CARACTERÍSTICAS IDENTIFICADORAS E DIFICULDADES NA APLICAÇÃO DE LISTAS PARA A ANOTAÇÃO DE ENTIDADES GEOGRÁFICAS MENCIONADAS

IDENTIFYING CHARACTERISTICS AND DIFFICULTIES WHEN USING GAZETTEERS TO ANNOTATE GEOGRAPHICAL NAMED ENTITIES

Afonso Xavier Canosa*
canosarodriguez@gmail.com

Na anotação automática de entidades geográficas mencionadas, as listas especializadas de topónimos têm que enfrentar ambiguidades e contextos em que o valor geográfico de uma expressão não é evidente. Neste artigo, estuda-se o caso prático de um índice de topónimos utilizado para criar um corpus anotado da *Peregrinação* de Mendes Pinto. As dificuldades achadas servem para classificar os tipos de erros que se produzem quando o topónimo é resolvido pela simples coincidência de expressões e introduzem critérios para a identificação das entidades geográficas, uma tarefa que deve preceder e tem um impacto direto nos resultados obtidos no processo de anotação automática.

Palavras-chave: Entidades Geográficas Mencionadas. REM. Topónimos. Anotação de corpus. Corpus histórico.

In order to annotate geographical named entities, gazetteers have to face ambiguities and contexts where the geographical value of a given expression is not clear. In this paper, an index of place names is used to examine the main problems encountered in the production of an annotated corpus of Mendes Pinto's *Pilgrimage*. The difficulties found serve to classify the types of errors that occur when the place name is solved by simple string match and introduce criteria for the identification of geographical entities, a task that should precede and has a direct impact on the results obtained in an automatic annotation approach.

Keywords: Geographical Named Entities. NERC. Toponyms. Corpus annotation. Historical corpus.

---

* Universidade de Santiago de Compostela, Espanha.

1. Introdução

Entidade Mencionada (EM) é o termo utilizado em Processamento da Linguagem Natural (PLN) para se referir mais comummente a nomes de pessoas, organizações e lugares (Amaral *et al*. 2014; Nadeau & Sekine 2007; Santos & Cardoso 2007). O termo propriamente aponta a um referente, objeto único no mundo real, porém, é frequentemente utilizado para se referir à expressão, isto é, a cadeia de caracteres que aparece num texto para designar uma entidade.

Neste artigo, *Entidade Geográfica Mencionada* (EGM) será aquele nome que refere um objeto geográfico único (pode haver vários lugares com um mesmo nome, mas quando o usamos estamos a nos referir a um lugar em concreto e só um) instância de uma classe (refere o indivíduo e não a classe, ex. *Lisboa* é uma cidade). O exemplo mais comum são os topónimos, ainda que também pode abranger entidades menores, tais como nomes de edifícios ou ruas e, menos frequentemente, os gentílicos, enquanto se considera o seu radical como expressão portadora das características semânticas do nome próprio. Quando quiser referir o objeto geográfico, espaço físico que ocupa uma posição determinada no planeta Terra, utilizarei mais frequentemente o termo *referente*.

Um recurso comum no processo de anotação de entidades geográficas mencionadas é o uso de listas de entidades geográficas (*gazetteers*) que provêm, para além do topónimo, informação complementar de utilidade para a desambiguação e georreferenciação (Leidner 2007, p. 51; Southall, Mostern & Berman 2011), particularmente as coordenadas geográficas em termos de latitude e longitude. Na aplicação de gazetteers para a anotação automática, quando um termo no texto coincide com um topónimo da lista, outorgamos o atributo de entidade geográfica e recuperamos a informação relevante disponível segundo os objetivos e o problema a resolver: quer simples reconhecimento e classificação das entidades mencionadas, quer labores mais específicos de resolução e análise geográfica (Gregory *et al*. 2013; 2015). Porém, mesmo quando se tiver uma lista específica, a simples aplicação dos topónimos produz ambiguidades (ex. *Carvalho* pode ser um topónimo, antropónimo ou nome comum).

