Proposta recebida em Abril 2018 e aceite para publicação em Junho 2019.

Uma Utilidade para o Reconhecimento de Topónimos em Documentos Medievais

A Tool for Toponym Recognition in Medieval Documents

Xavier Canosa
CiTIUS / Univ. de Santiago de Compostela
canosarodrigues@gmail.com

Pablo Gamallo
CiTIUS / Univ. de Santiago de Compostela
pablo.gamallo@usc.es

Xavier Varela
ILG / Univ. de Santiago de Compostela
xavier.varela@usc.es

José Ángel Taboada
CiTIUS / Univ. de Santiago de Compostela
joseangel.taboada@usc.es

Paulo Martínez Lema
ILS / Univ. de Santiago de Compostela
paulo.martinez.lema@edu.xunta.es

Marcos Garcia
Grupo LyS, Dpto. de Letras / Univ. da Corunha
marcos.garcia.gonzalez@udc.gal

Resumo

Este artigo apresenta o método de construção duma ferramenta para a anotação de entidades geográficas mencionadas em textos medievais. A nova ferramenta foi desenvolvida a partir dos módulos de língua contemporânea do LinguaKit, pacote multilingue de ferramentas de PLN. Uma coleção de corpora anotados manualmente serviu de recurso para elaborar uma lista de topónimos medievais (*gazetteers*) e observar padrões para a melhora e implementação de novas regras de reconhecimento dos nomes de lugar. Depois da lista de entidades geográficas, os ativadores contextuais (*triggers*) foram o recurso determinante na melhora da abrangência. Para o produto final, fizeram-se também ajustes menores na procura de recolher os elementos mais comuns do léxico e os contextos gramaticais das entidades geográficas mencionadas. Ainda que muito trabalho fica por fazer na elaboração de listas para entidades não geográficas, na construção dum modelo de língua medieval e um lexicon específico, o novo módulo pode ser utilizado para anotar textos e mostra uma melhora significativa a respeito dos módulos previamente existentes.

Palavras chave

entidades geográficas mencionadas, NERC, toponímia

Abstract

This paper describes a method to build a tool aimed at recognizing geographical named entities in medieval texts. The new tool has been developed using the corresponding modules for contemporary languages contained in LinguaKit, a suite of NLP tools. A collection of manually annotated corpora served as a resource to build a gazetteer of medieval toponyms and find patterns to improve and implement new rules for the recognition of place names. In addition to the gazetteer, a list of triggers was the most determinant factor to improve recall. Final adjustments considered the most frequent terms of the lexicon and grammatical contexts for geographical named entities. In the process of building a model of medieval language and a specific lexicon, the available tool can already be used to annotate texts and shows a significant improvement when compared with previous modules. However, most work remains to be done in terms of adding specific gazetteers for entities other than geographical.

Keywords

geographical named entities, NERC, place names

1 Introdução

O reconhecimento automático de topónimos foi atendido nas últimas duas décadas como parte do problema NERC (Named Entity Recognition and Classification) também chamado de REM (Reconhecimento de Entidades Mencionadas) dentro do Processamento da Linguagem Natural (PLN). Dado que os topónimos mais comuns aparecem sistematizados em nomenclatores e atlas já digitalizados, o uso de listas de entidades (*gazetteers*) sobre os que efetuar pesquisas por *string match* aparece como uma primeira solução para a anotação automática. Porém, a ambiguidade que se produz na língua (ex. *Santiago* como cidade ou como nome de pessoa) e a necessidade

DOI: 10.21814/lm.11.1.291

de marcar topónimos menores ou menos comuns (ex. microtopónimos, geografias exóticas e menos habituais) precisa de utilidades com capacidade de desambiguação e análise do contexto para prever os casos de formas desconhecidas nas listas. Duas aproximações contribuíram para dar o problema como resolvido com um nível satisfatório de eficácia: a aplicação de heurísticas (especialmente regras que especifiquem um contexto morfossintático) e o treino a partir de grandes volumes de corpora de onde se inferem regras de tipo estatístico. A anotação NERC passou assim a formar parte dos pacotes de utilidades de PLN mais comuns na atualidade, facilitando o processamento de textos para o reconhecimento de topónimos.

Dado que as ferramentas NERC foram desenvolvidas a partir de corpora contemporâneos (Won et al., 2018), a aplicação em variedades históricas da língua vê comprometido o desempenho em função da divergência a respeito dos usos linguísticos atuais. No caso galego-português medieval, ainda conservando uma estrutura gramatical e regularidade morfológica próxima às soluções contemporâneas, a dificuldade vem dada pelo grande número de variantes dos topónimos, limitando a aplicabilidade das listas (ex. as formas *Mondodnedo, Mondonedo, Mondonnedo, Mondoñedo* aparecem todas num mesmo *corpus*). O recurso a contextos sintáticos activados por palavras chave (*triggers*) que contribuem para a deteção da entidade geográfica com maior precisão, também se vê condicionado pelo fenómeno da variação (ex. *feegresia, figleſia, figressia, figresya, figrigia, figrisia...* até 438 variantes foram achadas para o mesmo tipo geográfico). Quanto menor seja a aplicabilidade de listas de entidades e de ativadores, mais limitação nos recursos da ferramenta e maior dependência na especificidade das regras ou no treino estatístico. Porém, para conseguir regras mais específicas, necessita-se um maior nível de PLN, particularmente lematização e etiquetagem morfossintática que, da sua parte, requer o uso de lexicons específicos para a variedade de língua. Numa solução estatística, o treino de modelos necessita de grandes corpora, com um volume suficiente como para serem estatisticamente relevantes. Para além de que este tipo de recursos são custosos e requerem atenção experta, existe ainda o problema de que a enorme produtividade da variação, acentuada por fatores tais qual época, área geográfica, tipologia textual e ainda usos individuais, faz com que as variantes se multipliquem e reduzam as frequências dos termos. Mesmo que um sistema NERC para textos medievais possa se desenhar com a mesma tecnologia e práticas utilizadas para a língua contemporânea, a adaptação duma ferramenta requer a disponibilidade de recursos adicionais que comprometem o desempenho do produto final. A nossa principal contribuição é a criação de recursos para o galego-português medieval e a sua integração e adaptação a uma ferramenta de NERC já existente para a língua contemporânea.

