

Part-of-Speech Tagging para um FAQ sobre Serviços de Telecomunicação

Rafael Lucien Bahr Arias¹, Cinthyan Renata Sachs C. de Barbosa¹,
Alan Salvany Felinto¹

¹Programa de Pós-Graduação em Ciência da Computação – Universidade Estadual de Londrina (UEL) – Londrina – PR – Brazil

rafarias94@hotmail.com, cinthyan@uel.br, alan@uel.br

Abstract. *This paper describes the development of a part-of-speech (POS) tagger for a frequently asked questions (FAQ) page about telecommunication services. POS tagging consists of a series of adjustments in context rules or adjacent words in text sentences for precise grammatical classification of a particular application. Lexical categories and associated attributes are summarized with presentation of relevant statistics from the system's Portuguese language dictionary, whose lexical processing was assisted by the open-source Flex library. This morphological lexical analyzer represents the initial part of a complete chatbot system, which can potentially replace the FAQ page and assist visitors in a more user-friendly and interactive way. This is possible through the evolution of Natural Language Processing (NLP) technologies, which can reduce operating costs in a number of areas, including Customer Services and sales.*

Resumo. *Este artigo descreve o desenvolvimento de um POS tagger para uma página de perguntas frequentes (FAQ) sobre serviços de telecomunicações. Esse consiste de uma série de ajustes nas regras de contexto ou palavras adjacentes nas frases de texto para uma classificação gramatical precisa de uma aplicação particular. Categorias lexicais e atributos associados são sumarizados com a apresentação de estatísticas relevantes do dicionário de português do sistema, cujo processamento léxico foi auxiliado pela biblioteca Flex de código aberto. Este analisador léxico morfológico representa a parte inicial de um sistema completo de chatbot, que pode potencialmente substituir a página de FAQ e auxiliar os visitantes de uma forma mais amigável e interativa. Isso é possível por meio da evolução das tecnologias de Processamento de Linguagem Natural (PLN), o que pode reduzir custos operacionais de várias áreas, incluindo o Serviço de Atendimento ao Consumidor (SAC) e vendas.*

1. Introdução

Com o surgimento de novas mídias e a popularização das redes sociais, a quantidade de informação produzida a cada segundo tem se tornado cada vez maior e, por conta disso, o uso de Processamento de Linguagem Natural (PLN) vem sendo cada vez mais comum em diversos contextos. Embora grande parte dessas informações se encontre em formato multimídia, uma enorme quantidade de informação se encontra em formato de texto, fazendo com que as abordagens baseadas em PLN sejam fortes candidatas para basear a construção de soluções que acompanhem essa crescente demanda de processamento de informações textuais, com um nível satisfatório de acurácia e desempenho [Thiele 2015].

O estilo de interação em linguagem natural é uma linguagem de comando cujas regras sintáticas são aquelas da língua natural usada pelo usuário e, portanto, dispensa o conhecimento de sintaxes muito específicas e inflexíveis, como é o caso das linguagens de programação [Barbosa 2004].

A análise lexical morfológica, frequentemente referenciada como POS tagging, é um dos componentes essenciais de qualquer aplicação de processamento de linguagem natural. É o processo de atribuir a parte apropriada da fala ou categoria lexical a cada palavra em uma frase, contribuindo na análise automática de textos como um passo essencial diante da análise sintática e semântica [Ghosh, Ghosh and Das 2016].

Em um mundo cada vez mais digital, o PLN desempenha um papel fundamental na comunicação. Chatbots são uma das aplicações que estão se tornando comuns no mercado por sua utilidade, permitindo a automação do suporte ao usuário em serviços *online*, reduzindo os custos de contratação de funcionários e aprimorando a experiência do usuário com uma comunicação mais direta de conteúdo [Leonhardt 2005]. Em um mercado competitivo, a presença de um chatbot em um *site* na internet, em vez de páginas de texto estáticas, também pode ser um recurso valioso, como o assistente de linguagem natural [Khurana *et al.* 2017] para responder automaticamente às perguntas frequentes sobre gerenciamento de seguro de saúde, evidenciando melhor experiência por parte dos clientes.