Para superar estas dificuldades, os sistemas de anotação automática introduzem listas auxiliares (de distintas categorias de EM e ativadores

da classe) e regras que permitam desfazer as ambiguidades a partir de padrões e regularidades linguísticas, do tipo: *se um nome próprio que coincide com um topónimo na lista vai precedido de antropónimo, será um apelido e não uma EGM* (ex. Paulo Carvalho) ou *se um nome próprio vai precedido de um nome comum da lista de ativadores de classe geográfica seguido pela preposição* de*, é uma EGM* (ex. rei de Portugal). Estas regras podem ser expressadas de modo explícito no código da ferramenta (sistema de regras) ou aprendidas a partir do treino em corpora (sistema de aprendizado de máquina). A vantagem do primeiro tipo é que, ao permitir otimizar os resultados mediante a depuração das regras, resulta facilmente adaptável para o trabalho com corpora históricos em que a modalidade de língua apresenta grandes diferenças com o padrão contemporâneo, caso do galego-português medieval (Canosa *et al*. 2018). Para serem convenientemente treinados, os sistemas de aprendizado requerem texto previamente anotado e estatisticamente relevante, num volume nem sempre disponível no caso de textos históricos. Porém, nas avaliações realizadas sobre textos em inglês dos séculos XVII e XVIII (modalidade de língua mais próxima ao padrão contemporâneo com que foram treinados) ofereceram os melhores resultados de desempenho (Won, Murrieta-Flores & Martins 2018).

2. Motivação e objetivos

Independentemente do sistema de anotação utilizado, o resultado final vem condicionado pelo que se entende como EGM. As regras, quer aprendidas automaticamente a partir de corpora já anotados, quer explícitas no código, necessitam de uma definição prévia do que se deve anotar. Para uma filóloga estudiosa da evolução de um topónimo, qualquer concordância serve para observar usos gráficos com que analisar possíveis alterações fonéticas, independentemente de se a expressão aparecer como apelido ou nome de lugar. Porém, para um historiador interessado em reconstruir redes de relações num momento dado, a origem toponímica de um apelido pode ser totalmente irrelevante. Há, portanto, um componente subjetivo importante à hora de definir o que deve ser uma EGM.

Doutra parte, dado que as ferramentas de que dispomos na atualidade são limitadas, faz-se necessário o trabalho em equipa para desenvolver ou melhorar as utilidades que facilitem o processo de anotação. O presente artigo pretende ilustrar as dificuldades que aparecem ao tratar de conciliar

os requerimentos próprios de quem elabora um corpus para a investigação humanística posterior com as limitações e possibilidades de automatização resultantes de aplicar uma lista que identifica as expressões coincidentes no texto (exemplo que serve de *baseline* para o desempenho de uma ferramenta de anotação mais específica).

Em textos históricos e obras comentadas com aparato crítico, é relativamente comum dispor de índices temáticos e glossários de entidades geográficas específicas ou muito próximas aos objetivos do corpus. A questão que se pretende responder aqui é, então, quais são as dificuldades que surgiram ao aplicar um destes índices para anotar automaticamente um texto duma modalidade linguística para a qual não há ferramentas específicas desenvolvidas. Que regras ou exceções necessitamos aplicar para anotar todas as entidades da lista? Quais foram os critérios utilizados para identificar uma expressão como EGM?

3. Materiais e métodos

Analiso a seguir as situações mais problemáticas achadas no caso prático de anotação do corpus da *Peregrinação* de Fernão Mendes Pinto (1614) a partir de uma lista específica da obra. Para este corpus em particular, temos índices e dicionários (Albuquerque 1994; Alves 2010; Flores, Gomes & Sousa 1983; Lagoa 1950–1953). O mais exaustivo e específico (Alves 2010) recolhe praticamente todas as formas toponímicas, ainda que não todas as variantes, motivo pelo qual elaborei uma lista própria, extraída manualmente em sucessivas leituras que acompanharam o estudo crítico do texto. Os glossários prévios foram de todos modos muito úteis para comprovar a qualidade da lista própria e um instrumento de trabalho imprescindível para a abordagem de problemas mais avançados de resolução geográfica.