Mais precisamente, neste artigo apresentamos uma metodologia que analisa os componentes do pacote de utilidades PLN LinguaKit (Gamallo & Garcia, 2017) com maior relevância para a anotação de topónimos em textos medievais do domínio galego-português. Partindo dum conjunto de corpora com topónimos anotados manualmente, preparamos uma série de testes para avaliar o desempenho da ferramenta conforme se foram adicionando ou modificando componentes, nomeadamente listas de entidades e ativadores, até desenvolvermos um novo módulo NERC de LinguaKit adaptado a textos medievais do galego-português. Mesmo se este módulo é apenas uma primeira versão a melhorar, oferece um incremento muito notável na abrangência e medida-F para as entidades geográficas a respeito dos módulos de língua contemporânea. O módulo contém também um tokenizador, um lematizador e um etiquetador morfossintático, ainda em fase de protótipo, todos eles adaptados para o galego-português medieval. O conjunto de módulos para a língua histórica, chamado de *histgz*, está integrado em LinguaKit, com licença livre GPLv3[1].

Para além da introdução, o resto do artigo está organizado como prossegue. Na Secção 2 mencionamos estratégias de aproximação a textos históricos para o reconhecimento de entidades mencionadas e revemos o desempenho de ferramentas NERC para o caso particular de entidades geográficas mencionadas em português. A Secção 3 descreve os corpora utilizados nos testes e introduz LinguaKit, a ferramenta sobre a que se obtém a nova utilidade, cujas fases de desenvolvimento são atendidas na Secção 4. A seguir, avaliam-se os resultados, com atenção particular a falsos positivos e falsos negativos que apontam para melhoras mais imediatas e trabalho futuro, recolhido, junto com as conclusões, na Secção 6.

2 Trabalho relacionado

A dificuldade de reconhecer entidades geográficas mencionadas em textos antigos favorece a anotação manual, mais ou menos auxiliada por ambientes que facilitem o labor experto e possibilitem o trabalho em equipa. Na procura de maior

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

automatização, tem-se recorrido a sistemas inicialmente concebidos para o processamento de textos contemporâneos, com duas linhas de atuação, seja pela adaptação dos recursos utilizados pela ferramenta, particularmente listas de entidades, ou adaptando o texto, aplicando estratégias de normalização com o objetivo de minimizar a variação linguística e aproximar a língua histórica ao padrão contemporâneo (Hendrickx & Marquilhas, 2011; Marquilhas & Hendrickx, 2014)

As métricas mais comuns para a avaliação de ferramentas NERC são a precisão, a abrangência e a medida-F que combina as duas primeiras (Santos et al., 2007; Pinto et al., 2016). Na década passada celebraram-se eventos em que corpora previamente anotados serviam de padrões dourados para medir o desempenho de utilidades NERC. Especialmente relevantes para o português contemporâneo foram as competições do HAREM (Santos & Cardoso, 2007; Mota & Santos, 2008; Freitas et al., 2010). Os melhores resultados para a categoria de Lugar situaram-se em 68,03% de precisão e 73% de abrangência no primeiro evento (Santos & Cardoso, 2007) e precisão 72,12%, abrangência 80,17% no segundo (Chaves, 2008).

Ainda no domínio do padrão contemporâneo, testado no corpus Bosque (Afonso et al., 2002), o sistema NERC que deu lugar aos módulos correspondentes de LinguaKit atingiu 85% de precisão e 57% de abrangência (68% medida-F) na categoria de Lugar (Gamallo & Garcia, 2011). Este mesmo sistema foi também adaptado para a variedade linguística galega, com resultados de medida-F de entre 74,5% e 80,4%, em função do tipo de avaliação realizada (Garcia et al., 2012).

Testes mais recentes, utilizando os mesmos corpora que o HAREM, ofereceram resultados mais baixos, de 62% precisão e 66% abrangência para a mesma categoria de Lugar (Amaral et al., 2014), o qual é indicativo de que não houve um avanço significativo na resolução do problema.

Em relação com o reconhecimento de entidades em textos antigos, a maior parte dos trabalhos têm implementado estratégias determinísticas que combinam listas de entidades com heurísticas para cada tipo de entidade. Tanto as caraterísticas deste tipo de documentos, muitas vezes digitalizados mediante OCR, como o seu tamanho fazem com que a implementação de sistemas estatísticos seja mais custosa. Existem, contudo, sistemas baseados em aprendizagem automática, tais como Byrne (2007), que treina um modelo de máxima entropia orientado principalmente à identificação de entidades que se sobrepõem.

Dentro dos métodos determinísticos as aproximações mais frequentes são centradas no documento (i.e., utiliza-se informação de todo o texto para classificar uma entidade, e não só o contexto da menção específica a analisar). Assim, Jones & Crane (2006) classificam 10 tipos de entidades num jornal americano do século XIX, com valores de precisão de entre 57% e 99% em função da classe. Borin et al. (2007) analisam as entidades mencionadas num corpus de literatura sueca do XIX, melhorando os resultados com um módulo de similaridade que computa a distância de edição entre as menções desconhecidas e as listas de entidades. Também mediante máquinas de estados finitos e um conjunto de listas (nomes, apelidos, etc.), Grover et al. (2008) identificam entidades geográficas e nomes de pessoa em textos parlamentares britânicos dos séculos XVII a XIX.

De modo similar a estes últimos, a metodologia empregada no presente trabalho adapta os conjuntos de ativadores e de listas de entidades e implementa regras específicas para melhorar o desempenho dum sistema NERC em texto medieval.

Ainda que o número de trabalhos de anotação de topónimos aumenta particularmente no campo das humanidades digitais, são poucos ainda os estudos específicos para a comparação do desempenho de ferramentas NERC em textos históricos. Canosa (2017) comparou o desempenho de ferramentas para um corpus em inglês do século XVII atingindo resultados do 68% na medida-F com uma ferramenta estatística treinada em corpora modernos e do 62% com um sistema de regras provisto duma lista específica. Resultados similares, com a melhor medida-F próxima a 70%, foram de novo obtidos na comparativa de cinco ferramentas NERC contemporâneas sobre córpora históricos também do inglês em Won et al. (2018). Estes mesmos autores apresentam um experimento novidoso em que se combinam as anotações das distintas ferramentas NERC para escolher o mais provável em caso de divergência, atingindo um 73,3% na medida-F como melhor resultado.

3 Materiais e ferramentas

Nesta secção, apresentamos os recursos textuais (corpora) e a ferramenta de processamento da língua natural utilizados na experimentação (LinguaKit).

3.1 Corpora

Como material de trabalho para o treino duma nova ferramenta e avaliação de resultados utilizou-se uma parte dos textos medievais recolhidos actualmente no *Corpus informatizado Galego-Português Antigo*[2] (CGPA) (Varela Barreiro et al., 2016).

Código corpus	Tokens	Topónimos
Mens	18711	381
Toxosoutos	44001	2945
Toxosoutos_gl	5138	295
CDMACM5	267965	2626
CDMACM5_gl	105172	844
Mens_TX_CD_gl	128992	5760

Tabela 1: Códigos dos *corpora* utilizados para extração de topónimos e tamanho dos corpora e listas obtidas.

