Universidade Federal do Ceará
Pró-Reitoria de Pesquisa e Pós-Graduação
PIBIC 2020/2021 - Edital Nº 1/2020

Técnicas em softwares livres para a linguística de corpus (12a Etapa)

Leonel Figueiredo de Alencar

Resumo

Esta pesquisa é mais uma etapa de projeto maior que objetiva a criação de ferramentas de processamento de linguagem natural distribuídas sob licença livre, voltadas para a anotação de corpora linguísticos. Falta, para as línguas indígenas brasileiras, corpora etiquetados morfossinticamente. Esse tipo de recurso é importante não apenas para a investigação da estrutura gramatical e documentação dessas línguas, mas constitui, também, fator de sobrevivência delas, na medida em que possibilita o desenvolvimento de aplicações de tecnologia da linguagem natural. O pré-requisito fundamental para a construção desse recurso é a existência de uma etiquetador morfossintático, ferramenta que atribui, a cada token do corpus, uma etiqueta indicativa da classe de palavra a que pertence (POS tag em inglês). Pelo que sabemos, no momento, nenhuma língua indígena brasileira dispõe de um corpus morfossinticamente anotado, resultado da inexistência de etiquetadores morfossintáticos para essas línguas. Na presente etapa, pretendemos contribuir para sanar essas deficiências, compilando um corpus da Língua Geral Amazônica (LGA), também conhecida como tupi moderno ou nheengatu, e implementando computacionalmente um etiquetador para sintagmas nominais, a ser aplicado nesse corpus. Para testar a ferramenta, extrairemos uma amostras aleatória de 10% das sentenças do corpus, que será corrigida manualmente. Conforme a hipótese que pretendemos verificar, esperamos obter nesse teste um índice F-Score igual ou superior a 0.95.

1. Introdução

Esta pesquisa é mais uma etapa de projeto maior, iniciado há alguns anos, que objetiva a criação de ferramentas de processamento de linguagem natural distribuídas sob licença de free software/open source software, voltadas para a anotação automática de corpora linguísticos. Na presente etapa, pretendemos contribuir para preencher uma lacuna na "paisagem de corpora", tradução do termo termo alemão "Korpuslandschaft" (LEMNITZER; ZINSMEISTER, 2006), das línguas faladas no Brasil, que é inexistência de corpora de línguas indígenas brasileiras etiquetados morfossinticamente. Nesse tipo de anotação, cada token do corpus é provido de uma etiqueta morfossintática (POS tag em inglês) que indica a sua classe de palavra, levando em conta propriedades do token, como a sua terminação, e seu contexto (VOUTILAINEN, 2004; GÜNGÖR, 2010). Diversos corpora com esse tipo de anotação estão disponíveis para diversas línguas, inclusive o português, como, por exemplo, o MacMorpho (FONSECA; ALUÍSIO; ROSA, 2015), o Tycho Brahe (GALVES; ANDRADE; FARIA, 2017) e o Bosque (PROJECTO, 2010).

Esse tipo de recurso é de fundamental importância não só para o desenvolvimento de aplicações de tecnologia da linguagem natural, mas também para as pesquisas em todo o campo das humanidades digitais, incluindo linguística, análise literária etc. (TAGNIN; VALE, 2008).

Também inexistem, no momento, pelo que sabemos, etiquetadores morfossintáticos para línguas indígenas brasileiras. Esse tipo de ferramenta constitui pré-requisito para a construção de corpora etiquetados morfossintaticamente. Todas as principais línguas majoritárias dispõem desse tipo de ferramenta, como, por exemplo, o Aelius (ALENCAR, 2013, 2015), voltado para o português.

Com o presente projeto, pretendemos disponibilizar às comunidades de desenvolvedores e de

pesquisadores, sob licença gratuita de free/open source software, um etiquetador de sintagmas nominais para a Língua Geral Amazônica (LGA), também conhecida como tupi moderno ou nheengatu (NAVARRO, 2011). Um subproduto do projeto será um corpus da LGA com sintagmas nominais parcialmente anotados com etiquetas morfossintáticas.

As aplicações tecnológicas desse tipo de ferramenta dividem-se em dois grupos (VOUTILAINEN, 2004). No primeiro estão, por exemplo, sistemas de extração de informações (information extraction), extratores de termos para a terminografia e tradutores automáticos estatísticos. No segundo, temos a anotação de corpora linguísticos, os quais permitem verificar automaticamente hipóteses linguísticas com base num grande volume de dados. Os etiquetadores que representam o estado da arte dessa tecnologia alcançam entre 96% e 97% de acurácia para línguas indo-europeias (GÜNGÖR, 2010, p. 207), o que minimiza o esforço de anotadores humanos, cuja atuação passa a limitar-se à correção dos erros do etiquetador.

