemos, a ferramenta desenvolvida nesta pesquisa é a primeira voltada para a resolução de ambiguidades, passo que é de fundamental importância na tarefa de etiquetagem morfossintática e em outras etapas do PLN.

O objetivo principal desta pesquisa foi implementar o desambiguador do Nheengatagger em Python. Considerando o exposto na seção 3.2, este objetivo foi cumprido completamente. Por outro lado, os objetivos relacionados à compilação dos textos em nheengatu a partir dos trabalhos de Navarro (2016), Navarro e Ávila (2017), Trevisan (2017) e Casasnovas (2006) foram alcançados parcialmente. Em pesquisas futuras, cumpre avançar nos seguintes aspectos: (i) a conclusão da compilação dos exercícios de Navarro (2016); (ii) a revisão das sentenças traduzidas para o português e o inglês; (iii) a revisão e o ajuste da ortografia utilizada por Casasnovas (2006) com base na ortografia utilizada por Navarro (2016), uma vez que esta última é a ortografia adotada no nosso trabalho; e (v) o aprimoramento do algoritmo do desambiguador.

Ao avaliarmos o desambiguador isoladamente, embora este tenha apresentado um desempenho inferior ao estado da arte (JURAFSKY; MARTIN, 2023a) nos três testes, a ferramenta representa uma melhoria considerável ao Nheengatagger. Com a integração do módulo, o etiquetador alcançou o índice de 95% de acurácia em dois testes, conforme os resultados apresentados na seção 4.3. A taxa de acerto proporcionada pelo desambiguador reduzirá significativamente o trabalho de etiquetagem e revisão de corpora do nheengatu anotados morfossintaticamente.

Ainda assim, o algoritmo pode ser aprimorado em alguns aspectos. Por exemplo, identificamos que a precisão da ferramenta pode ser melhorada levando em conta a morfologia dos verbos na resolução das ambiguidades, por meio da incorporação da palavra como uma das variáveis do cálculo e com a utilização de outro cálculo de frequência, como a probabilidade de transição, utilizada em modelos estatísticos, ou a frequência relativa, utilizada no *parser* desenvolvido no âmbito do projeto Transformations and Discourse Analysis (HARRIS, 1962; JURAFSKY; MARTIN, 2023a). Além disso, cumpre revisar a Tabela de Contexto para refletir os contextos gramaticais de cada classe e concluir a

etiquetagem e desambiguação de todas as sentenças dos corpora, de forma a expandir o nosso conjunto de dados e aumentar o desempenho do desambiguador.

Cumpre destacar, também, algumas limitações do trabalho do ponto de vista teórico que, por conta do tempo, não tivemos oportunidade de explorar. Um dos pontos identificados foi a ausência de uma definição detalhada e uma discussão aprofundada das etiquetas do Nheengatagger considerando os seus aspectos gramaticais. Especialmente para aquelas etiquetas que contêm subtipos, a análise por meio de exemplos teria fornecido maior clareza sobre sua aplicação no processo de etiquetagem e desambiguação. Ademais, a apresentação das propriedades distribucionais das classes de palavras do nheengatu fora do subdomínio do sintagma nominal poderia fundamentar de forma mais precisa a análise dos dados e enriqueceria a discussão dos resultados.

Em vista do exposto, aceitamos qualquer responsabilidade decorrente das limitações teóricas e metodológicas deste trabalho. Além disso, compreendemos que ainda há muito a ser feito no que diz respeito ao desenvolvimento de recursos computacionais para o nheengatu e à expansão dos corpora desta língua anotados morfossintaticamente.

Por fim, vale salientar que o desenvolvimento desta dissertação envolveu diferentes domínios teóricos e metodológicos e que, portanto, ainda que não tenhamos alcançado completamente os objetivos estabelecidos, uma vez que a ferramenta pode ser utilizada com diferentes *inputs*, este trabalho oferece uma contribuição significativa para o processamento computacional das línguas indígenas brasileiras em geral, bem como para uma compreensão mais aprofundada do nheengatu que sirva como ponto de partida para pesquisas subsequentes.

REFERÊNCIAS

ALENCAR, L. F. Técnicas em software livre para exploração de corpora do português livremente disponíveis na WWW. **Veredas On-Line**, Juiz de Fora, v. 2, p. 134-150, 2009. Disponível em: https://periodicos.ufjf.br/index.php/veredas/article/view/25156. Acesso em: 07 ago. 2021.

ALENCAR, L. F. de. Utilização de informações lexicais extraídas automaticamente de corpora na análise sintática computacional do português. **Revista Estudos da Linguagem**, Belo Horizonte, v. 19, n. 1, p. 7-85, 2011.

ALENCAR, L. F. de. Novos recursos do Aelius para o processamento computacional raso do português. *In*: LAPORTE, É.; SMARSARO, A.; VALE, O. A. (org.) **Dialogar é preciso:** linguística para o processamento de línguas. 1. Ed. Vitória: PPGEL; UFES, 2013a.

ALENCAR, L. F. de. BrGram: uma gramática computacional de um fragmento do português brasileiro no formalismo da LFG. *In*: BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY 2013 STIL, 9., 2013. Fortaleza. Proceedings. Fortaleza: Sociedade Brasileira de Computação, p. 183-188, 2013b.

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. Projeto de pesquisa **Técnicas em softwares livres para a linguística de corpus (13a Etapa)**. Fortaleza: Universidade Federal do Ceará, 2022. Não publicado.

ALENCAR, L. F. de. Aelius Brazilian Portuguese POS-Tagger and Corpus Annotation Tool, versão 0.9.7. Fortaleza: [s.n.], 2013c. Disponível em: http://aelius.sourceforge.net/. Acesso em: 18 fev. 2021.

ALENCAR, L. F. de. **Aelius**: uma ferramenta para anotação automática de corpora usando o NLTK. *In*: IBAÑOS, A. M. T.; MOTTIN, L. P.; SARMENTO, S.; BERBER SARDINHA, T. (Orgs.). Pesquisas e Perspectivas em Linguística de Corpus. Campinas: Mercado de Letras, 2015. p. 233-282.

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.

BIRD, S.; KLEIN, E.; LOPER, E. **Natural language processing with Python:** analyzing text with the Natural Language Toolkit. Sebastopol, CA: O'Reilly, 2009.

BUENO, L. *et al*. The late Pleistocene/early Holocene archaeological record in Brazil: A geo-referenced database. **Quaternary International**. v. 301, p. 74-93. 8 jul. 2013.

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: 04 jul. 2021.

EDELWEISS, F. G. **Estudos tupis e tupis-guaranis - confrontos e revisões**. Rio de Janeiro: Livraria Brasiliana Editora, 1969.

FRANCEZ, N.; WINTNER, S. **Unification grammars.** Cambridge: CUP, 2012. p. 2-3.

FREIRE, José Ribamar Bessa. Rio Babel: a história das línguas na Amazônia. 2.ed. Rio de Janeiro: EdUERJ, 2011.

GALVES, C.; ANDRADE, A. L. de; FARIA, P. Tycho Brahe Parsed Corpus of Historical Portuguese. Campinas: Universidade Estadual de Campinas, 2017. Disponível em: http://www.tycho.iel.unicamp.br/~tycho/corpus/ Acesso em 05. jul. 2021.

GOYVAERTS, J.; LEVITHAN, L. **Regular Expressions Cookbook**, 2. ed. CA: O'Reilly, 2012.

GRIES, S. T. ; MELLO, H. R. ; PAIXAO, C. A. ; SOUZA, A. L. E.; ZARA, J. (org.). **Estatística com R para a linguística**. 1. ed. Belo Horizonte: Faculdade de Letras - UFMG, 2019. 311p .

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, Vigo, p. 221-268, 2000.

GYNAN, S. Morphological Glossing Conventions for the Representation of Paraguayan Guaraní. *In*: ESTIGARRIBIA, B.; PINTA , J. (org.). **Guarani Linguistics in the 21st Century**. Leiden: Brill, 2017. p. 86-130.

HARRIS, Z. **String Analysis of Sentence Structure**. Haia: Mouton Publishers, 1962.

HEARST, M. TextTiling: segmenting text into multi-paragraph subtopic passages. *In*: **Computational Linguistics.** v. 22, p. 33-64, 1997.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **Censo Brasileiro de 2010.** Rio de Janeiro: IBGE, 2012. INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE).

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.

JURAFSKY, D. & MARTIN, H. J. Sequence Labeling for Parts of Speech and Named Entities. *In*: **Speech and Language Processing:** An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Stanford: Pearson, 2023a. Disponível em: https://web.stanford.edu/~jurafsky/slp3/ Acesso em: 01 maio 2023.

JURAFSKY, D. & MARTIN, H. J. Regular Expressions, Text Normalization, Edit Distance. *In*: **Speech and Language Processing:** An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Stanford: Pearson, 2023b. Disponível em: https://web.stanford.edu/~jurafsky/slp3/ Acesso em: 01 maio 2023.

KARTTUNEN, L. Comments on Joshi. *In*: A. Kornai, editor, **Extended Finite State Models of Language**. Proceedings of the ECAI 96 Workshop. Cambridge: Cambridge University Press, 1996, p. 24-25.

KOEHN, P. Europarl: A Parallel Corpus for Statistical Machine Translation. *In*: HUTCHINS, J. **The Tenth Machine Translation Summit Proceedings of Conference,** 2005, pp. 79-86.

KAY, M. Introduction. *In*: MITKOV, R. **The Oxford handbook of computational linguistics.** Oxford: Oxford University Press, 2004, p.xvii-xx.

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.

MAGER, M., GUTIERREZ-VASQUES, X., SIERRA, G., MEZA, I. Challenges of language technologies for the indigenous languages of the Americas. *In*: **Proceedings of the 27th International Conference on Computational Linguistics**, 2018, p. 55–69.

MARCUS, M. P.; SANTORINI, B. & MARCINKIEWICZ, M. A. **Building a large annotated corpus of English: The Penn Treebank.** *In*: Computational Linguistics, v. 19, n. 2, 1993.

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.

MOSELEY, C. **Atlas of the World's Languages in Danger**. 3. ed. Paris: UNESCO Publishing, 2010. Disponível em: https://unesdoc.unesco.org/ark:/48223/pf0000187026. Acesso em: 01 jun. 2023.

MÜLLER, S. **Grammatical theory:** From transformational grammar to constraint-based approaches. Berlim: Language Science Press, 2020.

NAVARRO, E. de A. A escravização dos índios num texto missionário em língua geral do século XVIII. **Revista USP,** v. 78, 2008, p. 105-114.

NAVARRO, E. de A. O corista europeu: tradução de um texto anônimo, em língua geral da Amazônia, do século XVIII. *In*: **Língua e Literatura**. São Paulo, FFLCH-USP, n. 27, 2009.

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. Narração que faz um sertanejo a um seu amigo de uma viagem que fez pelo sertão - Tradução de um texto anônimo, em língua geral amazônica, do século XVIII. *In*: **Revista USP**, v. 90, p. 181-192, 2011b.

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, 2016.

NAVARRO, E. de A.; AVILA, M. T. (org.). **Histórias em Língua Geral da Amazônia.** 1. ed. São Paulo: Centro Angel Rama da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, 2017. 112p.

NIVRE, J. *et al*. **Universal Dependencies v1:** A multilingual treebank collection. *In*: LREC. Portorož, SI: ELRA. 2016, p. 1659-1666.

OpenAI. 2023. GPT-4 Technical Report. ArXiv abs/2303.08774 (2023). Disponível em: https://arxiv.org/abs/2303.08774. Acesso em: 02 jun. 2023.

PEREIRA-NORIEGA, J. *et al*. Ship-LemmaTagger: Building an NLP Toolkit for a Peruvian Native Language. *In*: EKŠTEIN, K.; MATOUŠEK V. (org.) **Text, Speech, and Dialogue.** TSD 2017. Lecture Notes in Computer Science, v. 10415, 2017.

PYTHON Programming Language. Official Website. Disponível em: https://www.python.org/doc/. Acesso em: 20 de março de 2020.

ROBBINS, A.; BEEBE, N. H. F. **Classic Shell Scripting**. Sebastopol, CA: O'Reilly, 2005.

RODRIGUES, A. D. **Línguas Indígenas: 500 anos de descobertas e perdas**. D.E.L.T.A. 9.1:83-103. São Paulo, 1993.

RODRIGUES, A. D. **As línguas gerais sul-americanas**. Papia, São Paulo, v. 4, n. 2, p. 6-18, 1996.

SAG, I. A.; WASOW, T.; BENDER, E. M. **Syntactic theory:** A formal introduction. 2. ed. Stanford: CSLI, 2003.

SILVA, S. R.; VAZ FILHO, F. A. O Nheengatu no rio Tapajós: revitalização linguística e resistência política. *In*: SOUSA, Ivan Vale de. (org.). **A produção do conhecimento nas Letras, Linguística e Artes**. 1 ed. Ponta Grossa: Atena Editora, 2019, v. 3, p. 107-122.

SCHACHTER, P. Part-of-speech systems. *In*: SHOPEN, T. (org.). **Language Typology and Syntactic Description**, v. i. Cambridge: Cambridge University Press, 1985.

TREVISAN, Rodrigo Godinho. **Tradução comentada da obra Le Petit Prince, de Antoine de Saint-Exupéry, do francês ao nheengatu**. 2017. Dissertação (Mestrado) – Universidade de São Paulo, São Paulo, 2017. Disponível em: http://www.teses.usp.br/teses/disponiveis/8/8160/tde-07082017-124328/. Acesso em: 25 maio 2023.

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

APÊNDICE A – DETERMINANTES, INTERROGATIVOS E QUANTIFICADORES E SUA OCORRÊNCIA NAS ZONAS PREFIXAL (ZP) E NUCLEAR (ZN) DO SINTAGMA NOMINAL

Forma Lexical	Navarro (2016)	Cruz (2011)	ZP	ZN	Exemplos de Navarro (2016)	
amú	Pronome indefinido	Nome indicador de alteridade	SIM	SIM	ZP	*Kunhã usú **amú** piripanasawa ruka kití.* A mulher vai para **outra** loja.
					ZN	*Yepé uwapika igara gantĩme, **amú** uwapika yakumãme.* Um sentou na proa da canoa, **outro** sentou na popa.
awá	Pronome interrogativo Pronome indefinido	Nome genérico de humano	SIM	SIM	ZP	***Awá** ruka kití taá resú-putari?* *À casa **de quem** você quer ir?* (Magalhães, 64, adap.)
					ZN	***Awá** urikú turusú iwí umeẽ-kwáu **awá** nti urikú supé.* **Quem** tem muita terra pode dar para **quem** não tem.
bũa	Pronome quantificador Adjetivo de 1ª classe	Verbo estativo	SIM	NÃO	ZP	*Aikwé **bũa** timbiú penhẽ arama!* Há **muita** comida para vocês!
					ZN	
yepé yepé	Numeral Pronome quantificador	Indefinido	SIM	SIM	ZP	*Nhaã pituna suí, paá, **yepé yepé** kunhãmukú-itá usasá akití.* Contam que, desde aquela noite, as moças, **uma a uma**, passaram para lá.
					ZN	*Aikwé **yepé yepé** ukwáu waá upinaitika.* Há **alguns** que sabem pescar.
					ZP	*Aé umunuka **kwaíra** mirá.* Ele corta **poucas** árvores.

kwaíra	Pronome quantificador Adjetivo de 1ª classe	Nome autônomo	SIM	NÃO	ZN	***Kwaíra** ramé rẽ ixé, anheengari puranga.* *Quando eu (era) ainda **pequeno**, cantava bem.*
maã	Pronome indefinido Pronome interrogativo	Nome genérico (coisa)	SIM	SIM	ZP	***Maã** miapé taá rerikú?* **Qual** pão você tem?
					ZN	***Maã** i katú yambué-kwáu.* **O que** é bom devemos ensinar.
manungara ou *maãnungara*	Pronome indefinido	Nome autônomo	NÃO	SIM	ZP	
					ZN	***Manungara** amaã waá asú ambeú penhẽ arama.* **Algo** que vi vou contar para vocês.
muíri	Pronome indefinido Pronome interrogativo	Quantificador	SIM	SIM	ZP	***Muíri** kamixá taá Maria upiripana?* **Quantas** camisas Maria compra?
					ZN	***Muíri** rupí taá kwá kamixá?* Por **quanto** é esta camisa?
ne awá	Pronome indefinido	Negação contrastiva (Português) de nome genérico de humano	NÃO	SIM	ZP	
					ZN	***Ne awá** ukwáu awá umundá aintá ruka.* **Ninguém** sabe quem roubou a casa deles.
ne yepé	Pronome indefinido	Negação contrastiva (Português) de indefinido*	SIM	NÃO	ZP	***Ne yepé** apigawa uri uyenũ aé irũmu.* **Nenhum** homem veio deitar-se com ela.
					ZN	
ne maã	Pronome indefinido	Negação contrastiva (Português) de nome genérico	NÃO	SIM	ZP	
					ZN	***Ne maã** aintá urikú aintá umbaú arã.* **Nada** eles têm para comer (lit., para eles comerem).

nhaã	Pronome demonstrativo	Demonstrativo	SIM	NÃO	ZP	***Nhaã*** *xirura mukũi real yuíri.* **Aquela** calça é dois reais também.
					ZN	
nti awá	Pronome indefinido	Negação de nome genérico de humano	NÃO	SIM	ZP	
					ZN	***Nti awá*** *ukwáu awá nhaã mira-itá.* **Ninguém** sabia quem eram aquelas pessoas.
nti yepé	Pronome indefinido	Negação de indefinido	SIM	NÃO	ZP	*Mayawé* ***nti yepé*** *(...) apigawa usaisú nhaã kunhãmukú, aé nti upurasí.* Como **nenhum** homem amava aquela moça, ela não dançou.
					ZN	
nti maã	Pronome indefinido	Negação de nome genérico	NÃO	SIM	ZP	
					ZN	***Nti*** *yambeú-kwáu* ***maã*** *i xupé.* **Não** podemos contar-lhe **nada**.
pá	Pronome indefinido	NÃO CONSTA	SIM	SIM	ZP	*Ambaú* ***pá*** *pirá.* Comi todo o peixe.
					ZN	*Ambaú* ***pá****.* Comi tudo.
panhẽ	Pronome quantificador	Quantificador	SIM	NÃO	ZP	*Meẽsara upupeka* ***panhẽ*** *maã-itá.* O vendedor embrulha **todas** as coisas.
					ZN	
pawa	Pronome indefinido Verbo	Verbo intransitivo	SIM	SIM	ZP	*(...) yasí umuéu maã tatá u tatá umutipawa maã ií* ***pawa****.* (...) a lua apagaria o fogo ou o fogo secaria **toda** a água.

		dinâmico			ZN	*Kurumĩ umbaú **pawa**.* O menino come **tudo**.
siía	Pronome quantificador	Nome autônomo	SIM	NÃO	ZP	*Aé umunhã **siía** makira.* Ele faz **muitas** redes.
					ZN	
ti awá	Pronome indefinido	Negação de nome genérico de humano	NÃO	SIM	ZP	
					ZN	***Nti awá** ukwáu awá nhaã mira-itá.* **Ninguém** sabia quem eram aquelas pessoas.
waá	Pronome relativo	Relativizador				
upanhẽ	Pronome quantificador	Quantificador	NÃO	SIM	ZP	
					ZN	***Upanhẽ** urikú aintá piá surí.* **Todos** tinham seus corações felizes.
xinga	Pronome quantificador Advérbio intensificador	Atenuativo	NÃO	NÃO	Encontramos apenas ocorrências como advérbio: *Maria upurakí **xinga**.* Maria trabalha pouco.	