O CGPA é o resultado de reunir numa plataforma conjunta corpora históricos do galego, do português, do latim e do castelhano elaborados na Galiza, Portugal e Brasil. O núcleo de textos da Galiza forma-o o corpus plurilíngue *Xelmírez, Corpus lingüístico da Galiza medieval*[3] (Varela Barreiro, 2009) do Instituto da Lingua Galega (USC), em que estão integrados textos redigidos em galego-português (TMILG), em latim (TMILL) e em castelhano (TMILC). Os textos de Portugal e do Brasil não contêm obras em latim ou castelhano e estão representados por duas vias. Por parte portuguesa concorre o *Corpus Informatizado do Português Medieval*[4] (CIPM) (Xavier, 2000) e por parte brasileira o *Corpus Histórico do Português Tycho Brahe*[5] (Galves, 2018). Pelo momento o grande valor do CGPA é fazer possível, por meio de pesquisa única, o acesso à totalidade dos corpora integrados.

Os textos do CGPA selecionados para este projecto particular de desenvolvimento duma ferramenta de anotação de entidades geográficas mencionadas têm a particularidade de contarem com a etiquetagem dos topónimos, por quanto foram utilizados anteriormente no *Inventario Toponímico da Galiza Medieval*[6] (ITGM) (Varela Barreiro & Martínez Lema, 2009). O ITGM é um projecto lançado em 2005 com o intuito de fazer acessível de forma gradual a totalidade do material toponímico presente na documentação galega medieval, compilada e codificada no corpus *Xelmírez*. No processo de recuperação da informação, as agrupações de topónimos obtiveram-se pela aplicação de critérios linguísticos (relativos fundamentalmente ao processo de lematização) e/ou de critérios geográfico-administrativos. No seu estado actual, o ITGM acolhe 17.640 registos toponímicos, que remitem a um total de 3.086 topónimos e outros tantos lemas. Destes últimos, 2.876 (93% do conjunto) estão referenciados com maior ou menor margem de certeza, no entanto apenas para 7% (os 210 restantes) carecemos de qualquer parâmetro geográfico-administrativo de atribuição.

A tabela 1 mostra os textos procedentes do ITGM, caracterizados portanto por terem os topónimos anotados manualmente, que foram selecionados para a fase de recolhida de dados e testes no presente projecto. Cada corpus recolhe textos medievais com uma mesma origem documental, acessível no CGPA em que aparecem com código e referência bibliográfica individual. As obras do CGPA escolhidas para o desenvolvimento foram a *Colección diplomática do mosteiro de Santiago de Mens* (Zapico Barbeito, 2005), *Os documentos do tombo de Toxos Outos* (Rodríguez & Javier, 2004) e a *Colección diplomática medieval do Arquivo da Catedral de Mondoñedo* (Cal Pardo, 1999). Dado que nos corpora iniciais há também documentos em latim e castelhano, gerou-se uma segunda versão só com os textos em galego-português (identificada com a extensão *gl*). Nos experimentos finais agrupam-se todos os documentos num único arquivo (MensTXCDgl), a soma de selecionar só os documentos galego-portugueses dos três corpora. Na fase final de avaliação utilizou-se um novo corpus, o *Livro de Notas de Álvaro Pérez* (LNAP) (Tato Plaza, 1999), sem anotação de topónimos nenhuns na versão oferecida para este projeto.

3.2 LinguaKit

LinguaKit é um pacote livre de ferramentas multilíngues para o processamento da linguagem natural que pode aplicar-se ao português, galego, inglês e espanhol. Contém módulos de análise, extração, anotação e correção linguística. LinguaKit permite realizar um amplo conjunto de tarefas, entre as quais se encontram: segmentação em frases e tokenização, lematização, etiquetagem morfossintática, reconhecimento e classificação de entidades mencionadas (NERC), análise sintática de dependências, resolução de correferência a nível de entidade, extração de termos e de relações semânticas, análise de senti-

[2] http://ilg.usc.gal/CGPA
[3] http://sli.uvigo.gal/xelmirez/
[4] https://cipm.fcsh.unl.pt/
[5] http://www.tycho.iel.unicamp.br/~tycho/corpus/
[6] http://ilg.usc.gal/itgm

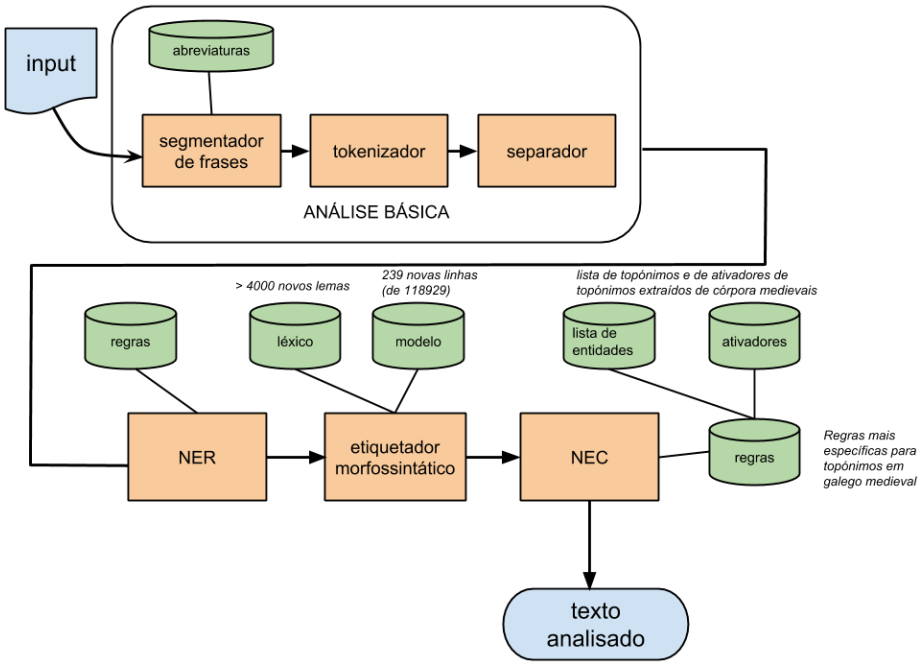

Figura 1: Arquitetura utilizada pelo Histgz e principais modificações a respeito dos módulos já existentes no LinguaKit.

mentos, anotação conceitual com ligação a recursos enciclopédicos (*entity linking*), correção e avaliação de léxico e sintaxe (só para o galego), conjugação verbal automática (exceto inglês), resumo automático, identificação de língua e visualização de concordâncias (palavras chave em contexto). O presente trabalho foca-se nas tarefas relacionadas com o NERC, tal e como mostra a figura 1, descrita mais à frente, na secção 4.

