Proposta recebida em January 2021 e aceite para publicação em November 2021.

Uso de tecnologias linguísticas para estudar a evolução dos sufixos -ÇOM e -VEL no galego-português medieval a partir de *corpora* históricos

Use of Linguistic Technologies for Analysing the Evolution of Sufixes -ÇOM and -VEL in Medieval Galician-Portuguese from Historical Corpora

Pablo Gamallo
CiTIUS
Univ. de Santiago de Compostela

José Ramom Pichel
CiTIUS, USC / imaxin software
Santiago de Compostela

José Martinho Montero Santalha
Universidade de Vigo

Marco Neves
Universidade Nova de Lisboa

Resumo

O trabalho apresentado neste artigo tem dois objectivos. Por um lado, descreve a adaptação de duas ferramentas de processamento da língua natural ao galego-português medieval, nomeadamente um analisador morfossintático e um reconhedor de variedades medievais, e por outro, visa testar hipóteses linguísticas sobre a evolução de sufixos medievais mediante o uso dessas ferramentas em *corpora* históricos. Apesar de o desempenho das ferramentas ser inferior do que quando utilizadas para variedades modernas mais estandardizadas e com menos variabilidade formal, mostramos que é possível usá-las com grande fiabilidade para estudos quantitativos baseados em *corpus*. O estudo linguístico baseado em *corpus* permite-nos conferir que, pela sua distribuição de frequências, a presença dos sufixos -CION e -BLE nos textos medievais da Galiza foi provavelmente influenciada pelo castelhano baixo medieval.

Keywords

etiquetagem morfossintática, reconhecedor de línguas, linguística histórica, humanidades digitais

Abstract

The work presented in this paper has two objetives. On the one hand, it describes how to adapt two natural language processing tools to medieval Galician-Portuguese, namely a morphosyntactic analyzer and a medieval language recognizer, and on the other hand, it verifies linguistic hypotheses about the evolution of medieval suffixes by using these tools by using historical corpora. Although the performance of the tools is inferior to those used for more standardized modern varieties with less formal variability, we show that it is possible to use them with great reliability for quantitative corpus-based studies. The corpus-based linguistic study allows us to verify that, on the basis of their frequency distribution, the presence of the suffixes -CION and -BLE in medieval Galician texts is probably influenced by medieval Castilian

Keywords

part-of-speech tagging, language recognizer, historical linguistics, digital humanities

1. Introdução

A principal dificuldade para analisar automaticamente os textos medievais que se conservam é a sua grande variabilidade e falta de estandarização, o que faz com que ainda não se disponha de mecanismos para lematizar ou normalizar os textos mediante formas canónicas. Um exemplo claro desta variabilidade aparece refletido na Tabela 1, onde são registadas 15 variantes gráficas diferentes do termo *exceção*, encontradas em textos medievais da Galiza entre os séculos XIII e XV.

Esta extrema variabilidade formal está por trás da falta de recursos lexicais ou dicionários relativos ao período medieval, exceto algumas tentativas apresentadas em glossários ou dicionários etimológicos incompletos (Fillo & Lopes, 2013), ou dicionários parciais em fase de construção, como o Dicionário de Verbos do Português Medieval — DVPM (Xavier, 2005). Também há escasez de *corpora* anotados para serem usados como treino de sistemas de etiquetagem morfossintática e lematização. O Corpus Informatizado de Textos Portugueses Medievais (CIPM) (Xavier, 2016) é um interessante recurso textual com

DOI: 10.21814/lm.13.2.347

exçeçon	2
exceiçon	2
exçeiçon	7
excepcion	1
exçepçion	1
exçepçom	3
exçepçon	132
excepçon	2
exçeyçon	1
exepçon	19
exeçon	4

Tabela 1: Variantes de *exceção* nos textos medievais da Galiza.

etiquetagem morfossintática revisada, mas não fornece nenhum tipo de lematização dos tokens etiquetados.[1]

Em datas recentes, foi desenvolvido um novo módulo da ferramenta LinguaKit para a etiquetagem morfossintática de textos do galego-português medieval (Canosa et al., 2019). Para garantir uma maior abrangência do sistema e minimizar as limitações do módulo estatístico e do modelo de língua, o sistema foi enriquecido com regras e heurísticas que resultaram num sistema híbrido simbólico-estatístico. Como se mostrará na experimentação, a prestação deste sistema híbrido resulta em valores de exatidão modestos, bastante inferiores aos valores obtidos por sistemas treinados para a análise sincrónica de línguas modernas, já normalizadas e estandardizadas.

O presente artigo tem dois objetivos perfeitamente entrelaçados, um focado na área do processamento da linguagem natural e um outro filológico, orientado para realizar um estudo diacrónico baseado na análise de corpus. Desde o ponto de vista do processamento da linguagem, o artigo tem como objetivo descrever a adequação de duas ferramentas linguísticas ao galego-português medieval, nomeadamente a adaptação do módulo de análise morfossintática da *suite* linguística LinguaKit, e o treino e adição de novos modelos medievais ao reconhecedor de línguas QueLingua. Tanto a LinguaKit como o QueLingua são ferramentas que desenhámos e implementámos há vários anos e nas quais continuamos a trabalhar, para adaptá-las a novos domínios. Desde a perspectiva da análise de corpus, tencionamos estudar a evolução da distribuição da frequência dos nomes e adjectivos terminados em -ÇOM ou -VEL, em conjunto com as variantes terminadas em -CION e -BLE, tomando como fonte de estudo um corpus representativo do galego medieval da Galiza.

[1] https://cipm.fcsh.unl.pt/

A partir das experiências realizadas, chegámos a dois tipos de conclusões, umas relacionadas com as ferramentas linguísticas e outras com o estudo filológico.

No tocante às ferramentas, concluimos que o módulo de etiquetagem morfossintática do galego-português medieval e o reconhecedor de variedades medievais são úteis e fiáveis para seren usados em tarefas de análise de corpus, mesmo com limitações importantes na prestação devido à grande variabilidade da língua. Chegamos a esta conclusão mediante uma avaliação extrínseca, que é a principal contribuição do presente trabalho. Além das avaliações intrínsecas com valores de exatidão obtidos a partir dum corpus de teste, neste artigo também realizámos uma avaliação extrínseca das ferramentas, ou seja, medimos o seu desempenho a partir de outra tarefa externa, que é a análise quantitativa da distribuições dos sufixos medievais. Mais concretamente, usámos primeiro um método semi-manual sem ferramentas de PLN para calcularmos a distribuição dos sufixos, muito custoso em tempo e trabalho. E a seguir realizámos a mesma tarefa com o uso exclusivo das ferramentas PLN adaptadas à língua medieval e sem revisão manual. Os resultados desta comparação mostranos que as duas ferramentas avaliadas são válidas para automatizar a compilação e análise de dados quantitativos fiáveis extraídos de textos medievais.

No que diz respeito ao estudo baseado em corpus, mostramos que a análise automática destes *corpora* medievais permite mesmo validar ou refutar hipóteses linguísticas sobre a *prolificidade* (Viaro, 2012) dos sufixos objeto de estudo.[2] O presente trabalho insere-se, desta forma, no âmbito das Humanidades Digitais, sendo a primeira vez que se pretende verificar mediante técnicas de linguística de corpus e processamento da língua as hipóteses filológicas formuladas sobre a evolução diacrónica dos sufixos objeto de estudo.