Este trabalho aborda o tratamento automático da linguagem natural, particularmente a descrição do conteúdo informacional de textos, para melhorar sua indexação e preencher os requisitos dos sistemas de informação documentária. É descrito o desenvolvimento de um POS tagger para um banco de dados de perguntas frequentes sobre os direitos do consumidor no *site* da ANATEL [Agência Nacional de Telecomunicações, 2018], que é a agência nacional de telecomunicações do Brasil, que lida com as regulamentações de serviços de Internet, telefonia e televisão paga. Este analisador léxico representa a etapa inicial de um processador de linguagem natural completo para um programa de chatbot que pode potencialmente substituir a página de FAQ (*frequently asked questions*), ou seja, *página de perguntas frequentes*, aprimorando a experiência do usuário interpretando suas consultas e fornecendo respostas apropriadas com informações associadas do banco de dados de respostas.

Este artigo está organizado da seguinte forma: a seção 2 descreve a aquisição do banco de dados de informações de telecomunicações, assim como as ferramentas empregadas no desenvolvimento do sistema; a seção 3 traz a explicação da estrutura do *POS tagging* com regras de contexto; na seção 4 são apresentadas as classes de dicionário e a especificação de atributos, fornecendo estatísticas com algumas amostras de execução para marcação ambígua de palavras; na seção 5 são elencadas as considerações finais.

2. Informações da Base de Dados e Análise Léxica

A base de dados deste projeto está disponível publicamente no *site* da Anatel [Agência Nacional de Telecomunicações 2018]. A aquisição foi realizada extraindo perguntas das páginas com ferramentas de manipulação de texto Unix como *sed* e *tr*, realizando os devidos ajustes nas linhas e espaços vazios e correção manual de alguns erros ortográficos no texto. A Figura 1 apresenta alguns exemplos de perguntas extraídas da página de perguntas frequentes.

Quais informações devem constar no histórico de demandas (reclamações, pedidos de informação, solicitações)?

Quais informações devem constar do relatório detalhado dos serviços e facilidades prestados pelas prestadoras de serviços de telecomunicações?

Qual o prazo que a prestadora tem para reativar o serviço suspenso por falta de pagamento?

Qual o tempo de espera máximo para atendimento pessoal, na loja da prestadora?

Quais os canais de atendimento a prestadora deve disponibilizar?

Qual o prazo para atendimento ao consumidor?

Qual o tempo de espera máximo para o contato com o atendente na central de atendimento telefônico?

Qual o prazo que a prestadora tem para disponibilizar a minha gravação?

Figura 1. Exemplos de perguntas extraídas do FAQ.

Um *analisador léxico-morfológico* tem a função de receber a frase do usuário e identificar palavras ou expressões isoladas na frase, transformando-a em uma lista de rótulos ou *tags*, para que possa ser processada pelo analisador sintático. Os delimitadores como pontuação e espaço em branco ajudam na identificação de palavras, as quais são categorizadas através do léxico em tipos de palavras [Barbosa 2004].

Tipos de palavras se referem a grupos de palavras classificadas conforme seu tipo de uso ou, em linguagem natural, *categoria gramatical*, onde uma instância da palavra em uma sentença gramaticalmente válida pode ser substituída por outra do mesmo tipo em uma sentença ainda válida (exemplo: substantivo, pronomes, verbos, etc). Dentro de um mesmo tipo de palavra, existem grupos de regras que caracterizam o comportamento de um subconjunto de vocábulos da linguagem (exemplo: formação do plural de substantivos terminados em *ão*, flexões dos verbos regulares terminados em *ar*, etc). Assim, a morfologia trata as palavras quanto à sua estrutura, forma, flexão e classificação, no que se refere a cada um dos tipos de palavras [Strube de Lima 1996].

Neste trabalho, o módulo *tokenizer* é o componente que especifica classes de tokens com expressões regulares, que podem ser compiladas em um único autômato finito determinístico que particiona strings de entrada [Guthrie *et al.* 1996]. Para implementação, a linguagem C++ foi usada com a biblioteca Flex (*The Fast Lexical Analyzer*) [The Fast Lexical Analyzer 2018]. Flex é um software livre e de código aberto do programa Lex da Unix, que permite a correspondência de tokens e expressões em texto e separa a lógica de varredura da configuração específica dos tokens da aplicação. Para instalá-lo no Linux basta utilizar o comando *sudo apt-get install flex*. Sua operação modular também permite uma integração eficiente com ferramentas de análise sintática como Yacc ou Bison. Neste aplicativo, ele é usado para extrair cada símbolo independente para pesquisa de categoria pelo dicionário e regras de contexto. A Figura 2 ilustra a arquitetura do sistema desenvolvido.