O LinguaKit está disponível como um serviço web[7] e é acessível via RESTful API4.[8] O código fonte está publicado sob licença GPLv3 e acessível desde repositório de GitHub (Cf. nota de rodapé 1).

4 Métodos e procedimento

O trabalho de adaptação do LinguaKit para o reconhecimento de topónimos medievais realizou-se aplicando uma metodologia experimental. Considerado um parámetro, aplicaram-se testes sobre os corpora para observar como contribui para a anotação dos topónimos. Primeiramente atendeu-se à incidência da lista de topónimos e a lista de ativadores. As regras e ajustes menores nos recursos de tipo morfossintático ocuparam o estádio final. Os corpora foram introduzidos no desenvolvimento gradualmente, de tal modo que uma vez concluída a necessidade de incluir um componente de melhora no módulo, se adicionava um novo corpus para a nova rolda de experimentos. As adaptações, ajustes, modificações e suplementos elaborados nos sucessivos ensaios foram implementados na arquitetura do LinguaKit como módulo independente, chamado de Histgz (galego-português histórico). A figura 1 mostra graficamente as principais modificações do novo módulo a respeito dos módulos prévios utilizados como base para o desenvolvimento. O Histgz inclui tanto tarefas básicas de análise (segmentação em frases, tokenização e quebra de contrações ou *separador*), quanto processos mais complexos: etiquetação morfossintática (*PoS tagging*) e NERC. As principais modificações foram realizadas no NEC (que forma parte do NER) e no etiquetador. O NEC é a tarefa final que classifica as entidades e permite reconhecer os topónimos.

4.1 Extração e elaboração de listas de topónimos

Para facilitar os labores de validação das anotações obtidas por procedimentos automáticos, criaram-se duas versões dos corpora: uma só texto, com os topónimos sem anotar, para ser usada como input pela ferramenta NERC, e outra com os topónimos marcados segundo a anotação manual facilitada pela

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

[8] https://market.mashape.com/linguakit/linguakit-natural-language-processing-in-the-cloud

Lista de topónimos	Sem lista medieval	Sem lista medieval	lista acrescentada com Lista_Mens	Lista acrescentada com Lista_Mens
Entidade geográfica mencionada	Verdadeiro positivo se coincide exatamente com a anotação manual	Verdadeiros positivos a partir do nome próprio	Verdadeiro positivo se coincide exatamente com a anotação manual	Verdadeiros positivos a partir do nome próprio
Precisão	57,53%	76,88%	72,97%	93,24%
Abrangência	17,95%	23,99%	45,3%	57,89%
Medida-F	27,36%	36,57%	55,9%	71,43%

Tabela 2: Comparativa de resultados da anotação do LinguaKit (com o módulo base em galego) sobre o corpus Mens segundo lista e critério de validação dos verdadeiros positivos das entidades geográficas mencionadas.

equipa do ITGM que servirá de padrão dourado e recurso para a extração de topónimos e listas para adicionar ao LinguaKit. As listas de topónimos obtidas relacionam-se na última coluna da tabela 1.

A lista última é a soma das listas de topónimos obtidas dos corpora envolvidos no desevolvimento do módulo NERC medieval, Lista_Mens_TX_CD (tabela 1) e aparece integrada na lista de entidades geográficas do Histgz acessível no repositório de GitHub (Cf. nota 1). Os testes realizados nas fases de desenvolvimento utilizaram as listas progressivamente, para distinguir o efeito da inclusão de cada nova lista segundo se processava e ensaiava sobre um novo corpus. Este efeito mostra-se na tabela 2 com os resultados do LinguaKit em galego condicionados pelo uso ou não da lista de entidades geográficas. As métricas mostram o limite de capacidade do módulo quando a lista contém todos os topónimos presentes no texto. Na avaliação dos resultados apareceu também como relevante o critério utilizado para definir o topónimo. O padrão dourado anota como topónimos frases do tipo "o rrío de Tallo", "no Esto" ou o artigo mesmo quando não foi grafado com maiúscula "o Esto", porém a lista de entidades limita o topónimo ao nome próprio. Nas primeiras avaliações discriminou-se entre os resultados que coincidiam plenamente com as anotações manuais (só se avalia como verdadeiro positivo quando se anota exatamente igual ao padrão, assim "o rrío de Tallo", "no Esto") face àqueles em que o nome próprio é suficiente para considerar a anotação como verdadeiro positivo ("Tallo", "Esto"). Este último critério, mais adequado às expectativas reais de desempenho duma ferramenta NERC, será o que se aplique nos sucessivos experimentos.

Os experimentos da tabela 2 mostram como, mesmo com uma lista que contém todos os topónimos presentes no corpus, ainda obtendo uma precisão muito alta, apenas se recuperam 57,89% das entidades geográficas mencionadas. As regras utilizadas pelos módulos de língua contemporânea precisam, portanto, melhoras para além das listas que, contudo, resultam determinantes para um bom desempenho (os resultados com a lista só de topónimos contemporâneos ficam em apenas 36,57% na medida-F face ao 71,43% obtido ao acrescentar a lista medieval).

4.2 Filtrado por língua

A análise dos primeiros testes, uma vez processados os textos e extraídos os topónimos anotados manualmente, mostraram as limitações da aplicação do módulo NERC mesmo com uma lista de topónimos específica. Porém, o primeiro factor a considerar para a melhora de resultados não é devido ao pacote PLN, mas aos próprios textos a anotar. Com efeito, ao trabalhar com o conjunto dos corpora, aparecem textos em latim e, em menor medida e em documentos mais tardios, espanhol, que necessariamente condicionam a efetivadade dos recursos e das heurísticas classificatórias, dependentes duma seleção linguística prévia (ex. *illa* é demonstrativo em latim, mas nome comum "terra rodeada por mar" em galego). Faz-se necessária, portanto, a discriminação por idioma. Os corpora utilizados nos vindouros experimentos (Toxosoutos e CDMACM5) foram processados para obter apenas o texto galego(-português), identificado com a extensão _gl nos códigos da tabela 1.

Precisão	**Abrangência**	**Medida-F**	**lista com os topónimos do corpus**	**Lista de ativadores**
82,26%	25,67%	39,13%	Não	Não
96,22	59,73%	73,71%	Sim	Não
77,5%	36,41%	49,54%	Não	Sim
90,31%	62,58%	73,93%	Sim	Sim

Tabela 3: Comparativa de resultados segundo o uso de lista de ativadores e topónimos no *corpus* Mens (596 entidades geográficas anotadas no texto padrão).