Mediante uma série de scripts que processam o corpus, utilizei a lista própria para anotar de modo automático os topónimos no texto completo da *Peregrinação* (*PR*) segundo a primeira edição. O procedimento e objetivos mais específicos de georreferenciação foram publicados com anterioridade (Canosa 2017). Os resultados concretos de corpus anotado e georreferenciado estão disponíveis para consulta pública *on-line*.[1]

1 https://www.pucau.org

4. Análise e discussão das principais dificuldades detetadas na anotação do *corpus* do caso prático

Os exemplos a continuação ilustram as principais dificuldades achadas e os critérios aplicados para anotar um *token* face a outros que, ainda coincidindo na mesma expressão, não são entidade geográfica mencionada. Cada dificuldade estudada é fechada com a característica mais relevante para determinar que a expressão em questão seja ou não uma EGM. As características identificadoras servem de regras que, em conjunto, provêem a definição de entidade geográfica aplicada para o corpus. A referência às concordâncias que ilustram os casos são feitas para o capítulo de modo que sejam recuperáveis em qualquer edição.

4.1. A entidade geográfica mencionada coincide com uma forma do vocabulário

Sejam as concordâncias:

a) "Como Antonio de Faria chegou ao rio de Tinacoreu, a que os nossos chamão **Varella**, & da informação que daquelle reyno lhe derão hũs mercadores." (*PR*, 41)

b) "& no mesmo dia se coroou por Rey de Péguu **na varella grande**" (*PR*, 193)

c) "Passados os dez dias deste encerramento, **as varellas** & pagodes, & brallas, que saõ os seus templos, amanheceraõ todos ornados de insignias de alegria" (*PR*, 184)

A forma normalizada "varela" tem vários significados. No contexto da *Peregrinação* achamos o recolhido por Pereira (1647) e Bluteau (1712–28):

> "Varela. Templum" *Thesouro*, (Pereira 1647, fól. 94 v.)

> "Varella, ou Varela. (Termo da India.) Templo de Idolos, ou mosteyro de Gentios."
> *Vocabulario*, (Bluteau 1712–28)

Estamos, portanto, perante um nome comum que se regista como tal num dicionário. No entanto, na primeira cita (a) achamos a forma precedida

por um topónimo, *Tinacoreu*, explicitando ademais o seu uso como nome de lugar, a forma portuguesa equivalente a uma outra asiática. Isto é, temos um topónimo transparente, um nome de lugar com um significado que se corresponde, para além da denotação de um espaço geográfico particular, com o de um nome comum, mas nome de lugar nesta concordância e, assim sendo, EGM.

Característica identificadora: a entidade geográfica mencionada começa por maiúscula. Assim (a) é EGM, (b) e (c), sendo a mesma expressão, não são entidades geográficas mencionadas.

4.2. A entidade geográfica mencionada coincide plenamente com uma forma do vocabulário

Uma dificuldade maior aparece na seguinte concordância:

d) "sendo tanto auante como o rio a que os naturaes da terra chamão Tinacoreu, & os nossos **a varella**" (*PR*, 41)

Citamos a forma transparente *a varella* como equivalente a um topónimo opaco, *Tinacoreu*, mas agora com uma particularidade, não se faz uso de maiúscula. No entanto o seu valor toponímico é tão claro como o citado anteriormente em (a). Temos um caso de variação ou dúvida na normalização, o editor mostra incoerência ou houve gralha na publicação, característica dos textos históricos, face ao processamento de textos contemporâneos em que aguardamos aderência a uma norma que nos permita sistematizar todos os casos e as suas exceções. Em (d) temos uma EGM, mas a sua codificação supõe uma exceção à primeira regra de identificação: não começa por maiúscula. No nosso caso anotamo-la igualmente como EGM e a lista regista-a como mais uma variante com a particularidade de aparecer em minúscula.

