Aplicação estendida de analisador computacional na extração de sintagmas nominais em textos antigos: um estudo de caso

Extended application of a computational parser in the extraction of noun phrases in old texts: a case study

César Nardelli Cambraia *[1] **e Ramon Cunha Sampaio Leite** †[1]

[1] *Universidade Federal de Minas Gerais, Faculdade de Letras, Belo Horizonte, MG, Brasil.*

Resumo

Este estudo teve como objetivo analisar a aplicação estendida do analisador sintático *LX-Parser* em um *corpus* composto do trecho inicial da obra *Peregrinação* (publicada em 1614), de Fernão Mendes Pinto (*ca.* 1510-1583). Fez-se extração manual e automática de sintagmas nominais dos dez primeiros capítulos da obra. Testou-se a hipótese de que as especificidades de textos antigos limitam a precisão dos resultados gerados pelo analisador sintático considerado. A hipótese foi confirmada, uma vez que os resultados dessa aplicação estendida não se mostraram produtivos em função da alta frequência de problemas na análise produzida. Identificou-se que os principais problemas relativos a textos antigos estão relacionados à questão da extensão das sentenças, da grafia, da variação e mudança linguística. Além disso, constataram-se também problemas que não são específicos de textos antigos, mas que, ainda assim, limitaram o desempenho, como é o caso da ambiguidade estrutural e das categorias linguísticas.

Palavras-chave: Tecnologia. Linguística Computacional. Linguística Histórica. Sintaxe.

Abstract

This study aimed to analyze the extended application of the *LX-Parser*, a syntatic parser, in a corpus composed of the initial passage from *Peregrinação* (published in 1614) written by Fernão Mendes Pinto (*ca.* 1510-1583). Manual and automatic extraction of noun phrases from the first ten chapters of the work were carried out. The hypothesis that the specificities of old texts limit the accuracy of the results generated by the considered parser was tested. The hypothesis was confirmed, since the results of this extended application did not prove to be productive due to the high frequency of problems in the produced analysis. It was identified that the main problems related to old texts are related to the issue of sentece extension, spelling and linguistic variation and change. In addition, there were also problems that are not specific to old texts, but still limited performance: the issues of structural ambiguity and linguistic categories.

Keywords: Technology. Computational Linguistics. Historical Linguistics. Syntax.

Textolíne
Linguagem e Tecnologia

DOI: 10.35699/1983-3652.2022.37557

Seção:
Artigos

Autor Correspondente:
César Nardelli Cambraia

Editor de seção:
Daniervelin Pereira
Editor de layout:
Carolina Garcia

Recebido em:
19 de dezembro de 2021
Aceito em:
10 de abril de 2022
Publicado em:
23 de maio de 2022

1 Introdução

O desenvolvimento da computação tem possibilitado novas formas de realizar análises linguísticas, uma vez que os recursos computacionais já disponíveis têm oferecido cada vez mais novas alternativas para executar tarefas inerentes ao processo de pesquisa em linguística, tais como formação de *corpora*, lematização do vocabulário, análise da estrutura sintática, dentre outros.

No presente trabalho[1] discute-se a aplicação estendida de um recurso computacional específico, o *LX-Parser*, para a extração de sintagmas nominais (SNs) de um texto do séc. XVI: a *Peregrinação*, de Fernão Mendes Pinto[2].

*Email: nardelli@ufmg.br
†Email: ramoncunhasl27@outlook.com

1 Os autores agradecem ao CNPq pelo financiamento e aos pareceristas anônimos por suas críticas, as quais permitiram um melhor desenvolvimento do texto como um todo.

2 Deve-se salientar aqui que a análise foi desenvolvida a partir do ponto de vista do linguista, e não do pesquisador em ciência da computação, razão pela qual não se adentra em questões propriamente técnicas de programação, sendo contempladas especificamente as questões de adequação da análise linguística produzida pelo recurso computacional examinado.

2 Fundamentação teórica

2.1 Linguística computacional

Conforme assinalam Othero e Menuzzi (2005, p. 25), a linguística computacional é "a área da ciência linguística voltada para o tratamento computacional da linguagem e das línguas naturais". Os primórdios dessa abordagem se situam na década de 1950, com o desenvolvimento da tradução automática.

Vieira e Lima (2001) esclarecem que a linguística computacional é passível de aplicação em diversos níveis do estudo da linguagem: no âmbito da *fonética* e da *fonologia*, através do desenvolvimento de sistemas de reconhecimento e síntese de fala; no âmbito da *morfologia*, através do desenvolvimento de sistemas de reconhecimento das categorias de palavras (etiquetadores de categorias gramaticais ou *POS* [*Parts-Of-Speech*] *taggers*); e no âmbito da *sintaxe*, através do desenvolvimento de sistemas de reconhecimento dos constituintes da frase (analisadores sintáticos ou *parsers*). No que se refere aos domínios da *semântica* e da *pragmática*, Vieira e Lima (2001) não citam recursos específicos desenvolvidos para a análise linguística, embora registrem questões importantes para essas áreas, como ontologias, ambiguidade, atos de fala, dentre outros.

Em breve revisão, Alencar (2011) analisou seis analisadores sintáticos de código aberto para o português[3] então disponíveis: *VISL* [*Visual Interactive Syntax Learning*], baseado no *Palavras* (BICK, 2000); *Curupira* (MARTINS; HASEGWA; NUNES, 2003); *Selva* (ALMEIDA et al., 2003); *Grammar Play* (OTHERO, 2004); *Linguístico* (CONTIER; PADOVANI; JOSÉ NETO, 2010); e *LX-Parser* (SILVA et al., 2010). Em sua análise, Alencar (2011) identificou, como problemas para o processamento adequado relativos às entradas lexicais os processos produtivos de formação de palavras, os nomes próprios e as grafias não padrão. Segundo a avaliação do autor, a questão ainda estava a exigir melhoramentos, mesmo especificamente no âmbito do modelo gerativo:

> Em suma, podemos afirmar que a análise sintática automática de textos irrestritos em PB [= *Português Brasileiro*], do ponto de vista do PLN [= *Processamento de Linguagem Natural*], é uma área que ainda demanda muitos esforços de pesquisa, dadas as limitações das ferramentas disponíveis, que ou constituem software proprietário ou são de acesso restrito, ou não são precisas o suficiente, gerando análises errôneas mesmo para exemplos aparentemente triviais. (ALENCAR, 2011, p. 30).

> Concluímos [...] apontando que a análise sintática automática do PB, do ponto de vista da gramática gerativa, constitui um vasto campo a ser explorado, pois inexiste um parser que implemente de forma consistente a teoria X-barra, não obstante a existência de descrições da língua com base nesse modelo que utilizam a CFG [= *Context-Free Grammar*] (e são, portanto, mais próximas de uma implementação computacional) [...] (ALENCAR, 2011, p. 32).

No presente estudo, pretende-se trabalhar com apenas um dos analisadores citados por Alencar (2011): o *LX-Parser*. A restrição a apenas este se baseia, em primeiro lugar, no objetivo de se fazer uma análise mais detalhada do grau de precisão do referido analisador, o que exige uma maior delimitação dos recursos computacionais investigados, e, em segundo lugar, no fato de se privilegiar um analisador com interface mais amigável, sem necessidade de instalação de *software* e acessível gratuitamente na internet.

O *LX-Parser* é um analisador sintático baseado no *Standford Parser* e foi desenvolvido por Patrícia Gonçalves e João Silva, sob a coordenação de António Branco, no NLX - Grupo de Fala e Linguagem Natural do Departamento de Informática da Universidade de Lisboa, tendo sido lançado em 2010 (SILVA et al., 2010) e estando atualmente na Plataforma Portulan Clarin[4], como parte de um amplo conjunto de recursos computacionais.

O *LX-Parser* trabalha com o seguinte inventário de etiquetas (Tabela 1)[5]:

3 Uma lista mais ampla de analisadores, incluindo aplicativos com código proprietário, foi apresentada por Menuzzi e Othero (2008).

4 Disponível em: https://portulanclarin.net/workbench. Acesso em: 04 dez. 2021.

5 Lista disponível em: https://portulanclarin.net/workbench/lx-parser. Acesso em: 04 dez. 2021. Nessa lista, no entanto, não constam as etiquetas ART (artigo) e PNT (sinal de pontuação), que aparecem nos resultados do processamento. Na lista, apresenta-se a etiqueta O (ordinal), mas, nos dados processados, constatou-se ORD.

