Aspectos do desenvolvimento de um etiquetador morfossintático da língua Asuriní do Trocará

Izabel Nunes Dias[1], Thiago Blanch Pires[1]

[1]Instituto de Letras – Universidade de Brasília (UnB)
Campus Universitário Darcy Ribeiro – 70910-900 – Brasília – DF – Brasil

izabeldnunes@gmail.com, pirestb@unb.br

Abstract. *This article addresses the preservation of indigenous languages and the development of Natural Language Processing (NLP) tools for linguistic analysis, with a focus on the Asuriní do Trocará language. Based on a compiled corpus, the research implemented a computational morphosyntactic tagger to identify arguments and predicates in the language. Linguistic patterns, manual labeling and data processing with Python were explored. After evaluation, the tagger was submitted to the F-Score calculation, suggesting the need for future optimization.*

Resumo. *Este artigo aborda a preservação das línguas indígenas e o desenvolvimento de ferramentas de Processamento de Linguagem Natural (PLN) para análise linguística, com foco na língua Asuriní do Trocará. Com base em um corpus compilado, a pesquisa implementou um etiquetador morfossintático computacional para identificar argumentos e predicados na língua. Nesta pesquisa, foram explorados padrões linguísticos, etiquetagem manual e processamento de dados com Python. Após avaliação, o etiquetador foi submetido ao cálculo do F-Score, sugerindo a necessidade de otimização futura.*

1. Introdução

Os povos e as línguas indígenas são considerados patrimônio imaterial da humanidade pela UNESCO (Organização das Nações Unidas para a Educação, a Ciência e a Cultura), pois a língua de um povo carrega consigo a identidade, os mais variados conhecimentos e características do grupo que a utiliza. Por isso, a UNESCO estabeleceu que de 2022 a 2032 será a década internacional das línguas indígenas, visando enfocar o olhar mundial para o rápido desaparecimento de muitas línguas indígenas e para a importância de preservá-las para as próximas gerações. Atualmente, o mundo está virtualmente hiperconectado, de forma que a presença de uma língua no espaço digital pode garantir a preservação e perpetuação dela. Sendo assim, esta pesquisa é parte de um projeto que tem por objetivo criar e desenvolver recursos voltados para a análise linguística, tais como corpora, etiquetadores e analisadores de diferentes níveis linguísticos em línguas indígenas.

Considerando a necessidade de preservação das línguas indígenas e reconhecendo também a importância da presença destas no ciberespaço, o presente trabalho buscou estabelecer um catálogo de etiquetas dos argumentos e predicados da língua indígena Asurini do Trocará, também conhecida como Asuriní do Tocantins, a partir do desenvolvimento computacional de um etiquetador morfossintático de textos

de um corpus já compilado do livro “Relatos Asurini 2" [Asuriní et al. 2007] em língua natural.

A língua Asuriní do Tocantins, enquanto língua flexiva, possui muitas derivações, incorporações e composições. Os seus diferentes morfemas e suas respectivas posições impactam significativamente nos conceitos e contextos. Dentre as categorias morfossintáticas de palavras, os argumentos e predicados são de suma importância para o processo de etiquetagem, pois modificam ou são modificados especialmente por nomes, verbos e posposições [Cabral et al. 2011]. A partir do conhecimento dessas classificações, foi possível pensar e colocar em prática o processo de digitalização da língua.

Assim, fez-se necessário contribuir para sanar essas deficiências, compilando um corpus da língua Asuriní do Tocantins, e implementando computacionalmente um etiquetador a ser aplicado nesse corpus. Para testar a ferramenta, foi extraída uma amostra aleatória das sentenças do corpus corrigida manualmente. Conforme a hipótese que queríamos confirmar, esperávamos obter nesse teste um índice *F-Score* igual ou próximo a 0.95. O *F-score* é uma métrica que combina precisão e cobertura, variando entre 0 e 1, sendo que valores próximos de 1 indicam alto desempenho. Assim, buscar um *F-score* de 0,95 significa almejar que o sistema acerte a maioria dos casos relevantes e evite erros de classificação. Essa meta reflete o padrão de qualidade esperado em sistemas de Processamento de Linguagem Natural [Bird et al. 2009].

2. Referencial teórico

Durante todo o processo de trabalho, foi necessária a consulta de diversas pesquisas anteriores sobre o assunto, tanto na parte relacionada à língua Asuriní quanto à parte computacional da etiquetagem.