A construção de um etiquetador morfossintático é uma tarefa complexa, que necessariamente precisa se desdobrar em várias etapas. De fato, a etiquetagem não pode se limitar a atribuir a um token a etiqueta correspondente em um dicionário. Uma vez que, no dicionário, as palavras geralmente não estão listadas em suas formas flexionadas, sendo, em vez disso, representadas por meio dos lemas, é preciso levar em conta, também, a flexão nominal e verbal. Por outro lado, nenhum dicionário esgota o léxico de uma língua, contendo, normalmente, apenas a palavras atestadas em textos. O léxico, porém, não se resume ao inventário das palavras existentes, mas incluí, igualmente, as palavras potenciais, capazes de ser geradas por meio dos processos de formação de palavras, como derivação, composição etc. (ROCHA, 2008; VILLALVA; SILVESTRE, 2014). Um etiquetador robusto, capaz de alcançar uma alta acurácia, deve ser capaz de classificar as palavras novas, não constantes do dicionário, por meio da modelação computacional desses processos morfológicos, com base em determinadas heurísticas ou por meio de técnicas de aprendizagem de máquina.

A estrutura sintática das línguas naturais é passível de representação por meio de estruturas de dados, como árvores, as quais possuem uma dimensão espacial. Desse modo, modelos topológicos têm sido propostos para descrever fenômenos sintáticos, ver, por exemplo, Sternefeld (2006). Dada a complexidade da tarefa de etiquetagem morfossintática, há que restringir o escopo deste projeto, a ser executado por dois estudantes de graduação no período de 12 meses. Para tanto, utilizamos noções topológicas para definir um subdomínio do sintagma nominal que possa constituir um objeto viável da pesquisa dentro dessas limitações.

Segundo Cruz (2011:282), o sintagma nominal da LGA apresenta uma %u201Czona prefixal%u201D em relação ao núcleo nominal N, a qual abriga quantificadores contínuos (QUANT), demonstrativos (DEM), quantificadores discretos (i.e. numerais, simbolizados como NUM), artigos indefinidos (INDF) e complementos nominais, sob a forma de pronome de segunda classe (PRON2) ou nome (N), conforme exemplificamos em (1)-(4). Imediatamente precedendo ou subseguindo o núcleo N, podem ocorrer, segundo Navarro (2011:12), adjetivos, os quais se classificam em primeira ou segunda classe (A1 e A2, respectivamente). Definimos esse tipo de modificação como zona nuclear do NP. Além disso, sintagmas preposicionais e orações relativas também podem funcionar como modificadores, abrigados no que denominamos zona periférica.

O presente projeto limita-se às zonas prefixal e nuclear do NP na LGA, tal como descritas por Cruz (2011) e Navarro (2011). O objetivo é implementar, na linguagem de programação Python (CHUN, 2006), um algoritmo que, dadas sentenças de entrada como (1) e (2), produza como saída, respectivamente, as sentenças anotadas (3) e (4). Como evidenciam os exemplos, o etiquetador a ser construído limitar-se-á a classificar as palavras que potencialmente integram essas zonas, ignorando verbos, posposições etc. Por outro lado, deverá etiquetar palavras dessas zonas mesmo quando ocorrem fora delas, como é o caso do PRON2 "ta" e do A2 "katu" em (4), os quais funcionam como predicativo do sujeito.

(1) Nhaã-itá pirasua kunhã mimbira umurári iepé uka kiá upé .
DEM.DIST:PL pobre mulher filho morar:3p um casa sujo em
'Aqueles filhos da mulher pobre moram numa casa suja.'

(2) Panh%u1EBD musapíri se mimbira ta katu uiku .
todo três 1s filho 3p bom estar:3p
'Todos os meus três filhos estão bem de saúde.'
(3) Nhaã-itá/DEM-PL pirasua/A1 kunhã/N mimbira/N umurári iepé/INDF uka/N kiá/A2 upé ./PUNCT
(4) Panh%u1EBD/QUANT musapíri/NUM se/PRON2 mimbira-itá/N-PL ta/PRON2 katu/A2 uiku.

Em seguida, avaliaremos a implementação, doravante referida por Nheentiquetador, aplicando-a ao conjunto de teste TEST-SET, uma amostra aleatória de 10% das sentenças de um corpus compilado a partir dos textos e exemplos de Navarro (2011), aferindo a acurácia por meio do índice F-Score.