Tabela 1. Etiquetas do LX-Parser.

A: adjetivo	**NP**: sintagma nominal [6]
AP: sintagma adjetival	**ORD**: ordinal
ADV: advérbio	**P**: preposição
ADVP: sintagma adverbial	**PNT**: pontuação
ART: artigo	**PP**: sintagma preposicional
C: complementador	**PPA**: particípio passado/adjetivo
CL: clítico	**POSS**: possessivo
CP: sintagma complementador	**PRS**: pronome pessoal
CARD: cardinal	**QNT**: quantificador
CONJ: conjunção	**REL**: relativo
CONJP: sintagma conjuncional	**S**: frase
D: determinante	**V**: verbo
DEM: demonstrativo	**VP**: sintagma verbal
N: nome	

Fonte: Elaboração dos autores.

Essas etiquetas estão associadas ao modelo teórico em que se baseia o analisador: trata-se da Teoria X-barra do modelo gerativista (CHOMSKY, 1970; JACKENDOFF, 1977), em que se trabalha, portanto, com níveis intermediários (X') entre a categoria básica (X) e o limite superior do sintagma (XP) projetado pela referida categoria.

2.2 Experiências prévias em extração automática de sintagmas nominais (SNs)

A extração automática de SNs de textos em língua portuguesa tem sido discutida já há algum tempo, especialmente do ponto de vista da ciência da informação, para identificação de descritores.

Miorelli (2001), com base na descrição de sintagma nominal (SN) de Perini (1995), propõe um método de extração de SNs, a que denomina *ED-CER*, nome que reflete a forma de processamento: "as palavras dos textos são lidas da esquerda (E) para a direita (D), existe um 'corte' (C), uma 'exclusão' de palavras (E) e 'reconhecimento' do SN (R)" (MIORELLI, 2001, p. 47). Esse método se baseia em dois módulos de processamento: *módulo seletor* e *módulo analisador*. O módulo seletor encontra candidatos a SN a serem encaminhados para o módulo analisador. Inicialmente, faz-se a etiquetagem manual dos itens do *corpus*. Em seguida, o módulo realiza o corte das sequências, tomando como referência as chamadas *etiquetas de corte* (referentes a itens que não fazem parte do SN, como verbos, sinais de pontuação e locuções prepositivas, adverbiais e conjuntivas) e faz exclusão desses itens. Já o módulo analisador extrai do *corpus* pré-processado apenas os candidatos a SN que estiverem em conformidade com a gramática de SN proposta, que se baseou na fórmula SN $\rightarrow$ DET PN NSN MOD[7]. O protótipo foi desenvolvido em linguagem "C" e o teste de validação demonstrou 95,6% de precisão (MIORELLI, 2001, p. 66). Os casos de SNs não extraídos envolviam SN precedido por advérbio em *-mente* ("inicialmente os principais métodos") e SN dentro de SP ("na tradução entre as línguas portuguesa e inglesa"). A pesquisadora concluiu que essas falhas poderiam ser resolvidas com a inclusão de novas regras na gramática do seu modelo.

Diferentemente do trabalho acima citado, em que houve a proposta de um recurso computacional, há trabalhos mais recentes que se ocuparam de testar recursos pré-existentes, como o de Silva (2014). O estudo de Silva (2014)[8] é especialmente relevante para a presente discussão, uma vez que contempla igualmente o *LX-Parser*.

Silva (2014) fez um levantamento do estado da arte da indexação automática por SNs para textos

6 NP (Noun Phrase) é sigla equivalente a SN (Sintagma Nominal) ao longo do presente texto.
7 DET = determinante; PN = pré-núcleo; NSN = núcleo do SN; e MOD = modificador.
8 O estudo de Silva (2014) contém uma interessante resenha de diversos trabalhos relativos a extração automática de SNs desenvolvidos entre 1995 e 2011, com elucidativo quadro às pp. 160-161. Recomenda-se consulta a esse trabalho ao leitor interessado em conhecer cada uma das experiências.

em português, avaliou e comparou as ferramentas *PALAVRAS*, *OGMA*[9] e *LX-Parser*, usando como referência a extração manual de SNs. Em sua análise, apurou as taxas de acerto e de revocação[10] para cada uma dessas ferramentas, as quais se apresentam na Tabela 2:

Tabela 2. Resultados da avaliação das ferramentas.

	PALAVRAS	OGMA	LX-Parser
Taxa de acerto	94%	72,5%	75%
Taxa de revocação	86,5%	36,5%	97,5%

Fonte: Adaptado de Silva (2014, p. 168-179).

Como se pode ver, o *Palavras* foi o que teve a melhor taxa de acerto, mas o *LX-Parser* foi o que apresentou a melhor taxa de revocação. No que se refere aos problemas encontrados em relação a cada uma das ferramentas, Silva (2014) esclarece que:

a. quanto ao *PALAVRAS*, os problemas estavam relacionados à marcação de adjetivos e verbos como SN, à exclusão de palavras do SN, à identificação de numerais isoladamente como SNs e à confusão do artigo "*os*" com o pronome oblíquo, sendo extraído como SN;
b. quanto ao *OGMA*, os problemas disseram respeito à marcação de adjetivos e verbos como SN, a SNs extraídos iniciados por preposição ou conjunção e à fragmentação do SN (deixando-o incompleto);
c. quanto ao *LX-Parser*, os problemas estavam relacionados à marcação de adjetivos, advérbios, numerais, siglas e letras isoladas como verbo, e à identificação de adjetivos, verbos, numerais, artigos e símbolos e letras isoladas como SNs.

3 Hipótese de trabalho

Lança-se, aqui, a hipótese de que as *especificidades dos textos antigos limitam a precisão dos resultados gerados pelo analisador sintático LX-Parser*, justificada pelo fato de que os analisadores sintáticos geralmente são concebidos a partir de *corpora* compostos de textos modernos, como é o caso do *LX-Parser*, que foi criado a partir de treino em dados modernos, já que o *corpus* de base, formado de 10.140 sentenças, apresentava 8.952 (88,3%) extraídas de jornais, 403 extraídas de romances (4%) e 785 (7,7%) como série de teste (*test suite*). Sendo confirmada a hipótese, cabe, em seguida, identificar quais são essas especificidades e que tipo de informação o analisador precisa incorporar para ampliar o grau de precisão dos resultados para textos de fases mais antigas da língua portuguesa.

É necessário enfatizar aqui que, como o *LX-Parser* foi concebido para textos modernos, não se está avaliando sua aplicação *regular* (ou seja, exatamente aquela para a qual foi concebido), mas sim sua aplicação *estendida* (isto é, diferente daquela para a qual foi concebido).

4 Metodologia

Este estudo se enquadra em um projeto em curso[11] de elaboração de uma gramática do português clássico a partir de uma descrição linguística integral da obra *Peregrinação* (1614), de Fernão Mendes Pinto (*ca.* 1510-1583). Essa obra teria sido redigida por volta do terceiro quartel do séc. XVI e revela o uso linguístico escrito, mas com traços de oralidade, de um autor culto da variante da região de Lisboa (do dialeto padrão, portanto) na faixa etária de 60 anos, em um estilo mais ou menos informal (CAMBRAIA, 2000). Seu interesse para os estudos linguísticos reside no fato de se tratar de texto narrativo em prosa extenso, permitindo assim um estudo mais aprofundado do nível sintático. Ademais, há evidências de que seu autor tinha grande sensibilidade linguística, empregando, no texto, marcas linguísticas que seriam próprias de diferentes grupos sociais (CAMBRAIA, 2003). A edição

9 Para a ferramenta *OGMA*, cf. Maia (2008).

10 A taxa de acerto é "a razão do total de expressões que de fato constituem SN sobre o total de expressões extraídas pelos programas" e a taxa de revocação é "a razão do número de expressões que de fato são SNs extraídos pelos softwares sobre o número total de SNs extraídos pela técnica manual" (SILVA, 2014, p. 168).