Ativadores	Lista de ativadores expandida por combinatória de caracteres	Lista de ativadores recuperada por inspeção de concordâncias no CGPA
Precisão	76,74%	76,16%
Abrangência	32,76%	33,06%
Medida-F	45,92%	46,11%

Tabela 4: Comparativa de resultados variando a lista de ativadores sobre o corpus Mens_TX_CDMACM5_gl (3.373 entidades geográficas anotadas no corpus padrão).

4.3 Lista de ativadores

O primeiro componente considerado para a melhora do desempenho do módulo uma vez comprovada a limitação da lista de topónimos foi a lista de ativadores para entidades geográficas (TwLOC). Da inspecção experta de concordâncias dos topónimos extraídos do primeiro corpus usado nos testes (Mens) obteve-se manualmente uma lista de 38 termos geográficos, contando como unidades distintas todas as variantes dum mesmo termo. Assim, na mesma lista aparecem todas as variantes achadas no *corpus* associáveis a *fregresia* (*fregesía, fijglesía, fijgresía, fjgresía, flegresía, frijguesía, frigresýas*), mosteiro (*moesteiro, moesteyro, mosteiro, mosteiros*), vila (*vila, villa, vjla*) junto com termos geográficos com uma única ocorrência (*tença*).

Dado que uma boa parte dos termos são variantes do mesmo tipo, experimentou-se com a expansão da lista de ativadores mediante a combinatória de caracteres equivalentes (ex. nasal palatal *nn*, *nh*, *nj*, *ni*, *jn*, *yn*, *in*, *ñ*, *gn*; vogais simples e geminadas *o*, *oo*; vogais nasais e terminações *õ*, *om*, *on*; sibilantes, lateral palatal, etc.).

Paralelamente fez-se um levantamento manual de termos geográficos a partir das listas de tipos extraídas do CGPA de onde se obtêm 1.900 termos geográficos (disponíveis na pasta de ativadores do próprio módulo Histgz de LinguaKit).

Os resultados da aplicação da lista de termos expandidos artificialmente não proporcionaram um incremento sobre a lista manual (tabelas 3 e 4) e, toda vez que esta contém as formas originais dos corpora, ficou esta última como a solução finalmente adoptada para o novo módulo.

Ao tempo que se elaborou a nova lista de ativadores com termos geográficos, recolheram-se termos adicionais com valor de ativador em contextos mais específicos, agrupados em uma lista, *nongeo*, composta principalmente por títulos, ex. *arcebispo*, *rei*. Recolhe 686 termos, com um alto número de variantes para o mesmo tipo.

4.4 Regras classificatórias

O módulo NERC aplica a lista de ativadores por meio de regras que priorizam a classificação numa classe dentro das quatro categorias de entidades mencionadas (PER Pessoa, LOC Lugar, ORG Organização e MISC Miscelânea). Um exemplo de regra é aquela que classifica um nome próprio como entidade geográfica mencionada quando não se achar em nenhuma das listas de entidades e o termo precedente for a preposição *em*. Outra classifica como nome de pessoa todo nome próprio ambíguo em ausência de outros condicionantes. Assim, se uma expressão aparece em ambas as listas de pessoas e topónimos, será considerada PER e não LOC por quanto os antropónimos são mais frequentes do que os nomes de lugar. Resulta óbvio que as regras têm um rendimento percentual e não representam o 100% dos casos (ex. *Penso em Ruy* daria erro com a primeira regra citada da preposição *em*, e *Santiago* seria classificado sempre como nome de pessoa se estiver em ambas as duas listas de entidades LOC e PER e não houvesse contexto nenhum para a desambiguação).

Regras complementares, como a aplicação de contextos gramaticais e as listas de ativadores, permitem corrigir parcialmente os erros derivados das regras mais genéricas. Como o problema é classificatório, quanto melhor se discriminem as outras categorias, melhores resultados se obterão na classe de entidades geográficas. Porém, dado que nesta adaptação de LinguaKit o foco de estudo foram os topónimos, as regras características do módulo consideram exclusivamente as entidades geográficas, desambiguando os contextos gramaticais e precisando formas específicas da língua medieval (caso das contrações da preposição *em,* ex. *enno*). O recurso a listas de categorias não geográficas faz-se quando é possível criar uma regra que melhore a recuperação de topónimos. Assim uma lista de ativadores para entidades da categoria pessoa (ex. *bispo*, *emperatriz*, *rrainha*, *rei*, etc.), extraída ao tempo que se recolheram os ativadores *geo*, permite classificar como nomes de pessoa os nomes próprios precedidos dos termos que assinalam a entidade pessoa (PER) em casos como:

bispo <PER>*Payo Rodrigues*</PER>

No entanto, ante preposição *em* ou *de* e, opcionalmente, artigo, as regras aplicam a lista para reconhecer uma entidade geográfica:

bispo de <LOC>*Mondonedo*</LOC>

Idealmente, o sistema devia também reconhecer e classificar *bispo de Mondonedo* como pessoa mas o trabalho presente está focado no reconhecimento de topónimos e a ferramenta de extração não foi configurada para identificar entidades dentro da expressão duma outra entidade.

4.5 Lexicon e modelo de língua

Para a etiquetagem morfossintática prévia à classificação das entidades mencionadas, LinguaKit usa um lexicon que regista o lema ou lemas e valores gramaticais de cada expressão. Na aussência dum corpus lematizado e etiquetado morfossintamente que permitisse capturar recursos e treinar um modelo específico de língua medieval, juntaram-se os lexicons de galego e português acrescentados com um novo dicionário criado a partir dos termos de maior frequência (mínimo 20 ocorrências) nos corpora usados na fase de desenvolvimento (tabela 1). Para a desambiguação da etiqueta morfossintática (ex. *era*: nome comum feminino singular ou verbo imperfeito indicativo da primeira ou terceira pessoa), utilizou-se o modelo de galego, preferido por quanto mantém os pronomes enclíticos ligados ao verbo sem marca nenhuma. Parte dos textos medievais anotados automaticamente nos testes mencionados nas secções anteriores foram revistos para serem utilizados num treino mínimo de bigramas de tokens adicionado ao modelo contemporâneo. O modelo criado é utilizado por um desambiguador bayesiano para levar a cabo a etiquetagem morfossintática tal como descreve Garcia & Gamallo (2015).Mais concretamente, este módulo é um classificador bayesiano baseado em bigramas de pares $< token, etiqueta >$. Para poder atribuir uma marca (ou etiqueta morfossintática) a um token, o classificador calcula a probabilidade de cada marca dado o token alvo tomando em conta o contexto à esquerda e à direita, nomeadamente tomando em conta as etiquetas imediatamente à esquerda e à direira do token alvo. O algoritmo desambigua de esquerda à direita, de tal maneira que o contexto esquerdo dum token ambíguo é um outro token já desambiguado, é dizer, ao qual já foi atribuído uma única etiqueta.