Seguindo o mesmo critério anotamos como EGM:

e) "A Quarta feira seguinte nos saimos logo deste **rio da varella** por nome Tinaçoreu" (*PR*, 42)

Porém, isto cria um novo problema, assim nas concordâncias:

f) "sem fazeres nenhũa detença te venhas logo com essas naos por junto do **baluarte do caez da varella**, onde me acharàs em pé esperãdo por ty" (*PR*, 148)

g) "E assi no tempo que o Rey Bramaa foy sobre o reyno de Sião, & após cerco à cidade de Odiaa, como atras fica dito, pregando o Xemindoo então **na varella do Comquiay de Pegù**, que he como See de todas as outras" (*PR*, 190)

h) "E com isto se partio logo para a cidade de Pegù, onde dos moradores della foy recebido com triumpho de Rey, & coroado por esse **na varella do Comquiay**, que he como See de todas as outras." (*PR*, 190)

Mesmo se todas as frases destacadas em (d), (e), (f), (g) e (h) referem um lugar concreto, que pode ser referenciado (e nesse sentido todas cinco são georreferências susceptíveis de lhe serem atribuídas umas coordenadas geográficas específicas), em (d) e (e) *a varella* é instância da classe *rio*, e não da classe *varela* (templo). Tem ademais o valor de unicidade, apenas há um *rio da Varela*. Propriedades, estas, de objeto único e indivíduo (e não classe) apontadas na introdução (§1) como características mais definitórias da EM.

No entanto em (f), (g), (h), a mesma expressão refere tanto o indivíduo quanto a classe (*varela*). Precisa, ademais, de modificadores para denotar uma individualidade (*"do Comquiay*" (g), (h)), ou aparece simplesmente como modificador duma individualidade ("*do baluarte do caez da varella*" (f)).

Característica identificadora: a entidade geográfica mencionada aparece em contexto de nome próprio. Assim em (d) e (e) temos nomes próprios que grafaríamos com maiúscula segundo as convenções atuais, no entanto em (f), (g) e (h) estamos perante nomes comuns. Para a identificação de (d) e (e) como EGM necessitamos, portanto, criar uma exceção à regra das maiúsculas (§4.2). As marcas identificadoras são agora os elementos lexicais que precedem a EGM (verbo *chamar* e termo geográfico *rio*).

4.3. A entidade geográfica mencionada é complexa e um dos seus elementos comporta-se como uma forma do vocabulário

Seja a concordância:

i) "Como nos partimos desta **ilha dos ladroẽs** para o porto de Liampoo, & do que passamos atè chegarmos a hum rio que se dizia Xingrau" (*PR*, 55)

Em (i) temos um exemplo similar a (e), um nome comum forma parte dum topónimo, todos os termos aparecem como elementos do vocabulário, no entanto em:

j1) "& com este concerto jurado & assinado por todos, se vieraõ surgir a hũa **ilha que se dezia dos ladroẽs"** (*PR*, 53)

o mesmo topónimo aparece com uma cláusula intercalada. Pelo critério usado em §4.2 para (e), (i) é também EGM, mas em (j) *ilha* comporta-se como um nome comum (de facto vem precedido do artigo indefinido para a singularizar, enfatizando o significado de não unicidade do termo). Consideramos três opções:

1) Simplificamos o topónimo e anotamos *dos ladroẽs* como variante.

De um ponto de vista estritamente linguístico, esta é a solução mais coerente com o estatuto gramatical das EGMs se quisermos manter uma cadeia única e contínua (não interrompida). Porém, da parte do processamento automático, surge mais uma dificuldade: ao aplicarmos a lista sobre o corpus obtemos também a concordância:

k) "mãdou tambem hum Naique com vinte Abexins que nos veyo guardando **dos ladroẽs**, & prouendonos de mãtimẽto & caualgaduras ate o porto de Arquico" (*PR*, 4)

Uma possível solução à ambiguidade criada por concordâncias como (k) requer processar não só as entidades geográficas mencionadas, mas o contexto em que aparecem. A anotação morfossintática mediante técnicas de PLN identifica um verbo (*guardando*) antes da frase preposicional (*dos ladrões*), uma regra simples indica que neste caso não se trata de uma EGM.