Para a compreensão da língua Asuriní algumas obras foram de extrema relevância, começando com o "Dicionário Asuriní do Tocantins - Português" [Cabral and Rodrigues 2003], obra essencial para a categorização e decisões de cada etiqueta dos termos encontrados no "Livro de Relatos Asuriní 2” [Asuriní et al. 2007]. Esse dicionário traz explicações sobre regras gramaticais da língua Asuriní, o significado de temas, palavras e aplicações de uso com diversos exemplos. Já o livro base para o corpus, "Livro de Relatos Asuriní 2” [Asuriní et. al. 2007], contém diversos relatos das histórias do povo Asuriní, contadas por eles mesmos e escritas em Asuriní e com tradução livre feita pelos Asuriní, tendo ao todo sete contos.

Outra obra que foi importante para o entendimento de argumento e predicado na língua Asuriní foi o artigo “Argumento e Predicado em Tupinambá” [Rodrigues 2011]. Sendo o Tupinambá uma língua irmã do Asuriní, foi possível extrair alguns conhecimentos aplicáveis também ao Asuriní, principalmente por explicar sobre as estruturas gramaticais que acompanham argumento e predicado e como identificá-los. Também foi utilizada a obra "Esboço Gramatical do Asuriní do Tocantins." [Cabral et al. 2011], que explica sobre aspectos morfológicos e sintáticos da língua.

Neste trabalho, argumento é considerado todo elemento nominal ou verbal que exerce funções nucleares na oração — sujeito de verbo transitivo (A) ou intransitivo (S), objeto direto (O) ou objeto de posposição — e que, morfologicamente, é identificado pela presença do caso argumentativo (-a ou Ø, que significa ausência). Essa marcação distingue elementos que funcionam como argumentos de circunstanciais e predicados.

Sem o caso argumentativo, nomes tendem a atuar como vocativos e verbos permanecem como predicados [Rodrigues 2001, Cabral et al. 2013].

O predicado, por sua vez, é o núcleo que expressa o evento, ação ou estado na oração. Verbos predicativos recebem prefixos pessoais que indexam o sujeito (e, em verbos transitivos, também o objeto direto), enquanto nomes podem funcionar como predicados nominais quando ocorrem sem marcação de caso argumentativo. Assim, a oposição entre argumento e predicado é estabelecida pela interação entre morfologia de caso, marcas pessoais e função sintática, não sendo determinada apenas pela classe lexical [Rodrigues 2001].

Com relação à parte computacional da pesquisa, os trabalhos de Pimentel (2021) e Silvério (2021), os quais integraram os grupos de pesquisa CompLin e GeLinC, e que criaram etiquetadores de sintagma nominal e verbal para a língua Asuriní, foram de suma importância para o desenvolvimento desta pesquisa. Também deve ser mencionada aqui a pesquisa realizada por Alexandre et al. (2021), o artigo "Nheentiquetador: um etiquetador morfossintático para o sintagma nominal do nheengatu", base para os métodos da presente pesquisa. No trabalho desses autores foi desenhado o primeiro etiquetador morfossintático para o sintagma nominal da Língua Geral Amazônica (LGA), ou Nheengatu, construído a partir de um corpus anotado nesta língua.

3. Métodos

Dentro do campo do processamento de linguagem natural, existem dois enfoques primordiais utilizados na construção dessa ferramenta: (i) a abordagem fundamentada em conhecimento e (ii) a abordagem orientada por dados [Voutilainen 2004, Duchier and Parmentier 2015]. O segundo método mencionado engloba o desenvolvimento de modelos estatísticos por meio do treinamento em corpora anotados por peritos humanos, utilizando algoritmos de aprendizado de máquina. Considerando a ausência desses dados para a língua Asuriní, a única alternativa é adotar a primeira abordagem, que se baseia na criação de regras com embasamento em descrições gramaticais da língua.

É importante salientar que a metodologia atual é fundamentada na abordagem utilizada para desenvolver o analisador morfossintático para a língua Nheengatu "NHEENTIQUETADOR: Um Etiquetador Morfossintático Para o Sintagma Nominal do Nheengatu." [Freitas et al. 2021]. Esta abordagem está sendo levada adiante por pesquisadores e estudantes da Universidade Federal do Ceará, pertencentes ao grupo de pesquisa denominado CompLin, que também abarca este projeto e seus participantes.