4.6 Ajustes finais

A aplicação do módulo sobre os corpora de desenvolvimento representa uma aproximação ao que seria o limite máximo de anotação do novo módulo quando operar nas condições idóneas de abrangência total da lista de entidades geográficas e textos com topónimos cujos tipos geográficos aparecem recolhidos na lista de ativadores. A tabela 5 mostra o resultado de variar estes componentes com as regras atualizadas. Nas melhores condições, mesmo com uma lista que contém todos os topónimos do corpus a anotar, a abrangência máxima apenas atinge o 60%.

Os últimos testes serviram para confirmar a configuração ótima a respeito dos parâmetros em que a diferença no desempenho resultou ser menor (expansão da lista de ativadores e regras específicas) e realizar correções menores e pontuais em casos de ambiguidades que, sendo muito específicas dos documentos utilizados no treino, podiam afetar negativamente o desempenho noutros textos.

5 Avaliação

Para a avaliação da configuração final do módulo Histgz utilizou-se o *corpus* LNAP, presente na coleção do CGPA, mas não utilizado nas fases de desenvolvimento. O texto também não tem nenhuma anotação de topónimos e apenas requereu um pré-processamento básico para ser enviado como input para o LinguaKit. Para a validação dos verdadeiros e falsos positivos e nega-

Topónimos	Sem lista de topónimos	Com lista de topónimos
Ativadores	Sem lista de ativadores	Com lista de ativadores
Precisão	86,15%	85,12%
Abrangência	14,2%	60,36%
Medida-F	24,38%	70,63%

Tabela 5: Configurações finais do Histgz para a anotação dos próprios corpora usados no treino Mens_TX_CDMACM5_gl (3.373 entidades geográficas anotadas no padrão)

FALSOS NEGATIVOS
morador morador NCMS000 ẽna en+a SPS00+DA freigresja fregresia NCFS000 de de SPS00 **Santa_Coõba_de_Rriãjo santa_coõba_de_rriãjo NP00SP0**
En en SPS00 **San_Tomé_de_o_Mar san_tomé_de_o_mar NP00SP0**
morador morador NCMS000 ẽna en+a SPS00+DA dita dita NCFS000 **Sã sã NP00O00** **Mjgell mjgell NP00V00**

Tabela 6: Exemplos de avaliações com falsos negativos sobre a anotação do Histgz. As etiquetas NP00SP0, NP00O00, NP00V00 representam respetivamente as classes PER(soa), ORG(anização) e MISC(eláneo).

tivos (tabela 6) reviram-se todos os outputs manualmente. Dado que este é um trabalho custoso, utilizaram-se apenas os documentos iniciais do corpus até superar os 2.000 tokens (12 documentos com 2.060 tokens em total). A tabela 6 oferece uma pequena amostra da saída de Histgz, onde cada unidade lexical do texto de entrada se divide em três colunas: *token*, lema e etiqueta. O conjunto de etiquetas empregado tem por volta de 250 etiquetas mofossintáticas e baseia-se nas recomendações do Grupo EAGLE.[9]

5.1 Resultados

Visto que não há ferramentas específicas para trabalho com língua medieval, a avaliação de resultados centra-se na melhora do novo módulo, Histgz, com os já existentes para língua contemporânea no próprio LinguaKit. O gráfico da figura 2 mostra os resultados obtidos pela validação manual das anotações obtidas com as configurações dos módulos Histgz (galego-português medieval), gl (galego) e pt (português) sobre o texto com os documentos do corpus LNAP. O teste mostra o incremento do desempenho a respeito dos módulos já existentes no pacote do LinguaKit para galego e português atuais (figura 2). A abrangência, a medida que mais se vê condicionada pela adequação da lista de topónimos ao corpus (cf. tabelas 3, 4 e 5), é a que mostra um incremento mais notável, ao nível mesmo da obtida nos testes mais favoráveis durante a fase de desenvolvimento (cf. tabela 5). A precisão também obtém um melhor rendimento a respeito das versões de língua contemporânea, porém o desempenho é menor que o obtido nos testes de desenvolvimento, mesmo nas situações mais adversas (cf. tabela 3). A comparação com outros sistemas fica fora dos objetivos presentes, por quanto o Histgz é um produto operativo mas em estado muito inicial e por serem os testes realizados sobre um texto ainda que não utilizado no treino, sim presente na mesma coleção do CGPA a que pertencem os corpora usados na fase de desenvolvimento. Apenas como referência das expectativas que se podem aguardar dum sistema de reconhecimento distinto a

[9] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/Portuguese-Tagset.html

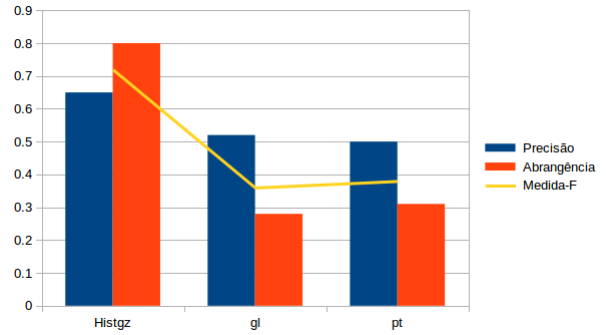

Figura 2: Comparativa do desempenho das configurações por língua do LinguaKit para um texto medieval não utilizado na fase de desenvolvimento. Os módulos *gl* e *pt* representam o Galego e Português contemporâneos, respectivamente.

LinguaKit, realizou-se um teste com os módulos de Português e Galego modernos de Freeling 4.1 sobre o mesmo texto de teste com o que foi avaliado o módulo Histgz. Os valores de medida-F obtidos foram de 44% (módulo do Português) e 39% (módulo de Galego). Estes resultados mostram um desempenho maior do que os módulos de Português e Galego modernos de LinguaKit, porém bem abaixo do nosso módulo histórico, Histgz, que, tal e como se observa na Figura 2, consegue uma medida-F superior ao 70%.

5.2 Discussão a análise de erros

A inspecção dos falsos positivos permite apontar melhoras imediatas pelo afinamento das regras e priorizar atuações nos componentes do Histgz.

5.2.1 Falsos negativos

Assim, como exemplo de falsos negativos corrigíveis por regras, um topónimo precedido por um termo geográfico (*freigresia*) é etiquetado como entidade pessoa por quanto contém também uma expressão que é internamente lematizada como ativador da classe PER (*santo*), que o classificador prioriza frente à classe LOC em casos de ambiguidade (por ex: *Santa_Coõba_de_Rriãjo*, falso negativo, tabela 6). Mesmo em contextos em que o classificador favorece a entidade geográfica, quando a ambiguidade se produz depois da preposição *em*, a presença dum token reconhecido como nome de pessoa (PER) na lista de entidades, faz com que o classificador dirima em favor desta classe (*San_Tomé_de_o_Mar*, falso negativo, tabela 6). Quando a entidade não é recuperável nem total nem parcialmente nas listas de ativadores ou entidades, a anotação já se vê afetada pela segmentação pois nenhum dos termos é reconhecido (*Sã Mjgell*, falso negativo, tabela 6).