O que resta do artigo está organizado do seguinte jeito. As hipóteses linguísticas sobre a evolução dos sufixos são introducidas na Secção 2. A seguir, na Secção 3, descrevem-se as duas ferramentas usadas: identificador de línguas medievais e analizador morfossintático do galego-português medieval. Estas ferramentas são utilizadas para a análise quantitativa dos dois pares de sufixos, -ÇOM/-CION e -VEL/-BLE, que é descrita na Secção 4. A avaliação das ferramen-

[2] Segundo Viaro (2012), a prolificidade refere-se à produtividade do ponto de vista diacrónico, e aponta, portanto, para o passado.

tas e da análise quantitativa é levada a cabo na Secção 5, para finalizarmos na Secção 6 com a enumeração das principais conclusões tiradas do estudo e de algumas novas ideias sobre trabalho futuro a desenvolver.

2. Hipóteses linguísticas sobre a evolução dos sufixos

2.1. Os nomes derivados de -TION

Segundo investigadores como Mariño (1998) e anteriormente Lorenzo (1985)[3], no tocante às vozes derivadas dos nomes latinos terminados em -TION, no galego sempre houve duas tendências, uma mais popular ou patrimonial, com a remoção do iode (-ÇOM), e uma outra mais culta, com a conservação ou reposição deste: -CION. No galego moderno, à diferença do português, com a penetração dos cultimos, triunfou sempre a forma culta. Trata-se portanto, segundo estes autores, duma evolução interna da língua da Galiza que se produziu independentemente de interferências externas.

Existem, porém, outros pesquisadores que consideram que o triunfo do sufixo culto -CION no galego moderno é devido às interferências do castelhano sobre a língua da Galiza (Freixeiro Mato, 1997; Ferreiro, 1997). No entanto, este processo não aconteceu na língua usada em Portugal, onde triunfaram as formas patrimoniais em -ÇOM.[4]

No presente estudo, iremos à procura de evidências no corpus diacrónico do galego-português medieval que nos permitam fortalecer ou rechaçar uma das duas hipóteses formuladas pela filologia: evolução interna ou interferência externa.

2.2. Os adjetivos derivados de -BĬLIS

Os sufixos descendentes de -BĬLIS sofrem dois tipos de evolução: a que mantém a vogal pós-tónica dando lugar a -vel(e) -uel(e), -uil(e), e a que se forma por síncope da vogal pós-tónica não final, dando lugar a -ble, -ule ou -bre. Segundo Mariño Paz (2005), as variantes sem síncope (variações de -VEL) são maioritárias até 1450 e sofrem uma rápida descida no seu uso a partir dessa data, a favor das variantes de -BLE. O próprio autor conclui que este fenómeno pode explicar-se por influência do castelhano, tal e come afirma em Mariño Paz (2005, p.111):

> [...] en todos os xéneros se percibe o predominio da opción -uel / -ueles (ou variantes) ata o ecuador do século XV. Paréceme fundada, por tanto, a sospeita de que a fulgurante expansión de -ble(s) na prosa notarial posterior a 1451 estivo en relación directa co aumento da familiaridade co castelán que se daría na actividade profesional de notarios e escribáns a partir do ecuador do século XV.

Como no caso anterior, procuraremos evidências baseadas em corpus que permitam conferir ou não esta hipótese filológica.

No entanto, devemos ter em conta uma limitação derivada das caraterísticas das fontes analisadas. Como bem explica Santalha & José-Martinho (2005), é preciso distinguir entre "edição paleográfica" e "edição filológica." A primeira é substancialmente fiel aos manuscritos, enquanto que a segunda procura reproduzir, de maneira regular, a língua que o escriba queria representar. No presente estudo, limitamo-nos a analisar os textos digitalizados das edições filológicas pois há muito poucas edições paleográficas digitalizadas.

No que resta do artigo, usamos as maiúsculas para representar a forma gráfica normalizada dos sufixos, enquanto escrevemos em minúsculas as variantes de cada um deles. Por exemplo -çon e -zom são variantes de -ÇOM, e -uel e -vil são variantes de -VEL.

3. Ferramentas PLN para a língua medieval

Para levar a cabo a análise linguística alvejada no presente trabalho, foram adaptadas e treinadas duas ferramentas: um etiquetador morfossintático e um identificador de línguas.

3.1. Etiquetador morfossintáctico para o galego-português medieval

Em Canosa et al. (2019), foi descrito o módulo de classificação e reconhecemento de entidades mencionadas para textos do galego-português medieval. Este módulo para textos históricos contém também um tokenizador, um lematizador e um

[3] Na sua edição da *Crónica Troiana*, Ramón Lorenzo escreve: "No galego, [...] sempre apareceron en xogo dúas tendencias: unha máis popular, coa supresión do iode, e outra máis culta, coa conservación ou resposición. Na lingua moderna, coa penetración dos cultimos, triunfou case sempre a forma culta" (Lorenzo, 1985, p.96).

[4] A nível lexical, pelo contrário, houve influência castelhana no Português (Messner, 2007; Silvestre, 2008; Venâncio, 2019).

etiquetador morfossintático, todos eles adaptados para o galego-português medieval, e integrados na suite de ferramentas linguísticas, LinguaKit (Gamallo & Garcia, 2017; Gamallo et al., 2018), com licença livre GPLv3.[5] Linguakit está também disponível como serviço web.[6]

O etiquetador e lematizador do galego-português medieval, ainda em fase de protótipo, contém os seguintes três elementos, dos quais o terceiro foi parcialmente adatado para o presente trabalho:

Léxico medieval: O etiquetador morfossintáctico de LinguaKit contém um léxico computacional de formas, onde cada forma é associada a um ou vários lemas e às correspondentes etiquetas morfossintácticas. O *tagset* empregado tem 255 etiquetas diferentes e baseia-se nas recomendações do Grupo EAGLES (Leach & Wilson, 1996). Este *tagset* é comum a outros sistemas de análise morfossintática, nomeadamente o FreeLing (Padró, 2012) e o TreeTagger para português.[7].

Em Canosa et al. (2019), descreve-se como foi construído um léxico específico medieval a partir dos termos mais frequentes dum corpus medieval de desenvolvimento. Este léxico foi inserido noutro maior, constituído pela reunião dos léxicos pertencentes aos módulos de galego e português contemporáneo de LinguaKit.

Modelo de língua: O classificador estatístico de LinguaKit está baseado num simples algoritmo bayesiano que trabalha com um modelo de língua constituído por bigramas de pares $< forma, etiqueta >$ e as suas probabilidades. O modelo usado para o processamento dos textos medievais foi desenvolvido a partir de duas fontes de dados: utilizou-se, por um lado, um modelo de galego moderno previamente treinado para o módulo correspondente do etiquetador (Garcia & Gamallo, 2015) e, por outro, treinou-se um novo modelo em base a um pequeno conjunto de textos medievais anotados automaticamente e revisados manualmente. Os dois modelos foram agrupados de tal maneira que se adicionaram ao modelo de galego moderno os novos pares do modelo medieval.

Regras linguísticas: O algoritmo de etiquetagem é um sistema híbrido, linguistico-estadístico, que contém, além dum modelo de probabilidades e um desambiguador bayesiano, um conjunto de regras que podem alterar a escolha do classificador. Dois tipos de regras são consideradas pelo sistema:

- Regras usadas para corrigir erros do classificador devido ao excesso de ambiguidade. Por exemplo, uma regra específica impede etiquetar as formas *as* ou *os* como pronome pessoal seguido de nome comum. Este tipo de regras atuam diretamente sobre o classificador bayesiano.
- Regras morfológicas pensadas para garantir uma maior abrangência do sistema, que alargam a cobertura do léxico. Estas regras, que se aplicam sobre o léxico, permitem associar etiquetas morfossintáticas a tokens desconhecidos quando estes contêm determinados afixos. Por exemplo, se um token desconhecido ou OOV (*out of vocabulary*) finaliza com a sequência *-ados*, é etiquetado como verbo em modo particípio, com os traços flexivos masculino e plural. As regras morfológicas são especialmente relevantes para melhorar a prestação de sistemas aplicados a línguas e variedades, como as dos textos medieviais, de grande variabilidade formal e nenhuma estandardização. Para o presente trabalho, foram desenvolvidas novas regras morfológicas com diferentes variantes de afixos do galego-português medieval. Nomeadamente, as 4 regras adicionadas ao módulo para levar a cabo as experiências do presente trabalho estão descritas na Tabela 2. Repare-se que as regras tomam em conta, não só informação morfológica, mas também informação sobre a presença ou não do token a etiquetar no dicionário de formas (OOV) e o facto de ser identificado ou não como nome próprio pelo módulo NER (Named Entity Recognition - Reconhecimento de Entidades Mencionadas).