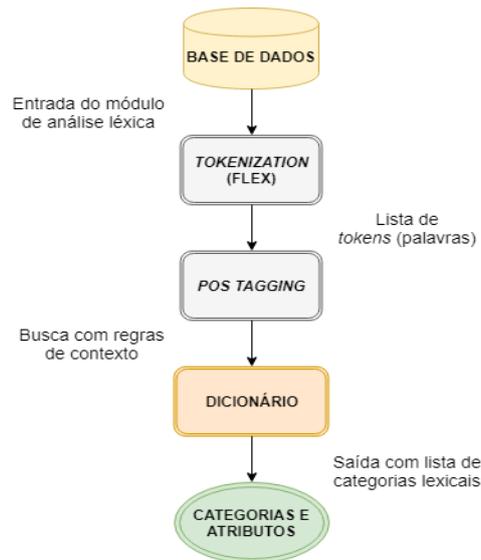


Figura 2. Representação da arquitetura do sistema.

3. Pos Tagging

POS (Part-of-Speech) tagger é um sistema que usa o contexto para atribuir partes da fala ou categorias léxicais às palavras [Cutting *et al.* 1992]. A Figura 3 ilustra esse processo para uma dada sentença de entrada da base de dados. Uma variedade de métodos pode ser utilizada para executar o procedimento de marcação, incluindo técnicas baseadas em refinamento iterativo de regras ou modelos estatísticos para realizar a classificação. Os melhores classificadores de *POS tagging* são treinados, cujos atributos incluem rótulos de palavras anteriores e posteriores, bem como várias palavras do contexto da frase e recursos manuais para lidar com casos especiais [Collobert *et al.* 2011]. Além disso, mesmo palavras desconhecidas podem ser corretamente marcadas por um classificador treinado baseado em sua morfologia e contexto [Fonseca and Aluísio 2016].

Cada palavra de uma frase de entrada (definida pelo ponto final ou símbolo de ponto de interrogação nesta aplicação) é inicialmente atribuída a uma categoria padrão e em seguida, regras baseadas em determinadas propriedades do contexto de frase são aplicadas para distinguir entre várias categorias da mesma palavra. Uma das maneiras mais simples de implementar regras de contexto é com condições na propriedade ou classe de palavras sentenciadas adjacentes para a categorização da palavra alvo, como “se uma palavra é a \underline{e} está no contexto c ou tem propriedade p , então altere o rótulo para b ” [Brill 1992].

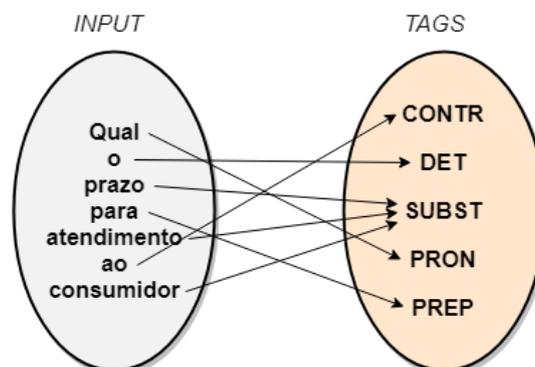


Figura 3. Procedimento de *POS tagging* para uma dada sentença de entrada.

Por exemplo, uma determinada palavra pode apresentar o rótulo de substantivo ou verbo, mas se essa aparecer no contexto de outras palavras, a ambiguidade pode ser geralmente reduzida ou mesmo eliminada para um determinado escopo. É importante desenvolver várias regras baseadas em linguagem com um dicionário de itens detalhados que pode ser usado para descobrir as regras do modelo. O processo de etiquetagem exige especialização dos linguistas, o que geralmente é caro e demorado. Em vez da categorização feita por abordagens manuais, os métodos empregados podem ser completamente automáticos, com foco no aprendizado de máquina [Scarton, Duran and Aluísio 2014]. Nesta aplicação, as regras podem não ser genéricas, mas classificam corretamente a maioria dos casos de banco de dados.