Fonte: Elaboração própria.

APÊNDICE B – *TAGSET* ORIGINAL DO NHEENGATAGGER

Etiqueta	Abreviatura no glossário[75]	Expansão da abreviatura
???	–	palavra desconhecida
A	adj.	adjetivo de 1ª cl.
A2	adj. 2ª cl.	adjetivo de 2ª cl.
ADP	posp.	posposição
ADV	adv.	advérbio
ADVA	adv. man.	advérbio de maneira
ADVC	adv. loc.	advérbio locativo
ADVD	adv. dem.	advérbio demonstrativo
ADVDI	adv. dem. dist.	advérbio demonstrativo distal
ADVDX	adv. dem. prox.	advérbio demonstrativo proximal
ADVG	adv. gr.	advérbio de grau
ADVJ	adv. conj.	advérbio conjuncional
ADVL	adv. rel.	advérbio relativo
ADVLA	adv. rel. man.	advérbio relativo de maneira
ADVLC	adv. rel. loc.	advérbio relativo locativo
ADVLT	adv. rel. temp.	advérbio relativo temporal
ADVM	adv. mod.	advérbio modal
ADVNC	adv. ind. loc.	advérbio indefinido locativo
ADVNT	adv. ind. temp.	advérbio indefinido temporal
ADVO	adv. ord.	advérbio ordinal
ADVP	adv. conj. opos.	advérbio conjuncional de oposição
ADVR	adv. interr.	advérbio interrogativo
ADVRA	adv. interr. man.	advérbio interrogativo de maneira
ADVRC	adv. interr. loc.	advérbio interrogativo locativo
ADVRT	adv. interr. temp.	advérbio interrogativo temporal
ADVRU	adv. interr. caus.	advérbio interrogativo causal
ADVS	adv. intensif.	advérbio de intensidade
ADVT	adv. temp.	advérbio temporal
AFF	part. afirm.	partícula de afirmação
ART	art. indef.	artigo indefinido
ASSUM	part. assum.	partícula de suposição
AUXFR	aux. flex. pré.	auxiliar flexionado pré-verbal
AUXFS	aux. flex. pós.	auxiliar flexionado pós-verbal

[75] Disponível em: https://github.com/CompLin/nheengatu/blob/main/data/glossary.json. Acesso em: 03 jun. 2023.

AUXN	aux. não flex.	auxiliar não flexionado
CARD	num. card.	numeral cardinal
CCONJ	cconj.	conjunção coordenativa
CERT	part. cert.	partícula de certeza
CLADP	posp. encl.	posposição enclítica
CLADV	adv. encl.	advérbio enclítico
COND	part. cond.	partícula de condicional
CONJ	conj.	conjunção
CONS	part. cons.	partícula de consentimento
COP	cop.	verbo de ligação
CQ	part. interr. cont.	partícula de pergunta de conteúdo
DEM	pron. dem.	pronome demonstrativo
DEMS	pron. dem. dist.	pronome demonstrativo distal
DEMSN	pron. dem. dist. não flex.	pronome demonstrativo distal não flexionado
DEMX	pron. dem. prox.	pronome demonstrativo proximal
EMP	pron. enf.	pronome de ênfase
EXST	part. exist.	partícula de existencial
FOC	part. foco	partícula de foco
FRUST	part. frust.	partícula de frustativo
FUT	part. fut.	partícula de futuro
IND	pron. indef.	pronome indefinido
INDQ	pron. quant.	pronome quantitativo indefinido
INT	pron. interr.	pronome interrogativo
INTJ	interj.	interjeição
N	s.	substantivo
NEC	part. neces.	partícula deôntica de necessidade
NEG	part. neg.	partícula de negação
NEGI	part. neg. imp.	partícula de imperativo negativo
ORD	num. ord.	numeral ordinal
PART	part.	partícula
PFV	part. perf.	partícula de perfectivo
PQ	part. interr. pol.	partícula de pergunta polar
PREF	pref.	prefixo
PREP	prep.	preposição
PRET	part. pret.	partícula de pretérito
PRON	pron.	pronome de 1ª classe
PRON2	pron. 2ª cl.	pronome de 2ª classe
PROPN	–	nome próprio

PROTST	part. prot.	partícula de protestivo
PRSV	part. pres.	partícula de presentativo
PUNCT	–	pontuação
REL	pron. relativo	pronome relativo
RELF	pron. rel. livre	pronome relativo livre
RPRT	part. report.	partícula de reportativo
SCONJ	sconj.	conjunção subordinativa
SCONJR	sconj. pré.	conjunção subordinativa pré-verbal
SUFF	suf.	sufixo
TOT	pron. quant. univ.	pronome quantitativo universal
TOTAL	part. tot.	partícula de totalitativo
V	v.	verbo de 1ª classe
V2	v. 2ª cl.	verbo de 2ª classe
V3	v. 3ª cl.	verbo de 3ª classe
VSUFF	v. suf.	verbo sufixal não flexionável

Fonte: Alencar (2020). Elaboração própria.

APÊNDICE C – *TAGSET* SIMPLIFICADO

Etiqueta	Abreviatura no glossário[76]	Expansão da abreviatura
???	–	palavra desconhecida
A	adj.	adjetivo de 1ª cl.
A2	adj. 2ª cl.	adjetivo de 2ª cl.
ADP	adp.	adposição
ADV	adv.	advérbio
AFF	part. afirm.	partícula de afirmação
ART	art. indef.	artigo indefinido
ASSUM	part. assum.	partícula de suposição
AUXFR	aux. flex. pré.	auxiliar flexionado pré-verbal
AUXFS	aux. flex. pós.	auxiliar flexionado pós-verbal
AUXN	aux. não flex.	auxiliar não flexionado
CARD	num. card.	numeral cardinal
CCONJ	cconj.	conjunção coordenativa
CERT	part. cert.	partícula de certeza
COND	part. cond.	partícula de condicional
CONJ	conj.	conjunção
CONS	part. cons.	partícula de consentimento
COP	cop.	verbo de ligação
CQ	part. interr. cont.	partícula de pergunta de conteúdo
DEM	pron. dem.	pronome demonstrativo
EMP	pron. enf.	pronome de ênfase
EXST	part. exist.	partícula de existencial
FOC	part. foco	partícula de foco
FRUST	part. frust.	partícula de frustrativo
FUT	part. fut.	partícula de futuro
IND	pron. indef.	pronome indefinido
INDQ	pron. quant.	pronome quantitativo indefinido
INT	pron. interr.	pronome interrogativo
INTJ	interj.	interjeição
N	s.	substantivo
NEC	part. neces.	partícula deôntica de necessidade
NEG	part. neg.	partícula de negação
NEGI	part. neg. imp.	partícula de imperativo negativo

[76] Disponível em: https://github.com/CompLin/nheengatu/blob/main/data/glossary.json. Acesso em: 03 jun. 2023.

ORD	num. ord.	numeral ordinal
PART	part.	partícula
PFV	part. perf.	partícula de perfectivo
PQ	part. interr. pol.	partícula de pergunta polar
PREF	pref.	prefixo
PRET	part. pret.	partícula de pretérito
PRON	pron.	pronome de 1ª classe
PRON2	pron. 2ª cl.	pronome de 2ª classe
PROTST	part. prot.	partícula de protestivo
PRSV	part. pres.	partícula de presentativo
PUNCT	-	pontuação
REL	pron. relativo	pronome relativo
RELF	pron. rel. livre	pronome relativo livre
RPRT	part. report.	partícula de reportativo
SCONJ	sconj.	conjunção subordinativa
SCONJR	sconj. pré.	conjunção subordinativa pré-verbal
SUFF	suf.	sufixo
TOT	pron. quant. univ.	pronome quantitativo universal
TOTAL	part. tot.	partícula de totalitativo
V	v.	verbo de 1ª classe
V2	v. 2ª cl.	verbo de 2ª classe
V3	v. 3ª cl.	verbo de 3ª classe
VSUFF	-	verbo sufixal não flexionável

Fonte: Alencar (2020). Elaboração própria.

APÊNDICE D – LISTA DE AMBIGUIDADES EXTRAÍDAS DE NAVARRO (2016) (*TAGSET* ORIGINAL)

N.	Ambiguidade	Classes	Frequência	(%)	Sentenças	(%)
1	A+A2	Adjetivo de 1ª classe + adjetivo de 2ª classe	3	0.29%	3	0,29%
2	A+A2+ADV	Adjetivo de 1ª classe + adjetivo de 2ª classe + advérbio	2	0.19%	2	0,19%
3	A+ADV+N	Adjetivo de 1ª classe + advérbio + substantivo	12	1.15%	12	1,15%
4	A+ADVA	Adjetivo de 1ª classe + advérbio de maneira	63	6.02%	63	6,05%
5	A+ADVA+N	Adjetivo de 1ª classe + advérbio de maneira + substantivo	9	0.86%	9	0,86%
6	A+ADVC	Adjetivo de 1ª classe + advérbio locativo	4	0.38%	4	0,38%
7	A+ADVC+N	Adjetivo de 1ª classe + advérbio locativo + substantivo	1	0.10%	1	0,10%
8	A+ADVS	Adjetivo de 1ª classe + advérbio de intensidade	3	0.29%	3	0,29%
9	A+ADVS+INDQ	Adjetivo de 1ª classe + advérbio de intensidade + pronome quantitativo indefinido	9	0.86%	9	0,86%
10	A+INDQ	Adjetivo de 1ª classe + pronome quantitativo indefinido	3	0.29%	3	0,29%
11	A+INTJ+N	Adjetivo de 1ª classe + interjeição + substantivo	1	0.10%	1	0,10%
12	A+N	Adjetivo de 1ª classe + substantivo	74	7.07%	64	6,15%
13	A+N+V2	Adjetivo de 1ª classe + substantivo + verbo de 2ª classe	2	0.19%	2	0,19%
14	A+PRET	Adjetivo de 1ª classe + partícula de pretérito	10	0.96%	10	0,96%
15	A+V	Adjetivo de 1ª classe + verbo de 1ª classe	3	0.29%	3	0,29%
16	A+V2	Adjetivo de 1ª classe + verbo de 2ª classe	1	0.10%	1	0,10%
17	A2+ADVS	Adjetivo de 2ª classe + advérbio de intensidade	6	0.57%	6	0,58%
18	A2+N	Adjetivo de 2ª classe + substantivo	7	0.67%	7	0,67%
19	ADP+ADVA+IND	Posposição + advérbio de maneira + pronome indefinido	12	1.15%	12	1,15%
20	ADP+ADVG	Posposição + advérbio de grau	15	1.43%	15	1,44%
21	ADP+FUT+SCONJ	Posposição + partícula de futuro + conjunção subordinativa pós-verbal	73	6.97%	66	6,34%
22	ADP+N	Posposição + substantivo	77	7.35%	71	6,82%
23	ADP+N+SCONJ	Posposição + substantivo + conjunção subordinativa pós-verbal	14	1.34%	14	1,34%

24	ADP+SCONJ	Posposição + conjunção subordinativa pós-verbal	98	9.36%	92	8,84%
25	ADV+CCONJ+V	Advérbio + conjunção coordenativa + verbo de 1ª classe	10	0.96%	10	0,96%
26	ADVDI+ADVJ	Advérbio demonstrativo distal + advérbio conjuncional	18	1.72%	18	1,73%
27	ADVDX+DEMX+V	Advérbio demonstrativo proximal + pronome demonstrativo proximal + verbo de 1ª classe	32	3.06%	32	3,07%
28	ADVJ+ADVT	Advérbio conjuncional + advérbio temporal	2	0.19%	2	0,19%
29	ADVJ+CCONJ	Advérbio conjuncional + conjunção coordenativa	8	0.76%	8	0,77%
30	ADVJ+SCONJR	Advérbio conjuncional + conjunção subordinativa pré-verbal	3	0.29%	3	0,29%
31	ADVLA+ADVRA+SCONJR	Advérbio relativo de maneira +advérbio interrogativo de maneira + conjunção subordinativa pré-verbal	14	1.34%	14	1,34%
32	ADVLC+ADVNC+ADVRC	Advérbio relativo locativo + advérbio indefinido locativo + advérbio interrogativo locativo	8	0.76%	8	0,77%
33	ADVLC+ADVRC	Advérbio relativo locativo + advérbio interrogativo locativo	14	1.34%	14	1,34%
34	ADVM+ADVT	Advérbio modal + advérbio temporal	1	0.10%	1	0,10%
35	ADVNT+ADVRT+SCONJR	Advérbio indefinido temporal + advérbio interrogativo temporal + conjunção subordinativa pré-verbal	13	1.24%	13	1,25%
36	ADVO+ORD	Advérbio ordinal + numeral ordinal	1	0.10%	1	0,10%
37	ADVS+INDQ	Advérbio de intensidade + pronome quantitativo indefinido	8	0.76%	8	0,77%
38	ADVT+CCONJ	Advérbio temporal + conjunção coordenativa	19	1.81%	19	1,83%
39	ART+CARD+FRUST+SCONJ	Artigo indefinido + numeral cardinal + partícula de frustrativo + conjunção subordinativa pós-verbal	66	6.30%	57	5,48%
40	CARD+INDQ	Numeral cardinal + pronome quantitativo indefinido	14	1.34%	14	1,34%
41	CERT+N	Partícula de certeza + substantivo	1	0.10%	1	0,10%
42	CLADP+PRON2	Posposição enclítica + pronome de 2ª classe	8	0.76%	8	0,77%
43	COND+IND+INT+N+REL+V	Partícula de condicional + pronome indefinido + pronome interrogativo + substantivo + pronome relativo + verbo de 1ª classe	26	2.48%	23	2,21%
44	DEMSN+PRON	Pronome demonstrativo distal não flexionado + pronome de 1ª classe	119	11.37%	110	10,57%
45	FOC+NEGI	Partícula de foco + partícula de imperativo negativo	6	0.57%	6	0,58%
46	FOC+PREP+SCONJR	Partícula de foco + preposição + conjunção subordinativa pré-verbal	21	2.01%	19	1,83%

47	IND+INT+RELF	Pronome indefinido + pronome interrogativo + pronome relativo livre	23	2.20%	19	1,83%
48	IND+NEG	Pronome indefinido + partícula de negação	1	0.10%	1	0,10%
49	INDQ+INT+TOT	Pronome quantitativo indefinido + pronome interrogativo + pronome quantitativo universal	6	0.57%	6	0,58%
50	N+REL	Substantivo + pronome relativo	6	0.57%	6	0,58%
51	N+V	Substantivo + verbo de 1ª classe	19	1.81%	19	1,83%
52	N+V2	Substantivo + verbo de 2ª classe	5	0.48%	5	0,48%
53	PRON+PRON2	Pronome de 1ª classe + pronome de 2ª classe	65	6.21%	52	5,00%
54	TOT+TOTAL+V	Pronome quantitativo universal + partícula de totalitativo + verbo de 1ª classe	2	0.19%	2	0,19%
55	V+V2	Verbo de 1ª classe + verbo de 2ª classe	6	0.57%	6	0,58%
		Total	**1047**	**100%**	**972**	**100,00%**

Fonte: Elaboração própria.