3.2. Identificador de línguas medievais

Nos textos medievais galegos, nomeadamente os de tipo notarial, encontramos um grande número de parágrafos escritos em castelhano. Isto tem uma influência dupla: por um lado restringe a efetivadade das ferramentas de processamento da língua e, por outro, distorce as conclusões que se consigam tirar dos dados quantitativos extraídos. É portanto necessário efetuar um processo automático de identificação da língua parágrafo a parágrafo.

[5] https://github.com/citiususc/Linguakit

[6] https://www.linguakit.com

[7] (https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/Portuguese-Tagset.html)

Regra 1	Se T é um token OOV não identificado como nome próprio e cuja forma termina numa variante do sufixo -ÇOM ou -CION, então T é um nome comúm masculino singular: NCMS000
Regra 2	Se T é um token OOV não identificado como nome próprio e cuja forma termina numa variante do sufixo -ÇONS ou -CIONS, então T é um nome comúm masculino plural: NCMP000
Regra 3	Se T é um token OOV não identificado como nome próprio e cuja forma termina numa variante do sufixo -VEL ou -BLE, então T é um adjetivo singular com género neutro: AQ0CS0
Regra 4	Se T é um token OOV não identificado como nome próprio e cuja forma termina numa variante do sufixo -VEIS ou -BLES, então T é um adjetivo plural com género neutro: AQ0CP0

Tabela 2: Regras morfológicas adicionadas ao módulo de etiquetagem morfossintática medieval. A notação morfossintática é a requirida pelo tagset da LinguaKit.

Com esse objetivo, elaborámos o primeiro identificador de idioma entre galego-português medieval e castelhano medieval. Treinamos a ferramenta QueLingua (Gamallo et al., 2014) com o corpus diacrónico Carvalho (Pichel et al., 2020),[8] escolhendo para treino as variantes do castelhano entre os séculos XIII e XV, e textos de galego-português de Portugal dos mesmo séculos. O corpus de treino foi tamém enriquecido com textos do TMILG - Tesouro Medieval Informatizado da Língua Galega (Varela Barreiro et al., 2016). Em total, escolhemos 1 milhão e meios de tokens para o castelhano medieval e um número semelhante para o galego-português medieval. O identificador está baseado num método simples que toma como fonte de informação um dicionário ordenado com as formas mais frequentes de cada idioma a identificar. Dado que a distribuição das frequências segue a lei de Zipf, um pequeno conjunto de palavras duma língua representa uma ampla proporção do total de ocorrências de tokens em qualquer corpus textual dessa língua. Na configuração do nosso treino, usámos as 200 formas mais frequentes por cada idioma, que tendem a ser palavras gramaticais. A maioria são palavras muito curtas (entre 1 e 3 caracteres), pois são principalmente palavras gramaticais: artigos, preposições, pronomes e conjunções. Foi escolhido este número como resultado das experiências realizadas no trabalho descrito em Gamallo et al. (2014). Nesse trabalho, foram feitas avaliações com diferentes partições dum dicionário com as 5000 formas mais frequentes e observou-se que o F1-Score deixava de crescer significativamente a partir dum tamanho pequeno: 200. Este reduzido número tem a vantagem de permitir uma revisão manual rápida e eficiente para limpar malformações frequentes derivadas dos textos de treino.[9]

4. Corpus histórico e estudo quantitativo

Nesta secção, o objetivo é levarmos a cabo um estudo quantitativo da distribuição dos sufixos -ÇOM/-CION e -VEL/-BLE mediante o uso das ferramentas descritas na secção precedente (3) aplicadas a um corpus histórico. Consideramos essencial que, para levar a cabo estudos filológicos sobre estes ou outros fenómenos linguísticos, é importante termos a nosso dispor dados quantitativos que espelhem a distribuição e evolução temporal das formas alvejadas. No presente estudo, compararemos a distribuição no corpus das formas terminadas em -ÇOM e -VEL com a distribuição das variantes em -CION e -BLE. A análise quantitativa ajudará a mostrar qual das hipóteses formuladas na Secção 2 condiz melhor com os dados quantitativos extraídos. Começamos a secção descrevendo a coleção de textos históricos.

4.1. O corpus histórico

O corpus medieval da Galiza utilizado para este estudo forma parte do TMILG (Varela Barreiro et al., 2016) e consta de 24 documentos datados entre os séculos XIII e XV com 1,5 milhões de tokens. Foi inicialmente construído para estudos sobre distâncias entre variedades diacrónicas (Pichel et al., 2020). Os documentos abrangem diferentes géneros textuais: líricos, ensaístas e no-

[8] O corpus Carvalho está disponível em `fegalaz.usc.es/~gamallo/resources/Carvalho.tgz`

[9] O reconhecedor de línguas, QueLingua, e os novos dicionários medievais construídos estão disponíveis em `https://github.com/gamallo/QueLingua`.

tariais. Entre as obras compiladas, incluem-se as Cantigas de Santa Maria, Crónica Geral de Castela, Cancioneiro de Ajuda, cantigas de Airas Nunes e numerosas atas notariais. A lista completa dos textos incluídos no corpus está no Anexo A. Para dar a conhecer a língua empregada nos textos deste corpus, deixamos uma breve amostra na Tabela 3.

"et este nombre munene quiere dezir en arauigo tanto como enel nuestro lenguage de castiella. lo que desseamos. et cuenta aquel sabio que esta duenna era de buen seso et de grand conseio. (General Estoria. Alfonso X)"

Tabela 3: Amostra de texto medieval em galego-português na parte galega.

O corpus medieval de Portugal inclui textos dos séculos XIII ao XV e forma parte do corpus diacrónico e multilingue Carvalho com licença livre (Pichel et al., 2019, 2020). Pode ver-se uma breve amostra dum texto deste corpus na Tabela 4. O sub-corpus português de Carvalho consta de 1,7 milhões de tokens (ligeiramente maior que o corpus medieval da Galiza), e inclui também textos de diferentes géneros, como Chronica de Dom João I, Cantigas de Dom Dinis e documentos notariais.

"A quantos esta carta uiren faço saber que Domingos perez filho de Maria. martjz dicta Daynha mj mostrou hũa mha carta que de mjn ten pola qual eu enprazei a el. e aa primeira molher con que fosse casado dous casaaes e hũu Moynho que eu ei na quintaa de Maceeira." (Chancelaria de Dom Afonso. Volume I)

Tabela 4: Amostra de texto medieval em galego-português na parte portuguesa.

O identificador de idioma foi aplicado aos dois *corpora* para separar os parágrafos em galego-português, que são a imensa maioria, dos parágrafos em castelhano medieval, presentes sobretudo nos textos notariais do corpus da Galiza. Uma vez realizada a selecção dos parágrafos, o texto em galego-português medieval foi processado com o etiquetador morfossintático medieval de LinguaKit.

No resto da secção, analisamos primeiro a distribuição dos sufixos -ÇOM/-CION, e a seguir apresentamos a mesma análise com o par -VEL/-BLE.