4. Especificação do Sistema

A seguir será descrito o dicionário e uma amostra de execução para classificação do sistema de POS *tagging*.

4.1. Dicionário

Os sistemas de PLN geralmente agrupam módulos associados aos níveis de processamento de linguagem, fazendo uso de uma estrutura fundamental que é o léxico ou dicionário. Um dicionário é uma estrutura de dados onde as palavras são armazenadas e, associadas a elas, algumas de suas informações [Barbosa 2004]. 708 palavras foram registradas no dicionário do banco de dados de perguntas, onde algumas estão presentes em mais de uma categoria. Dez categorias morfológicas foram consideradas, incluindo a maioria das classes principais, como adjetivo, advérbio, conjunção, determinante, substantivo, preposição, pronome e verbo, mas não a interjeição, pois não ocorre no texto. Foi incluída contração (combinação de determinante e preposição) e uma classe acrônimos para distinção de substantivos. A Figura 4 apresenta alguns exemplos de entradas registradas no dicionário do sistema.

```
23 Dic_adjetivo(anteriores, -, plural)
24 Dic_adjetivo(antiga, feminino, singular)
25 Dic_adjetivo(antigo, masculino, singular)

121 Dic_advérbio(livremente)
122 Dic_advérbio(mais)
123 Dic_advérbio(menos)

142 Dic_conjunção(porém, coordenativa)
143 Dic_conjunção(se, subordinativa)
144 Dic_contração(à, a, a)
145 Dic_contração(ao, a, o)

176 Dic_determinante(um, masculino, singular)
177 Dic_determinante(várias, feminino, plural)
178 Dic_preposição(a)
179 Dic_preposição(após)
180 Dic_preposição(com)

204 Dic_pronome(você, -, singular)
205 Dic_sigla(2G)
206 Dic_sigla(3G)

485 Dic_substantivo(virtude, feminino, singular)
486 Dic_substantivo(visitas, feminino, plural)
487 Dic_substantivo(vizinho, masculino, singular)
488 Dic_substantivo(vizinhos, masculino, plural)
489 Dic_verbo(acabando, acabar, gerúndio, -, -)
490 Dic_verbo(acabar, acabar, infinitivo, -, -)
491 Dic_verbo(acessar, acessar, infinitivo, -, -)
492 Dic_verbo(acontece, acontecer, presente do indicativo, -, singular)
```

Figura 4. Exemplos de entradas registradas no dicionário.

Dentro do mesmo tipo de palavra, existem grupos de regras que caracterizam o comportamento de um subconjunto de vocabulários de linguagem, como número, gênero e inflexões verbais. Os sistemas de *tagging* (marcação) geralmente usam essas propriedades adicionais para melhorar a análise sintática e semântica de aplicações específicas [Collobert 2011]. Os atributos de classe incluem gênero (masculino ou feminino) e número (singular ou plural) quando aplicável e podem apresentar valor indeterminado. As conjunções distinguem-se entre o grupo de coordenação ou de subordinação.

Nesta aplicação, a forma infinitiva dos verbos é associada, permitindo a compressão do modelo [Aluísio e Almeida 2006]. Além da forma infinitiva, o tempo verbal ou forma nominal é identificado sem pessoa de conjugação. Foram identificados cinco tempos verbais no léxico da aplicação: presente do indicativo, pretérito do indicativo, futuro do indicativo presente, presente do subjuntivo e futuro do subjuntivo, mais as três formas nominais: infinitivo, participio e gerúndio.

4.2. Amostra de Execução

A operação modular fornecida pela biblioteca Flex permite a adaptação de entrada com um arquivo de entrada(s) do usuário. Como exemplo, a entrada do usuário “*Posso transferir meu número de telefone para outra operadora?*” produziria a sequência de *tags* mostrada na Figura 5.

```
Posso transferir meu número de telefone para outra operadora?
verb(possso, poder, indicative present, -, singular)
verb(transferir, transferir, infinitive, -, -)
determinant(meu, male, singular)
noun(número, male, singular)
preposition(de)
noun(telefone, male, singular)
preposition(para)
determinant(outra, female, singular)
noun(operadora, female, singular)
```

Figura 5. Classificação do sistema de POS *tagging* para uma pergunta frequente sobre serviços de telefonia.