APÊNDICE E – LISTA DE AMBIGUIDADES EXTRAÍDAS DE NAVARRO (2016) (*TAGSET* SIMPLIFICADO)

N.	Ambiguidade	Classes	Frequência	(%)
1	A+A2	Adjetivo de 1ª classe + adjetivo de 2ª classe	3	0.30%
2	A+A2+ADV	Adjetivo de 1ª classe + adjetivo de 2ª classe + advérbio	2	0.20%
3	A+ADV	Adjetivo de 1ª classe + advérbio	70	6.97%
4	A+ADV+INDQ	Adjetivo de 1ª classe + advérbio + pronome quantitativo indefinido	9	0.90%
5	A+ADV+N	Adjetivo de 1ª classe + advérbio + substantivo	22	2.19%
6	A+INDQ	Adjetivo de 1ª classe + pronome quantitativo indefinido	3	0.30%
7	A+INTJ+N	Adjetivo de 1ª classe + interjeição + substantivo	1	0.10%
8	A+N	Adjetivo de 1ª classe + substantivo	74	7.37%
9	A+N+V2	Adjetivo de 1ª classe + substantivo + verbo de 2ª classe	2	0.20%
10	A+PRET	Adjetivo de 1ª classe + partícula de pretérito	10	1.00%
11	A+V	Adjetivo de 1ª classe + verbo de 1ª classe	3	0.30%
12	A+V2	Adjetivo de 1ª classe + verbo de 2ª classe	1	0.10%
13	A2+ADV	Adjetivo de 2ª classe + advérbio	6	0.60%
14	A2+N	Adjetivo de 2ª classe + substantivo	7	0.70%
15	ADP+ADV	Adposição + advérbio	15	1.49%
16	ADP+ADV+IND	Adposição + advérbio + pronome indefinido	12	1.20%
17	ADP+FUT+SCONJ	Adposição + partícula de futuro + conjunção subordinativa pós-verbal	73	7.27%
18	ADP+N	Adposição + substantivo	77	7.67%
19	ADP+N+SCONJ	Adposição + substantivo + conjunção subordinativa pós-verbal	14	1.39%
20	ADP+PRON2	Adposição + pronome de 2ª classe	8	0.80%
21	ADP+SCONJ	Adposição + conjunção subordinativa pós-verbal	98	9.76%
22	ADV+CCONJ	Advérbio + conjunção coordenativa	27	2.69%
23	ADV+CCONJ+V	Advérbio + conjunção coordenativa + verbo de 1ª classe	10	1.00%

24	ADV+DEM+V	Advérbio + pronome demonstrativo + verbo de 1ª classe	32	3.19%
25	ADV+INDQ	Advérbio + pronome quantitativo indefinido	8	0.80%
26	ADV+ORD	Advérbio + numeral ordinal	1	0.10%
27	ADV+SCONJR	Advérbio + conjunção subordinativa pré-verbal	30	2.99%
28	ART+CARD+FRUST+SCONJ	Artigo indefinido + numeral cardinal + partícula de frustrativo + conjunção subordinativa pós-verbal	66	6.57%
29	CARD+INDQ	Numeral cardinal + pronome quantitativo indefinido	14	1.39%
30	CERT+N	Partícula de certeza + substantivo	1	0.10%
31	COND+IND+INT+N+REL+V	Partícula de condicional + pronome indefinido + pronome interrogativo + substantivo + pronome relativo + verbo de 1ª classe	26	2.59%
32	DEM+PRON	Pronome demonstrativo + pronome de 1ª classe	119	11.85%
33	FOC+ADP+SCONJR	Partícula de foco + adposição + conjunção subordinativa pré-verbal	21	2.09%
34	FOC+NEGI	Partícula de foco + partícula de imperativo negativo	6	0.60%
35	IND+INT+RELF	Pronome indefinido + pronome interrogativo + pronome relativo livre	23	2.29%
36	IND+NEG	Pronome indefinido + partícula de negação	1	0.10%
37	INDQ+INT+TOT	Pronome quantitativo indefinido + pronome interrogativo + pronome quantitativo universal	6	0.60%
38	N+REL	Substantivo + pronome relativo	6	0.60%
39	N+V	Substantivo + verbo de 1ª classe	19	1.89%
40	N+V2	Substantivo + verbo de 2ª classe	5	0.50%
41	PRON+PRON2	Pronome de 1ª classe + pronome de 2ª classe	65	6.47%
42	TOT+TOTAL+V	Pronome quantitativo universal + partícula de totalitativo + verbo de 1ª classe	2	0.20%
43	V+V2	Verbo de 1ª classe + verbo de 2ª classe	6	0.60%
		Total	**1004**	**100%**

Fonte: Elaboração própria.

APÊNDICE F – CONTEXTOS DA AMBIGUIDADE ADP+SCONJ (POSPOSIÇÃO + CONJUNÇÃO SUBORDINATIVA)[77]

Etiqueta anterior	Classe da etiqueta anterior	Etiqueta posterior	Classe da etiqueta posterior
PROPN	Nome próprio	PUNCT	Pontuação
IND	Pronome indefinido	ADV	Advérbio
N	Substantivo	ADVJ	Advérbio conjuncional
PROPN	Nome próprio	N	Substantivo
A+N	Adjetivo de 1ª classe + substantivo	PUNCT	Pontuação
A+V	Adjetivo de 1ª classe + verbo de 1ª classe	PUNCT	Pontuação
IND+INT	Pronome indefinido + pronome interrogativo	PUNCT	Pontuação
N	Substantivo	PUNCT	Pontuação
PRON2	Pronome de 2ª classe	PUNCT	Pontuação
N	Substantivo	V	Verbo de 1ª classe
PRON2	Pronome de 2ª classe	V	Verbo de 1ª classe
A	Adjetivo de 1ª classe	ADV	Advérbio
A+N	Adjetivo de 1ª classe + substantivo	IND+INT+N+V	Pronome indefinido + pronome interrogativo + substantivo + verbo de 1ª classe
N	Substantivo	IND+INT+N+V	Pronome indefinido + pronome interrogativo + substantivo + verbo de 1ª classe
N	Substantivo	PFV	Partícula de perfectivo
N	Substantivo	PROPN	Nome próprio
N	Substantivo	PUNCT	Pontuação
N	Substantivo	V	Verbo de 1ª classe
NEG	Partícula de negação	IND	Pronome indefinido
NEG	Partícula de negação	V	Verbo de 1ª classe
PRON	Pronome de 1ª classe	PUNCT	Pontuação
V	Verbo de 1ª classe	ADV+ADVR+SCONJ	Advérbio + advérbio interrogativo + conjunção subordinativa
V	Verbo de 1ª classe	ADVD+DEMX+V	Advérbio demonstrativo + pronome demonstrativo proximal + verbo de 1ª classe
V	Verbo de 1ª classe	ART+CARD+FRUST+ SCONJ	Artigo indefinido + numeral cardinal + partícula de frustrativo + conjunção subordinativa
V	Verbo de 1ª classe	IND+INT+N+V	Pronome indefinido + pronome interrogativo + substantivo + verbo de 1ª classe

[77] Extraídos de Navarro (2016).

V	Verbo de 1ª classe	N	Substantivo
V	Verbo de 1ª classe	PRON	Pronome de 1ª classe
V	Verbo de 1ª classe	PUNCT	Pontuação
IND	Pronome indefinido	PUNCT	Pontuação
N	Substantivo	PUNCT	Pontuação
PRON2	Pronome de 2ª classe	PUNCT	Pontuação
PROPN	Nome próprio	PUNCT	Pontuação
SCONJ	Conjunção subordinativa	V	Verbo de 1ª classe
V	Verbo de 1ª classe	PUNCT	Pontuação
N	Substantivo	N	Substantivo
DEMS	Pronome demonstrativo distal	IND+INT	Pronome indefinido + pronome interrogativo

Fonte: Elaboração própria.

APÊNDICE G – CONTEXTOS DA ETIQUETA ADP (POSPOSIÇÃO)[78]

Etiqueta anterior	Classe da etiqueta anterior	Etiqueta posterior	Classe da etiqueta posterior
???	Palavra desconhecida	???	Palavra desconhecida
N	Substantivo	ADP	Posposição
None	Etiqueta atual no início da sentença	ADP	Posposição
PRON2	Pronome de 2ª classe	ADP	Posposição
N	Substantivo	ADV	Advérbio
N	Substantivo	ADV+CCONJ+V	Advérbio + conjunção coordenativa + verbo de 1ª classe
N	Substantivo	ADVL+ADVR	Advérbio relativo + advérbio interrogativo
ADP	Posposição	ADVS+INDQ	Advérbio de intensidade + pronome quantitativo indefinido
PROPN	Nome próprio	ART+CARD+FRUST+ SCONJ	Artigo indefinido + numeral cardinal + partícula de frustrativo + conjunção subordinativa
PROPN	Nome próprio	CARD	Numeral cardinal
IND+INT	Pronome indefinido + pronome interrogativo	CQ	Partícula de pergunta de conteúdo
IND+INT	Pronome indefinido + pronome interrogativo	DEMS	Pronome demonstrativo distal
ADVD+DEMX+V	Advérbio demonstrativo + pronome demonstrativo proximal + verbo de 1ª classe	EXST	Partícula de existencial
N	Substantivo	EXST	Partícula de existencial
N	Substantivo	IND+INT+N+V	Pronome indefinido + pronome interrogativo + substantivo + verbo de 1ª classe
PRON2	Pronome de 2ª classe	IND+INT+N+V	Pronome indefinido + pronome interrogativo + substantivo + verbo de 1ª classe
N	Substantivo	N	Substantivo
PROPN	Nome próprio	N	Substantivo
V	Verbo de 1ª classe	N	Substantivo
PRON+PRON2	Pronome de 1ª classe + pronome de 2ª classe	NEG	Partícula de negação
N	Substantivo	None	Etiqueta atual no final da sentença
N	Substantivo	PFV	Partícula de perfectivo
None	Etiqueta atual no início da sentença	PRON	Pronome de 1ª classe
N	Substantivo	PRON2	Pronome de 2ª classe

[78] Extraídos de Navarro (2016).

A+ADV	Adjetivo de 1ª classe + advérbio	PROPN	Nome próprio
N	Substantivo	PROPN	Nome próprio
???	Palavra desconhecida	PUNCT	Pontuação
A+ADV	Adjetivo de 1ª classe + advérbio	PUNCT	Pontuação
A+N	Adjetivo de 1ª classe + substantivo	PUNCT	Pontuação
ADP	Posposição	PUNCT	Pontuação
ADP+N	Posposição + substantivo	PUNCT	Pontuação
ADV	Advérbio	PUNCT	Pontuação
ADVD+DEMX+V	Advérbio demonstrativo + pronome demonstrativo proximal + verbo de 1ª classe	PUNCT	Pontuação
DEMS	Pronome demonstrativo distal	PUNCT	Pontuação
N	Substantivo	PUNCT	Pontuação
PRON+PRON2	Pronome de 1ª classe + pronome de 2ª classe	PUNCT	Pontuação
PRON2	Pronome de 2ª classe	PUNCT	Pontuação
PROPN	Nome próprio	PUNCT	Pontuação
V	Verbo de 1ª classe	PUNCT	Pontuação
N	Substantivo	RPRT	Partícula de reportativo
N	Substantivo	V	Verbo de 1ª classe
PRON+PRON2	Pronome de 1ª classe + pronome de 2ª classe	V	Verbo de 1ª classe
PRON2	Pronome de 2ª classe	V	Verbo de 1ª classe

Fonte: Elaboração própria.

APÊNDICE H – TABELA DE CONTEXTO DAS ETIQUETAS A E ADV GERADA A PARTIR DE NAVARRO (2016)

1	Etiqueta	Etiqueta do contexto	Tipo de contexto	Frequência
2	A	ADVR	Anterior	1
3	A	FUT	Anterior	1
4	A	N	Anterior	15
5	A	None	Anterior	5
6	A	PFV	Anterior	1
7	A	PRON	Anterior	3
8	A	PROPN	Anterior	2
9	A	V	Anterior	7
10	A	ADV	Posterior	1
11	A	ADVS	Posterior	2
12	A	DEMS	Posterior	1
13	A	N	Posterior	8
14	A	None	Posterior	1
15	A	PUNCT	Posterior	15
16	A	V	Posterior	3
17	ADV	A	Anterior	1
18	ADV	ADP	Anterior	4
19	ADV	ADV	Anterior	8
20	ADV	ADVD	Anterior	2
21	ADV	ADVR	Anterior	1
22	ADV	CCONJ	Anterior	1
23	ADV	CQ	Anterior	1
24	ADV	DEMS	Anterior	1
25	ADV	EXST	Anterior	1
26	ADV	FUT	Anterior	2
27	ADV	IND	Anterior	1
28	ADV	N	Anterior	9
29	ADV	NEG	Anterior	6
30	ADV	None	Anterior	36
31	ADV	PFV	Anterior	2
32	ADV	PRON	Anterior	1
33	ADV	PUNCT	Anterior	3

34	ADV	SCONJ	Anterior	1
35	ADV	V	Anterior	19
36	ADV	ADP	Posterior	2
3	ADV	ADV	Posterior	8
38	ADV	ADVD	Posterior	1
39	ADV	CQ	Posterior	1
40	ADV	DEMS	Posterior	1
41	ADV	FUT	Posterior	6
42	ADV	IND	Posterior	2
43	ADV	N	Posterior	9
44	ADV	None	Posterior	1
45	ADV	PQ	Posterior	1
46	ADV	PRON	Posterior	4
47	ADV	PRON2	Posterior	2
48	ADV	PROPN	Posterior	1
49	ADV	PUNCT	Posterior	41
50	ADV	V	Posterior	23

Fonte: Elaboração própria.

APÊNDICE I – CONJUNTO DE SENTENÇAS DOS TESTES PRELIMINARES

Ambiguidades	Conjuntos de sentenças
A+ADV	Aé/PRON **puranga/A+ADV** ./PUNCT
A+ADV	timbiú/N **sé/A+ADV**
A+ADV	ara/N **puranga/A+ADV**
A+ADV	Aé/PRON **puranga/A+ADV** uikú/V ./PUNCT
A+ADV	Maria/PROPN **puranga/A+ADV** ./PUNCT
A+ADV	Reyenũ/V **puranga/A+ADV** !/PUNCT
A+ADV	Kurumĩ/N **puranga/A+ADV** ./PUNCT
A+ADV	Apigawa/A+N **puranga/A+ADV** ./PUNCT
A+ADV	Ixé/PRON **puranga/A+ADV** ./PUNCT
A+ADV	Repuká/V **sé/A+ADV** ./PUNCT
ADP+FUT+SCONJ	Reruri/V timbiú/N ixé/PRON **arama/ADP+FUT+SCONJ** ./PUNCT
ADP+FUT+SCONJ	Mairamé/ADVR+SCONJ kurí/FUT bũa/A+INDQ ,/PUNCT ixé/PRON asú/V kurí/FUT ayuká/V indé/PRON **arama/ADP+FUT+SCONJ** kwá/ADVD+DEMX+V tukunaré/N ./PUNCT
ADP+FUT+SCONJ	Amunhã/V pindá-itá/N indé/PRON **arama/ADP+FUT+SCONJ** ./PUNCT
ADP+FUT+SCONJ	Apurakí/V indé/PRON **arama/ADP+FUT+SCONJ** ./PUNCT
ADP+FUT+SCONJ	Reruri/V aé/PRON ixé/PRON **arama/ADP+FUT+SCONJ** ./PUNCT
ADP+FUT+SCONJ	Remunhã/V ne/PRON2 kakurí/N nti/NEG **arã/ADP+FUT+SCONJ** rewatá/V remundá/V se/PRON2 kakurí/N ./PUNCT
ADP+FUT+SCONJ	Asarú/V remanũ/??? ambaú/V **arã/ADP+FUT+SCONJ** indé/PRON ./PUNCT
ADP+FUT+SCONJ	Se/PRON2 manha/N unheẽ/V kwera/A+PRET ixé/PRON **arama/ADP+FUT+SCONJ** :/PUNCT
ADP+FUT+SCONJ	Apitá/V ne/PRON2 ruka/N upé/ADP apurungitá/V **arama/ADP+FUT+SCONJ** ne/PRON2 irũmu/ADP+SCONJ ./PUNCT
ADP+FUT+SCONJ	Xukúi/V pirá/N indé/PRON rembaú/V **arã/ADP+FUT+SCONJ** ./PUNCT
ADP+SCONJ	Pedro/PROPN uikú/V Maria/PROPN **irũmu/ADP+SCONJ** ./PUNCT
ADP+SCONJ	Aé/PRON usú/V se/PRON2 **irũmu/ADP+SCONJ** ./PUNCT
ADP+SCONJ	Kwaíra/A **ramé/ADP+SCONJ** rẽ/ADV ,/PUNCT paá/RPRT ,/PUNCT ixé/PRON ,/PUNCT purangamirĩ/A+ADV aikú/V ./PUNCT
ADP+SCONJ	Kwaíra/A **ramé/ADP+SCONJ** rẽ/ADV ixé/PRON ,/PUNCT anheengari/V puranga/A+ADV ./PUNCT
ADP+SCONJ	Mairamé/ADVR+SCONJ arasú/V timbiú/N ,/PUNCT ixé/PRON nti/NEG apitá/V yumasisawa/N **irũmu/ADP+SCONJ** ./PUNCT
ADP+SCONJ	Aputari/V apurungitá/V ne/PRON2 **irũmu/ADP+SCONJ** ./PUNCT
ADP+SCONJ	Asú/V murakipí/N **ramé/ADP+SCONJ** ./PUNCT

ADP+SCONJ	Asú/V **ramé/ADP+SCONJ** ,/PUNCT arasú/V se/PRON2 mimbira/N ./PUNCT
ADP+SCONJ	Nti/NEG **ramé/ADP+SCONJ** repurakí/V ,/PUNCT indé/PRON repitá/V pirasua/A+N ./PUNCT
ADP+SCONJ	Apitá/V ne/PRON2 ruka/N upé/ADP apurungitá/V arama/ADP+FUT+SCONJ ne/PRON2 **irũmu/ADP+SCONJ** ./PUNCT
A+N	kunhã/**A+N** puranga/A+ADV
A+N	apigawa/**A+N** kirimbawa/**A+N**
A+N	Reyapukúi/V kirimbawa/**A+N** ./PUNCT[79]
A+N	Aé/PRON apigawa/**A+N** ./PUNCT
A+N	Apigawa/**A+N** puranga/A+ADV ./PUNCT
A+N	Indé/PRON kunhã/**A+N** ./PUNCT
A+N	yepé/ART+CARD+FRUST+SCONJ kunhã/**A+N**
A+N	yepé/ART+CARD+FRUST+SCONJ apigawa/**A+N**
A+N	Mayé/ADV+ADVR+SCONJ waá/REL kunhã/**A+N** taá/CQ usika/V ana/PFV ?/PUNCT
A+N	mukũi/CARD+INDQ apigawa/**A+N**
PRON+PRON2	Aintá/**PRON+PRON2** usú/V nhaã/DEMS tatatingawasú/N piterarupí/ADP …/PUNCT
PRON+PRON2	aintá/**PRON+PRON2** xirura-itá/N
PRON+PRON2	Anheẽ/V “/PUNCT puranga/A+ADV ara/N ”/PUNCT aintá/**PRON+PRON2** supé/ADP ./PUNCT
PRON+PRON2	Aintá/**PRON+PRON2** umuyeréu/V mirá-itawasú/N aintá/**PRON+PRON2** ararupí/ADP ./PUNCT
PRON+PRON2	aintá/**PRON+PRON2** rendawa/N
PRON+PRON2	Aintá/**PRON+PRON2** ruka/N puranga/A+ADV ./PUNCT
PRON+PRON2	aintá/**PRON+PRON2** uri/V ./PUNCT
PRON+PRON2	Aintá/**PRON+PRON2** rakú/A2 ./PUNCT
PRON+PRON2	Amú/IND ara/N ramé/ADP+SCONJ ana/PFV ,/PUNCT paá/RPRT ,/PUNCT aintá/**PRON+PRON2** upurandú/V :/PUNCT
PRON+PRON2	Aintá/**PRON+PRON2** upuká/V yané/PRON2 resé/ADP+SCONJ ./PUNCT

[79] Nesta sentença há um erro de etiquetagem que está presente na versão do corpus etiquetado utilizado nos testes preliminares. No lugar de A+N, a ambiguidade é A+ADV+N. O erro já foi corrigido na versão mais recente do glossário, disponível em: https://github.com/CompLin/nheengatu/blob/main/data/glossary.json. Acesso em: 11 jun. 2023.