4.2. Distribuição dos sufixos -ÇOM/-CION

Para levar a cabo o estudo sobre este par de sufixos, derivados do latino -TION, que constrói nomes de ação a partir de verbos, foram listadas primeiro todas as variantes gráficas identificadas nos textos e mostradas na Tabela 5. Foi tomada em conta a listagem publicada em Diéguez (2018). A variaçom é devida principalmente a três fenómenos de natureza morfofonológica: alomorfia da vogal temática, haplologia e fusão de vogais semelhantes (Cristine Prado & Massini-Cagliari, 2014).

-ÇOM:	-çom, -som, -zom, -çon, -son, -zon, -ção, -são, -sao, -çao, -çõ, -cón, -són, -zón, -zóm
-CION:	-cion, -sion, -siom, -çiom, -çion, -ción, -syon, -sióm, sión, -ción

Tabela 5: Variantes de -ÇOM e -CION em textos medievais.

Devido ao grande número de variantes para um mesmo sufixo, e para simplificar a tarefa, só as formas no singular foram consideradas.

4.2.1. Distribuição dos sufixos no corpus da Galiza

Uma vez definidas as variantes de cada sufixo, iniciamos o processo de extração a partir dos textos da Galiza etiquetados morfossintaticamente e identificados como sendo escritos em galego-português pelo reconhecedor de idioma. Deste jeito, foram extraídas todas as ocorrências das palavras etiquetadas mediante o tag NC (nome comum) e contendo alguma das variantes listadas na Tabela 5. Não foram considerados os nomes com iode desaparecido em estágios iniciais, nomeadamente *coraçom*, *razom* e *sazom*, cuja evolução é coincidente com a do castelhano.

Uma vez extraídas as palavras com os sufixos alvejados, contamos o número de ocorrências totais e número de palavras diferentes por variante e agregamos os resultados para obtermos o valor total de cada sufixo. As tabelas 6 e 7 apresentam a distribuição dos dados quantitativos de -ÇOM e -CION, respetivamente, em relação às suas variantes no corpus de textos medievais da Galiza.

Na última coluna das tabelas 6 e 7 mostra-se a probabilidade de cada variante do sufixo. Por exemplo, a probabilidade P de -çon, como variante de -ÇOM, é calculada na equação 1, onde f é a função que devolve a frequência de cada variante, v, pertencente ao conjunto {-ÇOM}.

-ÇOM	freq	formas	*P*
-sao	11	4	0,003
-som	80	18	0,022
-son	769	74	0,220
-são	1	1	0,0002
-són	24	8	0,006
-zom	19	4	0,005
-zon	61	21	0,017
-zóm	7	1	0,002
-zón	22	5	0,006
-çao	2	2	0,0005
-çom	67	23	0,019
-çon	**2119**	**226**	**0,606**
-ção	14	3	0,004
-són	61	29	0,017
-çõ	236	66	0,067
Total	3493	485	1

Tabela 6: Ocorrências (freq) das variantes de -ÇOM no corpus medieval da Galiza, junto com o número de formas de palavras diferentes por variante (formas) e probabilidade de cada variante (P).

-CION	freq	formas	*P*
-cion	15	11	0,087
-siom	2	2	0,011
-sion	33	16	0,191
-sióm	4	2	0,023
-sión	16	8	0,093
-çiom	1	1	0,005
çion	**78**	**40**	**0,453**
-ción	23	19	0,133
Total	172	99	1

Tabela 7: Ocorrências (freq) das variantes de -CION no corpus medieval da Galiza, junto com o número de formas de palavras diferentes por variante (formas) e a probabilidade da variante (P).

$$P(\text{-çon}) = \frac{f(\text{-çon})}{\sum_{v \in \{\text{-ÇOM}\}} f(v)} = \frac{2119}{3493} = 0.606 \quad (1)$$

A variante mais frequente do sufixo -ÇOM é -çon, com uma probabilidade de ocorrência de 0.606. É de salientar que a probabilidade de ocorrência desta variante é muito superior ao resto. No tocante, ao sufixo -CION, a variante mais frequente é çion, com uma probabilidade de 0.453. Segundo Santalha & José-Martinho (2005), nas edições paleográficas com os textos originais, o uso de 'õ' final era maioritário, mas este foi transcrito como 'n' final nas edições filológicas dos textos da Galiza, provavelmente por influência do castelhano.

-TION	freq	*P*	R_1	R_2
-ÇOM	3493	0,953	20,308	7,202
-CION	172	0,047	0,049	1,737

Tabela 8: Comparativa dos sufixos derivados de -TION (-ÇOM e -CION) no corpus medieval da Galiza, no tocante às ocorrências totais (freq), a probabilidade de ocorrência (P) e às duas ratios, R_1 e R_2.

A probabilidade dum sufixo dadas as duas alternativas derivadas do sufixo latino -TION é calculada dividindo a frequência agregada de todas as variantes do sufixo pelo total de ocorrências das duas alternativas. A equação 2 mostra o cálculo da probabilidade do sufixo -ÇOM dadas as alternativas (patrimoniais e cultas) derivadas do conjunto {-TION}, que reune todas as variantes de -ÇOM e -CION.

$$P(\text{-ÇOM}) = \frac{\sum_{v \in \{\text{-ÇOM}\}} f(v)}{\sum_{z \in \{\text{-TION}\}} f(z)} = \frac{3493}{3665} = 0.953 \quad (2)$$

A ratio R_1 devolve a razão entre os dois sufixos. As duas equações em 3 mostram a razão de -ÇOM para -CION e vice-versa.

$$\begin{aligned} R_1(\text{-ÇOM}, \text{-CION}) &= \frac{\sum_{v \in \{\text{-ÇOM}\}} f(v)}{\sum_{z \in \{\text{-CION}\}} f(z)} \\ &= \frac{3493}{172} = 20,308 \end{aligned} \quad (3)$$

$$\begin{aligned} R_1(\text{-CION}, \text{-ÇOM}) &= \frac{\sum_{v \in \{\text{-CION}\}} f(v)}{\sum_{z \in \{\text{-ÇOM}\}} f(z)} \\ &= \frac{172}{3493} = 0,049 \end{aligned}$$

A ratio R_2 devolve a razão entre a frequência total dum sufixo e o número de formas diferentes. É a ratio inversa à fórmula da *token/type ratio*, conhecida como TTR (Kettunen, 2014). Nas duas equações em 4, calcula-se a R_2 dos dois sufixos, onde T é a função que devolve o número de formas diferentes (ou *types*) associadas a todas as variantes dum sufixo.

$$R_2(\text{-ÇOM}) = \frac{\sum_{v \in \{\text{-ÇOM}\}} f(v)}{T(\text{-ÇOM})} = \frac{3493}{485} = 7,202 \quad (4)$$

$$R_2(\text{-CION}) = \frac{\sum_{v \in \{\text{-CION}\}} f(v)}{T(\text{-CION})} = \frac{172}{99} = 1,737$$

Em resumo: o sufixo -ÇOM tem uma frequencia total de 3493 ocorrências em 485 palavras ou formas diferentes, enquanto -CION ocorre só 172 vezes em 99 formas diferentes. A probabilidade de aparição do sufixo patrimonial é muito mais alta (0.953) que a do sufixo culto (0,047). No tocante à ratio de uso entre eles (R_1), o primeiro é 20 vezes mais usado do que o segundo. No que diz respeito à ratio R_2, cada forma diferente em -CION ocorre de média menos de 2 vezes no corpus, enquanto que as formas com variantes de -ÇOM tendem a ter uma frequência média muito superior (7,2) e, portanto, têm mais produtividade, é dizer *prolificidade* (vd. nota supra 2), e uso na época medieval. Esta informação comparativa pode consultar-se na Tabela 8, que mostra os resultados agregados de frequência e as medidas relativas: P, R_1 e R_2. Pode-se concluir, a partir destes cálculos, que o uso do sufixo patrimonial -ÇOM é quase hegemónico nos textos medievais da Galiza, sendo o uso do sufixo culto muito minoritário.