2. Perguntas de Partida

A pergunta de partida deste projeto refere-se à acurácia do Nheentiquetador quando aplicado na etiquetagem morfossintática do conjunto de sentenças TEST-SET:

Qual a acurácia do Nheentiquetador na etiquetagem morfossintática do conjunto de sentenças TEST-SET?

3. Hipóteses

Investigaremos a seguinte hipótese:

A acurácia do Nheentiquetador na etiquetagem morfossintática do conjunto de sentenças TEST-SET é de um F-Score de pelo menos 0.95.

4. Objetivos

O objetivo geral da pesquisa é implementar o Nheenquitador, um etiquetador morfossintático para as classes de palavras constitutivas das zonas prefixal e nuclear dos sintagmas nominais na LGA. Esse objetivo geral implica nos seguintes objetivos específicos:

(i) Compilar um corpus a partir dos textos de Navarro (2011).
(ii) Testar a hipótese de que a acurácia do Nheenquitador na etiquetagem morfossintática do conjunto de sentenças TEST-SET é de um F-Score de pelo menos 0.95.

5. Materiais e Métodos

Na linguística computacional, a construção de ferramentas para processamento em qualquer nível de descrição (fonológico, morfológico, sintático etc.) deve ocorrer de forma incremental, a partir da elaboração de modelos cada vez mais complexos, capazes de abarcar aspectos do fenômeno cada vez mais amplos (cf. FRANCEZ; WINTNER, 2012; SCHWARZE; ALENCAR, 2016), utilizando a técnica conhecida como desenvolvimento em espiral (ZELLE, 2004). Nesse processo, cada etapa do desenvolvimento deve ser testada com relação a um conjunto de teste, utilizando métricas apropriadas a cada tipo de tarefa.
O presente projeto objetiva implementar um etiquetador morfossintático para as classes de palavras constitutivas das zonas prefixal e nuclear dos sintagmas nominais na LGA. Como noutras tarefas de processamento de linguagem natural, há duas abordagens principais para construção dessa ferramenta: (i) abordagem baseada no conhecimento e (ii) abordagem baseada em dados (VOUTILAINEN, 2004; DUCHIER; PARMENTIER, 2015). Essa última abordagem consiste no

treinamento de modelos estatísticos a partir de corpora anotados por especialistas humanos, utilizando algoritmos de aprendizagem de máquina. Dada a inexistência desses dados para a LGA, resta aplicar a primeira abordagem, que consiste na implementação de regras com base em descrições gramaticais da língua.
O primeiro passo da pesquisa, portanto, é fazer revisão da literatura sobre a estrutura do NP na LGA, de modo a estabelecer o inventário de etiquetas das classes de palavras que podem ocorrer nas zonas prefixal e nuclear. O segundo passo é extrair do glossário de Navarro (2011) todos os itens pertencentes a essas classes. As entradas desse glossário são de três tipos: lemas, formas irregulares e variantes. No primeiro caso, exemplificado em (5)-(11), são fornecidos os seguintes tipos de informação sobre cada acepção do lema: (i) informações sobre o paradigma flexional no caso de formas irregulares, (ii) abreviatura da classe de palavra entre parênteses e (ii) traduções para o português. No caso das formas irregulares, como em (12), há uma remissão ao verbete correspondente. No terceiro tipo de entrada, indica-se a forma canônica, fornecendo-se explicações adicionais sobre a variante, ver (13).

(5) xári (v.) - deixar
(6) xibé (s.) - chibé (var. de bebida)
(7) xibuí (s.) - verme
(8) xinga (pron. quantif.) - 1. pouco; 2. um pouco; (adv. intensif.) - pouco
(9) xipu (s.) - cipó
(10) xirura (s.) - calça
(11) simiriku (rimiriku, simiriku) (s.) - esposa
(12) rimiriku - v. simiriku
(13) ximiriku - var. de simiriku (3a p. do sing.): esposa dele

As entradas das classes lexicais passíveis de ocorrer nas zonas prefixal e nuclear do NP precisam ser convertidas para uma tabela nos moldes de (14), de modo a poder ser processadas pelo algoritmo do etiquetador.

(14)
xibé N
xibuí N
simiriku N
rimiriku N
simiriku N
ximiriku N
xinga Adv

Em seguida, é preciso expandir essa lista com as formas flexionadas, por exemplo:

(15)
xibé-itá N-PL
xibuí-itá N-PL
simiriku-itá N-PL

O passo seguinte é converter essa última tabela para uma instância da estrutura de dados dictionary (dicionário) de Python, onde cada item da primeira coluna constitui uma chave cujo valor é a etiqueta da segunda coluna. Finalmente, é preciso implementar uma função capaz de aplicar o dicionário a cada token de um texto dado como entrada. Para cada um desses itens, a função retorna item/etiqueta se o item consta do dicionário, senão retorna o próprio item sem anotação.
O etiquetador assim construído será avaliado com base num conjunto de teste TEST-SET, uma amostra aleatória de 10% das sentenças de um corpus compilado a partir dos textos de Navarro (2011). A qualidade da ferramenta será calculada por meio da métrica F-Score, permitindo verificar

nossa hipótese. Essa métrica é baseada nos índices de precisão (P) e recall (R) por meio da fórmula (2*P*R)/(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; KLEIN; LOPER, 2009, p. 240).