APÊNDICE J – CONJUNTO DE SENTENÇAS DOS TESTES 1, 2 E 3

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
Testset 1 e Testset 2	**A2+ADVS**		
	A2+ADVS	t-yrl-por-eng-05-pos.txt:	Akwáu/V katú/**A2+ADVS** ixé/PRON nti/NEG ayupukwáu/V akití/ADVDI ./PUNCT # Sei bem que eu não me acostumo por ali.
	A2+ADVS	e-yrl-por-eng-01-pos.txt:	Mimi/ADVDI katú/**A2+ADVS** aé/DEMSN+PRON ./PUNCT # É bem ali.
	A2+ADVS	e-yrl-por-eng-11-pos.txt:	Awá/IND+INT+RELF usendú/V aintá/PRON+PRON2 upurungitá/V uruyari/V katú/**A2+ADVS** Piriripi/PROPN uikú/V ana/PFV aintá/PRON+PRON2 pu/ASSUM resé/ADP+SCONJ ./PUNCT # Quem as ouvia falar acreditava bem que Piriripi estava nas mãos deles.
	A2+ADVS	y-yrl-por-eng-08-pos.txt:	Remaã/V katú/**A2+ADVS** ,/PUNCT se/PRON2 pindá/N umutianha/V ne/PRON2 nambí/N ,/PUNCT # Olhe bem, (cuidado que) meu anzol fisga tua orelha,
	A2+ADVS	e-yrl-por-eng-13-pos.txt:	Maã/COND+IND+INT+N+REL+V i/PRON2 katú/**A2+ADVS** yambué-kwáu/??? ./PUNCT

[80] Disponíveis em: https://github.com/juliana-gurgel/nheengatu/tree/main/corpus/navarro-2016. Acesso em: 14 ago. 2023.

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
			# O que é bom devemos ensinar.
Testset 1 e Testset 2	**A2+N**		
	A2+N	t-yrl-por-eng-08-pos.txt:	Asuí/ADVT+CCONJ ,/PUNCT paá/RPRT ,/PUNCT wirawasú/N unheẽ/V :/PUNCT —/PUNCT Compadre/PROPN ,/PUNCT remunhã/V mayé/ADVLA+ADVRA+SCONJR se/PRON2 yawé/ADP+ADVA+IND :/PUNCT repisika/V suú/N+V sikwé**/A2+N** ,/PUNCT reyuká/V aé/DEMSN+PRON rembaú/V arã/ADP+FUT+SCONJ aé/DEMSN+PRON ./PUNCT # Depois, contam que o gavião disse: — Compadre, faça assim como eu: pegue animal vivo, mate-o para comê-lo.
	A2+N	t-yrl-por-eng-08-pos.txt:	Ape/ADVDI+ADVJ ,/PUNCT paá/RPRT ,/PUNCT urubú/N usuaxara/V :/PUNCT —/PUNCT Compadre/PROPN ,/PUNCT ixé/PRON nti/NEG aputari/V amunhã/V puxiwera/A+ADVA+N suú-itá/N supé/ADP ,/PUNCT ma/ADVJ+CCONJ kuíri/ADVT ixé/PRON se/PRON2 yumasí**/A2+N** retana/ADVS aikú/V ./PUNCT # Então, dizem que o urubu respondeu: — Compadre, eu não quero fazer mal aos animais, mas agora eu estou muito faminto.
	A2+N	t-yrl-por-eng-06-pos.txt:	Ixé/PRON nti/NEG se/PRON2 yumasí**/A2+N** ./PUNCT # Eu não estou faminto.
	A2+N	e-yrl-por-eng-05-pos.txt:	Se/PRON2 rikwé**/A2+N** aikú/V ./PUNCT # Eu estou vivo.

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
	A2+N	e-yrl-por-eng-05-pos.txt:	Kwá/ADVDX+DEMX+V wirá/N sikwé/**A2+N** uikú/V ./PUNCT # Este pássaro está vivo.
Testset 1 e Testset 2	**A+A2**		
	A+A2	e-yrl-por-eng-01-pos.txt:	Se/PRON2 pusé/**A+A2** ./PUNCT # Eu sou pesado.
	A+A2	e-yrl-por-eng-01-pos.txt:	Igara/N i/PRON2 pusé/**A+A2** ./PUNCT # A canoa é pesada.
	A+A2	e-yrl-por-eng-01-pos.txt:	Yandé/PRON yané/PRON2 pusé/**A+A2** ./PUNCT # Nós somos pesados.
Testset 1 e Testset 2	**A+A2+ADV**		
	A+A2+ADV	e-yrl-por-eng-05-pos.txt:	Uka/N sakú/**A+A2+ADV** ./PUNCT # A casa é quente.
	A+A2+ADV	e-yrl-por-eng-05-pos.txt:	Aé/DEMSN+PRON sakú/**A+A2+ADV** ./PUNCT # Ela é quente.
Testset 1 e Testset 2	**A+ADVA**		

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
	A+ADVA	e-yrl-por-eng-01-pos.txt:	Puranga/**A+ADVA** ara/N !/PUNCT # Bom dia!
	A+ADVA	e-yrl-por-eng-11-pos.txt:	Kunhãmukú-itá/N yepé/ART+CARD+FRUST+SCONJ aintá/PRON+PRON2 suí/ADP usú/V uikú/V uyumana/V aé/DEMSN+PRON ,/PUNCT amú-itá/IND upurungitá/V puranga/**A+ADVA** ./PUNCT # As moças, algumas delas estavam indo abraçá-lo, outras falavam bonito.
	A+ADVA	e-yrl-por-eng-10-pos.txt:	Kwá/ADVDX+DEMX+V kunhã/A+N unheengari/V puranga/**A+ADVA** mayé/ADVLA+ADVRA+SCONJR nhaã/DEMS yawé/ADP+ADVA+IND ./PUNCT # Esta mulher canta tão bem como aquela.
	A+ADVA	t-yrl-por-eng-02-pos.txt:	—/PUNCT Uií/ADVT ara/N nti/NEG puranga/**A+ADVA** pinaitikasara/N supé/ADP ./PUNCT # — Hoje o dia não é bom para os pescadores.
	A+ADVA	t-yrl-por-eng-01-pos.txt:	Indé/PRON puranga/**A+ADVA** !/PUNCT # Você é bonito!
Testset 1 e Testset 2	**A+ADVA+N**		

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
	A+ADVA+N	t-yrl-por-eng-08-pos.txt:	Asuí/ADVT+CCONJ ,/PUNCT paá/RPRT ,/PUNCT wirawasú/N unheẽ/V :/PUNCT —/PUNCT Compadre/PROPN ,/PUNCT puxiwera/**A+ADVA+N** asasá/V ./PUNCT # Depois, contam que o gavião disse: — Compadre, passo mal.
	A+ADVA+N	e-yrl-por-eng-03-pos.txt:	Kwá/ADVDX+DEMX+V tawa/N puranga/A+ADVA ,/PUNCT nhaã/DEMS tawa/N puxiwera/**A+ADVA+N** ./PUNCT # Esta cidade é bonita, aquela cidade é feia.
	A+ADVA+N	e-yrl-por-eng-09-pos.txt:	Watarampuá/??? ,/PUNCT nhaã/DEMS kurumiwasú/N kirimbawa/**A+ADVA+N** piri/ADP+ADVG waá/REL yandé/PRON retamawara/A+N aintá/PRON+PRON2 suí/ADP ,/PUNCT umendari/V arama/ADP+FUT+SCONJ waá/REL yepé/ART+CARD+FRUST+SCONJ xe/PRON2 irũmu/ADP+SCONJ ,/PUNCT nti/NEG rẽ/ADVT uyana/V kwá/ADVDX+DEMX+V kaxiwera/N ./PUNCT # Uatarampuá, aquele moço que é o mais valente dos que são de nossa terra, que era para casar-se comigo, não correu ainda esta cachoeira.
	A+ADVA+N	e-yrl-por-eng-12-pos.txt:	Nti/NEG nhuntu/ADV akwáu/V indé/PRON yepé/ART+CARD+FRUST+SCONJ pitua/**A+ADVA+N** ./PUNCT # Só não sabia que você era um covarde.

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
	A+ADVA+N	t-yrl-por-eng-08-pos.txt:	Ape/ADVDI+ADVJ ,/PUNCT paá/RPRT ,/PUNCT urubú/N usuaxara/V :/PUNCT —/PUNCT Compadre/PROPN ,/PUNCT ixé/PRON nti/NEG aputari/V amunhã/V puxiwera**/A+ADVA+N** suú-itá/N supé/ADP ,/PUNCT ma/ADVJ+CCONJ kuíri/ADVT ixé/PRON se/PRON2 yumasí/A2+N retana/ADVS aikú/V ./PUNCT # Então, dizem que o urubu respondeu: — Compadre, eu não quero fazer mal aos animais, mas agora eu estou muito faminto.
Testset 1 e Testset 2	**A+ADVC**		
	A+ADVC	e-yrl-por-eng-03-pos.txt:	Pedro/PROPN usú/V apekatú**/A+ADVC** pe/CLADP+PRON2 suí/ADP ./PUNCT # Pedro vai longe de vocês.
	A+ADVC	e-yrl-por-eng-08-pos.txt:	Mira/N puxí/A+ADVA uyumundú/V apekatú**/A+ADVC** kití/ADP ./PUNCT # Gente ruim é mandada para longe.
	A+ADVC	e-yrl-por-eng-10-pos.txt:	Mairamé/ADVNT+ADVRT+SCONJR usika/V ,/PUNCT apekatú**/A+ADVC** xinga/ADVS+INDQ upitá/V ./PUNCT # Quando chegou, ficou um pouco longe.
	A+ADVC	t-yrl-por-eng-03-pos.txt:	São/PROPN Gabriel/PROPN tawa/N puranga/A+ADVA ,/PUNCT apekatú**/A+ADVC** Barra/PROPN suí/ADP ./PUNCT # São Gabriel é uma cidade bonita, distante de Manaus.
Testset 1 e Testset 2	**A+ADVC+N**		

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
	A+ADVC+N	t-yrl-por-eng-09-pos.txt:	Uyeréu/V ,/PUNCT umaã/V iwí/N kití/ADP masuí/ADVLC+ADVNC+ADVRC uri/V iwaté**/A+ADVC+N** kití/ADP ./PUNCT # Virou, olhou para a terra donde veio para cima.
Testset 1 e Testset 2	**A+ADV+N**		
	A+ADV+N	t-yrl-por-eng-10-pos.txt:	Aé/DEMSN+PRON umurari/V amú/IND tendawa/N upé/ADP+N ,/PUNCT mirĩ**/A+ADV+N** piri/ADP+ADVG Pedro/PROPN yara/ADP+N suí/ADP ./PUNCT # Ele mora em outra comunidade, menor que a de Pedro.
	A+ADV+N	e-yrl-por-eng-02-pos.txt:	Uka-itá/N mirĩ**/A+ADV+N** ./PUNCT # As casas são pequenas.
	A+ADV+N	e-yrl-por-eng-03-pos.txt:	Kwá/ADVDX+DEMX+V garapá/N mirĩ**/A+ADV+N** ,/PUNCT nhaã/DEMS umbaá/NEG ./PUNCT # Este porto é pequeno, aquele não.
	A+ADV+N	y-yrl-por-eng-08-pos.txt:	Resendú/V maã/COND+IND+INT+N+REL+V ambeú/V indé/PRON arã/ADP+FUT+SCONJ ,/PUNCT Se/PRON2 mũ/N mirĩ**/A+ADV+N** :/PUNCT Resikari/V yupatí/N remunhã/V arã/ADP+FUT+SCONJ ne/PRON2 kakurí/N ,/PUNCT ti/NEG arã/ADP+FUT+SCONJ rewatá/V remundá/V se/PRON2 kakurí/N ./PUNCT # Escute o que lhe digo, meu irmãozinho: procure jupati para fazer seu cacuri para não andar furtando meu cacuri, para não andar furtando meu cacuri.

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
	A+ADV+N	e-yrl-por-eng-10-pos.txt:	Pedro/PROPN ruka/N mirĩ**/A+ADV+N** ./PUNCT Pe/CLADP+PRON2 yara/ADP+N turusú/A+ADVS+INDQ ./PUNCT # A casa de Pedro é pequena. A de vocês é grande.
Testset 1 e Testset 2	**A+ADVS**		
	A+ADVS	e-yrl-por-eng-10-pos.txt:	Aé/DEMSN+PRON upitá/V piaíwa/A2 reté**/A+ADVS** ne/PRON2 irũmu/ADP+SCONJ ./PUNCT # Ele fica muito bravo com você.
	A+ADVS	t-yrl-por-eng-01-pos.txt:	—/PUNCT Kwekatú/A+N reté**/A+ADVS** !/PUNCT # — Muito obrigado!
	A+ADVS	t-yrl-por-eng-03-pos.txt:	Maria/PROPN umukwekatú/V :/PUNCT —/PUNCT Kwekatú/A+N reté**/A+ADVS** !/PUNCT # Maria agradece: — Muito obrigada!
Testset 1 e Testset 2	**A+ADVS+INDQ**		
	A+ADVS+INDQ	y-yrl-por-eng-07-pos.txt:	Heitor/??? Villa-Lobos/PROPN nheengarisawa/N munhangara/A+N kwera/A+PRET turusú**/A+ADVS+INDQ** piri/ADP+ADVG waá/REL amú-itá/IND suí/ADP Brasil/PROPN upé/ADP+N ./PUNCT # Heitor Villa-Lobos foi o maior compositor de todos no Brasil.
	A+ADVS+INDQ	e-yrl-por-eng-04-pos.txt:	Kwá/ADVDX+DEMX+V mirá/N turusú**/A+ADVS+INDQ** ./PUNCT

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
			# Esta árvore é grande.
	A+ADVS+INDQ	e-yrl-por-eng-04-pos.txt:	Pirá/N sera/N waá/REL pirarukú/N ,/PUNCT turusú/**A+ADVS+INDQ** retana/ADVS ./PUNCT
			# O peixe cujo nome é pirarucu, é muito grande.
	A+ADVS+INDQ	e-yrl-por-eng-10-pos.txt:	Pedro/PROPN ruka/N mirĩ/A+ADV+N ./PUNCT Pe/CLADP+PRON2 yara/ADP+N turusú/**A+ADVS+INDQ** ./PUNCT
			# A casa de Pedro é pequena. A de vocês é grande.
	A+ADVS+INDQ	e-yrl-por-eng-04-pos.txt:	Aikwé/EXST yepé/ART+CARD+FRUST+SCONJ uka/N turusú/**A+ADVS+INDQ** waá/REL ./PUNCT
			# Há uma casa que é grande.
Testset 1 e Testset 2	**ADP+ADVA+IND**		
	ADP+ADVA+IND	e-yrl-por-eng-12-pos.txt:	Nti/NEG kurí/FUT yawé/**ADP+ADVA+IND** remunhã/V ./PUNCT
			# Você não fará assim.
	ADP+ADVA+IND	e-yrl-por-eng-13-pos.txt:	Akwera/ADVT yawewera/**ADP+ADVA+IND** aé/DEMSN+PRON ./PUNCT
			# Há muito tempo ele é acostumado assim.
	ADP+ADVA+IND	e-yrl-por-eng-09-pos.txt:	Yawé/**ADP+ADVA+IND** tẽ/FOC+NEGI aputari/V ./PUNCT
			# É assim mesmo que eu quero.
	ADP+ADVA+IND	t-yrl-por-eng-13-pos.txt:	Maria/PROPN unheẽ/V :/PUNCT —/PUNCT Kawera-itá/N yawé/**ADP+ADVA+IND** tẽ/FOC+NEGI ./PUNCT
			# Maria diz: — Bêbados são assim mesmo.