4.2.2. Distribuição dos sufixos no corpus de Portugal

Realizamos a mesma análise sobre os textos de Portugal etiquetados morfossintaticamente. Mostramos os resultados nas tabelas 9 e 10

-ÇOM	**freq**	**formas**	***P***
-sao	28	1	0.007
-som	470	77	0.124
-son	38	13	0.010
-são	38	1	0.018
-són	1	1	0.0002
-zom	275	13	0,072
-zon	37	14	0.009
-çao	14	13	0.003
-çom	**2239**	**386**	**0.591**
-çon	155	51	0.040
-ção	282	102	0.074
-són	1	1	0.0002
-çõ	175	90	0.046
Total	3784	770	1

Tabela 9: Ocorrências (freq) das variantes de -ÇOM no corpus medieval de Portugal, formas diferentes e probabilidade de cada variante (P).

A variante mais frequente do sufixo -ÇOM é -çom (terminada em 'm'), com uma probabilidade de ocorrência de 0,59 (ver Tabela 9). Repare-se que nos textos da Galiza a variante mais frequente temina em 'n', nomeadamente -çon.

A Tabela 11 mostra os resultados agregados de frequência e as medidas relativas: P, R_1 e R_2. Observamos que não há grandes diferenças entre os dois *corpora* no tocante aos resultados agregados, além do uso oposto do 'm/n' final e um uso de -CION ainda mais residual no corpus de Portugal. De resto, existe uma grande simetria no uso das variantes de -ÇOM nos textos da Galiza e de Portugal: encontramos quase as mesmas variantes gráficas e um número muito próximo de ocorrências de palavras com este sufixo. Além de mais, a probabilidade das duas variantes mais frequentes nos textos da Galiza, -çon e -son, é de 0,606 e 0,220, respetivamente, enquanto nos textos de Portugal as duas variantes mais frequentes são -çom e -som, com uma probabilidade de 0,591 e 0,124. Uma vez identificadas as diferenças no 'm/n' final, as distribuições no uso das variantes são muito semelhantes.

-CION	**freq**	**formas**	***P***
cion	8	5	0,307
siom	**9**	**3**	**0,346**
sion	6	3	0,230
çiom	2	2	0,076
çion	1	1	0,038
Total	26	14	1

Tabela 10: Ocorrências (freq) das variantes de -CION no corpus medieval de Portugal, formas diferentes e a probabilidade da variante (P).

Apesar destes fortes paralelos na língua medieval, o galego moderno padronizou a forma -CIÓN (como em espanhol) enquanto o português moderno reuniu em -ÇÃO as formas derivadas de -TION junto com as doutros sufixos latinos: -ANT, -UNT, -AN, -ON.

-TION	**freq**	***P***	**R_1**	**R_2**
-ÇOM	3784	0.993	145,53	4,914
-CION	26	0.007	0,006	1,857

Tabela 11: Comparativa dos sufixos derivados de **-TION** (-ÇOM e -CION) no corpus medieval de Portugal, no tocante às ocorrências totais (freq), à probabilidade de ocorrência (P) e a duas ratios, R_1 e R_2.

4.2.3. Indícios de castelhanização nos textos da Galiza

Dois novos testes foram conduzidos com o intuito de conferir se o uso do sufixo -CION está relacionado com a crescente castelhanização da língua medieval na Galiza.

Em primeiro lugar, foi realizada uma análise diacrónica, século a século, da distribuição dos dois sufixos na Galiza. Para este propósito, o corpus da Galiza foi dividido em três partições: textos do século XIII, do XIV e do XV. A Figura 1 mostra a evolução da ratio R_1 de -ÇOM para -CION ao longo dos três séculos. A figura revela que a proporção de palavras com o sufixo -ÇOM ao respeito de -CION é claramente menor no século XV que em séculos anteriores. Enquanto nos séculos XIII e XIV a ratio de palavras com sufixo em -ÇOM é por volta de 27 vezes superior, este valor baixa a 13 no século XV. Isto parece ser um indício de que a presência de -CION está relacionada com uma maior castelhanização, pois é no século XV onde a influência do castelhano sobre o galego é maior (Ferreiro, 1997).

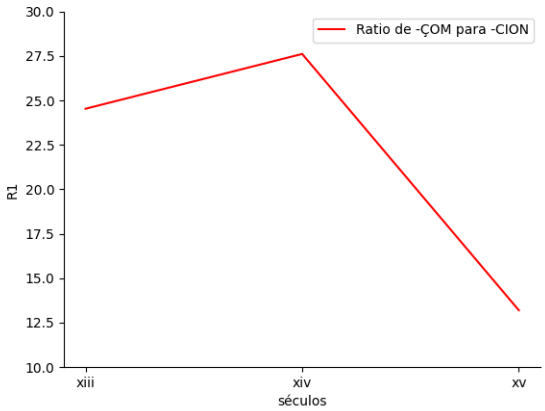

Figura 1: Evolução diacrónica ao longo de três séculos da ratio R_1 de -ÇOM para -CION, no corpus medieval da Galiza.

Em segundo lugar, foi levada a cabo a análise de distribuição dos dois sufixos utilizando só os excertos dos textos da Galiza identificados automaticamente como sendo escritos em castelhano pelo reconhecedor de língua (ver Tabela 12). Mesmo se a quantidade de texto é muito menor, as medidas relativas P, R_1 e R_2 mostram uma tendência contrária à encontrada nos textos em galego-português: maior presença de -CION que de -ÇOM e com maior ratio de uso.

-TION	**freq**	P	R_1	R_2
-ÇOM	62	0.319	0,469	1,589
-CION	132	0.681	2,129	2,425

Tabela 12: Comparativa dos sufixos derivados de **-TION** (-ÇOM e -CION) no corpus medieval da Galiza em excertos em castelhano, no tocante às ocorrências totais (freq), à probabilidade de ocorrência (P) e às duas ratios, R_1 e R_2.

4.3. Distribuição dos sufixos -VEL e -BLE

O mesmo tipo de experiência foi levada a cabo para analisar a distribuição do par -VEL/-BLE, derivados do sufixo latino -BĬLIS, que constrói adjetivos a partir de verbos. Baseamo-nos no trabalho de Mariño Paz (2005) para listar as variantes possíveis dos dois sufixos (ver Tabela 13). Foram extraídas todas as ocorrências das palavras etiquetadas mediante o tag AQ (adjetivo comum) e contendo alguma das variantes listadas na Tabela 13. Como no caso descrito na secção anterior, só as formas no singular foram consideradas.

-VEL:	-vel, -bel, -uel, -vil, -uil, -uel, -uele, -vele, -velle, -uelle
-BLE:	-ble, -vle, -ule

Tabela 13: Variantes de -VEL e -BLE em textos medievais.

4.3.1. Distribuição dos sufixos no corpus da Galiza

As tabelas 14, 15 e 16 mostram os resultados obtidos a partir dos textos da Galiza. Embora haja muitos mais casos de -VEL que de -BLE, as ratios são menores do que no par -ÇON/-CION.

Podemos observar também que as palavras com sufixos derivados de -BĬLIS são muito menos numerosas que as derivadas de -TION. Trata-se portanto dum sufixo pouco produtivo na Idade Média, aplicável a muito poucas formas lexicais.

-VEL	**freq**	**formas**	***P***
-bel	1	1	0,006
-uel	49	**17**	0,304
-uele	5	3	0,031
-uelle	6	4	0,037
-uil	31	14	0,192
-vel	**52**	8	**0,322**
-vele	3	1	0,018
-vil	14	9	0,086
Total	161	57	1

Tabela 14: Ocorrências (freq) das variantes de -VEL no corpus medieval da Galiza, formas diferentes por variante e probabilidade de cada variante (P).