11 Projeto "Para uma gramática do português clássico: o sintagma nominal e suas funções na *Peregrinação*, de Fernão Mendes Pinto [Fase I]", Produtividade em Pesquisa, CNPq, 2021-2024.

adotada para este estudo foi a edição digital da *Biblioteca Virtual dos Autores Portugueses* (BIBLIOTECA, 1998 (2 cd-roms))[12], conferida com a edição *princeps* de 1614[13] sempre que necessário, em função da constatação de alguns erros[14]: a edição digital adotou normas conservadoras de transcrição (CAMBRAIA, 2005), diferenciando-se da edição de 1614 basicamente pela transcrição de *s* longo como *s* curto, de *c*, como *ç*, de maiúscula após capitular como minúscula e da abreviatura *q̃* como *que*, pela união de palavras divididas em final de linha (com eliminação do hífen) e pela separação entre a indicação do capítulo e o texto do título do capítulo. Foram selecionados os 10 primeiros capítulos da obra, perfazendo um total de aprox. 9.000 palavras (*tokens*).

O *corpus* foi submetido a dois processos de extração de SNs: (a) manual[15], feito pelos autores deste estudo; e (b) automático, com base no analisador *LX-Parser*. A extração pelo *LX-Parser* contou com os seguintes passos: (1) criação de arquivo `.txt` com codificação UTF-8, contendo os 10 primeiros capítulos da *Peregrinação*, copiados da edição da BVAP, que apresenta o texto em formato `html`; (2) retificação do texto da edição em relação aos caracteres *ẽ* e *ũ* que estavam desconfigurados respectivamente como @ e ‰; (3) processamento do arquivo `.txt` no *LX-Tokenizer*, que faz separação de unidades lexicais (como no caso de contrações entre preposições e outras categorias)[16]; (4) processamento do arquivo `.txt` no *LX-Parser*, com geração de novo arquivo com dados no formato de visualização de parênteses; (5) processamento do novo arquivo `.txt` no *Tregex*[17] para seleção de SNs, utilizando o recurso de busca de padrão (*Search pattern*) com valor NP com a opção de exibição apenas da porção da árvore correspondente (*Show only matched portions of tree*); e (6) geração de novo arquivo `.txt` pelo *Tregex* com lista de apenas NPs[18].

Após essa coleta de dados, fez-se uma comparação entre os dois resultados (extração manual × automática) e analisaram-se as diferenças para verificar a natureza destas.

5 Descrição e discussão dos dados

A aplicação estendida do *LX-Parser* ao *corpus* com os 10 primeiros capítulos da *Peregrinação* de Fernão Mendes Pinto não apresentou resultado produtivo. Não se trata de resultado imprevisto, já que, como esclarecido, o *LX-Parser* foi concebido para ser aplicado a textos modernos, e não a textos antigos. O aspecto mais importante da experiência, portanto, não está nessa constatação, mas sim na possibilidade de identificar *quais são as especificidades de um texto do séc. XVI que limitam essa aplicação estendida do analisador sintático*.

Uma primeira análise dos resultados obtidos segundo os seis procedimentos descritos na metodologia permitiu perceber que o processamento pelo *LX-Tokenizer* criou um volume considerável de erros. Considerem-se os dados abaixo referentes ao texto do título do primeiro capítulo, em que (01a) contém o texto simples, (01b) contém o resultado de (01a) processado pelo LX-Tokenizer e (01c) contém o resultado (01b) processado pelo *LX-Parser*:

(01a) Do que passey em minha mocidade neste Reyno ate que me embarquey para a India.

(01b) <p> <s> de_ o que passey em minha mocidade em_ este Reyno ate que me embarquey para

12 Segundo esclarecido por Correia (1999, p. 185), a edição da BVAP baseia-se na edição de 1984 (PINTO, 1984), a qual, por sua vez, tomou a de 1829 (PINTO, 1829) como modelo (PINTO, 1984, p. VI), embora a de 1984 apresente normas de transcrição levemente diferentes.

13 Fac-símile digital do exemplar de cota RES-4409-V da BNP disponível em: https://purl.pt/82. Acesso em: 04 dez. 2021

14 Como erros mais graves apenas da edição de 1998, constam quatro repetições no lugar do texto correto: f. 51v (cap. 48) repetido no lugar do f. 59v (cap. 55); f. 72r (cap. 65) no lugar do f. 75v (cap. 68); f. 111r (cap. 97) no lugar do f. 112v (cap. 98) e f. 163v (cap. 136) no lugar do f. 165r (cap. 137). Esses trechos, no entanto, não fizeram parte do *corpus* do presente estudo, pois estão fora dos 10 primeiros capítulos.

15 A extração manual foi feita a partir de leitura do texto pelos autores deste estudo, seleção de sequências que consistiam em SN e transcrição em uma planilha. Nessa extração, não foram coletados os SNs compostos por pronomes pessoais, pronomes demonstrativos neutros e pronomes relativos (*que*, *o*(*s*) *que*, *a*(*s*) *que*, (*o*)*s qual*(*is*) e (*a*)*s qual*(*is*)). Esta decisão decorreu do objetivo do estudo dos SNs em andamento, que é o de compreender sua organização interna: nos casos citados, a estrutura apresenta pouca complexidade, não sendo, portanto, de interesse prioritário.

16 Em uma primeira rodada de processamento, o *LX-Tokenizer* travou em 97% de processamento e não terminou a tarefa. Em uma segunda rodada, no entanto, concluiu a tarefa.

17 Disponível em: https://nlp.stanford.edu/software/tregex.shtml. Acesso em: 04 dez. 2021.

18 Os autores agradecem ao Dr. João Silva (da Universidade de Lisboa), um dos desenvolvedores do *LX-Parser*, pelas orientações fornecidas para possibilitar a realização desse conjunto de procedimentos.

a India .*/ </s>

(01c) (ROOT (S (VP (PP (P de) (VP (V _) (NP (ART o) (N' (CP (NP (REL que)) (S (VP (V passey) (PP (P em) (NP (N' (POSS minha) (N' (N mocidade) (PP (P em) (NP (N _)))))))) (NP (DEM este) (N' (N' (N' (N Reyno) (A ate)) (CP (NP (REL que)) (S (V' (NP (CL me)) (V embarquey)) (PP (PP (P para) (NP (ART a) (N India))) (PNT .))))) (N' (N *) (N /)))))))))))))

Como se vê em (01b), o *LX-Tokenizer* inseriu marca de início de parágrafo ( ), de início de sentença (<s>), de separação de preposição (<_>), de pontuação (<*/>), de final de sentença (</s>) e de final de parágrafo ( ). Entretanto, o *LX-Parser* não reconheceu as marcas de separação e de pontuação como tal e considerou-as como parte do texto, atribuindo a elas etiquetas sintáticas: `V` para a primeira ocorrência de _, `N` para a segunda ocorrência de _ e `N` para *. Dada a alta frequência dessas marcas, o volume de erros tornou o resultado fortemente improdutivo.

Tendo ficado claro que foi justamente o *LX-Tokenizer* que deu origem ao problema das marcas acima mencionado, considerou-se então a possibilidade de fazer o processamento do texto sem passá-lo inicialmente por esse recurso, ou seja, eliminando o procedimento (3) descrito na metodologia. Realizado esse novo processamento, o resultado foi melhor, como se pode ver pelo dado em (01d), referente ao mesmo título de capítulo tratado aqui antes:

(01d) (ROOT (S (S (VP (PP (P de_) (NP (ART o) (N' (CP (NP (REL que)) (S (VP (V passey) (PP (P em) (NP (N' (POSS minha) (N' (N mocidade) (PP (P em_) (NP (DEM este) (N' (N' (N Reyno) (A ate)) (CP (NP (REL que)) (S (VP (V' (NP (CL me)) (V embarquey)) (PP (P para) (NP (ART a) (N India))))))))))))))))))))) (PNT .)))

O resultado permitiu perceber que o próprio *LX-Parser* fez a maioria das separações entre preposição e outras categorias (demonstrativos e artigos) — como no caso de neste > em este no exemplo em (01a) — que o *LX-Tokenizer* faria, exceto no caso da preposição em e do artigo definido (cf. *no*, *na*, *nos*, *nas*), o que limitou a precisão dos resultados.

A discussão que se segue toma por base essa segunda rodada de processamento, sem a aplicação do *LX-Tokenizer*.