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
	ADP+ADVA+IND	e-yrl-por-eng-10-pos.txt:	Kwá/ADVDX+DEMX+V apigawa/A+N pirasua/A+N mayé/ADVLA+ADVRA+SCONJR nhaã/DEMS yawé/**ADP+ADVA+IND** ./PUNCT # Este homem é tão pobre como aquele.
Testset 1 e Testset 2	**ADP+ADVG**		
	ADP+ADVG	t-yrl-por-eng-07-pos.txt:	Antônio/PROPN ururi/V ana/PFV piri/**ADP+ADVG** yepeawa/N memũitawa/N kití/ADP ./PUNCT # Antônio já trouxe mais lenha para o fogão.
	ADP+ADVG	y-yrl-por-eng-07-pos.txt:	Suasú/N manha/N ,/PUNCT usenũi/V ramé/ADP+SCONJ ,/PUNCT uri/V i/PRON2 mimbira/N piri/**ADP+ADVG** ./PUNCT # A mãe do veado, quando ele chamou, veio para junto de seu filho.
	ADP+ADVG	t-yrl-por-eng-11-pos.txt:	Aé/DEMSN+PRON usú/V kutara/ADVA piri/**ADP+ADVG** Paranãwasú/N rupí/ADP ./PUNCT # Ela vai mais rápido pelo rio Negro.
	ADP+ADVG	y-yrl-por-eng-07-pos.txt:	Heitor/??? Villa-Lobos/PROPN nheengarisawa/N munhangara/A+N kwera/A+PRET turusú/A+ADVS+INDQ piri/**ADP+ADVG** waá/REL amú-itá/IND suí/ADP Brasil/PROPN upé/ADP+N ./PUNCT # Heitor Villa-Lobos foi o maior compositor de todos no Brasil.

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
	ADP+ADVG	e-yrl-por-eng-09-pos.txt:	Watarampuá/??? ,/PUNCT nhaã/DEMS kurumiwasú/N kirimbawa/A+ADVA+N piri**/ADP+ADVG** waá/REL yandé/PRON retamawara/A+N aintá/PRON+PRON2 suí/ADP ,/PUNCT umendari/V arama/ADP+FUT+SCONJ waá/REL yepé/ART+CARD+FRUST+SCONJ xe/PRON2 irũmu/ADP+SCONJ ,/PUNCT nti/NEG rẽ/ADVT uyana/V kwá/ADVDX+DEMX+V kaxiwera/N ./PUNCT # Uatarampuá, aquele moço que é o mais valente dos que são de nossa terra, que era para casar-se comigo, não correu ainda esta cachoeira.
Testset 1 e Testset 2	**ADP+FUT+SCONJ**		
	ADP+FUT+SCONJ	t-yrl-por-eng-08-pos.txt:	Asuí/ADVT+CCONJ ,/PUNCT paá/RPRT ,/PUNCT wirawasú/N unheẽ/V :/PUNCT —/PUNCT Compadre/PROPN ,/PUNCT remunhã/V mayé/ADVLA+ADVRA+SCONJR se/PRON2 yawé/ADP+ADVA+IND :/PUNCT repisika/V suú/N+V sikwé/A2+N ,/PUNCT reyuká/V aé/DEMSN+PRON rembaú/V arã**/ADP+FUT+SCONJ** aé/DEMSN+PRON ./PUNCT # Depois, contam que o gavião disse: — Compadre, faça assim como eu: pegue animal vivo, mate-o para comê-lo.
	ADP+FUT+SCONJ	e-yrl-por-eng-13-pos.txt:	Maria/PROPN umeẽ/V kawĩrana/N aintá/PRON+PRON2 supé/ADP nti/NEG arã**/ADP+FUT+SCONJ** aintá/PRON+PRON2 ukaú/V ./PUNCT # Maria deu a eles pinga fraca para eles não se embebedarem.

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
	ADP+FUT+SCONJ	e-yrl-por-eng-03-pos.txt:	Reruri/V aé/DEMSN+PRON ixé/PRON arama/**ADP+FUT+SCONJ** ./PUNCT # Traga-o para mim.
	ADP+FUT+SCONJ	t-yrl-por-eng-03-pos.txt:	Ariré/ADVT kurí/FUT apiripana/V maã-itá/N+REL ixé/PRON arama/**ADP+FUT+SCONJ** ./PUNCT # Depois comprarei coisas para mim.
	ADP+FUT+SCONJ	t-yrl-por-eng-11-pos.txt:	Maria/PROPN anama-itá/N amuramé/ADVT aintá/PRON+PRON2 usemu/V uwatá-watá/V arama/**ADP+FUT+SCONJ** ./PUNCT # Os familiares de Maria às vezes saem para passear.
Testset 1 e Testset 2	**ADP+N**		
	ADP+N	t-yrl-por-eng-04-pos.txt:	Aikwé/EXST mira-itá/N upurungitá/V waá/REL amú/IND nheenga-itá/N São/PROPN Gabriel/PROPN upé/**ADP+N** :/PUNCT Baniwa/N ,/PUNCT Tukano/N ,/PUNCT Desana/N ,/PUNCT Yanomami/N ./PUNCT # Há pessoas que falam outras línguas em São Gabriel: baniwa, tucano, desana, ianomami.
	ADP+N	e-yrl-por-eng-09-pos.txt:	Tá/PRON+PRON2 uyuká/V yepé/ART+CARD+FRUST+SCONJ tatú/N paranã/N upé/**ADP+N** ./PUNCT # Mataram um tatu no rio.
	ADP+N	t-yrl-por-eng-06-pos.txt:	Marantaá/ADVR indé/PRON reikupukú/V ana/PFV São/PROPN Gabriel/PROPN upé/**ADP+N** ?/PUNCT # Por que você demorou em São Gabriel?
	ADP+N	e-yrl-por-eng-08-pos.txt:	Reyana/V taína-itá/N rakakwera/**ADP+N** ./PUNCT # Corres atrás das crianças.

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
	ADP+N	e-yrl-por-eng-06-pos.txt:	Kunhã/A+N usú/V senundé/**ADP+N** ./PUNCT # A mulher foi antes dele.
Testset 1 e Testset 2	**ADP+N+SCONJ**		
	ADP+N+SCONJ	e-yrl-por-eng-10-pos.txt:	Asú/V amupinima/V sesewara/**ADP+N+SCONJ** ./PUNCT # Vou escrever o que é relativo a ela.
	ADP+N+SCONJ	e-yrl-por-eng-08-pos.txt:	Ukiri/V renundé/**ADP+N+SCONJ** ,/PUNCT aé/DEMSN+PRON uú/V xibé/N ./PUNCT # Antes de dormir, ele bebe chibé.
	ADP+N+SCONJ	t-yrl-por-eng-08-pos.txt:	Sesewara/**ADP+N+SCONJ** ixé/PRON apitá/V sasiara/A ./PUNCT # Por causa disso eu fico triste.
	ADP+N+SCONJ	e-yrl-por-eng-06-pos.txt:	Ixé/PRON apurakí/V sesewara/**ADP+N+SCONJ** ./PUNCT # Eu trabalho por ele.
	ADP+N+SCONJ	e-yrl-por-eng-06-pos.txt:	Asú/V aikú/V se/PRON2 renundé/**ADP+N+SCONJ** kití/ADP ./PUNCT # Estou indo para adiante de mim.
Testset 1 e Testset 2	**ADP+SCONJ**		

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
	ADP+SCONJ	t-yrl-por-eng-02-pos.txt:	Aé/DEMSN+PRON upitá/V yepé/ART+CARD+FRUST+SCONJ igara/N mirĩ/A+ADV+N upé/ADP+N Maria/PROPN irũmu**/ADP+SCONJ** ./PUNCT # Ele fica em uma canoa pequena com Maria.
	ADP+SCONJ	y-yrl-por-eng-07-pos.txt:	Yamaã/V iké/ADVDX yepé/ART+CARD+FRUST+SCONJ nheengarisawa/N aé/DEMSN+PRON umunhã/V waá/REL 1952/??? ramé**/ADP+SCONJ** ./PUNCT # Vemos aqui uma música que ele fez em 1952.
	ADP+SCONJ	e-yrl-por-eng-06-pos.txt:	Ixé/PRON amandwari/V Maria/PROPN resé**/ADP+SCONJ** ./PUNCT # Eu me lembro de Maria.
	ADP+SCONJ	e-yrl-por-eng-08-pos.txt:	Maria/PROPN upurakari/V tipití/N maniaka/N kitika/A+V irũmu**/ADP+SCONJ** ./PUNCT # Maria enche o tipití com a mandioca ralada.
	ADP+SCONJ	t-yrl-por-eng-08-pos.txt:	Umbaú/V riré**/ADP+SCONJ** ,/PUNCT Maria/PROPN mimbira/N usú/V ukiri/V ./PUNCT # Depois de comer, o filho de Maria vai dormir.
Testset 1 e Testset 2	**ADV+CCONJ+V**		
	ADV+CCONJ+V	t-yrl-por-eng-03-pos.txt:	Aé/DEMSN+PRON upiripana/V kurí/FUT maã-itá/N+REL i/PRON2 mimbira/N supé/ADP ,/PUNCT i/PRON2 mena/N supé/ADP yuíri**/ADV+CCONJ+V** ./PUNCT # Ela vai comprar coisas para seu filho, para seu marido também.
	ADV+CCONJ+V	t-yrl-por-eng-07-pos.txt:	Aé/DEMSN+PRON umunhã/V kurí/FUT kwaíra/A uí/N yuíri**/ADV+CCONJ+V** ./PUNCT

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
			# Ela fará um pouco de farinha também.
	ADV+CCONJ+V	t-yrl-por-eng-11-pos.txt:	Maria/PROPN umaã/V uikú/V arara-itá/N ,/PUNCT tukana-itá/N ,/PUNCT amú-itá/IND wirá/N kaapura/A+N yuíri**/ADV+CCONJ+V** i/PRON2 igara/N suí/ADP ./PUNCT
			# Maria está vendo araras, tucanos e outras aves silvestres de sua canoa.
	ADV+CCONJ+V	t-yrl-por-eng-07-pos.txt:	—/PUNCT Antônio/PROPN ,/PUNCT indé/PRON reputari/V ixé/PRON amunhã/V meyú/N yuíri**/ADV+CCONJ+V** ?/PUNCT
			# — Antônio, você quer que eu faça biju também?
	ADV+CCONJ+V	e-yrl-por-eng-06-pos.txt:	Nhaã/DEMS apigawa/A+N umbaú/V ana/PFV pirá/N meyú/N yuíri**/ADV+CCONJ+V** ./PUNCT
			# Aquele homem comeu peixe e biju.
Testset 1	**ADVDI+ADVJ**		
	ADVDI+ADVJ	t-yrl-por-eng-09-pos.txt:	Ape**/ADVDI+ADVJ** i/PRON2 kwema/N+V2 ara/N ./PUNCT
			# Então amanheceu o dia.
	ADVDI+ADVJ	e-yrl-por-eng-08-pos.txt:	Ape**/ADVDI+ADVJ** ,/PUNCT paá/RPRT ,/PUNCT usika/V sesé/ADP wirawasú/N ./PUNCT
			# Então, dizem que chegou a ele o gavião.
	ADVDI+ADVJ	t-yrl-por-eng-08-pos.txt:	Ape**/ADVDI+ADVJ** ,/PUNCT paá/RPRT ,/PUNCT urubú/N usú/V sakakwera/ADP+N merupí/ADVA ,/PUNCT té/FOC+PREP+SCONJR mairamé/ADVNT+ADVRT+SCONJR umaã/V wirawasú/N ./PUNCT
			# Então, dizem que o urubu foi atrás dele devagar, quando viu o gavião.

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
	ADVDI+ADVJ	t-yrl-por-eng-09-pos.txt:	Ape/**ADVDI+ADVJ** yautí/N kwera/A+PRET upupuka/V pá/IND ./PUNCT # Então o jabuti (que "já era") rebentou todo.
	ADVDI+ADVJ	t-yrl-por-eng-03-pos.txt:	Aé/DEMSN+PRON usika/V ape/**ADVDI+ADVJ** ,/PUNCT usú/V yepé/ART+CARD+FRUST+SCONJ piripanasawa/N ruka/N kití/ADP ./PUNCT # Ela chega lá, vai a uma loja.
Testset 1 e Testset 2	**ADVDX+DEMX+V**		
	ADVDX+DEMX+V	t-yrl-por-eng-10-pos.txt:	—/PUNCT Maria/PROPN ,/PUNCT kwá/**ADVDX+DEMX+V** pururé/N se/PRON2 mũyara/??? ./PUNCT # — Maria, esta enxada é do meu irmão.
	ADVDX+DEMX+V	e-yrl-por-eng-10-pos.txt:	Kwá/**ADVDX+DEMX+V** kunhã/A+N unheengari/V puranga/A+ADVA piri/ADP+ADVG nhaã/DEMS suí/ADP ./PUNCT # Esta mulher canta melhor que aquela.
	ADVDX+DEMX+V	t-yrl-por-eng-05-pos.txt:	—/PUNCT Kuxiíma/ADVT ,/PUNCT reyuri/V ramé/ADP+SCONJ kwá/**ADVDX+DEMX+V** kití/ADP ,/PUNCT mayé/ADVLA+ADVRA+SCONJR taá/CQ reyuri/V ?/PUNCT # — Antigamente, quando você veio para cá, como você veio?
	ADVDX+DEMX+V	t-yrl-por-eng-10-pos.txt:	Aé/DEMSN+PRON nti/NEG kurí/FUT upuú/V siía/INDQ kumandamirĩ/N kwá/**ADVDX+DEMX+V** akayú/N nhaãsé/ADVJ+SCONJR amana/N nti/NEG uwari/V retana/ADVS ./PUNCT # Ele não vai colher muito feijão este ano porque a chuva não caiu muito (i.e., não choveu muito).

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
	ADVDX+DEMX+V	e-yrl-por-eng-10-pos.txt:	Kwá/**ADVDX+DEMX+V** kunhã/A+N unheengari/V puranga/A+ADVA mayé/ADVLA+ADVRA+SCONJR nhaã/DEMS yawé/ADP+ADVA+IND ./PUNCT # Esta mulher canta tão bem como aquela.
Testset 1	**ADVJ+ADVT**		
	ADVJ+ADVT	e-yrl-por-eng-08-pos.txt:	Aramé/**ADVJ+ADVT** té/FOC+PREP+SCONJR ,/PUNCT paá/RPRT ,/PUNCT usasá/V yepé/ART+CARD+FRUST+SCONJ wiramirĩ/N ./PUNCT # Nesse momento mesmo, dizem que passava um passarinho.
	ADVJ+ADVT	t-yrl-por-eng-08-pos.txt:	Aramé/**ADVJ+ADVT** té/FOC+PREP+SCONJR ,/PUNCT paá/RPRT ,/PUNCT usasá/V yepé/ART+CARD+FRUST+SCONJ wiramirĩ/N ./PUNCT # Nesse momento mesmo, contam que passou um passarinho.
Testset 1 e Testset 2	**ADVJ+CCONJ**		
	ADVJ+CCONJ	t-yrl-por-eng-08-pos.txt:	Ape/ADVDI+ADVJ ,/PUNCT paá/RPRT ,/PUNCT urubú/N usuaxara/V :/PUNCT —/PUNCT Compadre/PROPN ,/PUNCT ixé/PRON nti/NEG aputari/V amunhã/V puxiwera/A+ADVA+N suú-itá/N supé/ADP ,/PUNCT ma/**ADVJ+CCONJ** kuíri/ADVT ixé/PRON se/PRON2 yumasí/A2+N retana/ADVS aikú/V ./PUNCT # Então, dizem que o urubu respondeu: — Compadre, eu não quero fazer mal aos animais, mas agora eu estou muito faminto.