As entradas “*Qual é a velocidade?*” e “*Posso continuar a ter acesso?*” são exemplos em que a mesma palavra “*a*” pode apresentar categorias distintas: artigo no primeiro caso e preposição no segundo, como pode ser observado pelas saídas do sistema na Figura 6. A classificação correta requer uma combinação de regras incluindo classes adjacentes e propriedades como gênero e número da próxima palavra.

```
Qual é a velocidade?
pronoun(qual, -, singular)
verb(é, ser, indicative present, -, singular)
determinant(a, female, singular)
noun(velocidade, female, singular)

Posso continuar a ter acesso?
verb(possso, poder, indicative present, -, singular)
verb(continuar, continuar, infinitive, -, -)
preposition(a)
verb(ter, ter, infinitive, -, -)
noun(acesso, male, singular)
```

Figura 6. Palavra comum “a” com regras dependentes do contexto rotulada como determinante ou preposição.

A Figura 7 ilustra um caso mais simples para diferenciar substantivo e pronome para o vocábulo “*minha*”, no qual a diferenciação pelo contexto em “*minha cidade*” e “*maior que a minha*” requer uma única condição implementada.

```

Na minha cidade...
contraction(na, em, a)
determinant(minha, female, singular)
noun(cidade, female, singular)

Meu vizinho tem velocidade maior que a minha.
determinant(meu, male, singular)
noun(vizinho, male, singular)
verb(tem, ter, indicative present, -, singular)
noun(velocidade, female, singular)
adjective(maior, -, singular)
pronoun(que, -, -)
determinant(a, female, singular)
pronoun(minha, female, singular)

```

Figura 7. Palavra “*minha*” com regras dependentes do contexto rotulada como substantivo ou pronome.

Outro exemplo comum é a diferenciação entre adjetivos e verbos na forma de particípio, como na palavra “*instalado*”, mas esse caso pode ser facilmente diferenciado porque o adjetivo é geralmente amarrado com um substantivo, enquanto o particípio geralmente tem outro verbo; a regra de marcação só pode especificar casos adjetivos. De maneira semelhante, outros casos especiais podem ser distinguidos pela presença de um substantivo adjacente, como na palavra “*mesmo*”, que pode assumir as classes de adjetivo e advérbio no dicionário.

A Tabela 1 apresenta dados referentes ao número de instâncias de cada classe encontradas no texto e aquelas do dicionário que não contém repetições. Existem 237 frases de banco de dados com pouco menos de 700 palavras únicas, indicando um escopo de aplicativo relativamente fechado, em que algumas palavras apresentam mais de uma *tag*. Como esperado, verbos e substantivos ocupam a maior parte do dicionário, mas as palavras mais comuns são a preposição “*de*” e os determinantes “*a*” e “*o*” com aproximadamente 150 ocorrências cada, seguidos pelo substantivo “*prestadora*” com 69 ocorrências.

Tabela 1. Classificação no Banco de Dados e Estatísticas do Dicionário

Classe	Quantidade (texto)	Quantidade (dicionário)
Acrônimos	56	16
Adjetivos	248	91
Advérbios	121	25
Conjunções	80	8
Contrações	218	15
Artigos	401	19
Substantivos	894	268
Preposições	371	11
Pronomes	152	16
Verbos	552	238
Total	3093	707

A Tabela 2 mostra as grandezas em relação aos atributos da classe. Pode-se observar que: a distribuição entre palavras masculinas e femininas é bem equilibrada; a

ocorrência de palavras singulares é muito maior que as do plural; e uma alta ocorrência de verbos no presente indicativo, seguida pelas formas nominais do infinitivo e do particípio.