A etapa inicial da pesquisa consistiu em conduzir uma análise abrangente da literatura que explorou a estrutura linguística da língua Asuriní do Tocantins, com o suporte do Laboratório de Línguas e Literaturas Indígenas - LALLI/UnB. Esse processo teve como objetivo estabelecer um conjunto de etiquetas que correspondiam aos argumentos e predicados que podem se apresentar por meio de prefixos pessoais ou relacionais, a depender do modo de maneira sufixal, de acordo com os tempos, entre outros. O próximo passo envolveu a extração dos elementos pertencentes a essas classes e a criação de um glossário a partir do Livro de Relatos Asuriní 2 [Asuriní et al. 2007].

Quando necessário, foram utilizados dicionários, glossários e gramáticas disponíveis da língua, seguindo a metodologia delineada por Alencar (2021) e Alexandre et al. (2021). Isso permitiu a execução do passo seguinte, que consiste em

adicionar as informações aos itens de acordo com a transitividade dos verbos, modos, tempos, pessoas, e afixos que formam os componentes linguísticos que se pretende analisar.

Posteriormente, esses itens passaram por um processo de tokenização e normalização. Em seguida, a pesquisa definiu os tipos de informações linguísticas que as entradas desses recursos apresentaram como conjunto de dados (corpus). A partir daí, iniciou-se a criação de uma estrutura de dados no estilo de um dicionário (*dictionary*) dentro do contexto da linguagem de programação *Python*. Por fim, esssa pesquisa desenvolveu uma função essencial para aplicar o dicionário a cada *token* presente em um texto de entrada. Para cada um desses elementos, a função forneceu o item/etiqueta caso o item estivesse incluído no dicionário; caso contrário, retornava o próprio item sem qualquer anotação.

Finalmente, o etiquetador que havia sido desenvolvido foi avaliado por meio da análise de um conjunto de teste denominado *TEST-SET*. Esse conjunto consistiu em uma seleção aleatória das sentenças que compunham um corpus originado a partir dos textos presentes no Livro de Relatos Asuriní 2 [Asuriní et al. 2007]. Para medir a eficácia da ferramenta, utilizamos a métrica *F-Score*, comumente utilizada para avaliar a qualidade e o desempenho de sistemas de classificação. Essa métrica é baseada nos índices de precisão (P) e recall (R) por meio da formula (2PR)/(P+R), onde P=TP/(TP +FP) e R=TP/(TP+FN), sendo TP=total de positivos verdadeiros, FP=total de falsos positivos, FN=total de falsos negativos [Bird et al. 2009]. Nesta pesquisa foi utilizada a função *f1_score()* da biblioteca *Sklearn.metrics*.

Dentro da primeira etapa foi possível definir alguns padrões de classificação de argumentos e predicados. Todos estes padrões foram estabelecidos de acordo com as explicações do "Dicionário Asuriní do Tocantins - Português" [Cabral and Rodrigues 2003], e do artigo “Argumento e Predicado em Tupinambá” [Rodrigues 2011]. Primeiramente, esses padrões foram anotados em planilha no *MS Excel*, de acordo com a Tabela 1, e posteriormente implementados em código *Python*. O sistema foi implementado a partir de 17 padrões de reconhecimento construídos com base nas formas e terminações observadas no corpus Asuriní. Esses padrões foram concebidos para classificar cada token em três categorias:

ARG (argumento) — tokens que apresentavam sufixos ou padrões compatíveis com o caso argumentativo (-a ou Ø) ou que, pelo uso atestado no corpus, exerciam função nuclear.

PRED (predicado) — tokens que, pela ausência de marca argumentativa e/ou pela presença de radicais verbais recorrentes no corpus, indicavam função de núcleo predicativo.

NONE — tokens que não se enquadram nas categorias anteriores, como partículas, advérbios ou elementos sem marcação morfológica compatível.

Optou-se por uma abordagem padrão-lexical, sem análise morfológica completa ou verificação automática de prefixos pessoais e dependências sintáticas, em razão do tamanho reduzido do corpus e do objetivo exploratório da pesquisa. Dessa forma, as regras implementadas representam uma simplificação prática dos critérios morfossintáticos estabelecidos na literatura, permitindo avaliar a viabilidade da tarefa de

rotulação automática e identificar pontos de aprimoramento para futuras implementações.