A grande presença de topónimos hagiográficos nos textos medievais aconselha uma maior especificação na aplicação desta regra de desambiguação.

5.2.2 Falsos positivos

No caso dos falsos positivos, a anotação dos numerais romanos como nomes próprios é uma mostra da necessidade de melhorar o lexicon, factível de modo mais imediato neste caso dos numerais, frequentes e sistematizáveis (*XXVJ*, falso positivo, tabela 7). A carência duma lista de nomes de pessoa medievais provoca a classificação como entidade geográfica dos antropónimos (*Rroý_Bouçón*, falso positivo, tabela 7) quando aparecem perto de termos geográficos (uso da lista de ativadores ao não ser reconhecido o termo na lista de entidades). Dado que os corpora de desenvolvimento anotam também os antropónimos, a elaboração duma lista de nomes de pessoa medieval é também suscetível de melhora imediata.

5.2.3 Critérios para a consideração dos topónimos como verdadeiros ou falsos positivos

A análise dos falsos positivos e negativos durante a validação dos resultados mostrou que as métricas de desempenho vêm muito condicionadas pelos critérios utilizados para a definição da entidade geográfica mencionada. Nos corpora usados na fase de desenvolvimento, nomes próprios que seguem um antropónimo aparecem ocasionalmente anotados como topónimos e, muito mais frequentemente, quando precedidos pela preposição *de*. Assim, em casos como *Domingo_**Vidal*** e *Diego_Sanches_de_**Ribadeneyra***, considerados em âmbitos NERC como entidade mencionada de pessoa no seu conjunto, as formas em negrito vieram anotadas como topónimos nos corpora de treino e, consequentemente, foram avaliados como falsos negativos caso de não serem reconhecidos como geográficos nos resultados do LinguaKit. Porém, o critério utilizado para a elaboração das regras do Histgz utiliza a definição mais standard dos sistemas NERC, comum com os módulos das distintas línguas contemporâneas já presentes no pacote, de considerar uma única entidade mencionada multipalavra, mesmo se um dos elementos for também em origem classificável como entidade pertencente a outra classe. A aplicação dum critério que maximize o reconhecimento de topónimos indepen-

FALSOS POSITIVOS
dorna dorna NCFS000 de de SPS00 **XXVJ xxvj NP00G00** canadas canada NCFP000
morador morador NCMS000 ẽno en+o SPS00+DA0MS0 dito dito AQ0MS0 porto porto NCMS000 , , Fc a o DA0FS0 **Rroý_Bouçón rroý_bouçón NP00G00** , , Fc morador morador NCMS000 ẽna en+a SPS00+DA freigresja fregresia NCFS000
Testigos testigos NP00V00 : : Fd Vasco vasco AQ0MS0 de de SPS00 **Lees lees NP00G00**

Tabela 7: Exemplos de avaliações com falsos positivos sobre a anotação do Histgz. A etiqueta NP00G00 representa a classe LOC(alização) ou topónimo.

dentemente de qual for o referente principal ou, pela contra, a minimização do número de entidades em favor dum único referente, é discutível e varia em função dos interesses particulares da anotação. Em efeito, e mais particularmente nos textos medievais, uma mesma estrutura sintática pode ter tanto valor duma única entidade mencionada (<PER>*Vasco de Lees*</PER>) quanto de duas (<PER>*Vasco*</PER> *de* <LOC>*Lees*</LOC>). No caso da validação da anotação do Histgz, com o fim de aplicar um mesmo critério para todos os casos, entende-se que o nome próprio precedido por antropónimo mais preposição *de* deve ser validado como entidade mencionada de pessoa em todos os casos. Consequentemente, a anotação deste exemplo na tabela 7 foi avaliada como falso positivo. Esta divergência no critério do que é ou não uma entidade geográfica mencionada influi, portanto, nas métricas, e deve ser tida em conta à hora de valorar os resultados.

6 Conclusões e trabalho futuro

A adaptação para o trabalho com textos medievais duma ferramenta NERC inicialmente concebida para labores PLN com textos contemporâneos ofereceu uns resultados que melhoram notavelmente o desempenho a respeito dos módulos existentes. O labor de configuração consistiu na avaliação do desempenho da utilidade sobre corpora previamente anotados, modificando parâmetros e adicionando recursos segundo se ia experimentado com os textos. A principal dificuldade para a melhora dos resultados na medida-F é o incremento da abrangência sem comprometer excessivamente a precisão. A lista de topónimos foi considerada o recurso mais determinante, porém, dada a alta variação gráfica e a tipologia textual, que favorece a aparição de microtopónimos, qualquer lista moderna aparece comprometida na abrangência. A aplicação duma lista com termos geográficos permitiu melhorar os resultados, ainda que em menor grau do que a lista de topónimos, contudo, os ativadores contribuem também para a desambiguação de entidades de mais difícil classificação. O critério para a definição do que é ou não um topónimo e quando se deve reconhecer como entidade geográfica mencionada, influi na definição de regras e na validação dos resultados. Com a necessidade de salientar que a avaliação vem condicionada pela consideração de entidade geográfica mencionada mais conforme às práticas NERC do que a uma definição mais abrangente de topónimo, e que o corpus utilizado é reduzido para permitir uma validação manual, a análise do teste final mostra que o produto obtido melhora os resultados dos módulos NERC prévios. O Histgz é,

contudo, apenas uma versão inicial necessitada de melhoras. Ao atender preferentemente as entidades geográficas, ficaram desatendidos ou minimamente considerados outros recursos que contribuem tanto para o rendimento da anotação NERC quanto para a expansão das capacidades PLN dentro do amplo abano de utilidades que o LinguaKit oferece. O módulo no seu estado atual é maiormente dependente do treino sobre textos contemporâneos que serviram de base de desenvolvimento, com exceção da lista de ativadores e em menor medida da lista de entidades geográficas. O trabalho futuro consiste na ampliação do lexicon e o treino dum modelo de língua representativo da variedade medieval, labor que, pela dependência que tem na validação experta, requer de recursos notavelmente superiores aos utilizados para a obtenção do produto atual. Porém, o próprio módulo Histgz pode ser já aplicado para facilitar a preparação dos textos e produzir corpora que simplifiquem e agilizem o trabalho de reconhecimento e anotação.