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
	ADVJ+CCONJ	t-yrl-por-eng-08-pos.txt:	Sakakwera/ADP+N wirawasú/N usú/V ,/PUNCT ma**/ADVJ+CCONJ** uyutuká/V :/PUNCT mirá/N rumitera/N pisãwera/V uwiké/V i/PRON2 putiá/N upé/ADP+N ./PUNCT # Atrás dele foi o gavião, mas se chocou: um pedaço de tronco de árvore entrou no seu peito.
	ADVJ+CCONJ	t-yrl-por-eng-09-pos.txt:	Ixé/PRON amaã/V arama/ADP+FUT+SCONJ yepé/ART+CARD+FRUST+SCONJ nhaã/DEMS murasí/N ,/PUNCT ma**/ADVJ+CCONJ** nti/NEG arikú/V se/PRON2 pepú/N awewé/V arã/ADP+FUT+SCONJ ./PUNCT # Era para eu ver aquele baile, mas não tenho (minhas) asas para voar.
	ADVJ+CCONJ	e-yrl-por-eng-04-pos.txt:	Se/PRON2 paya/N ,/PUNCT paá/RPRT ,/PUNCT usú/V ana/PFV kaá/N kití/ADP ,/PUNCT ma**/ADVJ+CCONJ** aé/DEMSN+PRON nti/NEG usú/V ./PUNCT # Dizem que meu pai foi para a mata, mas ele não foi.
	ADVJ+CCONJ	t-yrl-por-eng-04-pos.txt:	Maria/PROPN umbaú/V pirá/N i/PRON2 pu/ASSUM irũmu/ADP+SCONJ ,/PUNCT ma**/ADVJ+CCONJ** uú/V ií/N i/PRON2 kuya/N irũmu/ADP+SCONJ ./PUNCT # Maria come o peixe com suas mãos, mas bebe água com sua cuia.
Testset 1 e Testset 2	**ADVJ+SCONJR**		
	ADVJ+SCONJR	t-yrl-por-eng-04-pos.txt:	Aé/DEMSN+PRON uwapika/V ,/PUNCT asuí/ADVT+CCONJ umbaú/V pirá/N uí/N irũmu/ADP+SCONJ nhaãsé**/ADVJ+SCONJR** aé/DEMSN+PRON i/PRON2 yumasí/A2+N ./PUNCT

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
			# Ela senta-se e come peixe com farinha porque ela está faminta.
	ADVJ+SCONJR	t-yrl-por-eng-10-pos.txt:	Aé/DEMSN+PRON nti/NEG kurí/FUT upuú/V siía/INDQ kumandamirĩ/N kwá/ADVDX+DEMX+V akayú/N nhaãsé**/ADVJ+SCONJR** amana/N nti/NEG uwari/V retana/ADVS ./PUNCT # Ele não vai colher muito feijão este ano porque a chuva não caiu muito (i.e., não choveu muito).
	ADVJ+SCONJR	t-yrl-por-eng-11-pos.txt:	Maria/PROPN surí/A2 uikú/V nhaãsé**/ADVJ+SCONJR** aé/DEMSN+PRON umaã/V kurí/FUT amú-itá/IND tetama/N ./PUNCT # Maria está feliz porque ela vai ver outras regiões.
Testset 1 e Testset 2	**ADVLA+ADVRA+SCONJR**		
	ADVLA+ADVRA+SCONJR	e-yrl-por-eng-10-pos.txt:	Se/PRON2 manha/N puranga/A+ADVA mayé**/ADVLA+ADVRA+SCONJR** ne/PRON2 yawé/ADP+ADVA+IND ./PUNCT # Minha mãe é tão bonita como tu.
	ADVLA+ADVRA+SCONJR	e-yrl-por-eng-10-pos.txt:	Kwá/ADVDX+DEMX+V apigawa/A+N pirasua/A+N mayé**/ADVLA+ADVRA+SCONJR** nhaã/DEMS yawé/ADP+ADVA+IND ./PUNCT # Este homem é tão pobre como aquele.
	ADVLA+ADVRA+SCONJR	e-yrl-por-eng-10-pos.txt:	Se/PRON2 manha/N puranga/A+ADVA mayé**/ADVLA+ADVRA+SCONJR** ne/PRON2 yara/ADP+N yawé/ADP+ADVA+IND ./PUNCT # Minha mãe é tão bela como a sua.
	ADVLA+ADVRA+SCONJR	e-yrl-por-eng-05-pos.txt:	Kuxiíma/ADVT ,/PUNCT reyuri/V ramé/ADP+SCONJ kwá/ADVDX+DEMX+V kití/ADP ,/PUNCT mayé**/ADVLA+ADVRA+SCONJR** taá/CQ reyuri/V ?/PUNCT

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
			# Antigamente, quando você veio para cá, como você veio?
	ADVLA+ADVRA+SCONJR	e-yrl-por-eng-09-pos.txt:	Mayé**/ADVLA+ADVRA+SCONJR** taá/CQ asú/V awatá/V se/PRON2 retama/N kití/ADP ?/PUNCT
			# Como vou andar lá para a minha terra?
Testset 1	**ADVLC+ADVNC+ADVRC**		
	ADVLC+ADVNC+ADVRC	t-yrl-por-eng-09-pos.txt:	Iwí/N kití/ADP uwari/V ,/PUNCT poh/INTJ !/PUNCT Iwí/N kití/ADP ,/PUNCT kwá/ADVDX+DEMX+V kití/ADP ,/PUNCT makití**/ADVLC+ADVNC+ADVRC** yaikú/V ./PUNCT
			# Caiu para a terra, poh! Para a terra, para cá, para onde estamos.
	ADVLC+ADVNC+ADVRC	e-yrl-por-eng-02-pos.txt:	Masuí**/ADVLC+ADVNC+ADVRC** taá/CQ reyuri/V kuxiíma/ADVT ?/PUNCT
			# Donde você veio antigamente?
	ADVLC+ADVNC+ADVRC	t-yrl-por-eng-09-pos.txt:	Uyeréu/V ,/PUNCT umaã/V iwí/N kití/ADP masuí**/ADVLC+ADVNC+ADVRC** uri/V iwaté/A+ADVC+N kití/ADP ./PUNCT
			# Virou, olhou para a terra donde veio para cima.
	ADVLC+ADVNC+ADVRC	e-yrl-por-eng-02-pos.txt:	Nti/NEG akwáu/V masuí**/ADVLC+ADVNC+ADVRC** Pedro/PROPN usika/V ./PUNCT
			# Não sei donde Pedro chegou.
	ADVLC+ADVNC+ADVRC	e-yrl-por-eng-02-pos.txt:	Asú/V makití**/ADVLC+ADVNC+ADVRC** aputari/V ./PUNCT
			# Vou aonde quero.
Testset 1	**ADVLC+ADVRC**		
	ADVLC+ADVRC	t-yrl-por-eng-10-pos.txt:	Mamé**/ADVLC+ADVRC** taá/CQ uikú/V se/PRON2 yara/ADP+N ?/PUNCT

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
			# Onde está a minha?
	ADVLC+ADVRC	e-yrl-por-eng-01-pos.txt:	Mamé/**ADVLC+ADVRC** Barcelos/PROPN ?/PUNCT # Onde é Barcelos?
	ADVLC+ADVRC	e-yrl-por-eng-12-pos.txt:	Remaã/V mamé/**ADVLC+ADVRC** nhuntu/ADV reikú/V !/PUNCT # Olhe só onde você está!
	ADVLC+ADVRC	e-yrl-por-eng-08-pos.txt:	Mamé/**ADVLC+ADVRC** taá/CQ puranga/A+ADVA uyumunhã/V arã/ADP+FUT+SCONJ yané/PRON2 ruka/N ?/PUNCT # Onde é bom para se fazer nossa casa?
	ADVLC+ADVRC	y-yrl-por-eng-07-pos.txt:	Kurumĩwasú/N ramé/ADP+SCONJ ,/PUNCT aé/DEMSN+PRON uwatá-watá/V siía/INDQ tetama/N rupí/ADP Brasil/PROPN rupí/ADP ,/PUNCT mamé/**ADVLC+ADVRC** aé/DEMSN+PRON uyumbué/V siía/INDQ kawoka/N nheengarisawa/N ./PUNCT # Quando era jovem, ele viajou por muitas regiões pelo Brasil, onde ele aprendeu muitas canções de caboclos.
Testset 1	**ADVM+ADVT**		
	ADVM+ADVT	y-yrl-por-eng-07-pos.txt:	Aé/DEMSN+PRON kuité/**ADVM+ADVT** uyumũ/V yuíri/ADV+CCONJ+V suasumirĩ/N manha/N ./PUNCT # Ele (o caçador), então, flechou também a mãe do veadinho.
Testset 1 e Testset 2	**ADVNT+ADVRT+SCONJR**		

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
	ADVNT+ADVRT+SCONJR	t-yrl-por-eng-12-pos.txt:	Aé/DEMSN+PRON upurandú/V i/PRON2 mena/N suí/ADP :/PUNCT —/PUNCT Pedro/PROPN ,/PUNCT mairamé/**ADVNT+ADVRT+SCONJR** taá/CQ indé/PRON remupinima/V ukapí/N mamé/ADVLC+ADVRC Rute/PROPN usú/V ukiri/V ?/PUNCT # Ela pergunta a seu marido: — Pedro, quando você pintará o quarto em que Rute vai dormir?
	ADVNT+ADVRT+SCONJR	t-yrl-por-eng-13-pos.txt:	Ne/PRON2 mairamé/**ADVNT+ADVRT+SCONJR** aintá/PRON+PRON2 rurí/A2 ./PUNCT # Nunca são felizes.
	ADVNT+ADVRT+SCONJR	t-yrl-por-eng-11-pos.txt:	Aintá/PRON+PRON2 umundá/V ana/PFV Maria/PROPN ruka/N mairamé/**ADVNT+ADVRT+SCONJR** aintá/PRON+PRON2 uikú/V ukara/N kití/ADP ./PUNCT # Roubaram a casa de Maria quando eles estavam fora.
	ADVNT+ADVRT+SCONJR	t-yrl-por-eng-11-pos.txt:	Ne/PRON2 mairamé/**ADVNT+ADVRT+SCONJR** aintá/PRON+PRON2 umaã/V nhaã/DEMS mimi/ADVDI ./PUNCT # Nunca viram isso ali.
	ADVNT+ADVRT+SCONJR	e-yrl-por-eng-10-pos.txt:	Mairamé/**ADVNT+ADVRT+SCONJR** usika/V ,/PUNCT apekatú/A+ADVC xinga/ADVS+INDQ upitá/V ./PUNCT # Quando chegou, ficou um pouco longe.
Testset 1 e Testset 2	**ADVO+ORD**		
	ADVO+ORD	t-yrl-por-eng-12-pos.txt:	Yepesawa/ORD ukapí/N Antônio/PROPN yara/ADP+N ,/PUNCT mukũisawa/**ADVO+ORD** ukapí/N mamé/ADVLC+ADVRC Maria/PROPN ukiri/V i/PRON2 mena/N irũmu/ADP+SCONJ ./PUNCT # O primeiro quarto é o de Antônio e o segundo quarto é onde Maria dorme com seu marido.
Testset 1 e	**ADVS+INDQ**		

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
Testset 2			
	ADVS+INDQ	e-yrl-por-eng-10-pos.txt:	Maria/PROPN upurakí/V xinga/**ADVS+INDQ** ./PUNCT # Maria trabalha pouco.
	ADVS+INDQ	e-yrl-por-eng-11-pos.txt:	Kurí/FUT mirĩ/A+ADV+N xinga/**ADVS+INDQ** ,/PUNCT nti awá/IND ukwáu/V masuí/ADVLC+ADVNC+ADVRC uyukwáu/V mira/N ./PUNCT # Um pouquinho depois, ninguém sabe donde apareceu gente.
	ADVS+INDQ	e-yrl-por-eng-10-pos.txt:	Mairamé/ADVNT+ADVRT+SCONJR usika/V ,/PUNCT apekatú/A+ADVC xinga/**ADVS+INDQ** upitá/V ./PUNCT # Quando chegou, ficou um pouco longe.
	ADVS+INDQ	e-yrl-por-eng-10-pos.txt:	Senundé/ADP+N kití/ADP xinga/**ADVS+INDQ** aé/DEMSN+PRON usuantí/V yepé/ART+CARD+FRUST+SCONJ tatú/N ./PUNCT # Um pouco à frente dele, ele encontrou um tatu.
	ADVS+INDQ	e-yrl-por-eng-10-pos.txt:	Aé/DEMSN+PRON tepusimanha/A+N xinga/**ADVS+INDQ** ./PUNCT # Ele está um pouco sonolento.
Testset 1 e Testset 2	**ADVT+CCONJ**		
	ADVT+CCONJ	t-yrl-por-eng-11-pos.txt:	Aintá/PRON+PRON2 uwiyé/V yepé/ART+CARD+FRUST+SCONJ yupirisawa/N paranã/N ruakí/ADP ,/PUNCT asuí/**ADVT+CCONJ** aintá/PRON+PRON2 uyuruari/V yepé/ART+CARD+FRUST+SCONJ igara/N puranga/A+ADVA upé/ADP+N ./PUNCT # Eles descem uma escada perto do rio, e embarcam numa bonita canoa.

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
	ADVT+CCONJ	t-yrl-por-eng-09-pos.txt:	Uyeréu/V ,/PUNCT uwari/V asuí**/ADVT+CCONJ** ,/PUNCT poh/INTJ !/PUNCT # Virou, caiu dali, poh!
	ADVT+CCONJ	y-yrl-por-eng-03-pos.txt:	Asuí**/ADVT+CCONJ** ,/PUNCT pituna/N ramé/ADP+SCONJ ,/PUNCT reyúri/??? remusakú/V se/PRON2 putiá/N ./PUNCT # Depois, de noite, Venha aquecer o meu peito.
	ADVT+CCONJ	t-yrl-por-eng-06-pos.txt:	Asuí**/ADVT+CCONJ** ixé/PRON amusarái/V paranã/N upé/ADP+N ./PUNCT # Depois eu brinquei no rio.
	ADVT+CCONJ	t-yrl-por-eng-10-pos.txt:	Asuí**/ADVT+CCONJ** ,/PUNCT aé/DEMSN+PRON ukapiri/V pá/IND ,/PUNCT uyuka/V siía/INDQ tukandira/N ,/PUNCT xibuí/N ,/PUNCT yandú/N maniwatiwa/N suí/ADP ./PUNCT # Depois, ele capina tudo, tira muitas tocandiras, vermes e aranhas do mandiocal.
Testset 1 e Testset 2	**A+INDQ**		
	A+INDQ	e-yrl-por-eng-03-pos.txt:	Mairamé/ADVNT+ADVRT+SCONJR kurí/FUT bũa**/A+INDQ** ,/PUNCT ixé/PRON asú/V kurí/FUT ayuká/V indé/PRON arama/ADP+FUT+SCONJ kwá/ADVDX+DEMX+V tukunaré/N ./PUNCT # Quando eu for grande, eu vou matar para você este tucunaré.
	A+INDQ	t-yrl-por-eng-13-pos.txt:	Aikwé/EXST bũa**/A+INDQ** timbiú/N penhẽ/PRON arama/ADP+FUT+SCONJ !/PUNCT # Há muita comida para vocês!
	A+INDQ	t-yrl-por-eng-13-pos.txt:	Aikwé/EXST bũa**/A+INDQ** meyú-itá/N yurá/N árupi/ADP pembaú/V arama/ADP+FUT+SCONJ !/PUNCT

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
			# Há grandes bijus sobre o jirau para vocês comerem!
Testset 1 e Testset 2	**A+INTJ+N**		
	A+INTJ+N	e-yrl-por-eng-04-pos.txt:	Aé/DEMSN+PRON kwera/A+PRET ,/PUNCT taité/**A+INTJ+N** !/PUNCT # Ele já era, coitado!
Testset 1 e Testset 2	**A+N**		
	A+N	e-yrl-por-eng-03-pos.txt:	Aikwé/EXST raẽ/ADV será/PQ tuxawa/N tuyuwé/**A+N** ?/PUNCT # Existe ainda o velho tuxaua?
	A+N	e-yrl-por-eng-06-pos.txt:	Kunhã/**A+N** usú/V i/PRON2 mena/N renundé/ADP+N+SCONJ ./PUNCT # A mulher foi antes de seu marido.
	A+N	e-yrl-por-eng-11-pos.txt:	Mayawé/SCONJR nti yepé/IND apigawa/**A+N** usaisú/V nhaã/DEMS kunhãmukú/**A+N** ,/PUNCT aé/DEMSN+PRON nti/NEG upurasí/V ./PUNCT # Como nenhum homem amava aquela moça, ela não dançou.
	A+N	t-yrl-por-eng-11-pos.txt:	Maria/PROPN umaã/V uikú/V arara-itá/N ,/PUNCT tukana-itá/N ,/PUNCT amú-itá/IND wirá/N kaapura/**A+N** yuíri/ADV+CCONJ+V i/PRON2 igara/N suí/ADP ./PUNCT

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
			# Maria está vendo araras, tucanos e outras aves silvestres de sua canoa.
	A+N	e-yrl-por-eng-04-pos.txt:	Asuiwara/**A+N** kwera/A+PRET ,/PUNCT mukũi/CARD+INDQ apigawa/**A+N** nhũ/ADV upitá/V umanhana/V uikú/V yané/PRON2 itá/N uka/N ./PUNCT
			# Desde então, somente dois homens ficaram vigiando nossa casa de pedra.
Testset 1 e Testset 2	**A+N+V2**		
	A+N+V2	e-yrl-por-eng-03-pos.txt:	Kwá-itá/DEMX mirá/N santá/A ,/PUNCT nhaã-itá/DEMS membeka/**A+N+V2** ./PUNCT
			# Estas madeiras são duras, aquelas são moles.
	A+N+V2	t-yrl-por-eng-04-pos.txt:	Aé/DEMSN+PRON urasú/V pirá/N suka/N+V suí/ADP sukwera/N waá/REL membeka/**A+N+V2** ./PUNCT
			# Ela leva peixe da sua casa, cuja carne é mole.
Testset 1 e Testset 2	**A+PRET**		
	A+PRET	e-yrl-por-eng-05-pos.txt:	Se/PRON2 manha/N unheẽ/V kwera/**A+PRET** ixé/PRON arama/ADP+FUT+SCONJ :/PUNCT —/PUNCT Nhaã/DEMS se/PRON2 kurumĩ/N ./PUNCT
			# Minha mãe dizia para mim: Aquele é meu menino.
	A+PRET	e-yrl-por-eng-04-pos.txt:	Asuiwara/A+N kwera/**A+PRET** ,/PUNCT mukũi/CARD+INDQ apigawa/A+N nhũ/ADV upitá/V umanhana/V uikú/V yané/PRON2 itá/N uka/N ./PUNCT

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
			# Desde então, somente dois homens ficaram vigiando nossa casa de pedra.
	A+PRET	e-yrl-por-eng-04-pos.txt:	Aé/DEMSN+PRON kwera**/A+PRET** ./PUNCT
			# Ela já era.
	A+PRET	t-yrl-por-eng-09-pos.txt:	Ape/ADVDI+ADVJ yautí/N kwera**/A+PRET** upupuka/V pá/IND ./PUNCT
			# Então o jabuti (que "já era") rebentou todo.
	A+PRET	y-yrl-por-eng-07-pos.txt:	Heitor/??? Villa-Lobos/PROPN nheengarisawa/N munhangara/A+N kwera**/A+PRET** turusú/A+ADVS+INDQ piri/ADP+ADVG waá/REL amú-itá/IND suí/ADP Brasil/PROPN upé/ADP+N ./PUNCT
			# Heitor Villa-Lobos foi o maior compositor de todos no Brasil.
Testset 1 e Testset 2	**ART+CARD+FRUST+SCONJ**		
	ART+CARD+FRUST+SCONJ	e-yrl-por-eng-13-pos.txt:	Ixé/PRON amaã/V arama/ADP+FUT+SCONJ yepé**/ART+CARD+FRUST+SCONJ** nhaã/DEMS murasí/N ,/PUNCT arikú/V ramé/ADP+SCONJ yepé**/ART+CARD+FRUST+SCONJ** se/PRON2 pepú/N awewé/V arama/ADP+FUT+SCONJ ./PUNCT
			# Era para eu ver aquele baile, se tivesse minhas asas para voar.
	ART+CARD+FRUST+SCONJ	e-yrl-por-eng-11-pos.txt:	Upukwari/V aintá/PRON+PRON2 yepé**/ART+CARD+FRUST+SCONJ** amú-itá/IND resé/ADP+SCONJ ./PUNCT
			# Amarraram-nos uns nos outros.
	ART+CARD+FRUST+SCONJ	e-yrl-por-eng-03-pos.txt:	Amú/IND pituna/N piterarupí/ADP ,/PUNCT asendú/V yepé**/ART+CARD+FRUST+SCONJ** nheenga/N ./PUNCT
			# Pelo meio da outra noite, ouvi uma voz.