-CION	**freq**	**formas**	***P***
ble	16	7	0,80
ule	2	1	0,10
vle	2	11	0,10
Total	20	9	1

Tabela 15: Ocorrências (freq) das variantes de -BLE no corpus medieval da Galiza, formas diferentes por variante e a probabilidade de cada variante (P).

-BĬLIS	**freq**	***P***	**R_1**	**R_2**
-VEL	161	0,889	8,050	2,824
-BLE	20	0,111	0,124	2,222

Tabela 16: Comparativa das ocorrências totais (freq), da probabilidade de ocorrência (P) e de duas ratios (R_1 e R_2) dos sufixos derivados de **-BĬLIS** (-VEL e -BLE) no corpus medieval da Galiza.

4.3.2. Distribuição dos sufixos no corpus de Portugal

As tabelas 17, 18 e 19 mostram os resultados obtidos a partir dos textos de Portugal. Como no caso de -CION, as variantes de -BLE são marginais ao respeito das de -VEL. Observamos que o número de variantes diferentes dos sufixos -VEL e -BLE são menores que nos textos da Galiza. Parece, portanto, que há um maior grau de estandarização nos textos de Portugal em relação a este tipo de adjetivos.

-VEL	**freq**	**formas**	***P***
-bel	3	3	0,008
-uel	**120**	**44**	**0,335**
-uil	94	16	0,262
-vel	118	54	0,329
-vil	23	12	0,064
Total	358	129	1

Tabela 17: Ocorrências (freq) das variantes de -VEL no corpus medieval de Portugal, formas diferentes e probabilidade de cada variante (P).

-BLE	**freq**	**formas**	***P***
ule	2	2	1
Total	2	2	1

Tabela 18: Ocorrências (freq) das variantes de -BLE no corpus medieval de Portugal, junto com o número de formas de palavras diferentes por variante (formas) e a probabilidade da variante (P).

-BĬLIS	**freq**	***P***	**R_1**	**R_2**
-VEL	358	0,994	179.0	2,0
-BLE	2	0,006	0,005	1.0

Tabela 19: Comparativa das ocorrências totais (freq), da probabilidade de ocorrência (P) e de duas ratios (R_1 e R_2) dos sufixos derivados de **-BĬLIS** (-VEL e -BLE) no corpus medieval de Portugal.

4.3.3. Indícios de castelhanização nos textos da Galiza

Como no caso do par anterior, foram realizadas experiências suplementares em busca de indícios que demostrem ou não que o uso do sufixo -BLE está relacionado com a castelhanização da língua medieval na Galiza.

A Figura 2 mostra uma maior proporção (ratio R_1) de variantes de -VEL sobre variantes de -BLE nos séculos XIII e XIV, frente ao século XV. Esta mesma tendência também se pode observar no trabalho de Mariño Paz (2005), que defende, para este par de sufixos, a hipótese da influência castelhana.

Por outro lado, não há maior número de ocorrências de -VEL frente a -BLE nos parágrafos identificados automaticamente como sendo escritos em castelhano. Como no caso do par anterior, encontramos a tendência inversa (embora com muito poucos casos e valores mais igualados), tal e como se mostra na Tabela 20.

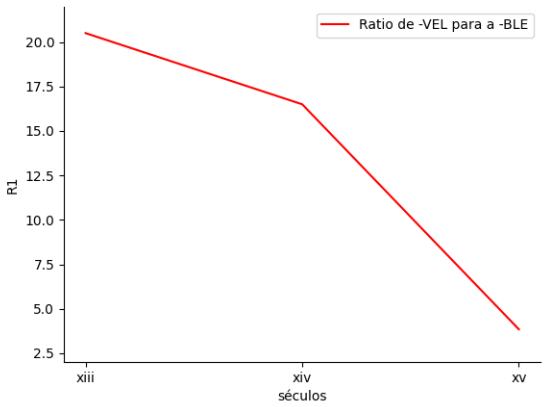

Figura 2: Evolução diacrónica ao longo de três séculos da ratio R_1 de -VEL para -BLE, no corpus medieval da Galiza.

-BĬLIS	**freq**	***P***	$\mathbf{R_1}$	$\mathbf{R_2}$
-VEL	3	0,429	0,750	1,00
-BLE	4	0,571	1,333	1,333

Tabela 20: Comparativa das ocorrências totais (freq), da probabilidade de ocorrência (P) e de duas ratios (R_1 e R_2) dos sufixos derivados de **-BĬLIS** (-VEL e -BLE) no corpus medieval da Galiza em excertos em castelhano.

4.4. Discusão

As experiências descritas parecem demostrar que as vozes cultas das palavras com sufixos derivados de -TION e -BĬLIS, é dizer, -CION e -BLE, respetivamente, são muito minoritárias nos textos da Galiza e residuais nos de Portugal. Os testes suplementares mostram que a maior presença na Galiza destas formas cultas correlaciona com indicadores de castelhanização, o que contradiz a hipótese de Lorenzo (1985) e Mariño (1998) sobre o par -ÇON/-CION, que sustentam que se trata duma evolução interna da língua, alheia a influências externas. Aliás, os indícios de castelhanização são muito mais conclusivos no caso de -CION que no de -BLE, mesmo se no segundo caso é admitida a influência castelhana (Mariño Paz, 2005).

Por último, se houve desde a Idade Media um registo culto interno e próprio da língua medieval da Galiza que favoreceu as formas em -CION, então é difícil explicar como é possível que a forma "popular," é dizer, patrimonial, apareça em muitas vozes de registo culto, tais como: *absoluçon, alegaçon, anexaçon, comutaçon, disençon, diçon, encarnaçon, encorporaçon, interposiçom, precisson, procuraçom, restituyçon, validaçón*, etc. Os dados achados no corpus semelham contradizer a principal hipótese filológica na que se baseia a norma moderna do galego para defender o uso de -CION frente a -ÇOM.

5. Avaliações das ferramentas

Nesta secção, o objetivo é avaliarmos as ferramentas PLN de duas maneiras diferentes:

Avaliação intrínseca: mediante o uso de corpus de teste anotados.

Avaliação extrínseca: mediante a comparação dos resultados obtidos automaticamente com as ferramentas e os obtidos mediante uma análise tradicional de corpus sem ferramentas PLN, só com expressões regulares e revisão manual. Para levar a cabo esta avaliação, os resultados da análise tradicional deram lugar à criação dum *gold standard* composto pelos dados de referência revisados manualmente.

5.1. Avaliação intrínseca

A Tabela 21 mostra os valores de exatidão (*accuracy*) obtidos na avaliação do reconhecedor de língua e do etiquetador de galego-português medieval.

O reconhecedor de língua foi avaliado a partir dum conjunto de 300 parágrafos extraídos de textos dos três séculos: 100 parágrafos do século XIII, 100 do XIV e 100 do XV. Foram extraídos aleatoriamente e alguns deles representam só uma pequena frase de poucas palavras, o que dificulta a identificação da língua. O sistema atingiu um 94,3% de exatidão (283 parágrafos identificados corretamente de 300). Os resultados foram avaliados manualmente. Devemos ter em conta que, apesar de não se poder fazer uma comparação direta, uma vez que os dados são diferentes, as experiências reportadas por Zampieri et al. obtiveram menos de 95% de exatidão para variantes e línguas similares.

	exatidão
Reconhecedor de língua	94,3%
Etiquetador morfossintático	87,4%

Tabela 21: Valores de exatidão do reconhecedor de línguas medievais e do etiquetador morfossintácico do galego-português medieval.

Para avaliarmos o etiquetador, utilizamos como corpus de teste o documento *Pedro Rodrigiz notario publico del Rey en Trasancos* datado

de 1257, que consta de 564 tokens. O valor de exatidão do sistema é de 87,4%, considerando só a etiqueta principal e sem tomar em conta o lema nem a informação morfológica. Portanto, a avaliação foi levada a cabo tomando em conta a primeira letra de cada etiqueta, pois é a que codifica a classe de palavra (nome, adjetivo, verbo, etc). O número total de classes de palavra do *tagset* é 11.

