Compilação de um corpus etiquetado da Língua Geral Amazônica

Dominick M. Alexandre, Juliana L. Gurgel, Leonel F. de A. Araripe

Universidade Federal do Ceará (UFC) – Fortaleza, CE – Brasil

domimaia@alu.ufc.br, julianalgurgel@alu.ufc.br, leonel@daad-alumni.de

Resumo. *Este trabalho apresenta as etapas de compilação de um corpus da Língua Geral Amazônica (LGA), ou nheengatu, desenvolvido para a posterior implementação de um etiquetador morfossintático para o sintagma nominal dessa língua. O estudo representa um avanço na construção de banco de dados para línguas indígenas e na inclusão dessas línguas minoritárias no atual contexto científico e tecnológico. Os resultados confirmam a aplicabilidade do corpus compilado para etiquetadores e outros algoritmos de processamento de linguagem natural.*

Abstract. *This work presents the stages of compilation of a corpus of the Amazonian Lingua Franca (LGA), or nheengatu, developed for the implementation of a part-of-speech tagger for the noun phrase of this language. The study represents an advance in the construction of a database for indigenous languages and the inclusion of these minority languages in the current scientific and technological context. The results confirm the applicability of the compiled corpus to POS taggers and other natural language processing algorithms.*

1. Introdução

Combinando Linguística e Ciência da Computação, o processamento de linguagem natural (PLN) é tradicionalmente considerado uma subárea da Inteligência Artificial no Brasil. A princípio, essa tecnologia permite o tratamento computacional dos diversos níveis da linguagem humana e o aperfeiçoamento da comunicação entre humanos e computadores [Guinovart 2000]. Contudo, o trabalho descritivo empreendido pelo PLN pode, ainda, contribuir para a preservação de línguas em risco de extinção, como a Língua Geral Amazônica (LGA), ou nheengatu, cujos bancos de dados disponíveis e inserção no cenário tecnológico são ínfimos ou inexistentes. Nesse contexto, o presente trabalho apresenta as etapas da compilação de um *corpus* da LGA, visando sua implementação no desenvolvimento de um etiquetador morfossintático para o sintagma nominal desta língua [Alencar 2020].

Em seu período de máxima difusão, em meados do século XVII, a LGA era falada do Maranhão à fronteira brasileiro-peruana. Atualmente, o nheengatu é falado pelos habitantes da região do Alto Rio Negro, na Amazônia [Navarro 2012]. Segundo o banco de dados *Ethnologue*, existem aproximadamente 6.000 falantes de nheengatu no Brasil, 8.000 na Colômbia e um número ínfimo na Venezuela [Eberhard, Simons e Fennig 2021]. Em apenas cinco anos, o número total de falantes diminuiu de 19.060

falantes em 2016 [Lewis, Simons e Fennig 2016], para 14.000 em 2021 [Eberhard, Simons e Fennig 2021].

Uma vez que a existência de corpora anotados morfossintaticamente é condição obrigatória para o desenvolvimento de tecnologias de processamento automático de textos, esta pesquisa, que faz parte de um projeto de Iniciação Científica, ainda em andamento, representa um primeiro passo na construção de um banco de dados do nheengatu voltado para o PLN [Alencar 2020]. Com este trabalho, pretendemos contribuir para a visibilidade do nheengatu e fornecer, para a comunidade científica, um *corpus* útil ao desenvolvimento de ferramentas de PLN e de pesquisas em diferentes áreas das ciências humanas, especialmente a linguística e a literatura [Alencar 2021].

2. Fundamentação teórica

Este trabalho envolve duas áreas da linguística: a linguística computacional e a linguística descritiva. Uma vez que o objetivo principal deste trabalho é a compilação de um *corpus* do nheengatu para a etiquetagem morfossintática do sintagma nominal (SN) de sentenças em nheengatu, uma parte do arcabouço teórico norteador da pesquisa tem caráter descritivo e documental, enquanto outra parte tem caráter computacional.