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
	ART+CARD+FRUST+SCONJ	e-yrl-por-eng-09-pos.txt:	Watarampuá/??? ,/PUNCT nhaã/DEMS kurumiwasú/N kirimbawa/A+ADVA+N piri/ADP+ADVG waá/REL yandé/PRON retamawara/A+N aintá/PRON+PRON2 suí/ADP ,/PUNCT umendari/V arama/ADP+FUT+SCONJ waá/REL yepé**/ART+CARD+FRUST+SCONJ** xe/PRON2 irũmu/ADP+SCONJ ,/PUNCT nti/NEG rẽ/ADVT uyana/V kwá/ADVDX+DEMX+V kaxiwera/N ./PUNCT # Uatarampuá, aquele moço que é o mais valente dos que são de nossa terra, que era para casar-se comigo, não correu ainda esta cachoeira.
	ART+CARD+FRUST+SCONJ	e-yrl-por-eng-11-pos.txt:	Yepé**/ART+CARD+FRUST+SCONJ** ara/N Tupana/N usú/V uwatá/V ,/PUNCT uxari/V tatá/N uka/N ukara/N kití/ADP ./PUNCT # Um dia Tupã foi andar e deixou o fogo para fora de casa.
Testset 1 e Testset 2	**A+V**		
	A+V	e-yrl-por-eng-08-pos.txt:	Maria/PROPN upurakari/V tipití/N maniaka/N kitika**/A+V** irũmu/ADP+SCONJ ./PUNCT # Maria enche o tipití com a mandioca ralada.
	A+V	t-yrl-por-eng-03-pos.txt:	I/PRON2 xirura-itá/N suruka**/A+V** ./PUNCT # As calças dele estão rasgadas.
	A+V	t-yrl-por-eng-07-pos.txt:	Ariré/ADVT ,/PUNCT aé/DEMSN+PRON upurakari/V tipití/N maniaka/N kitika**/A+V** irũmu/ADP+SCONJ ./PUNCT # Depois, ela enche o tipití com a mandioca ralada.
Testset 1 e Testset 2	**A+V2**		

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
	A+V2	e-yrl-por-eng-05-pos.txt:	Mayawé/SCONJR ixé/PRON maraari/**A+V2** aikú/V yepé/ART+CARD+FRUST+SCONJ ,/PUNCT ayenũ/V se/PRON2 mimbira/N ruakí/ADP ,/PUNCT akiri/V ana/PFV ./PUNCT # Como eu estava cansado, deitei-me perto de meu filho e dormi.
Testset 1 e Testset 2	**CARD+INDQ**		
	CARD+INDQ	e-yrl-por-eng-08-pos.txt:	Mukũi/**CARD+INDQ** apigawa/A+N uyuyuká/V ./PUNCT # Os dois homens se mataram.
	CARD+INDQ	e-yrl-por-eng-07-pos.txt:	Mukũi/**CARD+INDQ** itá/N umanũ/??? ana/PFV ./PUNCT # Os dois já morreram.
	CARD+INDQ	e-yrl-por-eng-04-pos.txt:	Asuiwara/A+N kwera/A+PRET ,/PUNCT mukũi/**CARD+INDQ** apigawa/A+N nhũ/ADV upitá/V umanhana/V uikú/V yané/PRON2 itá/N uka/N ./PUNCT # Desde então, somente dois homens ficaram vigiando nossa casa de pedra.
	CARD+INDQ	t-yrl-por-eng-07-pos.txt:	Aé/DEMSN+PRON upisika/V mukũi/**CARD+INDQ** sapukaya/N ,/PUNCT uyuká/V aintá/PRON+PRON2 i/PRON2 pu/ASSUM irũmu/ADP+SCONJ umemũi/V arama/ADP+FUT+SCONJ aintá/PRON+PRON2 ./PUNCT # Ela pegou duas galinhas, matou-as com suas mãos para cozinhá-las.
	CARD+INDQ	t-yrl-por-eng-03-pos.txt:	Aé/DEMSN+PRON urikú/V mukũi/**CARD+INDQ** akayú/N ./PUNCT # Ele tem dois anos.

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
Testset 1 e Testset 2	**CERT+N**		
	CERT+N	t-yrl-por-eng-13-pos.txt:	Kuíri/ADVT supí**/CERT+N** aintá/PRON+PRON2 uyumusurí-kwáu/V ./PUNCT # Agora de fato eles podem se divertir.
Testset 1 e Testset 2	**CLADP+PRON2**		
	CLADP+PRON2	e-yrl-por-eng-03-pos.txt:	Pedro/PROPN usú/V apekatú/A+ADVC pe**/CLADP+PRON2** suí/ADP ./PUNCT # Pedro vai longe de vocês.
	CLADP+PRON2	e-yrl-por-eng-04-pos.txt:	Atuká/V pe**/CLADP+PRON2** rukena/N ./PUNCT # Bati à porta de vocês.
	CLADP+PRON2	e-yrl-por-eng-09-pos.txt:	Penhẽ/PRON pe**/CLADP+PRON2** akanhemu/V+V2 ./PUNCT # Vocês se assustam.
	CLADP+PRON2	e-yrl-por-eng-03-pos.txt:	pe**/CLADP+PRON2** ruka/N # casa de vocês
	CLADP+PRON2	e-yrl-por-eng-09-pos.txt:	Penhẽ/PRON pe**/CLADP+PRON2** resarái/V2 ./PUNCT # Vocês se esquecem.
Testset 1 e Testset 2	**COND+IND+INT+N+REL+V**		

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
	COND+IND+INT+N+REL+V	e-yrl-por-eng-02-pos.txt:	Ukwáu/V maã**/COND+IND+INT+N+REL+V** i/PRON2 murakí/N ./PUNCT # Sabe qual é seu trabalho.
	COND+IND+INT+N+REL+V	e-yrl-por-eng-11-pos.txt:	Ne/PRON2 maã**/COND+IND+INT+N+REL+V** aintá/PRON+PRON2 urikú/V aintá/PRON+PRON2 umbaú/V arã/ADP+FUT+SCONJ ./PUNCT # Nada eles têm para eles comerem.
	COND+IND+INT+N+REL+V	e-yrl-por-eng-13-pos.txt:	Nti/NEG yambeú-kwáu/??? maã**/COND+IND+INT+N+REL+V** i/PRON2 xupé/ADP ./PUNCT # Não podemos contar nada a ele.
	COND+IND+INT+N+REL+V	e-yrl-por-eng-03-pos.txt:	Apurandú/V ne/PRON2 suí/ADP maã**/COND+IND+INT+N+REL+V** aputari/V akwáu/V ./PUNCT # Pergunto de você o que quero saber.
	COND+IND+INT+N+REL+V	t-yrl-por-eng-07-pos.txt:	—/PUNCT Maã**/COND+IND+INT+N+REL+V** taá/CQ reputari/V ixé/PRON amunhã/V ?/PUNCT # — Que você quer que eu faça?
Testset 1 e Testset 2	**DEMSN+PRON**		
	DEMSN+PRON	t-yrl-por-eng-04-pos.txt:	Asuí/ADVT+CCONJ aé**/DEMSN+PRON** usú/V uyuruari/V i/PRON2 igara/N mirá/N suiwara/ADP upé/ADP+N ./PUNCT # Depois, ela vai embarcar em sua canoa de madeira.
	DEMSN+PRON	e-yrl-por-eng-04-pos.txt:	Aé**/DEMSN+PRON** kwera/A+PRET ,/PUNCT taité/A+INTJ+N !/PUNCT # Ele já era, coitado!

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
	DEMSN+PRON	t-yrl-por-eng-11-pos.txt:	Maria/PROPN surí/A2 uikú/V nhaãsé/ADVJ+SCONJR aé/**DEMSN+PRON** umaã/V kurí/FUT amú-itá/IND tetama/N ./PUNCT # Maria está feliz porque ela vai ver outras regiões.
	DEMSN+PRON	e-yrl-por-eng-05-pos.txt:	Nti/NEG aé/**DEMSN+PRON** uputari/V será/PQ indé/PRON ?/PUNCT # Ele não quis você?
	DEMSN+PRON	e-yrl-por-eng-01-pos.txt:	Aé/**DEMSN+PRON** apigawa/A+N ./PUNCT # Ele é homem.
Testset 1 e Testset 2	**FOC+NEGI**		
	FOC+NEGI	t-yrl-por-eng-06-pos.txt:	—/PUNCT Ambaú/V ana/PFV kuíri/ADVT tẽ/**FOC+NEGI** ./PUNCT # — Comi agora mesmo.
	FOC+NEGI	e-yrl-por-eng-03-pos.txt:	Aiwã/ADVT tẽ/**FOC+NEGI** kurí/FUT ./PUNCT # Logo será (ou daqui a pouco).
	FOC+NEGI	e-yrl-por-eng-09-pos.txt:	Yawé/ADP+ADVA+IND tẽ/**FOC+NEGI** aputari/V ./PUNCT # É assim mesmo que eu quero.
	FOC+NEGI	t-yrl-por-eng-01-pos.txt:	—/PUNCT Puranga/A+ADVA tẽ/**FOC+NEGI** asasá/V ./PUNCT # — Passo bem mesmo.
	FOC+NEGI	t-yrl-por-eng-13-pos.txt:	Maria/PROPN unheẽ/V :/PUNCT —/PUNCT Kawera-itá/N yawé/ADP+ADVA+IND tẽ/**FOC+NEGI** ./PUNCT # Maria diz: — Bêbados são assim mesmo.

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
Testset 1 e Testset 2	**FOC+PREP+SCONJR**		
	FOC+PREP+SCONJR	e-yrl-por-eng-02-pos.txt:	Mamé/ADVLC+ADVRC taá/CQ té/**FOC+PREP+SCONJR** remaã/V se/PRON2 manha/N ?/PUNCT # Onde mesmo você viu minha mãe?
	FOC+PREP+SCONJR	t-yrl-por-eng-07-pos.txt:	—/PUNCT Té/**FOC+PREP+SCONJR** remburi/V kiínha/N pirá/N resé/ADP+SCONJ ./PUNCT # — Não ponha pimenta no peixe.
	FOC+PREP+SCONJR	t-yrl-por-eng-08-pos.txt:	Ape/ADVDI+ADVJ ,/PUNCT paá/RPRT ,/PUNCT urubú/N usú/V sakakwera/ADP+N merupí/ADVA ,/PUNCT té/**FOC+PREP+SCONJR** mairamé/ADVNT+ADVRT+SCONJR umaã/V wirawasú/N ./PUNCT # Então, dizem que o urubu foi atrás dele devagar, quando viu o gavião.
	FOC+PREP+SCONJR	e-yrl-por-eng-07-pos.txt:	Té/**FOC+PREP+SCONJR** yapitá/V iké/ADVDX !/PUNCT # Não fiquemos aqui!
	FOC+PREP+SCONJR	e-yrl-por-eng-09-pos.txt:	Ixé/PRON té/**FOC+PREP+SCONJR** amunhã/V ana/PFV timbiú/N ./PUNCT # Eu mesmo fiz a comida.
Testset 1 e Testset 2	**IND+INT+RELF**		
	IND+INT+RELF	e-yrl-por-eng-10-pos.txt:	Awá/**IND+INT+RELF** igara/N taá/CQ nhaã/DEMS ?/PUNCT # Canoa de quem é aquela?

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
	IND+INT+RELF	t-yrl-por-eng-05-pos.txt:	Aikwé/EXST awá/**IND+INT+RELF** ururi/V indé/PRON ú/V reputari/V reyuri/V ne/PRON2 retama/N suí/ADP ne/PRON2 rupí/ADP ?/PUNCT # Houve quem a trouxesse ou você quis vir da sua terra por você (mesma)?
	IND+INT+RELF	e-yrl-por-eng-02-pos.txt:	Awá/**IND+INT+RELF** taá/CQ penhé/??? suí/ADP usú-putari/??? ?/PUNCT # Qual de vocês quer ir?
	IND+INT+RELF	e-yrl-por-eng-04-pos.txt:	Awá/**IND+INT+RELF** kwera/A+PRET ?/PUNCT # Quem era?
	IND+INT+RELF	t-yrl-por-eng-01-pos.txt:	Awá/**IND+INT+RELF** taá/CQ indé/PRON ?/PUNCT # Quem é você?
Testset 1 e Testset 2	**IND+NEG**		
	IND+NEG	e-yrl-por-eng-13-pos.txt:	Mira/N ramé/ADP+SCONJ maã/COND+IND+INT+N+REL+V indé/PRON ,/PUNCT indé/PRON nti maã/**IND+NEG** rexari/V ixé/PRON amanũ/??? ,/PUNCT indé/PRON resú-kwáu/??? maã/COND+IND+INT+N+REL+V reyuka/V meyú/N ixé/PRON ambaú/V arama/ADP+FUT+SCONJ ./PUNCT # Se você fosse gente, você não me deixaria morrer, você poderia ir arranjar beiju para eu comer.
Testset 1 e Testset 2	**INDQ+INT+TOT**		
	INDQ+INT+TOT	t-yrl-por-eng-05-pos.txt:	—/PUNCT Indé/PRON muíri/**INDQ+INT+TOT** akayú/N taá/CQ remurari/V iké/ADVDX kwá/ADVDX+DEMX+V tawa/N upé/ADP+N ?/PUNCT

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
	INDQ+INT+TOT	t-yrl-por-eng-03-pos.txt:	# — Você há quantos anos mora aqui nesta cidade? —/PUNCT Muíri**/INDQ+INT+TOT** rupí/ADP taá/CQ kwá/ADVDX+DEMX+V xirura/N ?/PUNCT # — Por quanto é esta calça?
	INDQ+INT+TOT	e-yrl-por-eng-02-pos.txt:	Muíri**/INDQ+INT+TOT** kuya/N taá/CQ aé/DEMSN+PRON uú/V ?/PUNCT # Quantas cuias ele bebeu?
	INDQ+INT+TOT	t-yrl-por-eng-12-pos.txt:	Nhaã/DEMS pukusawa/ADP+SCONJ muíri**/INDQ+INT+TOT** ara/N Maria/PROPN umupuranga/V suka/N+V ./PUNCT # Enquanto isso, cada dia Maria enfeita sua casa.
	INDQ+INT+TOT	e-yrl-por-eng-03-pos.txt:	Muíri**/INDQ+INT+TOT** rupí/ADP taá/CQ kwá/ADVDX+DEMX+V kamixá/N ?/PUNCT # Por quanto é esta camisa?
Testset 1 e Testset 2	**N+REL**		
	N+REL	e-yrl-por-eng-05-pos.txt:	Asaisú/V se/PRON2 maã-itá**/N+REL** ./PUNCT # Sovino minhas coisas.
	N+REL	y-yrl-por-eng-07-pos.txt:	Aé/DEMSN+PRON usaisú/V retana/ADVS Brasil/PROPN maã-itá**/N+REL** ./PUNCT # Ele amava muito as coisas do Brasil.
	N+REL	t-yrl-por-eng-03-pos.txt:	Ariré/ADVT kurí/FUT apiripana/V maã-itá**/N+REL** ixé/PRON arama/ADP+FUT+SCONJ ./PUNCT # Depois comprarei coisas para mim.