O valor de exatidão obtido é baixo, mesmo se comparado com outros etiquetadores adatados às variedades históricas. No trabalho descrito em Sánchez-Marco et al. (2011) para o espanhol antigo, a melhor configuração do etiquetador atinge o 94,5% de exatidão. Em Rögnvaldsson & Helgadóttir (2008), para textos medievais do antigo nórdico (Old Norse), língua da que derivam o norueguês, islandês, danês e sueco, o sistema atinge um 92% de exatidão.

5.2. Avaliação extrínseca

Mesmo se os resultados do etiquetador não são muito encorajadores, na seguinte avaliação medimos a qualidade dos resultados obtidos automaticamente com o etiquetador, contrastando os erros e acertos com um conjunto de dados de referência revisados manualmente e criados a partir do corpus de textos da Galiza.

O conjunto de dados de referência é a lista correta de formas com os sufixos estudados (-ÇOM/-CION e -VEL/-BLE). Esta lista foi construida selecionando todos os tokens possíveis cuja terminação coincide com as variantes dos sufixos, gerando assim uma lista ampla de candidatos. Para gerar esta lista de candidatos, utilizamos um simples tokenizador e expressões regulares. Posteriormente, a lista de candidatos foi revisada manualmente e só os casos corretos foram selecionados. A partir desta revisão manual foram criados dois léxicos de frequências, um com todas as variantes terminadas em -COM/-CION e outro com as terminadas em -VEL/-BLE, que estão disponíveis para descarga livre.[10]

A Tabela 22 mostra os valores de precisão, abrangência (*recall*) e F1 obtidos comparando a lista de referência ou *gold standard* com as listas extraídas mediante o uso do etiquetador morfossintático. Estamos, de facto, a avaliar o desempenho da estratégia automática (sem revisão manual) de extração das formas com os sufixos objeto de estudo e descrita na Secção 4.

Nesta tarefa, o método automático baseado na etiquetagem atinge um F1 muito aceitável (95,54%) e muito superior ao desempenho do etiquetador (87,4%). Isto é devido a que o etiquetador tende a ser mais preciso na etiquetagem de nomes e adjetivos, as etiquetas pertinentes na nossa tarefa filológica, que na etiquetagem doutro tipo de formas com significado gramatical. Estas formas tendem a ser mais ambíguas e frequentes, como é o caso dalguns determinantes, pronomes e conjunções, que as categorias lexicais.

Na última coluna da Tabela 22 mostramos o valor da correlação de *Pearson*, calculado tomando em conta a relação de valores da coluna de frequências da lista extraída automaticamente e a da lista de referência. A correlação média é muito alta: 0.9838. Repare-se que o valor para -VEL/-BLE é menor (mesmo tendo maior F1) porque a frequência de casos é muito mais baixa e este é um parámetro determinante no cálculo da correlação.

5.3. Análise de erros

A seguir, após analisar os falsos positivos da última avaliação, apresentamos o tipo de erro mais comum que comete este método de extração automática:

- Erros de -ÇOM:
  - Nomes próprios terminados em variantes de -ÇOM, que podem ir em maiúscula ou não. Por exemplo: *Jaason, iaason, iasom.*
  - Variantes do verbo ser, por exemplo: *sao, ssom, sson.*
  - Variantes não esperadas dos termos *razom*, *coraçom* e *sazom*, com iode que desaparece também em castelhano, e que não foram considerados na contagem. Por exemplo: *rrason, rrazon, rrazóm.*
- Erros de -VEL:
  - Nomes próprios terminados em variantes de -VEL: *Çentule.*
  - Verbos em forma terminada em 'u', seguido pelo pronome *le* enclítico, por exemplo: *doule, deule, outorgoule.*

No tocante aos falsos negativos, o mais habitual é encontrarmos substantivos etiquetados como adjetivos, e viceversa. Como estes casos são mais frequentes que os falsos positivos, a abrangência do sistema é menor que a sua precisão.

[10] http://fegalaz.usc.es/~gamallo/resources/sufixos_medievais.zip

	prec.	abrang.	F1	ρ
-ÇOM/-CION	94,37%	92,54%	93,45%	0,9972
-VEL/-BLE	99,39	95,93	97,63	0,9705
Média	96,85	94,23	95,54	0,9838

Tabela 22: Precisão, abrangência e F1 dos resultados obtidos pelo sistema (etiquetador), junto com a correlação *Pearson* entre os resultados do sistema e o *gold standard*.

6. Conclusões

No presente artigo, descrevemos a adaptação, uso e avaliação de ferramentas de etiquetagem e de identificação de varieadades medievais, cujo desempenho é inferior às ferramentas utilizadas para variedades modernas mais estandardizadas. As dificuldades para etiquetar textos medievais derivam principalmente da grande variabilidade de formas não estandardizadas.

Usamos estas ferramentas para um estudo filológico concreto cujo objetivo é verificar hipóteses linguísticas sobre a evolução de sufixos medievais. Conferimos que, pela sua distribuição de frequências, o uso dos sufixos -CION e -BLE nos textos da Galiza já está influenciado pelo castelhano baixo medieval. Os indícios de castelhanização são claros: a ratio de -ÇOM para -CION e de -VEL para -BLE desce consideravelmente nos textos do século XV, quando a influência do castelhano e mais evidente, e esta ratio inverte-se nos excertos escritos em castelhano ou mais castelhanizados. É pertinente considerar que o triunfo das soluções castelhanas dos dois sufixos, -CION e -BLE, no galego moderno viu-se favorecido pelo facto de se aplicarem, em geral, a palavras cultas, muitas delas herdadas diretamente do castelhano a partir do XV e séculos posteriores com a necessidade de incorporarmos na língua termos técnicos e cultismos.

No estudo quantitativo, observamos que as variantes dos sufixos patrimoniais, -ÇOM e -VEL, são claramente maioritárias tanto nos textos da Galiza como de Portugal. É também importante sublinhar que a produtividade (ou *prolificidade*) de -ÇOM é muito maior que a de -VEL, tanto nos textos medievais de Galiza como de Portugal. Também mostramos como a proporção do uso de -VEL na Galiza frente ao castelhanismo -BLE (7 vezes maior) é claramente menor que a de -ÇOM frente a -CION (20 vezes maior). Isto põe em questão a polémica decisão de recuperarmos para o galego moderno o uso de formas em -VEL mas não em -ÇOM.

Os resultados obtidos automaticamente (etiquetagem e identificaço de língua) foram avaliados mediante a construção dum *gold standard* com revisão manual. Verificamos que os resultados da análise automática estão muito próximos do *gold standard*, mostrando que a abordagem automática é fiável e pode ser utilizada para este ou outros estudos filológicos que precissem de informação quantitativa extraída de corpus.

Como trabalho futuro, desenharemos um método automático de normalização de variantes para a língua medieval galego-portuguesa tomando em conta regras fonológicas, como já se fez para uma parte do léxico do Inglês Antigo (Sáenz, 2015). O normalizador/lematizador assim construido permitirá alargar o modelo de língua do nosso etiquetador, com base no corpus anotado CIPM previamente enriquecido com os lemas automaticamente gerados. Finalmente, continuaremos desenvolvendo o modelo híbrido do etiquetador com a definição de mais regras morfológicas e de correção, que enriquecerão o classificador estatístico.

Agradecementos

Este trabalho foi financiado pelo projeto NÓS, da Xunta de Galicia, pelos projetos DOMINO (PGC2018-102041-B-I00 , MCIU/AEI/FEDER, UE) e eRisk (RTI2018-093336-B-C21), e também pela Consellería de Cultura, Educación e Ordenación Universitaria (acreditação 2016-2019, ED431G/08 e Programa de Formación Posdoutoral da Xunta de Galicia 2016) e European Regional Development Fund (ERDF).