2) Outra solução passa por anotar todo o sintagma como variante. Isto é:

j2) "& com este concerto jurado & assinado por todos, se vieraõ surgir a hũa **<LUGAR>ilha que se dezia dos ladroẽs</LUGAR>**" (*PR*, 53)

3) Uma solução intermédia adiciona um módulo no sistema de aplicação da lista ao corpus que identifica o início do topónimo (*ilha*) e resto dos componentes (*dos ladroẽs*), obviando os elementos alheios (*que se dezia*). A dificuldade desta proposta é termos de processar a lista de variantes de entidades geográficas identificando os seus componentes. Neste exemplo, a marca <B> assinala o começo da entidade mencionada, <O> o segmento a omitir e <I> a parte final.

j3) "& com este concerto jurado & assinado por todos, se vieraõ surgir a hũa **<LUGAR><B>ilha<B> <O>que se dezia</O> </I>dos ladroẽs</I></LUGAR>**" (*PR*, 53)

Optamos pela solução 2, operativamente mais eficaz no processamento do corpus, já que requer unicamente adicionar mais uma variante na lista. As soluções alternativas guardam uma maior homogeneidade nas variantes do lexema mas dificultam o processamento, ao necessitarem de uma regra para solucionar um único caso no corpus.

Característica identificadora: a entidade geográfica mencionada incorpora uma forma genérica de identificativo geográfico (ex. rio de, cidade de, ilha de) e mesmo sintagmas verbais adicionais quando a forma mais característica do nome próprio fica, de outro modo, incompleta.

4.4. A entidade mencionada é ambígua na expressão de referente geográfico

No exemplo:

l1) "E sendo sua alteza certificado da sua morte, proueo segunda vez na mesma capitania a hum Diogo Cabral da ilha da **Madeyra**, a quem Martim Afonso de **Sousa** a tirou por justiça, por se dizer que praguejara delle sendo Gouernador, & a deu a hum Ieronymo de **Figueiredo** fidalgo do Duque de **Bargança"** (*PR*, 20)

temos quatro entidades geográficas mencionadas com um elemento comum: servirem de complemento a um nome próprio antropónimo, com uma função similar à dos apelidos, mas precedidas de uma preposição que permite a interpretação da frase como lugar de procedência. Cada caso apresenta alguma particularidade que não têm os outros. Aceitando a definição de entidade mencionada como portadora do princípio de unicidade (§1) estamos perante um referente de pessoa, isto é, uma solução por exemplo do tipo:

l2) "E sendo sua alteza certificado da sua morte, proueo segunda vez na mesma capitania a hum **<PESSOA> Diogo Cabral </PESSOA>** da ilha da **<LUGAR> Madeyra </LUGAR>**, a quem **<PESSOA>Martim Afonso de Sousa</PESSOA>** a tirou por justiça, por se dizer que praguejara delle

> sendo Gouernador, & a deu a hum **<PESSOA>Ieronymo de Figueiredo</PESSOA>** fidalgo do **<PESSOA>Duque de Bargança</PESSOA>**" (*PR*, 20)

Importa agora notar como quatro entidades do tipo PESSOA aparecem, dentro duma estrutura sintática similar (Frase Nominal + Frase Preposicional), ligadas a entidades geográficas de modo distinto. No primeiro caso, *Diogo Cabral da ilha da Madeyra*, (óbvio agora o facto de o segundo nome próprio, *Cabral*, ser também expressão de um topónimo) a presença de um termo do domínio geográfico (ilha) evidencia que estamos perante uma entidade geográfica na frase preposicional: é indício claro de referente geográfico, a pessoa está a ser referida como procedente de um lugar concreto. No nosso corpus é anotada como EGM com a marca <LUGAR> em (l_2).

No segundo caso, *Martim Afonso de Sousa*, temos uma forma (*Sousa*) apelido comum, mas também topónimo, nome de rio em N 41° 5' 48", W 8° 30' 6", onde também dá nome a freguesia, para além de ser topónimo noutras localidades de Portugal. A expressão é ambígua e temos de optar por uma interpretação. A questão a responder é, optamos por marcar Sousa como um nome de lugar ou deixamo-lo como apelido (tal e como se faz com *Cabral* em *Diogo Cabral*, independentemente de que o apelido tenha tido originalmente uma origem toponímica)? Uma solução é usarmos critérios que determinem a escolha. Considerando que a forma candidata a entidade geográfica aparece em todas as ocorrências como denotadora de pessoa, sem acharmos uso nenhum como entidade geográfica independente, optamos por deixar sem anotar como EGMs estes casos, isto é, o critério de avaliação de candidatos ao usarmos uma lista de termos geográficos penaliza a ambiguidade e favorece aquelas formas que aparecem em contextos em que o referente é mais inequivocamente geográfico.