Neste trabalho, consideramos as descrições da estrutura do SN e das classes de palavras conforme Cruz (2011), mas adotamos a terminologia das classes do nheengatu segundo Navarro (2011), devido ao seu aspecto formal e simplificador. À luz das duas descrições gramaticais, inventariamos as seguintes classes do SN do nheengatu: nomes, adjetivos, pronomes pessoais, indefinidos, quantificadores, demonstrativos e numerais. Combinados, estes trabalhos oferecem caminhos para lidar com desafios que emergem da classificação de palavras da língua, como a ocorrência da posposição *upé* (*em*, em português), que, em alguns casos, assume as formas variantes *-pe* e *-me* (esta sempre após vogal nasal), afixadas aos substantivos, por exemplo: (i) *kunhã-itá uiku paranã upé* (*as mulheres estão no rio*, em português); e (ii) *paranãme igara upitá ana* (*no rio a canoa ficou*, em português). Além disso, algumas palavras podem ter mais de uma etiqueta morfossintática, como *iuaté* (*alto*, em português), que pode ocorrer como substantivo ou adjetivo. Uma maneira de lidar com estes e outros desafios é implementar regras que modelem computacionalmente esses fenômenos.

A etapa computacional do trabalho diz respeito ao pré-processamento de textos eletrônicos e à construção de uma versão piloto do etiquetador. A primeira tarefa na construção de um *corpus* para etiquetagem consiste na tokenização, isto é, a segmentação do texto em unidades processáveis por máquinas, denominadas *tokens* [Mikheev 2004]. Um etiquetador morfossintático é uma ferramenta de PLN cuja função é atribuir uma etiqueta morfossintática a cada *token* (ou palavra) de um texto dado como entrada e retorna como saída o mesmo texto anotado morfossintaticamente [Jurafsky e Martin 2019]. Em geral, a construção dessa ferramenta pode ser feita a partir de duas abordagens: (i) baseada em dados ou probabilística, a mais comum, que consiste na utilização de corpora anotados para o treinamento de modelos com base no contexto e na frequência das etiquetas; e (ii) baseada em regras, que consiste na implementação de regras com base nas descrições gramaticais da língua [Voutilainen 2004]. Neste trabalho, utilizamos uma abordagem baseada em regras devido à inexistência de corpora anotados do nheengatu [Alencar 2020].

3. Metodologia

A compilação do *corpus* foi dividida em duas etapas: (i) a compilação dos textos e exemplos das lições do *Curso de Língua Geral* [Navarro 2011] a partir de um arquivo tokenizado do livro; e (ii) a compilação do glossário de Navarro (2011) em uma tabela passível de conversão para a estrutura de dados *Dictionary*, da linguagem de programação Python. Devido à complexidade da segunda etapa, esta foi subdividida em três partes: (i) revisão das classes de palavras; (ii) pré-processamento para extração; (iii) extração e finalização.

A compilação dos textos e exemplos de Navarro (2011) consistiu na extração manual das sentenças selecionadas e na sua compilação em arquivos de texto à parte, utilizando a codificação UTF-8. Em relação à compilação do glossário, na parte (i), revisamos as entradas lexicais do glossário e listamos verbetes cujas classificações eram ausentes ou incompatíveis com a descrição gramatical do livro ou com a versão mais recente do livro-texto de Navarro, publicada em 2016. Na parte (ii), realizamos a conversão de caracteres especiais, por meio de um programa, denominado "replace-char.py". Em seguida, modificamos ou adicionamos as classes de palavras nas entradas lexicais listadas na parte (i). Na parte (iii), extraímos as entradas lexicais que ocorrem no sintagma nominal do nheengatu. Em seguida, geramos tabelas de duas colunas por meio de um programa, denominado "tag-words.py". A primeira coluna contém a entrada lexical, que constitui uma chave e, a segunda, a etiqueta correspondente à sua classe gramatical, que constitui um valor atribuído à chave (ver Figura 1). Depois, expandimos a lista de nomes com as formas flexionadas no plural por meio de um programa, o qual denominamos "nominal-flexionizer-yrl.py". Assim, adicionamos o sufixo de plural a todos os substantivos definidos por Navarro (2011), isto é, neutros, masculinos e femininos.

Na Figura 1, temos um exemplo de como uma tabela de nomes se parece antes e depois do processamento pelo programa flexionador. Do lado esquerdo, apresentamos a tabela gerada pelo programa "tag-words.py" e, do lado direito, apresentamos a tabela gerada pelo "nominal-flexionizer-yrl.py".

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

Figura 1. Tabela de nomes do nheengatu antes e após o processamento pelo programa nominal-flexionizer-yrl.py.