Tabela 1. Amostra de identificação de argumento e predicado

Afixo/tema	*Padrão*	*Categoria*	*Etiqueta*	*Obs.*
Sufixo	-PE	Caso locativo pontual	NONE	Não será nem argumento, nem predicado
Sufixo	-IMO	Caso locativo difuso	NONE	
Sufixo	-I	Caso locativo situacional	NONE	
Sufixo	- y'ým	Negação	PRED	
Sufixo	- e'ýma	Negação de gerúndio	PRED	
Sufixo	- y'ýw	Negação	PRED	
Sufixo	-A	caso argumentativo	ARG	
Prefixo	A-	Série 1 do modo indicativo	PRED	
Prefixo	O-	Série 1 do modo indicativo	PRED	
Prefixo	PE-	Série 1 do modo indicativo	PRED	
Prefixo	ERE-	Série 1 do modo indicativo	PRED	
Prefixo	SA-	Série 1 do modo indicativo	PRED	
Prefixo	ORO-	Série 1 do modo indicativo	PRED	
Sufixo	- IHÍ	Negação de indicativo	PRED	
Sufixo	- RAPO	Proibitivo/Advertência	PRED	
Sufixo	- REME	Proibitivo/Advertência	PRED	
Prefixo	- MO -	Causativo aumento de valência	PRED	

Ao observar a tabela, percebe-se que os três primeiros padrões apenas estabelecem aquilo que não será argumento nem predicado, recebendo, portanto, a etiqueta *NONE*, que dentro do código *Python* irá indicar tudo aquilo que não foi possível identificar, pela dificuldade de abranger toda a língua, ou não se cabe identificar, por não se tratar de argumento ou predicado, que é o escopo deste trabalho.

Em seguida, a tabela apresenta os sufixos de negação que acompanham os verbos, sendo "-e'ýma" para verbos no gerúndio. Por sempre acompanhar um verbo, esse padrão recebe a etiqueta "PRED", que indica predicado. Depois, fica estabelecido que as palavras com o sufixo –a em sua maioria devem ser classificadas como argumento, recebendo, assim, a etiqueta "ARG".

Seguindo a ordem da Tabela 1, logo após vem a série 1 do modo indicativo, que marca as pessoas dentro do tema verbal [Cabral and Rodrigues 2003]. Portanto, esses prefixos indicam o verbo e recebem a etiqueta "PRED". Após isso, a Tabela 1 mostra sufixos de negação de indicativo e sufixos proibitivos - advertências que irão acompanhar predicados, recebendo, então, a etiqueta "PRED".

Por fim, temos o padrão "-MO-", que aumenta a valência dos verbos [Cabral et al. 2011]. Por isso, recebe a etiqueta "PRED" de predicado. O passo seguinte, depois da anotação dos padrões e etiquetas, foi desenvolver o etiquetador utilizando a linguagem

Python, por possuir uma vasta e pública documentação. Foram utilizadas as bibliotecas *Pandas, NLTK* e *RE*.

Utilizando o corpus compilado do *Livro Relatos Asuriní 2* [Asuriní et al. 2007] em formato *.txt*, através da biblioteca *Pandas*, o conteúdo do livro foi padronizado através da remoção de caracteres especiais e utilização de todo o texto em letras minúsculas. Após a limpeza, o corpus foi dividido e armazenado em uma lista de palavras em ordem alfabética. O código pode ser visto na Figura 1 a seguir.

```
import nltk
import re
from nltk.tag import RegexpTagger
from nltk.tag import DefaultTagger


text = open('relatos_asurini.txt', 'r', encoding='utf-8')
text = text.read()
text_lower = text.lower()
text_clean = re.sub('[.;,-:!?—()]','',text_lower)
set_list = list(set(text_clean.split()))
set_list.sort()
set_list
```

Figura 1. Processamento e limpeza do corpus

Os padrões e etiquetas mencionados anteriormente foram implementados por meio de código utilizando o padrão *RegexpTagger* da biblioteca *NLTK*, apresentado na Figura 2. Em seguida, foi utilizada a função *tag()* da biblioteca *NLTK* para implementar a etiquetagem propriamente dita do livro, utilizando o método da Figura 3.

```
patterns = [
            (r'.*pe$', 'None'),
            (r'.*imo$', 'None'),
            (r'.*i$', 'None'),
            (r".*y'ým.*", 'PRED'),
            (r".*e'ým.*", 'PRED'),
            (r".*y'ýw.*", 'PRED'),
            (r'.*a$', 'ARG'),
            (r'a.*','PRED'),
            (r'o.*','PRED'),
```