Como já foi dito, o módulo foi integrado em LinguaKit e tanto o léxico como as listas de entidades e ativadores de Histgz estão disponíveis com licença livre.[10]

Agradecimentos

Este trabalho foi desenvolvido no marco da rede galega de investigação TECANDALI, ED341D R2016/011, financiada pela Consellaria de Educación e Ordenación Universitaria da Xunta de Galicia, e do European Regional Development Fund (ERDF).

Referências

Afonso, Susana, Eckhard Bick, Renato Haber & Diana Santos. 2002. Floresta sintá(c)tica: a treebank for portuguese. Em *3rd International Conference on Language Resources and Evaluation (LREC)*, 1698–1703.

Amaral, Daniela, Evandro Fonseca, Lucelene Lopes & Renata Vieira. 2014. Comparative analysis of portuguese named entities recognition tools. Em *9th International Conference on Language Resources and Evaluation (LREC)*, 2554–2558.

Borin, Lars, Dimitrios Kokkinakis & Leif-Jöran Olsson. 2007. Naming the past: Named entity and animacy recognition in 19th century Swedish literature. Em *Workshop on Language Technology for Cultural Heritage Data (LaTeCH)*, 1–8.

Byrne, Kate. 2007. Nested named entity recognition in historical archive text. Em *International Conference on Semantic Computing (ISCS)*, 589–596.

Cal Pardo, Enrique (ed.). 1999. *Colección diplomática medieval do arquivo da Catedral de Mondoñedo. Transcrición íntegra dos documentos*. Santiago de Compostela: Consello da Cultura Galega.

Canosa, Afonso Xavier. 2017. *A identificacão e referenciacão de entidades geográficas mencionadas: o caso da 'Peregrinação' de Fernão Mendes Pinto*: Universidade de Santiago de Compostela. Tese de Doutoramento.

Chaves, Marcírio. 2008. Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o SEI-Geo no segundo HAREM. Em Cristina Mota & Diana Santos (eds.), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, Linguateca.

Freitas, Cláudia, Cristina Mota, Diana Santos, Hugo Gonçalo Oliveira & Paula Carvalho. 2010. Second HAREM: Advancing the state of the art of named entity recognition in Portuguese. Em *7th International Conference on Language Resources and Evaluation (LREC)*, 3630–3637.

Galves, Charlotte. 2018. Tycho brahe parsed corpus of historical Portuguese. Universidade de Campinas. http://www.tycho.iel.unicamp.br/~tycho/corpus/texts/psd.zip.

Gamallo, Pablo & Marcos Garcia. 2011. A resource-based method for named entity extraction and classification. Em *Progress in Artificial Intelligence, 15th Portuguese Conference on Artificial Intelligence (EPIA)*, vol. 7026, 610–623. 10.1007/978-3-642-24769-9.

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.

Garcia, Marcos & Pablo Gamallo. 2015. Yet another suite of multilingual NLP tools. Em *Languages, Applications and Technologies (SLATE)*, vol. 563 Communications in Computer and Information Science, 65–75. 10.1007/978-3-319-27653-3_7.

[10] https://github.com/citiususc/Linguakit/tree/master/tagger/histgz

Garcia, Marcos, Iria Gayo & Isaac González López. 2012. Identificação e classificação de entidades mencionadas em galego. *Estudos de Lingüística Galega* 4. 13–25.

Grover, Claire, Sharon Givon, Richard Tobin & Julian Ball. 2008. Named entity recognition for digitised historical texts. Em *6th International Conference on Language Resources and Evaluation (LREC)*, 1343–1346.

Hendrickx, Iris & Rita Marquilhas. 2011. From old texts to modern spellings: An experiment in automatic normalisation. *Journal for Language Technology and Computational Linguistics* 26(2). 65–76.

Jones, Alison & Gregory Crane. 2006. The challenge of virginia banks: an evaluation of named entity analysis in a 19th-century newspaper collection. Em *6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, 31–40.

Marquilhas, Rita & Iris Hendrickx. 2014. Manuscripts and machines: the automatic replacement of spelling variants in a Portuguese historical corpus. *International Journal of Humanities and Arts Computing* 8(1). 65–80. 10.3366/ijhac.2014.0120.

Mota, Cristina & Diana Santos. 2008. Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O segundo HAREM. Em Cristina Mota & Diana Santos (eds.), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, Linguateca.

Pinto, Alexandre, Hugo Gonçalo Oliveira & Ana Oliveira Alves. 2016. Comparing the performance of different NLP toolkits in formal and social media text. Em *5th Symposium on Languages, Applications and Technologies (SLATE)*, vol. 51, 3:1–3:16. 10.4230/OASIcs.SLATE.2016.3.

Rodríguez, Pérez & Francisco Javier (eds.). 2004. *Os documentos do tombo de Toxos Outos*. Santiago de Compostela: Consello da Cultura Galega.

Santos, Diana & Nuno Cardoso (eds.). 2007. *Reconhecimento de entidades mencionadas em português: Documentação e actas do harem, a primeira avaliação conjunta na área*. Linguateca.

Santos, Diana, Nuno Cardoso & Nuno Seco. 2007. Avaliação no HAREM: Métodos e medidas. Em Diana Santos & Nuno Cardoso (eds.), *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*, 245–282. Linguateca.

Tato Plaza, Fernando R. (ed.). 1999. *Libro de notas de Álvaro Pérez, notario da Terra de Rianxo e Postmarcos*. Santiago de Compostela: Consello da Cultura Galega.

Varela Barreiro, Xavier (ed.). 2009. *Xelmírez. Corpus Lingüístico da Galiza Medieval*. Santiago de Compostela: Instituto da Lingua Galega (USC).

Varela Barreiro, Xavier & Paulo Martínez Lema. 2009. *Inventario Toponímico da Galiza Medieval*. Santiago de Compostela: Instituto da Lingua Galega.

Varela Barreiro, Xavier, Maria Francisca Xavier & Charlotte Galves. 2016. *Corpus informatizado Galego-Português Antigo*. Santiago de Compostela / Lisboa / Campinas: Instituto da Lingua Galega / Centro de Lingüística da Universidade Nova de Lisboa / Universidade de Campinas.

Won, Miguel, Patricia Murrieta-Flores & Bruno Martins. 2018. Ensemble Named Entity Recognition (NER): Evaluating NER Tools in the Identification of Place Names in Historical Corpora. *Frontiers in Digital Humanities* 5. 2. 10.3389/fdigh.2018.00002.

Xavier, Maria Francisca (ed.). 2000. *Corpus Informatizado do Português Medieval*. Lisboa: Centro de Lingüística da Universidade Nova de Lisboa.

Zapico Barbeito, Pilar (ed.). 2005. *Colección diplomática do mosteiro de Santiago de Mens. Edición e estudo*. Noia: Toxosoutos.