5.1 Resultados quantitativos

Aplicados os procedimentos descritos na metodologia (sem o do *LX-Tokenizer*, como já esclarecido), o processamento pelo *LX-Parser* levou à identificação de 2.340 SNs. Desse total, foram excluídos 238, que eram os compostos de pronomes pessoais, pronomes demonstrativos neutros e pronomes relativos (que não fizeram parte da extração manual), deixando um total de 2.002 SNs para comparação. Levando em conta que a extração manual resultou em um total de 2.157 SNs e a automática 2.002 SNs, poder-se-ia pensar que o desempenho foi de precisão de 92,8%. Entretanto, um exame detalhado mostra que o desempenho foi muito inferior. Confrontando-se os SNs da extração manual e os da extração automática, apenas 540 foram corretamente identificados pelo *LX-Parser*, portanto, seu desempenho foi de 25%, o que é um valor muito baixo. Trata-se de valor muito inferior ao de 97,5% da taxa de revocação obtido por Silva (2014, p. 170), com base em *corpus* de textos modernos.

5.2 Fatores específicos que limitaram o desempenho

Dentre os fatores que limitaram o desempenho do *LX-Parser*, há primeiramente aqueles que se consideraram que são específicos de textos antigos, a saber: a questão da extensão das sentenças, a questão gráfica e a questão da variação e da mudança linguística.

5.2.1 A questão da extensão das sentenças

O primeiro obstáculo constatado para o adequado processamento do *corpus* foi a extensão das sentenças. O *LX-Parser* excluiu da análise 7 sentenças, as quais foram apresentadas em arquivo separado pelo programa. Informou-se a mensagem "*Error: processing this sentence was taking too long*" e também o tempo após o qual se decidiu pela interrupção do processamento, que foi entre 41 e 69 segundos (p. ex. "*Gave up after 69 seconds*"). Essas frases apresentavam extensão entre 197 e 369

palavras (segundo o critério gráfico).

A exclusão dessas 7 sentenças pelo *LX-Parser* levou a uma perda de 403 SNs que deveriam ter sido extraídos, o que significou uma queda de 18,7 pontos percentuais na taxa de desempenho (ou de revocação).

Embora sentenças longas não sejam propriamente uma especificidade de textos dos sécs. XVI e XVII, já que, em textos medievais, também se faziam presentes, ainda assim constituem um aspecto muito particular do estilo desse autor quinhentista, em cuja obra não são raras[19]. Nessa obra, é possível encontrar SNs pesadíssimos, estando um deles justamente dentre as sentenças que foram excluídas (cf. trecho em itálico):

(02) Mas por outra parte quãdo vejo que do meyo de todos estes perigos & trabalhos me quis Deos tirar sempre em saluo, & porme em seguro, acho que não tenho tanta razão de me queixar por todos os males passados, quãta de lhe dar graças por este só bẽ presente, pois me quis conseruar a vida, paraque eu pudesse fazer *esta rude & tosca escritura, que por erança deixo a meus filhos (porque só para elles he minha tenção escreuella) paraque elles vejão nella estes meus trabalhos, & perigos da vida que passei no discurso de vinte & hũ ãnos em que fuy treze vezes catiuo, & dezasete vendido, nas partes da India, Etiopia, Arabia felix, China, Tartaria, Macassar, Samatra, & outras muitas prouincias daquelle oriental arcipelago, dos confins da Asia, a que os escritores Chins, Siames, Gueos, Elequios nomeão nas suas geografias por pestana do mũdo, como ao diante espero tratar muito particular, & muito diffusamente, & daqui por hũa parte tomem os homẽs motiuo de se não desanimarem cos trabalhos da vida para deixarem de fazer o que deuem, porque não ha nenhũs, por grandes que sejão, com que não possa a natureza humana, ajudada do fauor diuino & por outra me ajudem a dar graças ao Senhor omnipotente por vsar comigo da sua infinita misericordia, a pesar de todos meus peccados, porque eu entendo & cõfesso que delles me nacerão todos os males que por mim passarão, & della as forças, & o animo para os poder passar, & escapar delles com vida.*

5.2.2 A questão gráfica

O segundo obstáculo foi a questão gráfica: o sistema de representação gráfica da *Peregrinação* apresenta diferenças em relação aos textos modernos, o que resulta em interpretação inadequada dos itens lexicais pelo *LX-Parser*.

Primeiramente, há a questão dos diacríticos. No que se refere ao til, por um lado, há os caracteres *ẽ* e *ũ*[20], combinações ausentes do sistema moderno; por outro lado, há os caracteres *ã* e *õ*, combinações presentes no sistema moderno, mas usados na *Peregrinação* em casos em que se empregam formas alternativas atualmente.

Nos casos de *ẽ* e *ũ*, parece ter havido mudança de codificação no curso do processamento: tanto o arquivo *.txt* inserido com o *corpus* quanto o arquivo .txt gerado com visualização das árvores sintáticas em parênteses estavam na codificação UTF-8, mas houve a substituição desses dois caracteres por sinal de interrogação. Parece, então, que se trabalhou intermediariamente, no curso do processamento, com a codificação ANSI, incompatível com os dois citados caracteres. A consequência dessa desconfiguração foi a de análise do sinal de interrogação efetivamente como um sinal de pontuação, gerando interpretação equivocada (o substantivo *tẽpo* foi analisado de forma fragmentada: o *t* foi etiquetado como N, o *ẽ* substituído por *?* e etiquetado como PNT, e o *po* etiquetado como N)[21]:

(03) ... *pella mayor parte, & milhor tẽpo da minha vida* ... ⇒ ... (NP (N' (N *pella*) (N *mayor*) (N *parte*))))))))))) (S (PNT ,) (S (S (NP (NP (N' (N &) (N *milhor*) (N *t*))) (NP (PNT *?*) (N' (N *po*) (PP (P *de_*) (NP (ART *a*) (N' (POSS *minha*) (N *vida*))))) ...

No caso de *ã* e *õ*, não houve a substituição desses caracteres por sinal de interrogação, mas o fato de a forma gráfica em que ocorrem ser diferente da moderna também levou à interpretação

19 Segundo Savoy (2020, p. 7), constata-se uma diminuição da extensão das sentenças ao longo das épocas.
20 No trecho do *corpus* analisado, não ocorre *ĩ*.
21 Nos exemplos, os itens linguísticos serão apresentados em itálico para facilitar a leitura.

equivocada (o adjetivo *grãdes* foi etiquetado como N e a preposição *cõ* como V):

(04) ... *os muitos & grãdes trabalhos & infortunios* ... ⇒ ... (NP (ART *os*) (N' (CP (NP (QNT *muitos*) (N' (N' (N &) (N *grãdes*) (N *trabalhos*) (N &) (N *infortunios*)) ...

(05) ... *ficando cõ isto as nossas fustas de todo mancas* ... ⇒ ... (VP (VP (V *ficando*) (VP (V' (V *cõ*) (NP (DEM *isto*))) (NP (ART *as*) (N' (POSS *nossas*) (N' (N *fustas*) (PP (P *de*) (NP (QNT *todo*) (N mancas)))))))) ...

Mas a questão se estende também a outros diacríticos empregados de forma diferente da moderna: a preposição *atê* (mod. *até*) foi etiquetada como N, o advérbio *câ* ou *cà* (mod. *cá*) como V, dentre outros.

Curiosamente, certos termos com formas diferentes na escrita moderna do português (em função do diacrítico) receberam etiquetagem correta: o verbo *pòr* (mod. *pôr*) foi etiquetado como V, os verbos de perfeito *forão* e *viraõ* (mod. *foram* e *viram*) igualmente como V, o adjetivo *nùs* (mod. *nus*) como A, dentre outros.

Em segundo lugar, há a questão dos *grafemas*. Existem algumas diferenças bem evidentes entre como se escrevia nos sécs. XVI-XVII e hoje. Assim, o processamento pelo analisador gerou erros relativos também às palavras com grafia diferente da moderna, tais como *Qvando*[22] e *prouue*:

(06) *Qvando às vezes ponho diante dos olhos* ... ⇒ (NP (N' (N *Qvando*) (PP (P *a_*) (NP (ART *as*) (N vezes))))) (VP (V' (V *ponho*) (ADV *diante*)) (PP (P *de_*) (NP (ART *os*) (N *olhos*))))) ...

(07) ... *sua derrota prouue a nosso Senhor* ... ⇒ ... (NP (NP (N' (POSS *sua*) (N' (N *derrota*) (A *prouue*)))) (NP (ART *a*) (N' (POSS *nosso*) (N *Senhor*)))))) ...