Figura 2. Padrões de etiquetas

```
nom_tagger = nltk.RegexpTagger(patterns)
tags = nom_tagger.tag(set_list)
tags = dict(tags)

for key, value in tags.copy().items():
    if value is None:
        tags[key] = 'NONE'
```

Figura 3. Funções de etiquetagem

Já na última etapa, o primeiro conto, “Cobra Coral”, do Livro de Relatos Asuriní 2 [Asuriní et al. 2007] foi etiquetado manualmente, criando assim um pequeno corpus "padrão ouro", com 42 *tokens*. O corpus total da pesquisa incluía os sete relatos em Asuriní, sem suas traduções, totalizando 433 *tokens*.

4. Resultados

Tendo os padrões regulares estabelecidos, o corpus devidamente limpo, foi possível fazer a etiquetagem de 433 *tokens*. Ao todo foram definidos dezessete padrões com três etiquetas, sendo "ARG" para argumento (sendo 145 casos = 33.5% do corpus), "PRED" para predicado (sendo 132 casos = 30.5% do corpus) e "NONE" para tudo o que não for argumento ou predicado (sendo 156 casos = 36% do corpus). Os padrões estabelecidos foram escolhidos por meio de estudos em gramáticas [Harrison 1975, Cabral 2011] e dicionário linguístico [Cabral and Rodrigues 2003] da língua Asuriní do Tocantins.

A partir do pequeno corpus nomeado "padrão ouro", do conto “Cobra Coral”, foi possível calcular o *F-score*. O resultado obtido, como mostra a Figura 4, foi de 0.64. O esperado era obter um *F-Score* próximo de 0.95, contudo o valor obtido foi bem abaixo do esperado. A simplificação metodológica, embora necessária para viabilizar a implementação do etiquetador com os recursos disponíveis, contribuiu para que o desempenho do etiquetador ficasse abaixo do F-score esperado para sistemas de etiquetagem em tarefas semelhantes (≈ 0.95 em corpora amplos e para línguas com mais recursos). A ausência de verificação sistemática de todos os marcadores de caso argumentativo, a não detecção automática de prefixos pessoais e a limitação da cobertura a formas atestadas no corpus aumentaram a ocorrência de falsos positivos e falsos negativos, refletindo-se diretamente nas métricas de avaliação.

```
f_score = f1_score(padrao_ouro, tags, average='weighted')
print("F-Score:", f_score)
```

```
F-Score: 0.6437543133195308
```

Figura 4. Obtenção do F-Score

Os arquivos com os códigos desenvolvidos neste trabalho, os corpora utilizados, assim como os resultados e as tabelas com as etiquetas desenvolvidas estão disponibilizados no seguinte repositório online: https://github.com/IzabelDiasNunes/EtiquetagemAsurini/

5. Considerações finais

Em suma, esta pesquisa atingiu o objetivo de estabelecer um catálogo de etiquetas dos argumentos e predicados da língua Asuriní do Tocantins, a partir do desenvolvimento computacional de um etiquetador morfossintático de textos em língua natural. O que foi compreendido como argumento ou predicado dentro dos parâmetros desta investigação foi determinado com base em análises e pesquisas conduzidas principalmente por meio de estudos em gramáticas [Harrison 1975, Cabral 2011] e um dicionário linguístico [Cabral and Rodrigues 2003] que descreve a língua de estudo. Consequentemente, após examinar essa descrição, a decisão foi considerar palavras que contivessem afixos que sinalizassem a ideia de predicação, como por exemplo a palavra "opotáreté" que é um verbo cujo sufixo o- indica predicação, ou a presença de um caso argumentativo em sua composição, como por exemplo a palavra "kosoá" que significa “mulher” tendo o sufixo –a indicando o caso argumentativo, sendo registradas conforme as anotações presentes no dicionário que foi objeto de pesquisa.

A etiquetagem foi realizada manualmente e, posteriormente, foi efetuado um processamento computadorizado por meio de um código implementado em *Python*, que

se fundamentou em expressões regulares e suas etiquetas. Contudo, não foi possível alcançar o valor almejado para o *F-Score* da etiquetagem deste trabalho. O valor alcançado foi 0.64. Por isso, recomenda-se para trabalhos futuros o melhoramento das etiquetas e a otimização do etiquetador para se aproximar da métrica esperada.