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
	N+REL	t-yrl-por-eng-06-pos.txt:	—/PUNCT Ixé/PRON aikupukú/V xinga/ADVS+INDQ apiripana/V arama/ADP+FUT+SCONJ maã-itá/**N+REL** indé/PRON arama/ADP+FUT+SCONJ ./PUNCT # — Eu demorei um pouco para comprar coisas para você.
	N+REL	t-yrl-por-eng-03-pos.txt:	Aé/DEMSN+PRON upiripana/V kurí/FUT maã-itá/**N+REL** i/PRON2 mimbira/N supé/ADP ,/PUNCT i/PRON2 mena/N supé/ADP yuíri/ADV+CCONJ+V ./PUNCT # Ela vai comprar coisas para seu filho, para seu marido também.
Testset 1 e Testset 2	**N+V**		
	N+V	e-yrl-por-eng-05-pos.txt:	Pedro/PROPN umaã/V kurí/FUT indé/PRON suka/**N+V** upé/ADP+N ./PUNCT # Pedro vai ver-te na casa dele.
	N+V	e-yrl-por-eng-08-pos.txt:	Upitá/V ,/PUNCT paá/RPRT ,/PUNCT suka/**N+V** upé/ADP+N ./PUNCT # Contam que ficou na sua casa.
	N+V	t-yrl-por-eng-12-pos.txt:	Maria/PROPN usú/V umupitá/V i/PRON2 amũ/N suka/**N+V** upé/ADP+N té/FOC+PREP+SCONJR akayú/N pisasú/A yupirungawa/N ./PUNCT # Maria vai hospedar sua irmã em sua casa até o começo do ano novo.
	N+V	e-yrl-por-eng-04-pos.txt:	Suka/**N+V** puranga/A+ADVA ./PUNCT # A casa dele é bonita.
	N+V	t-yrl-por-eng-12-pos.txt:	Suka/**N+V** urikú/V mukũi/CARD+INDQ ukapí/N ./PUNCT # Sua casa tem dois quartos.
Testset 1 e	**N+V2**		

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
Testset 2			
	N+V2	e-yrl-por-eng-09-pos.txt:	I/PRON2 kwema/**N+V2** ./PUNCT # Amanhece.
	N+V2	e-yrl-por-eng-09-pos.txt:	Ee/??? ruí/**N+V2** ./PUNCT # Eu sangro.
	N+V2	e-yrl-por-eng-09-pos.txt:	Pedro/PROPN tuí/**N+V2** ./PUNCT # Pedro sangra.
	N+V2	e-yrl-por-eng-09-pos.txt:	Aintá/PRON+PRON2 ruí/**N+V2** ./PUNCT # Eles sangram.
	N+V2	t-yrl-por-eng-09-pos.txt:	Ape/ADVDI+ADVJ i/PRON2 kwema/**N+V2** ara/N ./PUNCT # Então amanheceu o dia.
Testset 1 e Testset 2	**PRON+PRON2**		
	PRON+PRON2	t-yrl-por-eng-13-pos.txt:	Kuíri/ADVT supí/CERT+N aintá/**PRON+PRON2** uyumusurí-kwáu/V ./PUNCT # Agora de fato eles podem se divertir.
	PRON+PRON2	e-yrl-por-eng-11-pos.txt:	Upanhẽ/TOT urikú/V aintá/**PRON+PRON2** piá/N surí/A2 ./PUNCT # Todos tinham seus corações felizes.
	PRON+PRON2	t-yrl-por-eng-11-pos.txt:	Aintá/**PRON+PRON2** uwiyé/V yepé/ART+CARD+FRUST+SCONJ yupirisawa/N paranã/N ruakí/ADP ,/PUNCT asuí/ADVT+CCONJ aintá/**PRON+PRON2** uyuruari/V yepé/ART+CARD+FRUST+SCONJ igara/N puranga/A+ADVA upé/ADP+N ./PUNCT # Eles descem uma escada perto do rio, e embarcam numa bonita canoa.
	PRON+PRON2	e-yrl-por-eng-11-pos.txt:	Kunhã-itá/N tá/**PRON+PRON2** usemu/V tá/**PRON+PRON2** uikú/V kaá/N suí/ADP kuíri/ADVT ./PUNCT

Testsets	Ambiguidades	Arquivos[80]	Sentenças etiquetadas pelo Nheengatagger e sua tradução para o português
			# As mulheres estão saindo da mata agora.
	PRON+PRON2	e-yrl-por-eng-09-pos.txt:	Aintá/**PRON+PRON2** tá/**PRON+PRON2** akanhemu/V+V2 ./PUNCT
			# Eles se assustam.
Testset 1 e Testset 2	**TOT+TOTAL+V**		
	TOT+TOTAL+V	e-yrl-por-eng-07-pos.txt:	Kurumĩ/N umbaú/V pawa/**TOT+TOTAL+V** ./PUNCT
			# O menino come tudo.
	TOT+TOTAL+V	e-yrl-por-eng-11-pos.txt:	Suá/N i/PRON2 kiá/A2 pawa/**TOT+TOTAL+V** ./PUNCT
			# A cara dele está toda suja.
Testset 1 e Testset 2	**V+V2**		
	V+V2	e-yrl-por-eng-09-pos.txt:	Aintá/PRON+PRON2 tá/PRON+PRON2 akanhemu/**V+V2** ./PUNCT
			# Eles se assustam.
	V+V2	e-yrl-por-eng-09-pos.txt:	Penhẽ/PRON pe/CLADP+PRON2 akanhemu/**V+V2** ./PUNCT
			# Vocês se assustam.
	V+V2	e-yrl-por-eng-09-pos.txt:	Aé/DEMSN+PRON i/PRON2 akanhemu/**V+V2** ./PUNCT
			# Ele se assusta.
	V+V2	e-yrl-por-eng-09-pos.txt:	Ixé/PRON se/PRON2 akanhemu/**V+V2** ./PUNCT
			# Eu me assusto.
	V+V2	e-yrl-por-eng-09-pos.txt:	Yandé/PRON yané/PRON2 akanhemu/**V+V2** ./PUNCT
			# Nós nos assustamos.

APÊNDICE L – RESULTADOS DO TESTE 1

Ambiguidades	Sentenças	Ocorrências no *Testset*	Resolvidas	*Output* correto	VP	VN	FP	FN	Acurácia
A2+ADVS	5	5	4	4	4	1	0	0	0.02
A2+N	5	5	5	0	0	0	5	0	0.00
A+A2	3	3	3	3	3	0	0	0	0.01
A+A2+ADV	2	2	1	0	0	1	1	0	0.00
A+ADVA	5	5	5	3	3	0	2	0	0.01
A+ADVA+N	5	5	4	1	1	1	3	0	0.01
A+ADVC	4	4	0	0	0	4	0	0	0.02
A+ADVC+N	1	1	0	0	0	1	0	0	0.00
A+ADV+N	5	5	4	1	1	1	3	0	0.01
A+ADVS	3	3	1	1	1	2	0	0	0.01
A+ADVS+INDQ	5	5	3	3	3	2	0	0	0.02
ADP+ADVA+IND	5	5	4	1	1	1	3	0	0.01
ADP+ADVG	5	5	0	0	0	5	0	0	0.02
ADP+FUT+SCONJ	5	5	0	0	0	5	0	0	0.02
ADP+N	5	5	5	2	2	0	3	0	0.01
ADP+N+SCONJ	5	5	0	0	0	5	0	0	0.02
ADP+SCONJ	5	5	0	0	0	5	0	0	0.02
ADV+CCONJ+V	5	5	0	0	0	5	0	0	0.02
ADVDI+ADVJ	5	5	2	1	1	3	1	0	0.02
ADVDX+DEMX+V	5	5	3	0	0	2	3	0	0.01
ADVJ+ADVT	2	2	0	0	0	2	0	0	0.01
ADVJ+CCONJ	5	5	0	0	0	5	0	0	0.02

ADVJ+SCONJR	3	3	0	0	0	3	0	0	0.01
ADVLA+ADVRA+SCONJR	5	5	0	0	0	5	0	0	0.02
ADVLC+ADVNC+ADVRC	5	5	0	0	0	5	0	0	0.02
ADVLC+ADVRC	5	5	0	0	0	5	0	0	0.02
ADVM+ADVT	1	1	0	0	0	1	0	0	0.00
ADVNT+ADVRT+SCONJR	5	5	0	0	0	5	0	0	0.02
ADVO+ORD	1	1	0	0	0	1	0	0	0.00
ADVS+INDQ	5	5	2	1	1	3	1	0	0.02
ADVT+CCONJ	5	5	0	0	0	5	0	0	0.02
A+INDQ	3	3	3	1	1	0	2	0	0.00
A+INTJ+N	1	1	0	0	0	1	0	0	0.00
A+N	5	7	6	4	4	1	2	0	0.02
A+N+V2	2	2	2	0	0	0	2	0	0.00
A+PRET	5	5	0	0	0	5	0	0	0.02
ART+CARD+FRUST+SCONJ	5	6	0	0	0	6	0	0	0.03
A+V	3	3	3	0	0	0	3	0	0.00
A+V2	1	1	1	1	1	0	0	0	0.00
CARD+INDQ	5	5	3	0	0	2	3	0	0.01
CERT+N	1	1	0	0	0	1	0	0	0.00
CLADP+PRON2	5	5	0	0	0	5	0	0	0.02
COND+IND+INT+N+REL+V	5	5	0	0	0	5	0	0	0.02
DEMSN+PRON	5	5	0	0	0	5	0	0	0.02
FOC+NEGI	5	5	0	0	0	5	0	0	0.02
FOC+PREP+SCONJR	5	5	0	0	0	5	0	0	0.02
IND+INT+RELF	5	5	0	0	0	5	0	0	0.02
IND+NEG	1	1	1	0	0	0	1	0	0.00

INDQ+INT+TOT		5	5	0	0	0	5	0	0	0.02
N+REL		5	5	5	5	5	0	0	0	0.02
N+V		5	5	3	1	1	2	2	0	0.01
N+V2		5	5	3	0	0	2	3	0	0.01
PRON+PRON2		5	8	6	5	5	2	1	0	0.03
TOT+TOTAL+V		2	2	0	0	0	2	0	0	0.01
V+V2		5	5	3	3	3	2	0	0	0.02
	Total	**224**	**230**	**85**	**41**	**41**	**145**	**44**	**0**	**0.809**

APÊNDICE M – RESULTADOS DO TESTE 2

Ambiguidades	Sentenças	Ocorrências no *Testset*	Resolvidas	*Output* correto	VP	VN	FP	FN	Acurácia
A2+ADVS	5	5	5	5	5	0	0	0	0.02
A2+N	5	5	5	0	0	0	5	0	0.00
A+A2	3	3	3	3	3	0	0	0	0.01
A+A2+ADV	2	2	1	0	0	1	1	0	0.00
A+ADVA	5	5	5	5	5	0	0	0	0.02
A+ADVA+N	5	5	5	2	2	0	3	0	0.01
A+ADVC	4	4	3	0	0	1	3	0	0.00
A+ADVC+N	1	1	1	0	0	0	1	0	0.00
A+ADV+N	5	5	4	1	1	1	3	0	0.01
A+ADVS	3	3	1	1	1	2	0	0	0.01
A+ADVS+INDQ	5	5	4	3	3	1	1	0	0.02
ADP+ADVA+IND	5	5	4	1	1	1	3	0	0.01
ADP+ADVG	5	5	5	3	3	0	2	0	0.01
ADP+FUT+SCONJ	5	5	5	4	4	0	1	0	0.02
ADP+N	5	5	5	2	2	0	3	0	0.01
ADP+N+SCONJ	5	5	5	1	1	0	4	0	0.00
ADP+SCONJ	5	5	4	3	3	1	1	0	0.02
ADV+CCONJ+V	5	5	5	0	0	0	5	0	0.00
ADVDI+ADVJ	5	5	3	3	3	2	0	0	0.02
ADVDX+DEMX+V	5	5	3	2	0	2	3	0	0.01
ADVJ+ADVT	2	2	0	0	0	2	0	0	0.01
ADVJ+CCONJ	5	5	3	2	2	2	1	0	0.02

ADVJ+SCONJR	3	3	3	2	2	0	1	0	0.01
ADVLA+ADVRA+SCONJR	5	5	5	0	0	0	5	0	0.00
ADVLC+ADVNC+ADVRC	5	5	5	4	4	0	1	0	0.02
ADVLC+ADVRC	5	5	4	4	4	1	0	0	0.02
ADVM+ADVT	1	1	1	1	1	0	0	0	0.00
ADVNT+ADVRT+SCONJR	5	5	1	1	2	3	0	0	0.02
ADVO+ORD	1	1	0	0	0	1	0	0	0.00
ADVS+INDQ	5	5	4	2	2	1	2	0	0.01
ADVT+CCONJ	5	5	2	1	1	3	1	0	0.02
A+INDQ	3	3	3	1	1	0	2	0	0.00
A+INTJ+N	1	1	0	0	0	1	0	0	0.00
A+N	5	7	6	4	4	1	2	0	0.02
A+N+V2	2	2	2	0	0	0	2	0	0.00
A+PRET	5	5	3	1	1	2	2	0	0.01
ART+CARD+FRUST+SCONJ	5	6	6	1	1	0	5	0	0.00
A+V	3	3	3	0	0	0	3	0	0.00
A+V2	1	1	1	1	1	0	0	0	0.00
CARD+INDQ	5	5	3	1	1	2	2	0	0.01
CERT+N	1	1	1	0	0	0	1	0	0.00
CLADP+PRON2	5	5	0	0	0	5	0	0	0.02
COND+IND+INT+N+REL+V	5	5	5	1	1	0	4	0	0.00
DEMSN+PRON	5	5	3	3	3	2	0	0	0.02
FOC+NEGI	5	5	5	3	3	0	2	0	0.01
FOC+PREP+SCONJR	5	5	4	1	1	1	3	0	0.01
IND+INT+RELF	5	5	4	2	2	1	2	0	0.01
IND+NEG	1	1	1	0	0	0	1	0	0.00

INDQ+INT+TOT		5	5	5	3	3	0	2	0	0.01
N+REL		5	5	5	5	5	0	0	0	0.02
N+V		5	5	3	1	1	2	2	0	0.01
N+V2		5	5	3	0	0	2	3	0	0.01
PRON+PRON2		5	8	6	4	4	2	2	0	0.03
TOT+TOTAL+V		2	2	2	0	0	0	2	0	0.00
V+V2		5	5	3	3	3	2	0	0	0.02
	Total	**224**	**230**	**181**	**91**	**90**	**48**	**92**	**0**	**0.600**

APÊNDICE N – RESULTADOS DO TESTE 3

Ambiguidades	Sentenças	Ocorrências no *Testset*	Resolvidas	*Output* correto	VP	VN	FP	FN	Acurácia
A+A2	3	3	3	3	3	0	0	0	0.01
A+A2+ADV	2	2	1	0	0	1	1	0	0.00
A+ADV	12	12	9	7	7	3	2	0	0.05
A+ADV+INDQ	5	5	4	3	3	1	1	0	0.02
A+ADV+N	11	11	10	4	4	1	6	0	0.02
A+INDQ	3	3	3	2	2	0	1	0	0.01
A+INTJ+N	1	1	0	0	0	1	0	0	0.00
A+N	5	7	6	1	4	1	2	0	0.02
A+N+V2	2	2	2	0	0	0	2	0	0.00
A+PRET	5	5	3	1	1	2	2	0	0.01
A+V	3	3	3	0	0	0	3	0	0.00
A+V2	1	1	1	1	1	0	0	0	0.00
A2+ADV	5	5	5	5	5	0	0	0	0.02
A2+N	5	5	5	0	0	0	5	0	0.00
ADP+ADV	5	5	5	5	5	0	0	0	0.02
ADP+ADV+IND	5	5	4	1	1	1	3	0	0.01
ADP+FUT+SCONJ	5	5	5	5	5	0	0	0	0.02
ADP+N	5	5	5	4	4	0	1	0	0.02
ADP+N+SCONJ	5	5	5	1	1	0	4	0	0.00
ADP+PRON2	5	5	5	3	3	0	2	0	0.01
ADP+SCONJ	5	5	4	3	3	1	1	0	0.02
ADV+CCONJ	10	10	6	1	1	4	5	0	0.02

ADV+CCONJ+V	5	5	5	0	0	0	5	0	0.00
ADV+DEM+V	5	5	3	0	0	2	3	0	0.01
ADV+INDQ	5	5	4	4	4	1	0	0	0.02
ADV+ORD	1	1	1	0	0	0	1	0	0.00
ADV+SCONJR	13	13	11	6	6	2	5	0	0.04
ART+CARD+FRUST+SCONJ	5	6	6	1	1	0	5	0	0.00
CARD+INDQ	5	5	3	1	1	2	2	0	0.01
CERT+N	1	1	1	0	0	0	1	0	0.00
COND+IND+INT+N+REL+V	5	5	5	1	1	0	4	0	0.00
DEM+PRON	5	5	3	2	2	2	1	0	0.02
FOC+ADP+SCONJR	5	5	4	0	0	1	4	0	0.00
FOC+NEGI	5	5	5	3	3	0	2	0	0.01
IND+INT+RELF	5	5	4	2	2	1	2	0	0.01
IND+NEG	1	1	1	0	0	0	1	0	0.00
INDQ+INT+TOT	5	5	5	3	3	0	2	0	0.01
N+REL	5	5	5	5	5	0	0	0	0.02
N+V	5	5	3	1	1	2	2	0	0.01
N+V2	5	5	4	0	0	1	4	0	0.00
PRON+PRON2	5	8	6	4	4	2	2	0	0.03
TOT+TOTAL+V	2	2	2	0	0	0	2	0	0.00
V+V2	5	5	4	3	3	1	1	0	0.02
Total	**206**	**212**	**179**	**86**	**89**	**33**	**90**	**0**	**0,575**