Era comum, nos textos da época de publicação da *Peregrinação*, o uso alternado de *v* e *u* de forma diferente da moderna (cf. mod. *Quando* e *prouve*). No primeiro caso, o *LX-Parser* etiquetou erroneamente a conjunção *Qvando* como N e, no segundo caso, o verbo *prouue*, perfeito de *prazer*, como A.

Outro exemplo interessante é a forma verbal *he*, presente de *ser*, que foi etiquetada como N:

(08) ... *mas ja que he forçado ser assi* ... ⇒ ... (NP (CONJ *mas*) (NP (N' (N *ja*) (CP (NP (REL *que*)) (S (NP (N' (N *he*) (A *forçado*))) (VP (V *ser*) (AP (A *assi*)

Um aspecto também relevante é a forma *&*, historicamente um composto dos grafemas *e* e *t* em nexo, ou seja, com o *t* usando parte dos traços do *e*. Esse composto, baseado na conjunção coordenativa latina *et*, passou por uma estilização formal e foi adotado como representação de conjunção coordenativa nos sistemas gráficos baseados no alfabeto latino, como é o caso do português. No exemplo em (04) acima, vê-se que ele foi etiquetado equivocadamente como N, e não como CONJ.

A grande diversidade na direção do erro de etiquetagem chama a atenção para uma certa imprevisibilidade em relação ao critério de decisão do *LX-Parser* para a escolha de determinada etiqueta.

Em terceiro lugar, há a questão da *separação vocabular*[23]. No português dos sécs. XVI-XVII, o processo de mudança da separação vocabular do sistema de vocábulo fonológico (baseado em grupo acentual) para o de morfológico (baseado em classe de palavra) já estava bastante avançado, mas, ainda assim, persistia registro em certos casos segundo o sistema mais antigo (cf. *perseguirme*, *paraque*, *aquem* etc.) e ocorriam também separações que não vingaram (cf. *com nosco* etc.), o que causou interpretação equivocada pelo analisador (*paraque* foi etiquetado como N em vez de P + CONJ; e *com nosco* como P + NP (N) em vez de P + NP (PRS)):

22 Como já assinalado, na edição da BVAP, as maiúsculas que se seguiam às capitulares foram convertidas em minúsculas: assim, no texto impresso da *Peregrinação*, estava na verdade *QVando* (com *Q* capitular), o que foi convertido na referida edição em *Qvando* (com *Q* não capitular).

23 A questão da separação vocabular é um aspecto bastante complicador para processamento automático de dados: como já assinalado por Cambraia (2007), o próprio fato de editores de textos antigos adotarem diferentes critérios em relação a este aspecto faz com que a coleta de dados em bases digitais de textos exija cuidados especiais para se evitarem conclusões equivocadas sobre o comportamento dos dados.

(09) ... *paraque em nome daquelle pouo fizesse* ... ⇒ ... (N' (N *paraque*) (PP (P *em*) (NP (N' (N *nome*) (N *daquelle*) (N *pouo*) (N *fizesse*) ...

(10) ... *fosse com nosco* ... ⇒ ... (S (VP (V *fosse*) (PP (P *com*) (NP (N *nosco*))))))) ...

Por fim, ainda em relação à questão gráfica há o tema da *pontuação*. O *LX-Parser* usa como referência para limite de frase o ponto final, mas, no texto da *Peregrinação*, esse sinal aparece com mais duas outras funções: sinal abreviativo (cf. *S. Tome*)[24] e identificador de algarismo[25] (cf. *1538.*). No primeiro caso, o item *S.* não foi analisado corretamente, tendo sido etiquetado como N e não como A (pois é uma abreviatura para o adjetivo *São*), e, além disso, foi considerado como final de fronteira de NP, mesmo não o sendo. Ademais, a questão da ausência do diacrítico inviabilizou a interpretação correta de *Tome* como N, tendo sido etiquetado como V (provavelmente considerado forma flexionada do verbo *tomar*):

(11) ... *trazião fretada de S. Tome* ... ⇒ ... (N *trazião*) (N *fretada*))))) (PP (P *de*) (NP (N *S.*))))) (VP (V' (V *Tome*) ...

No segundo caso, o sinal de ponto final foi interpretado como marcador de limite de período (e não como identificador de algarismo), o que fez com que a oração principal que se seguia à subordinada temporal fosse considerada um novo período (marcado na anotação por ROOT):

(12) *E surgindo as tres na barra de Diu a cinco de Setembro do mesmo anno de 1538. Antonio da Sylueira* ... ⇒ (ROOT (S (S (CONJ *E*) (S (VP (VP (V *surgindo*) (NP (ART *as*) (N' (N' (N *tres*) (A *na*)) (N' (N *barra*) (PP (P *de*) (NP (N' (N *Diu*) (PP (P *a*) (NP (CARD *cinco*))))))))))) (PP (P *de*) (NP (N' (N *Setembro*) (PP (P *de_*) (NP (ART *o*) (N' (A *mesmo*) (N' (N *anno*) (PP (P *de*) (NP (N *1538*))))))))))))) (PNT .))) (ROOT (S (S (NP (NP (N' (N *Antonio*) (PP (P *de_*) (NP (ART a) (N' (N' (N *Sylueira*) ...

É importante assinalar aqui que a dita questão gráfica se refere tanto ao fato de uma dada palavra ser grafada sistematicamente de forma diferente da moderna, como no caso de *daly*, registrada apenas com *y* na *Peregrinação*, quanto ao fato de uma mesma palavra ser grafada de diferentes formas, apresentando assim variação gráfica: cf. p. ex, *abstinẽcia* e *abstinencia*. Considera-se aqui que, nesses casos, a diferença gráfica não reflete diferenças no nível fonológico.

5.2.3 A questão da variação e da mudança linguística

Com o processo de normatização da língua portuguesa através de sua codificação em gramáticas prescritivas a partir do séc. XVI, a variação linguística, constitutiva de qualquer língua natural, passou paulatinamente a se tornar menos visível na escrita do que era antes, por exemplo, em registros escritos da Idade Média. Embora esse processo já estivesse em curso na época de redação da *Peregrinação* (séc. XVI) e de sua publicação (séc. XVII), ainda assim esse texto se mostrava permeável a uma maior expressão da variação linguística. Nesse caso, trata-se da variação gráfica, que refletiria variação em outros níveis linguísticos, como fonológico, morfológico, sintático etc. (cf. a variação fonológica em *razão* × *rezão*). Como, na escrita moderna, a adoção da ortografia limita a expressão da variação linguística, disso resulta que o *LX-Parser*, treinado em dados de textos modernos, encontra dificuldade para lidar com esse tipo de variação. Assim, por exemplo, o adjetivo comparativo *milhor* (na ortografia moderna, *melhor*) foi inadequadamente etiquetado como N:

(13) ... *para milhor fortuna* ... ⇒ ... (PP (P *para*) (NP (N' (N *milhor*) (N *fortuna*))))) ...

24 Para verificar como o *LX-Parser* analisa abreviatura em texto moderno, testou-se uma sentença criada para essa finalidade (*João mora na r. Camões*) e constatou-se que *r.* foi analisado corretamente como N.

25 Na escrita medieval, algarismos romanos geralmente apareciam entre pontos como forma de indicar que se tratava de número e não de outro tipo de classe de palavra: assim, *.vi.* deveria ser lido como algarismo *6* e não como o perfeito do verbo *ver*. A presença de ponto após número na *Peregrinação* no interior de uma sentença parece ser resquício desse sistema, ainda que, nessa obra, já se tenham usado algarismos arábicos, além dos romanos (estes, na numeração de capítulo).

A dificuldade também se verifica no caso de formas linguísticas que passaram por mudanças, como foi o caso do advérbio *despois* em locução conjuncional na *Peregrinação* (no português padrão moderno, *depois*), que foi etiquetado como N:

(14) ... *despois de auer sobre isto* ... ⇒ ... (S (NP (N' (N *despois*) (PP (P *de*) (NP (N' (N *auer*) (PP (P *sobre*) (NP (DEM *isto*) ...