Tabela 2. Estatística dos Atributos de Classe

Classe	Quantidade (texto)	Quantidade (dicionário)
Conjunção (coordenação)	53	4
Conjunção (subordinação)	27	4
Gênero (feminino)	760	196
Gênero (masculino)	797	216
Número (plural)	355	151
Número (singular)	1640	389
Verbo (futuro do presente do indicativo)	23	14
Verbo (pretérito perfeito do indicativo)	6	6
Verbo (presente do indicativo)	237	61
Verbo (futuro do subjuntivo)	11	5
Verbo (presente do subjuntivo)	7	5
Verbo (gerúndio)	6	5
Verbo (infinitivo)	180	85
Verbo (particípio)	82	57

Algumas exceções são mais difíceis de tratar, por exemplo, a palavra “*como*”, que na Língua Portuguesa pode assumir o papel de conjunção, advérbio, verbo, preposição ou interjeição [Oliveira e Freitas 2006]. Na maioria dos casos deste banco de dados, usa-se o rótulo de advérbio em perguntas como “*Como posso?*” ou “*Como fazer?*”, mas também pode ser marcado como conjunção ou preposição. A Figura 8 ilustra regras de contexto descobertas para classificação da palavra “*como*” baseada nas categorias de palavras adjacentes (próxima e anterior), como uma cascata de condições *if-then-else* formando uma estrutura de árvore de decisão.

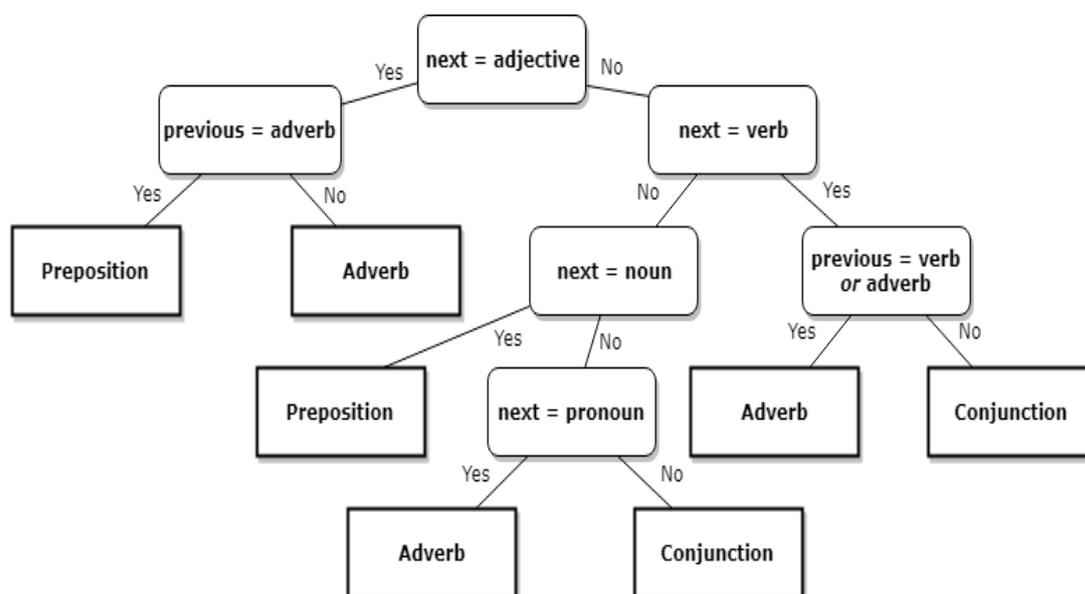


Figura 8. Regras de contexto baseadas em classes de palavras adjacentes para a palavra "como".

5. Considerações Finais

Um POS tagger foi desenvolvido como a parte inicial de um projeto de chatbot com o objetivo de substituir uma página de FAQ da Internet. Isso empregou a descoberta iterativa de regras de contexto específicas baseadas nos dados iniciais do dicionário, o que permitiu obter a classificação precisa das entradas do sistema.

Como o método empregado permite a categorização única dentro do escopo específico, o suporte a uma quantidade maior de entradas depende da inclusão de outras amostras ou uma variação das existentes. Portanto, há também possibilidades de melhorar o sistema léxico feito até agora através da expansão de domínio de consulta com uma tabela de sinônimos como feita por Amorim, Cury e Menezes (2012), permitindo que perguntas sejam feitas de várias maneiras diferentes.

Outra técnica é a inserção de palavras do dicionário com erros ortográficos que são comumente escritas pelos usuários. A elaboração de um PLN com analisadores sintáticos e semânticos automáticos baseados em técnicas de aprendizado de máquina também é deixada como trabalho futuro.