6. Referências Bibliográficas

ALENCAR, L. F. de . Novos recursos do Aelius para o processamento computacional raso do português. In: LAPORTE, Éric; SMARSARO, Aucione; VALE, Oto Araújo. (Org.). Dialogar é preciso: linguística para o processamento de línguas. 1. ed. Vitória: PPGEL/UFES, 2013. p. 7-20.

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.; SARDINHA, T. B.. (Org.). Pesquisas e perspectivas em linguística de corpus. 1. ed. Campinas: Mercado de Letras, 2015. p. 233-282.

ALUÍSIO, S. et al. 2003. An account of the challenge of tagging a reference corpus for brazilian portuguese. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE %u2013 PROPOR, n. 6, 2003, Faro. Proceedings... Berlin: Springer, 2003.

FONSECA, E.R.; ROSA, J.L.G. Mac-morpho revisited: Towards robust part-of-speech. In: BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY %u2013 STIL, n. 9, 2013, Fortaleza. Proceedings... Fortaleza: SBC, 2013.

FONSECA, E.R.; ALUÍSIO, Sandra Maria; ROSA, J.L.G. Evaluating word embeddings and a revised corpus for part-of-speech tagging in Portuguese. Journal of the Brazilian Computer Society, v. 21, n. 2, 2015.

BRANCO, A.; SILVA, J. Evaluating Solutions for the Rapid Development of State-of-the-Art POS Taggers for Portuguese. In: LINO, M. T. et al. (Org.). Proceedings of the 4th International Conference on Language Resources and Evaluation. Paris: ELRA, pp. 507-510, 2004.

BRANDTS, T. TnT %u2013 A Statistical Part-of-Speech Tagger. [S.l.]: [s.n.], 2000. Disponível em:<http://acl.ldc.upenn.edu/A/A00/A00-1031.pdf> Acesso em: 2. fev. 2011.

BIRD, S.; KLEIN, E.; LOPER, E. Natural language processing with Python: analyzing text with the Natural Language Toolkit. Sebastopol, CA: O%u2019Reilly, 2009.

CHUN, W. J. Core Python programming. 2. ed. Upper Saddle River, NJ: Prentice Hall, 2006.

CRUZ, A. Fonologia e Gramática do Nheengatú: A língua falada pelos povos Baré, Warekena e Baniwa. Utrecht: LOT, 2011.

CONJUNTO de etiquetas (tagset) do Mac-Morpho. São Carlos: Universidade de São Paulo, 2003. Disponível em: <http://www.nilc.icmc.usp.br/lacioweb/manuais.htm> Acesso em: 16 abr. 2012.

DOMINGUES, M. L.; FAVERO, E. L.; MEDEIROS, I. P.. O desenvolvimento de um etiquetador morfossintático com alta acurácia para o português. In: TAGNIN, S. E. O.; VALE, O. A. (Org.). Avanços da Linguística de Corpus no Brasil. São Paulo: Humanitas, 2008. p. 267-286.

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

FELDMAN, A. e HANA, J. A resource-light approach to morpho-syntactic tagging. Amsterdam; New York: Rodopi, 2010.

FRANCEZ, N.; WINTNER, S. Unification grammars. Cambridge: CUP, 2012.

GARCIA, M.; GAMALLO, P. Análise morfossintática para português europeu e galego: problemas, soluções e avaliação. LinguaMÁTICA, Braga, vol. 2, n. 2, p. 59-67, 2010.

GALVES, Charlotte; ANDRADE, Aroldo Leal de; FARIA, Pablo. Tycho Brahe Parsed Corpus of Historical Portuguese. Campinas: Unicamp, 2017. Disponível em: <http://www.tycho.iel.unicamp.br/~tycho/corpus/texts/psd.zip>. Acesso em: 20. mar. 2020.

GÜNGÖR, T. Part-of-Speech Tagging. In: INDURKHYA, N.; DAMERAU, F. J. (Org.). Handbook of Natural Language Processing. 2. ed. Boca Raton, FL: Chapman & Hall/CRC, 2010. p. 205-235.