Embora o próprio analisador tenha feito separação de certas contrações presentes na escrita moderna (cf. *das* > *de_as*, *neste* > *em_este* etc.), falta-lhe incluir as contrações comuns nos sécs. XVI-XVII, como *cos* (= *com os*), equivocadamente interpretado como `N`, em vez de `P + ART`:

(15) ... *cos mercadores que nella vinhão,* ... ⇒ ... (N' (N' (N *cos*) (A *mercadores*)) (CP (NP (REL *que*)) (S (VP (V *nella*) (A *vinhão*))))) (PNT ,))) ...

Não houve análise correta também nos casos de elisão que não envolvessem apenas palavras gramaticais, como é o caso de *douro*, interpretado erroneamente como N, e não como P + NP (N):

(16) ... *mandara de esmolla trezentas oqueas douro* ... ⇒ ... (NP (N' (N' (N *mandara*) (PP (P *de*) (NP (N' (N *esmolla*) (N *trezentas*) (N *oqueas*) (N *douro*))))) ...

Não é muito claro, no entanto, o quanto a variação limita o processamento correto pelo *LX-Parser*, já que, em certos casos, mesmo com diferenças linguísticas em relação ao português moderno, a etiquetagem foi correta, como N + A para *citim cramesim* (cf. mod. *cetim carmesim*):

(17) ... *hum chapeo forrado de citim cramesim* ... ⇒ ... (N' (N' (N &) (N *hum*) (N *chapeo*) (N *forrado*)) (PP (P *de*) (NP (N' (N *citim*) (A *cramesim*))))) ...

5.3 Fatores gerais que limitaram o desempenho

Dentre os fatores que limitaram o desempenho do *LX-Parser*, há também aqueles que foram constatados, mas que não eram específicos de textos antigos, a saber: a questão da ambiguidade estrutural e a das categorias linguísticas.

5.3.1 A questão da ambiguidade estrutural

Como Vieira e Lima (2001) já tinham assinalado, um aspecto difícil para o processamento automático é a ambiguidade estrutural. Uma frase como *O homem viu o menino com o telescópio* permite diferentes interpretações "o menino com o telescópio" ou "viu com o telescópio") e uma frase como *Cachorros e gatos felizes vivem na fazenda* é ambígua em relação ao alcance do adjetivo *felizes* (pode dizer respeito apenas aos gatos ou ainda aos cachorros e aos gatos).

No caso da *Peregrinação*, as estruturas ambíguas estão presentes, embora não sejam tão comuns (perfazem aprox. 50 casos em um universo de aprox. 2.157 SNs obtidos por extração manual nos 10 primeiros capítulos, ou seja, aprox. 2,3% dos dados). Há diversos tipos de situações, como nos exemplos a seguir:

a) SP (*neste Reyno*) associável a V (*passey*) ou a N (*mocidade*):

(18) ... *passey em minha mocidade neste Reyno* ... ⇒ ... (S (VP (V *passey*) (PP (P *em*) (NP (N' (POSS *minha*) (N' (N *mocidade*) (PP (P *em_*) (NP (DEM *este*) (N' (N' (N *Reyno*) ...

b) Elementos na margem esquerda de SN com coordenação, associáveis ao primeiro N ou a ambos os Ns (cf. *muytas* no primeiro exemplo[26] e *grandes* no segundo):

(19) ... *muytas gritas & apupadas* ... ⇒ ... (N *muytas*) (N *gritas*) (N &) (N *apupadas*) ...

(20) ... *grandes gritas & tangeres* ... ⇒ ... (A *grandes*) (N' (N *gritas*) (N &) (N *tangeres*)) ...

26 No exemplo (19), há também o problema de *muytas* ter sido etiquetado como N em vez de QNT, talvez por causa da grafia antiga com *y*.

c) Elementos na margem direita de SN com coordenação, associáveis ao último N ou a ambos os Ns (cf. o SP *da pobre casa de meu pay na villa de Montemór o velho*):

(21) ... *na miseria & estreiteza da pobre casa de meu pay na villa de Montemór o velho* ... ⇒ ... (N *na*) (N *miseria*) (N *&*) (N *estreiteza*))) (VP (PP (P *de_*) (NP (ART *a*) (N' (A *pobre*) (N' (N *casa*) (PP (P *de*) (NP (N' (POSS *meu*) (N' (N *pay*) (A *na*)))))))))) (VP (V *villa*) (PP (P *de*) (NP (N *Montemór*)))))) (S (NP (ART *o*) (N' (A *velho*) ...

d) Relativa (*que os Portugueses nos tinhão insinado*) associável a diferentes Ns precedentes (*cerimonias*, *cortesia* ou *vso*):

(22) ... *com mais outras cerimonias de cortesia ao seu vso que os Portugueses nos tinhão insinado* ... ⇒ ... (ADV *mais*))) (NP (ART *outras*) (N' (N *cerimonias*) (PP (P *de*) (NP (N' (N *cortesia*) (PP (P *a_*) (NP (ART *o*) (N' (POSS *seu*) (N *vso*))))))))))) (CP (NP (REL *que*)) (S (NP (ART *os*) (N' (N *Portugueses*) (A *nos*))) (VP (V *tinhão*) (A *insinado*)))))))))))))) ...

É interessante notar que o analisador não informa a existência ambiguidade, mas simplesmente propõe uma interpretação: no dado em (18), interpretou que o SP *neste Reyno* se liga ao N (*mocidade*) e não ao V (*passey*); no dado em (20), considerou que o adjetivo *grandes* se refere aos dois Ns (*gritas* e *tangeres*), e não apenas ao primeiro (*gritas*).

Nos casos de ambiguidade comentados acima, não há dados no contexto em que ocorrem que sejam capazes de solucionar a pluralidade de interpretações, por isso não se poderia esperar que o analisador apresentasse uma solução incontroversa para eles. Entretanto, não está claro quais são os critérios que o analisador utiliza para propor sua interpretação e, como ele não assinala que se trata de dados com ambiguidade estrutural, o usuário fica com a impressão de que se trata de dado de análise pacífica, o que não é o caso.

5.3.2 A questão das categorias linguísticas

Como apresentado no final da seção 2, o *LX-Parser* trabalha com uma tipologia específica de etiquetas e, portanto, de categorias linguísticas.

Por um lado, a hipercategorização de certas classes pode ser considerada um benefício, porque aumenta o nível de informação fornecida: em vez apenas da tradicional classe de pronomes, as categorias com que se trabalha são clítico (CL), demonstrativo (DEM), possessivo (POSS), pronome pessoal [tônico] (PRS), quantificador (QNT) e relativo (REL).

Por outro lado, a hipocategorização apresenta o efeito inverso, com perda de informação relevante: assim, há apenas a classe de artigos (ART), sem diferenciação de definidos e indefinidos.

Outra questão é a da realocação de diversos itens da classe dos pronomes (segundo a classificação tradicional) em outras classes: *tal* como determinante (D); *outro* como artigo (ART) ou adjetivo (A); e *mesmo* como adjetivo (A).

Ademais, as categorias com que trabalha o *LX-Parser* permite perceber que se baseia em um estágio mais antigo da teoria X-barra do modelo gerativo, em que, por exemplo, não se trabalhava com DPs (sintagmas determinantes) nem com IPs (sintagmas flexionais), o que parece corresponder ao estado da questão segundo trabalhos anteriores à 2ª metade da década de 1980 (RAPOSO, 1992). Nesse sentido, para a aplicação dos resultados segundo os modelos mais recentes, mesmo dentro do próprio quadro do gerativismo, seria necessária uma nova análise dos dados, a ser feita manualmente pelo pesquisador[27].

27 Othero (2009) aborda uma série de problemas que o processamento computacional apresenta no quadro do modelo gerativo para lidar com a estrutura do SN, tais como o advérbio como modificador nominal (p. ex., *até*, *eis*, *inclusive* etc.) e os elementos pesados no SN apenas em posição pós-nominal.

5.4 A otimização do analisador sintático

Como se viu acima, há cinco questões principais que devem ser enfrentadas para a otimização do *LX-Parser*, a fim de que possa ser aplicado a textos antigos, ou, pelo menos, a textos do séc. XVI: *extensão das sentenças, grafia, variação e mudança, ambiguidade estrutural e categorias linguísticas*.