Agradecimentos

Agradecemos a valiosa contribuição e apoio dos membros do Laboratório de Línguas e Literatura Indígenas – LALLI/UnB, e dos recursos advindos da FAP-DF e CNPq.

Referências

Alencar, L. F. de. (2024). Uma gramática computacional de um fragmento do nheengatu. *Revista de Estudos da Linguagem*, [S. l.], v. 29, n. 3, p. 1717–1777.

Alexandre, D. M.; Araripe, L. F. A.; Gurgel, J. L. (2021). Nheentiquetador: um etiquetador morfossintático para o sintagma nominal do nheengatu. *Revista Encontros Universitários da UFC*, XL Encontro de Iniciação Científica, v. 6, n. 2.

Asuriní, M.; Rodrigues, A. D.; Cabral, A. S. A. C. (2007). *Livro de Relatos Asurini 2*. Brasília: LALI/UnB.

Bird, S.; Klein, E.; Loper, E. (2009). *Natural language processing with Python: analyzing text with the Natural Language Toolkit*. Sebastopol, CA: O'Reilly.

Cabral, A. S. A. C. (2016). Prefixos relacionais no Asuriní do Tocantins. *Moara – Revista Eletrônica do Programa de Pós-Graduação em Letras*, v. 2, n. 8, p. 7–24.

Cabral, A. S. A. C.; Silva, A. P. do C. e; Sousa, S. A. (2013). Expressão do caso argumentativo em três línguas Tupí-Guaraní: Asuriní do Tocantins, Avá-Canoeiro e Zo'é. In: SILEL, v.3, n.1. *Anais*. Uberlândia: EDUFU.

Cabral, A. S. A. C. et al. (2011). Esboço gramatical do Asuriní do Tocantins. In: *Contribuições para o inventário da língua Asuriní do Tocantins*. [S.l.: s.n.], p. 25–35.

Cabral, A. S. A. C.; Rodrigues, A. D. (2003). *Dicionário Asuriní do Tocantins–Português*. Belém: Universidade Federal do Pará.

Duchier, D.; Parmentier, Y. (2015). High-level methodologies for grammar engineering, introduction to the special issue. *Journal of Language Modelling*, v. 3, n. 1, p. 5-19.

Freitas, L.; Alexandre, D.; Gurgel, J.; Alencar, L. (2021). Nheentiquetador: um etiquetador morfossintático para o sintagma nominal do nheengatu. *Encontros Universitários da UFC*, [S. l.], v. 6, n. 2, p. 1481. Disponível em: https://periodicos.ufc.br/eu/article/view/74604 .

GLOBAL action plan of the International Decade of Indigenous Languages (IDIL 2022–2032) – UNESCO Digital Library. (2023). Disponível em: https://unesdoc.unesco.org/ark:/48223/pf0000379851 . Acesso em: 14 jul.

Harrison, C. H. (1975). *Gramática Asurini: aspectos de uma gramática transformacional e discursos monologados da língua Asurini, família tupi-guarani*. Série Linguística, v. 4. Brasília: Summer Institute of Linguistics.

Pimentel, C. (2022). Avaliação de um etiquetador automático para sintagmas verbais da língua Asuriní do Tocantins. In: *Congresso de Iniciação Científica da UnB e Congresso de Iniciação Científica do DF*, Brasil. Disponível em: https://conferencias.unb.br/index.php/iniciacaocientifica/28CICUnB19df/paper/view/43455 . Acesso em: 14 jul. 2023.

Rodrigues, A. D. (2011). Argumento e predicado em Tupinambá. *Revista Brasileira de Linguística Antropológica*, v. 3, n. 1, p. 93–102.

Rodrigues, A. D. (2001). *Sobre a natureza do caso argumentativo*. In: QUEIXALOS, Francesc (resp.), Des Noms et des Verbes en Tupí-Guaraní: État de la Question. Studies in Native America Linguistics. München: LINCOM.

Silvério, P. (2023). Um etiquetador para sintagmas nominais da língua Asuriní do Tocantins. In: *Congresso de Iniciação Científica da UnB e Congresso de Iniciação Científica do DF*, Brasil. Disponível em: https://conferencias.unb.br/index.php/iniciacaocientifica/28CICUnB19df/paper/view/45063 . Acesso em: 14 jul. 2023.

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