Uma vez geradas, cada uma dessas tabelas pode ser convertida para uma estrutura de dados *Python Dictionary*. A construção da versão beta do etiquetador foi feita por meio da implementação de uma função capaz de aplicar essa estrutura de dados aos *tokens* de um texto recebido como entrada no etiquetador. Para cada *token*, é atribuída uma etiqueta, desde que a palavra esteja contida no dicionário; caso contrário, o programa retorna a palavra, sem anotação [Alencar 2020]. Para testar a ferramenta, realizamos um primeiro teste a fim de verificar a aplicabilidade do *corpus* compilado para utilização em tarefas de PLN e identificar os aspectos do algoritmo a serem

aperfeiçoados. No teste, a versão beta do etiquetador recebeu como entrada um arquivo contendo as sentenças do texto da primeira lição de Navarro (2011), que perfazem um total de 16 sentenças e 74 palavras, entre itens que ocorrem ou não no sintagma nominal.

4. Conclusões

Além do *corpus*, composto por sentenças do nheengatu extraídas de Navarro (2011), esta etapa da pesquisa resultou em mais dois produtos úteis à aplicação em tarefas de PLN, conforme Tabela 1: (i) um dicionário do tipo *Python Dictionary*, contendo as entradas lexicais do nheengatu e suas respectivas etiquetas morfossintáticas; e (ii) um conjunto de etiquetas das classes que ocorrem no sintagma nominal da LGA.

Tabela 1. Produtos da pesquisa

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

Como resultado do teste, a versão beta do etiquetador obteve uma acurácia de 100%. Vale ressaltar, contudo, que todos os itens pertencentes ao sintagma nominal presentes no arquivo de teste constavam no dicionário utilizado e que a parcela do *corpus* testada representa um escopo bastante limitado do banco de dados. Portanto, a acurácia alcançada neste primeiro teste é meramente ilustrativa, servindo apenas ao propósito de verificar a aplicabilidade do *corpus* compilado para a construção de ferramentas voltadas para o PLN e de identificar os erros do algoritmo que precisam ser corrigidos a seguir.

Para trabalhos futuros, cumpre ampliar o banco de dados do nheengatu através da compilação de textos de outras obras, como Casasnovas (2006). Além disso, é preciso aperfeiçoar o algoritmo do etiquetador e testá-lo com relação ao restante do *corpus* compilado. Paralelamente, uma pesquisa de mestrado, ainda em andamento, objetiva a construção de uma ferramenta capaz de etiquetar sentenças inteiras do nheengatu [Gurgel 2021]. Todos os produtos da presente pesquisa, por sua vez, estão sendo gradualmente disponibilizados sob licença livre para comunidade acadêmica na plataforma GitHub.

Referências

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

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

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

Cruz, A. (2011). “Fonologia e Gramática do Nheengatú: A língua falada pelos povos Baré, Warekena e Baniwa”. Utrecht: LOT.

Eberhard, D. M.; Simons, G. F.; Fennig, C. D. (org.). (2021). “Ethnologue: Languages of the World”. 24. ed. Dallas: SIL International. Disponível em: http://www.ethnologue.com. Acesso em: 04 jul. 2021.

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

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

Jurafsky, D.; Martin, J. H. (2009). “Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition”. 2. ed. Upper Saddle River: Prentice Hall.

Lewis, M. P.; Simons, G. F.; Fennig, C. D. (org.). (2016). “Ethnologue: Languages of the World”. 19. ed. Dallas: SIL International. Disponível em: http://www.ethnologue.com. Acesso em: 04 jul. 2021.

Mikheev, A. (2004). “Text segmentation”. In: Mitkov, R. (Org.). *The Oxford handbook of computational linguistics*. Oxford, Oxford University Press, p. 209-221.

Navarro, E. D. A. (2011). “Curso de Língua Geral (Nheengatu ou Tupi moderno): A Língua das origens da civilização amazônica”. São Bernardo do Campo: Paym Gráfica e Editora.

Navarro, E. D. A. (2012). “O último refúgio da língua geral no Brasil”. *Estudos Avançados*, v. 26, p. 245-254.

Voutilainen, A. (2004). “Part-of-speech tagging”. In: Mitkov, R. (Org.). *The Oxford handbook of computational linguistics*. Oxford, Oxford University Press, p. 219-232.