Dessas cinco questões, a que parece mais simples de ser otimizada é a das *categorias linguísticas*. Considerando que a maioria das classes de itens gramaticais é de inventário fechado, há pouca dificuldade para fazer com que o analisador atribua etiquetas mais precisas: assim, é facilmente implementável uma etiquetagem que registre a diferença entre artigo definido (*o, os, a, as*), com uma etiqueta como ARTD, e artigo indefinido (*um, uns, uma, umas*), com etiqueta ARTI. Outra inovação auspiciosa seria passar a considerar como uma única forma as partes de estruturas gramaticalizadas como no caso das locuções conjuncionais (*para que, a fim de que* etc.) e prepositivas (*diante de, acima de* etc.), não as desdobrando em partes menores. Boa parte das gramáticas tradicionais apresenta extensa lista desse tipo de estruturas, o que permite inseri-las no *thesaurus*, que serve de base à etiquetagem no *LX-Parser*. Naturalmente, há expressões que não mais estão presentes em textos modernos e, por isso, escapam aos inventários das gramáticas tradicionais, como *so color de*, usada na *Peregrinação* com o sentido de "na condição de" (cf. *E cometeome se queria eu lá yr, porque leuaria nisso muyto gosto, para so color de Embaixador yr visitar de sua parte o Rey dos Batas, & yr tambem com elle ao Achem* [...]), mas especificidades como estas podem ir sendo paulatinamente incluídas no *thesaurus*, o que faz pensar que um modelo de analisador com possibilidade de personalização do *thesaurus* pelo usuário[28], com inclusão de elementos, permitiria adequar o analisador à necessidade de cada interessado[29].

A segunda questão que também não é tão difícil de ser implementada é a da *grafia*. Como já assinalado aqui antes, a equipe do *LX-Parser* tem em andamento a elaboração de um normalizador ortográfico, embora não estejam disponíveis informações mais precisas sobre o projeto. A questão da normalização gráfica para processamento computacional tem sido discutida já há algum tempo, e opções para a superação desse obstáculo têm sido propostas, como a construção de um dicionário de variantes gráficas (GIUSTI et al., 2007) ou processamento através de *softwares* específicos com conversão de formas como o VARD2 (BARON; RAYSON, 2008), testados por Hendrickx e Marquilhas (2011) e Marquilhas e Hendrickx (2014) para o português.No que se refere a textos antigos, uma fonte relevante para o desenho de um tal sistema são os trabalhos no âmbito da crítica textual, nos quais não raramente se apresenta um conjunto de normas para uniformização gráfica de textos no processo editorial: cf., p. ex., as normas de uniformização gráfica para edições interpretativas apresentadas por Cambraia (2005, p. 131-132). Para confirmação do peso dessa questão no processamento de dados pelo *LX-Parser*, aplicou-se a normalização gráfica nos dados já citados — cf. dados (03) a (07) —, e em quase todos eles, o analisador realizou a etiquetagem correta após a normalização: (03) *tempo*, N; (04) *grandes*, A; (05) *com*, P; (06) *Quando*, CONJ; (07) *prouve*, V; e (08) *é*, V. Entretanto, a questão parece ser mais complicada em relação a alguns dados: no caso de (09), a normalização de *paraque* em para que melhorou a análise de N para (PP (P) (NP (REL))), mas não foi exatamente para CONJ, o que seria o mais adequado levando em conta a gramaticalização dessa locução conjuntiva; e no caso de (10), mesmo a normalização de *com nosco* em conosco não resolveu a situação, porque passou de (PP (P) (NP (N))) para V, e não para (PP (P) (PRS)) como seria de esperar[30]. Vê-se, assim, que a normalização gráfica tem peso relevante para a produção de análises corretas, mas há outros aspectos a serem considerados, como a alomorfia do pronome pessoal na forma *conosco*.

Já a questão da variação linguística é mais complexa, porque há certa imprevisibilidade em relação às formas possíveis em textos antigos. Se, no que se refere à questão gráfica, há um escopo limitado de variações possíveis, o que torna as formas possíveis de certa forma previsíveis (cf. p. ex. a

28 Os autores agradecem a um dos pareceristas por assinalar que uma das vantagens dos analisadores de código aberto é justamente a possibilidade de modificar a base de dados, resultando em ganho no desempenho sem comprometer seu processamento de textos modernos. É necessário ponderar, no entanto, que se trata de uma atividade passível de ser realizada por linguistas-programadores, e não por linguistas em geral.

29 Para uma discussão sobre o tratamento de neologismos em uma perspectiva de processamento morfológico, cf. Alencar (2009).

30 Para verificar como o *LX-Parser* analisa conosco em texto moderno, testou-se uma sentença criada para essa finalidade (*João falou conosco*) e constatou-se que *conosco* foi analisado equivocadamente como N.

representação de ditongo nasal final como *–am*, *–ão* ou *–aõ* na *Peregrinação*), por outro lado, há uma ampla gama de variações que refletem aspectos fonológicos difíceis de serem previstos, como no caso de *razão* × *rezão*, sem contraparte no caso de *naçaõ* (ou seja, não se constata *neçaõ*). Por essa razão, criar módulos de regras para processamento de formas com variação linguística é bem mais complicado. É fato, porém, que a normalização sobre os casos de variação linguística também aumenta a precisão do *LX-Parser*, pois a normalização de *despois* em *depois* levou a etiquetagem de N como ADV no dado em (14) e de *cos* em *com os* de N para P + ART no dado em (15), mas de *douro* em *d'ouro* no dado em (16) não resolveu, porque se passou de N para N + S, já que o apóstrofo foi interpretado como sinal de pontuação e, portanto, marcação de final de sentença.

É importante assinalar que o resultado gerado pelo analisador não pode ser com as formas normalizadas, já que se trataria de afastamento dos dados autênticos: é necessário que se trate de um módulo intermediário de processamento, que realiza a operação de normalização para a etiquetagem e análise sintática, mas que apresente os resultados com as formas originais que foram inseridas como *input*.

Quanto à questão da extensão das sentenças, não está muito clara a razão de terem sido excluídas as muito longas. Em função da mensagem gerada como retorno ("*Error: processing this sentence was taking too long*"), parece que a limitação decorre do código de programação, que deve conter uma instrução para interromper o processamento a partir de certo tempo de duração, provavelmente para não ocupar o processador por tempo demasiado. Em se tratando de textos antigos, em que a grande extensão dos períodos não é rara, o tempo de processamento aceito precisa naturalmente ser mais longo.

Por fim, há a questão mais complexa de todas, que são os casos de ambiguidade estrutural. Há, por lado, a possibilidade de desambiguação com base em informações precedentes no contexto em certos casos, mas, em outros, essa desambiguação é simplesmente impossível, já que, muitas vezes, depende de conhecimento pressuposto pelo autor da produção linguística em relação aos leitores, e um processador não teria como buscar essa informação. Portanto, não se trata propriamente de um problema relacionado ao modelo do analisador sintático. Entretanto, o fato de o analisador não identificar para o usuário que há, nos dados processados, ambiguidade estrutural é certamente um problema, porque o usuário poderia considerar tratar-se de uma proposta de análise isenta de dúvida, mesmo não o sendo. Assim, por exemplo, para o dado em (18) seria conveniente que o analisador informasse ao usuário as duas possibilidades de análise (associação do SP ao V ou ao N), deixando a este a decisão de optar por uma ou outra ou então tratar separadamente dados com ambiguidade, opção esta que parece ser a mais adequada no caso, para se evitar arbitrariedade na opção por uma das análises possíveis.

As soluções propostas têm como objetivo tornar o analisador compatível com especificidades de textos antigos (mas não apenas), melhorando seu desempenho. Assim, não haverá má etiquetagem em função de diferenças gráficas, por exemplo, e, no caso de ambiguidade estrutural, o usuário será alertado para o fato, dando-lhe a oportunidade de tomar a decisão de como tratar o dado em questão.

6 Considerações finais

Neste estudo, analisou-se a aplicação estendida do analisador sintático *LX-Parser* em um *corpus* composto do trecho inicial da *Peregrinação* de Fernão Mendes Pinto, obra redigida no séc. XVI e publicada no séc. XVII. Os resultados dessa aplicação não se mostraram produtivos em função da alta frequência de problemas na análise produzida. Vê-se, portanto, que a afirmativa de Alencar (2011) sobre analisadores sintáticos anteriormente citada continua válida para o caso de aplicação estendida do *LX-Parser*, ou seja, trata-se de uma área que ainda demanda muitos esforços de pesquisa, uma vez que ainda não se obtêm resultados suficientemente adequados. Mais concretamente, deve-se reconhecer que essa tecnologia atualmente disponível ainda não serve para a finalidade de elaboração de uma gramática do português clássico a partir de uma descrição linguística integral da obra *Peregrinação*.