Referências

Canosa, Xavier, Pablo Gamallo, Xavier Canosa, José Ángel Taboada, Paulo Martínez Lema & Marcos Garcia. 2019. Uma utilidade para o reconhecimento de topónimos em documentos medievais. *Linguamática* 2(11). 3–15. doi 10.21814/lm.11.1.291.

Cristine Prado, Natália & Gladis Massini-Cagliari. 2014. Formação de nomes deverbais nas cantigas de Santa Maria: Um estudo morfofonológico. *Revista Do GEL* 11(2). 71–96.

Diéguez, Ignacio Vázquez. 2018. Sobre algúns sufixos galegos medievais. *Estudios de Lingüística del Español* 39. 241–277.

Ferreiro, Manuel. 1997. *Gramática histórica da lingua galega. ii. lexicoloxía*. Santiago de Compostela: Lailovento.

Fillo, Machado & Américo Venâncio Lopes. 2013. *Dicionário etimológico do português arcaico: Projeto DEPARC*. Salvador: Edufba.

Freixeiro Mato, Xosé Ramón. 1997. *Lingua galega: normalidade e conflito*. Santiago de Compostela: Lailovento.

Gamallo, Pablo & Marcos Garcia. 2017. LinguaKit: uma ferramenta multilingue para a análise linguística e a extração de informação. *Linguamática* 9(1). 19–28. 10.21814/lm.9.1.243.

Gamallo, Pablo, Marcos Garcia, Cesar Pineiro, Rodrigo Martinez-Castano & Juan C. Pichel. 2018. LinguaKit: a big data-based multilingual tool for linguistic analysis and information extraction. Em *5th International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 239–244. 10.1109/SNAMS.2018.8554689.

Gamallo, Pablo, Susana Sotelo & José Ramom Pichel. 2014. Comparing ranking-based and naive bayes approaches to language detection on tweets. Em *Workshop TweetLID: Twitter Language Identification Workshop at SEPLN 2014*, n/p.

Garcia, Marcos & Pablo Gamallo. 2015. Yet another suite of multilingual NLP tools. Em *Languages, Applications and Technologies*, 65–75. 10.1007/978-3-319-27653-3_7.

Kettunen, Kimmo. 2014. Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics* 21. 223–245. 10.1080/09296174.2014.911506.

Leach, Geoffrey & Andrew Wilson. 1996. Recommendations for the morphosyntactic annotation of corpora. Em *Technical Rapport*, Expert Advisory Group on Language Engineering Standard (EAGLES).

Lorenzo, Ramón. 1985. *Crónica troiana. introducción e texto*. A Coruña: Fundación Pedro Barrié de la Maza, Conde de Fenosa.

Mariño, Ramón. 1998. Notas sobre a historia das terminacións -ión / -ón en galego. Em D. Kremer (ed.), *Homenaxe a Ramón Lorenzo*, 735–760. Vigo, Galaxia, vol. 2.

Mariño Paz, Ramón. 2005. Forma e función do sufixo -uel no galego medieval. *Cadernos de Lingua* 27. 155–193.

Messner, Dieter. 2007. Os dicionários portugueses, devedores da lexicografia espanhola. *Península, Revista de Estudos Ibéricos* 4. 141–151.

Padró, Lluís. 2012. Analizadores multilingües en FreeLing. *Linguamática* 3(2). 13–20.

Pichel, José Ramom, Pablo Gamallo, Iñaki Alegria & Marco Neves. 2020. A methodology to measure the diachronic language distance between three languages based on perplexity. *Journal of Quantitative Linguistics* 28(4). 306–336. 10.1080/09296174.2020.1732177.

Pichel, José Ramom, Pablo Gamallo & Inaki Alegria. 2019. Measuring diachronic language distance using perplexity: Application to english, portuguese, and spanish. *Natural Language Engineering* 26(4). 433–454. 10.1017/S1351324919000378.

Rögnvaldsson, Eiríkur & Sigrún Helgadóttir. 2008. Morphological tagging of old norse texts and its use in studying syntactic variation and change. Em *2nd Workshop on Language Technology for Cultural Heritage Data*, 40–46.

Sánchez-Marco, Cristina, Gemma Boleda & Lluís Padró. 2011. Extending the tool, or how to annotate historical language varieties. Em *5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 1–9.

Santalha, Montero & José-Martinho. 2005. Documentos medievais galegos (3). *Agália* 81–82. 255–264.

Silvestre, João Paulo. 2008. *Bluteau e a origens da lexicografia moderna*. Lisbon: Imprensa Nacional – Casa da Moeda: Colecção filologia portuguesa.

Sáenz, Marta. 2015. The lemmatization of Old English verbs from the second weak class on a lexical database. *Journal of English Studies* 13. 135. 10.18172/jes.2861.

Varela Barreiro, Xavier, Maria Francisca Xavier & Charlotte Galves. 2016. Corpus informatizado Galego-Portugués antigo. Instituto da Lingua Galega / Centro de Lingüística da Universidade Nova de Lisboa / Universidade de Campinas. http://ilg.usc.gal/tmilg.

Venâncio, Fernando. 2019. *Assim nasceu uma língua. sobre as origens do português*. Lisbon: Guerra e Paz.

Viaro, Mário Eduardo. 2012. A produtividade dos sufixos do ponto de vista diacrônico. Em T. Lobo, Z. Carneiro, J. Soledade, A. Almeida & S. Ribeiro (eds.), *Rosae: linguística histórica, história das línguas e outras histórias*, 275–292. SciELO Books.

Xavier, Maria Francisca. 2005. A caminho de um dicionário do português medieval. Em *Des(a)fiando discursos: Homenagem a Maria Emília Ricardo Marques*, 667–686. Lisboa: Universidade Aberta, Língua, Literatura e Cultura Portuguesas.

Xavier, Maria Francisca. 2016. O CIPM — corpus informatizado do português medieval, fonte de um dicionário exaustivo. Em Carlota de Benito Moreno Johannes Kabatek (ed.), *Lingüística de corpus y lingüística histórica iberorrománica*, 137–156. De Gruyter.

Zampieri, Marcos, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer & Noëmi Aepli. ???? Findings of the VarDial evaluation campaign 2017. Em *4th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, 1–15.

Appendices

A. Lista de obras do corpus galego

Nome	**Tamanho**
Cantigas de Afonso Eanes do Coton	7,1K
Cantiga de Afonso Gomes de Sarria	0,706K
Cantigas de Airas Nunes	5,9K
Arte de Trovar	11K
BMSEH (Tui século XV)	334K
Cancioneiro de Ajuda	77K
Cantigas de Amigo	20K
CDMO (Santa María de Oseira)	1,3M
CI (Crónica de Santa María de Iria - Rui Vasques)	79K
CSMp (Cantigas de Santa María: Pauta - Afonso X)	89K
CSMr (Cantigas de Santa María: Rúbricas - Afonso X)	45K
CT (Crónica troiana - Benoît de Saint-Maure)	1,2M
DAG (Documentos antigos de Galicia)	179K
FDUSC (Fontes documentais da Universidade de Santiago de Compostela)	1,2M
FR (fragmento do Foro Real - Afonso X)	4,7K
HGPg (Historia do galego-portugués - Galicia)	134K
HT (Historia troiana - Î de Saint-Maure)	528K
LCS (Libro do Concello de Santiago)	680K
LNAP (Libro de notas de álvaro Pérez)	237K
MSCDR (San Clodio do Ribeiro)	1,2M
MSPT (San Salvador de Pedroso)	260K
OMOM (San Martiño de Vilourente)	1,1M