KEPLER, F. N. Um etiquetador morfo-sintático baseado em cadeias de Markov de tamanho variável. 2005. 70 p. Dissertação (Mestrado) %u2013 Programa de Pós-Graduação em Ciência da

Computação, Universidade de São Paulo, São Paulo. Disponível em: <http://www.ime.usp.br/~kepler/msc/kepler2005MSc.pdf> Acesso em: 15 abr. 2010.
JURAFSKY, D.; MARTIN, J. H. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. 2. ed. Londres: Pearson International, 2009.
LEMNITZER, L.; ZINSMEISTER, H. Korpuslinguistik: eine Einführung. Tübingen: Narr, 2006.
MACAMBIRA, J. R. A estrutura morfo-sintática do português: aplicação do estruturalismo lingüístico. 5. ed. São Paulo: Pioneira, 1987.
NAVARRO, E. A. Curso de Língua Geral (Nheengatu ou Tupi moderno): A Língua das origens da civilização amazônica. São Bernardo do Campo: Paym Gráfica e Editora, 2011.
NAVARRO, E. A.; TWARDOWSKY, M. T.; TREVISAN, R. G. O Nheengatu, entre a vida e a morte: a tradução literária como possível instrumento de sua revitalização lexical. Revista Letras Raras, Campina Grande, v. 6, n. 2, p. 9-29, 2017.
PALMER, M.; XUE, N. Linguistic annotation. In: CLARK, A; FOX, C.; LAPPIN, S. (Org.). The handbook of computational linguistics and natural language processing. Malden: Wiley & Blackwell, 2010. p. 238-270.
PIRINEN, T. et al. Introduction. In: INTERNATIONAL WORKSHOP FOR COMPUTATIONAL LINGUISTICS OF URALIC LANGUAGES, n. 3, 2017, St. Petersburg. Proceedings%u2026 Stroudsburg, USA: Association for Computational Linguistics, 2017. p. iii.
PROJECTO Floresta Sintá(c)tica. [s.l.]: [s.n], 2010. Disponível em: <http://www.linguateca.pt/Floresta/> Acesso em: 1. mar. 2018.
PRAÇA, W. N.; MAGALHÃES, M. M. S.; CRUZ, A. Indicativo II da família Tupi-Guaraní: uma questão de modo? Liames, Campinas, vol. 17, n. 1, p. 39-58, 2017.
ROCHA, L. C. de A. Estruturas morfológicas do português. 2. ed. São Paulo: Martins Fontes, 2008.
SCHWARZE, C.; ALENCAR, Leonel F. de. Lexikalisch-funktionale Grammatik: eine Einführung am Beispiel des Französischen mit computerlinguistischer Implementierung. 1. ed. Tübingen: Stauffenburg, 2016.
STERNEFELD, W. Syntax: eine morphologisch motivierte generative Beschreibung des Deutschen. Tübingen: Stauffenburg, 2006. 2 v.
TAGNIN, S. E. O.; VALE, O. A. (Org.). Avanços da Linguística de Corpus no Brasil. São Paulo: Humanitas, 2008.
VILLALVA, A.; SILVESTRE, J. P. Introdução ao estudo do léxico: descrição e análise do português. Petrópolis: Vozes, 2014.
VOUTILAINEN, A. Part-of-speech tagging. In: MITKOV, R. (Org.). The Oxford handbook of computational linguistics. Oxford, Oxford University Press, 2004. p. 219-232.
ZELLE, J. M. Python programming: an introduction to computer science. Wilsonville: Franklin, Beedle & Associates, 2004.

7. Plano de Atividades

Mês	Bolsista 1	Bolsista 2
1	revisão da literatura	revisão da literatura
2	revisão da literatura	revisão da literatura
3	revisão da literatura	revisão da literatura
4	compilação do dicionário do etiquetador	revisão do dicionário do etiquetador
5	compilação do dicionário do etiquetador	revisão do dicionário do etiquetador
6	compilação do dicionário do etiquetador	revisão do dicionário do etiquetador
7	implementação do algoritmo de etiquetagem	compilação do corpus para testagem da ferramenta
8	revisão do corpus	compilação do corpus para testagem da ferramenta

9	apicação do etiquetador no conjunto de teste	revisão da etiquetagem
10	revisão da etiquetagem	revisão da etiquetagem
11	elaboração do relatório	revisão do relatório
12	elaboração do relatório	revisão do relatório

Alencar, Leonel Figueiredo de. Técnicas em softwares livres na linguística de corpus (12ª etapa): projeto de pesquisa submetido ao Edital PIBIC 2020/2021. Fortaleza, 2020. Não publicado.