Referências

ALENCAR, L. F. de. Produtividade morfológica e tecnologia do texto: aspectos da construção de um transdutor lexical do português capaz de analisar neologismos. *Calidoscópio*, v. 7, n. 3, p. 199–220, 2009. Disponível em: http://revistas.unisinos.br/index.php/calidoscopio/article/view/4874. Acesso em: 10 mai. 2022.

ALENCAR, L. F. de. Utilização de informações lexicais extraídas automaticamente de corpora na análise sintática computacional do português. *Revista de Estudos da Linguagem*, v. 19, n. 1, p. 7–85, 2011. DOI: 10.17851/2237-2083.19.1.7-85. Disponível em: http://periodicos.letras.ufmg.br/index.php/relin/article/view/2553. Acesso em: 10 mai. 2022.

ALMEIDA, S. de et al. Selva: a new syntactic parser for Portuguese. In: MAMEDE, N. J. et al. (Ed.). *Computational Processing of the Portuguese Language*. Faro, Portugal: PROPOR 2003, jun. 2003. p. 102–109.

BARON, A.; RAYSON, P. A tool for dealing with spelling variation in historical corpora. In: POSTGRADUATE Conference in corpus linguistics. Birmingham: Aston University, mai. 2008.

BIBLIOTECA, Virtual dos Autores Portugueses. Coordenação científica de Ivo Castro, Teresa Amado, Cristina Almeida Ribeiro e Paula Mourão. 1998 (2 cd-roms).

BICK, E. *The parsing system "palavras": automatic grammatical analysis of portuguese in a constraint grammar famework*. 2000. Tese (Doutorado em Linguística) – Aarhus University, Aarhus.

CAMBRAIA, C. N. Contributo para uma gramática do português clássico: a linguagem da *Peregrinação* de Fernão Mendes Pinto. In: CONGRESSO NACIONAL DA ABRALIN. Florianópolis: UFSC, 2000. (2), p. 1355–1362.

CAMBRAIA, C. N. Mudança interrompida na história do português: *nós outros* e *vós outros*. In: CONGRESSO INTERNACIONAL DA ABRALIN. Fortaleza: UFC, mar. 2003. (2), p. 112–114.

CAMBRAIA, C. N. *Introdução à crítica textual*. São Paulo: Martins Fontes, 2005.

CAMBRAIA, C. N. Edições digitais como base para análises lingüísticas: revisão crítica de experiências. In: SEMINÁRIO DE ESTUDOS FILOLÓGICOS. Salvador: Quarteto, 2007. v. 1. (II), p. 13–24.

CHOMSKY, N. Remarks on nominalization. In: JACOBS, R.; ROSENBAUM, P. (Ed.). *English transformational grammar*. Washington: Georgetown University Press, 1970.

CONTIER, A.; PADOVANI, D.; JOSÉ NETO, J. Tecnologia adaptativa aplicada ao processamento da linguagem natural. In: MEMÓRIAS... São Paulo: EPUSP, 2010. p. 35–42.

CORREIA, J. D. P. A construção do colectivo na *Peregrinação*: percursos e significado. In: SEIXO, M. A.; ZURBACH, C. (Org.). *O discurso literário da Peregrinação: aproximações*. Lisboa: Cosmos, 1999. p. 169–212.

GIUSTI, R. et al. Automatic detection of spelling variation in historical corpus: an application to build a Brazilian Portuguese spelling variants dictionary. In: DAVIES, M. et al. (Ed.). *Proceedings of the Corpus Linguistics Conference (CL2007)*. Birmingham: University of Birmingham, 2007.

HENDRICKX, I.; MARQUILHAS, R. From old texts to modern spellings: an experiment in automatic normalisation. *Journal for Language Technology and Computational Linguistics*, v. 26, n. 2, p. 65–76, 2011.

JACKENDOFF, R. *X syntax: a study of phrase structure*. Cambridge, Mass: MIT Press, 1977. (Linguistic inquiry monographs, 2).

MAIA, L. C. G. *Uso de sintagmas nominais na classificação automática de documentos eletrônicos*. 2008. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte. Disponível em: https://repositorio.ufmg.br/handle/1843/ECID-7NXJKZ. Acesso em: 10 mai. 2022.

MARQUILHAS, R.; HENDRICKX, I. Manuscripts and machines: the automatic replacement of spelling variants in a Portuguese historical corpus. *International Journal of Humanities and Arts Computing*, v. 8, n. 1, p. 65–80, abr. 2014. DOI: 10.3366/ijhac.2014.0120. Disponível em: https://www.euppublishing.com/doi/10.3366/ijhac.2014.0120. Acesso em: 11 mai. 2022.

MARTINS, R. T.; HASEGWA, R.; NUNES, M. G. V. Curupira: a functional parser for Brazilian Portuguese. In: MAMEDE, N. J. et al. (Ed.). *Proceedings of Computational Processing of the Portuguese Language, 6th International Workshop, PROPOR 2003*. Berlin: Springer, jun. 2003. p. 179–183.

MENUZZI, S. de M.; OTHERO, G. de Á. Sintaxe X-barra: uma aplicação computacional. *Working Papers em Linguística*, p. 15–29, 2008. DOI: 10.5007/1984-8420.2008v9nespp15. Disponível em: https://periodicos.ufsc.br/index.php/workingpapers/article/view/1984-8420.2008v9nespp15. Acesso em: 11 mai. 2022.

MIORELLI, S. T. *ED-CER: extração do sintagma nominal em sentenças em português*. 2001. Dissertação (Mestrado em Ciências da Computação) – Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre.

OTHERO, G. de Á. *Grammar play: um parser sintático em Prolog para a língua portuguesa*. 2004. Dissertação (Mestrado em Linguística Aplicada) – Faculdade de Letras, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre.

OTHERO, G. de Á. *A gramática da frase em português: algumas reflexões para a formalização da estrutura frasal em português*. Porto Alegre: EdiPUCRS, 2009.

OTHERO, G. de Á.; MENUZZI, S. de M. *Linguística computacional: teoria e prática*. São Paulo: Parábola, 2005.

PERINI, M. A. *Gramática descritiva do português*. São Paulo: Ed. Ática, 1995.

PINTO, F. M. *Peregrinação de Fernão Mendez Pinto*. Lisboa: Typographia Rollandiana, 1829.

PINTO, F. M. *Peregrinação de Fernão Mendes Pinto e Itinerário de António Tenreiro, Tratado das Cousas da China, Conquista do Reino de Pegu*. Porto: Lello & Irmão, 1984.

RAPOSO, E. P. *Teoria da gramática: a faculdade da linguagem*. Lisboa: Caminho, 1992.

SAVOY, J. *Machine learning methods for stylometry: authorship attribution and author profiling*. Cham: Springer International Publishing, 2020. DOI: 10.1007/978-3-030-53360-1. Disponível em: https://link.springer.com/10.1007/978-3-030-53360-1. Acesso em: 11 mai. 2022.

SILVA, J. et al. Out-of-the-box robust parsing of portuguese. In: PARDO, T. A. S. et al. (Ed.). *Computational processing of the Portuguese language, 9th International Conference, PROPOR 2010, Porto Alegre, RS, Brazil*. Berlin: Springer, 2010.

SILVA, T. J. da. *Indexação automática por meio da extração e seleção de sintagmas nominais em textos em língua portuguesa*. 2014. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal de Pernambuco, Recife. Disponível em: https://repositorio.ufpe.br/handle/123456789/12950. Acesso em: 23 mai. 2022.

VIEIRA, R.; LIMA, V. L. S. de. Linguística computacional: princípios e aplicações. In: MARTINS, A. T.; BORGES, D. L. et al. (Org.). *As tecnologias da informação e a questão social*. Fortaleza: SBC, 2001. v. 3, p. 47–88.

Contribuições dos autores

César Nardelli Cambraia: Conceituação, Curadoria de dados, Recursos, Investigação, Metodologia, Visualização, Escrita – revisão e edição; **Ramon Cunha Sampaio Leite**: Curadoria de dados, Investigação, Escrita – rascunho original.