Os resultados da pesquisa apontam para a melhoria que o aprofundamento à análise semântica do processador de linguagem natural oferece para a seleção automática de descritores e conseqüentemente para a recuperação de informação. Essa área carece, como salienta Camara Junior (2016), de resultados amadurecidos na área de PLN.

Referências

- Agência Nacional de Telecomunicações (2018). <http://www.anatel.gov.br/consumidor/perguntas-frequentes?view=faq>
- Aluísio, S. e Almeida, G. (2006). O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa linguística. *Calidoscópio* 4(3), pages 156–178.

- Amorim, M. T. C. F., Cury, D. e Menezes, C. (2012). Um Helpdesk Inteligente baseado em Ontologias. In: *Anais do 23º Simpósio Brasileiro de Informática na Educação*. Rio de Janeiro: CBIE.
- Barbosa, C. R. S. C. de. (2004) *Técnicas de Parsing para Gramática Livre de Contexto Lexicalizada da Língua Portuguesa*. São José dos Campos: CPG da Engenharia Eletrônica e Computação do Instituto Tecnológico de Aeronáutica. Tese de Doutorado. 171p.
- Brill, E. (1992). A Simple rule-based Part of Speech Tagger. In: *Proceedings of the Third Conference on Applied Natural Language Processing*, pages 152–155. Association for Computational Linguistics, Trento.
- Camara Junior, A. T. (2016). Processamento de Linguagem Natural para Indexação Automática Semântico-ontológica. *Revista Ibero-Americana de Ciência da Informação*. 9(2). p.569. jul/dez.
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. and Kuksa, P. (2011). Natural Language Processing (almost) from Scratch. *Journal of Machine Learning Research*. 12, pages 2493–2537.
- Cutting, D., Kupiec, J., Pedersen, J. and Sibun, P. (1992). A practical part-of-speech tagger. In: *Proceedings of the Third Conference on Applied Natural Language Processing*, pages 133–140. Association for Computational Linguistics, Trento.
- Fonseca, E. and Aluísio, S. (2016). Improving POS Tagging across Portuguese Variants with Word Embeddings. In: *Proceedings of the International Conference on Computational Processing of the Portuguese Language*, pages 227-232. Springer, Cham.
- Ghosh, S., Ghosh, S. and Das, D. (2016). Part-of-speech Tagging of Code-Mixed Social Media Text. In: *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 90–97. Association for Computational Linguistics, Austin.
- Guthrie, L., Pustejovsky, J., Wilks, Y. and Slator, M. (1996). The Role of Lexicons in Natural Language Processing. *Communications of the ACM*. 39(1), pages 63–72.
- Khurana, P., Agarwal, P., Shroff, G., Vig, L. and Srinivasan, A. (2017). Hybrid BiLSTM-Siamese network for FAQ Assistance. In: *Proceedings of the ACM on Conference on Information and Knowledge Management*, pages 537–545. ACM, Singapore.
- Leonhardt, M. D. (2005). *Doroty: um Chatterbot para Treinamento de Profissionais Atuantes no Gerenciamento de Redes de Computadores*. Porto Alegre: CGCC da Universidade Federal do Rio Grande do Sul. Dissertação de Mestrado. 110p.
- Oliveira, C. e Freitas, M. (2006). Classes de Palavras e Etiquetagem na Linguística Computacional. *Calidoscópio* 4(3), pages 179–188.
- Scarton, C., Duran, M. and Aluísio, S. (2014). Using cross-linguistic knowledge to build VerbNet-style lexicons: results for a (Brazilian) Portuguese VerbNet. In: *Proceedings of the International Conference on Computational Processing of the Portuguese Language*, pages 149-160. Springer, Cham.
- Strube de Lima, V. L. (1996). Processamento da Linguagem Natural - premissas e desafios. In: *4º Anais da Escola Regional de Informática -SBC/Regional Sul*, pages 110-124. SBC, Canoas/Londrina.

The Fast Lexical Analyzer (2018). <https://www.gnu.org/software/flex/>

Thiele, P. F. O. (2015) *Desambiguação de anotações morfossintáticas feitas por MTMDD*. Porto Alegre: PPGCC da Pontifícia Universidade Católica do Rio Grande do Sul. Dissertação de Mestrado. 58p.