O terceiro caso, *Ieronymo de Figueiredo*, também contém um nome de várias freguesias em Portugal. Tem a particularidade de ser transparente, isto é, leva um significado explícito associado com um morfema derivativo associado à marca de lugar (*-edo,* expressão de fitotopónimo com o significado de lugar em que abunda uma espécie). Mas sintaticamente aparece numa estrutura típica de antropónimo. Aceitando a situação de ambiguidade (um estudo filológico mais detido poderia desfazê-la com relativa facilidade) usamos os critérios de simplificação e frequência no corpus (apenas ocorre em nome de pessoa), e deixamos *Figueiredo* como mais uma expressão sem função de referente geográfico.

Finalmente temos *Duque de Bargança*. O conjunto é uma entidade mencionada de pessoa, mas agora aparece uma ligação a um referente

geográfico de forma não ambígua. Nos casos precedentes, um ou vários nomes próprios de pessoa vão seguidos de mais um nome próprio como resultado de uma codificação histórica ou cultural (por exemplo, sistemas romano, português ou inglês para o nome completo de uma pessoa) que pode, como mais uma possibilidade, ser ocupado por um nome de lugar (em função dos usos administrativos numa jurisdição, período, circunstâncias individuais mesmo). Nomes associados a estruturas governativas e territoriais requerem semanticamente um âmbito geográfico e, portanto, desaparece a ambiguidade que achamos no apelido (*de Sousa* e *de Figueiredo*). Mais ainda, em *Duque de Bargança*, a frase nominal núcleo não contém um nome próprio, mas um comum com um significado genérico. Nestes casos sim optamos por atribuir a marca de EGM, pois o primeiro termo da entidade pessoa aponta para uma entidade geográfica de forma não ambígua. Aliás, usamos o critério de frequência, porquanto a mesma expressão tem uma ocorrência (m) em que aparece como EGM independente:

m) "hum Portuguez que andaua com elles, por nome Christouão Sarmento natural de **Bargança**" (*PR*, 195)

Característica identificadora: a entidade geográfica mencionada tem como referente primeiro uma entidade geográfica.

Isto é, *Cabral* em *Diogo Cabral*, *Sousa* em *Martim Afonso de Sousa* e *Figueiredo* em *Ieronymo de Figueiredo* são, primeiramente, apelido, e aparecem no corpus unicamente como apelido: o seu valor referencial é o de um antropónimo e não o de um topónimo. Mesmo querendo atribuir-lhe um valor de nome de lugar, são expressões de grande ambiguidade geo / geo (topónimos muito comuns). Quando se quiser explicitar a procedência geográfica como modificador do antropónimo, necessitaremos mais algum elemento explicativo (*da ilha da Madeyra* em *Diogo Cabral da ilha da Madeyra*). No entanto, em *Duque de Bargança* temos uma expressão que de modo inequívoco está a especificar um espaço geográfico, o próprio núcleo nominal requer que o modificador seja uma entidade geográfica. Enquanto o objetivo principal de anotação do corpus é unicamente o estudo da geografia (as entidades de pessoa não são anotadas), e uma boa parte dos topónimos do corpus, particularmente na Ásia, vêm precedidos por um nome comum que ativa um topónimo (termos com significado de autoridade sobre um território facilmente sistematizáveis numa lista) considera-se o referente de *Bargança* como uma entidade geográfica e a expressão é, portanto, anotada como EGM.

4.5. A expressão da entidade geográfica mencionada é uma entidade não geográfica

No exemplo:

n) "& hum destes Portugueses era hum Christouão Doria, que nesta terra foy despois mandado por capitão a **São Tomè**, & os outros dous erão Luys Taborda, & Simão de Brito, todos homẽs honrados & mercadores ricos" (*PR,* 147)

temos uma situação inversa a §4.4, estamos perante uma EGM cuja expressão, *São Tomè*, tem também valor de hagiónimo. Neste caso o referente é claramente a entidade geográfica. Seguindo este mesmo critério, de anotarmos a expressão segundo o referente primeiro, no exemplo:

o) "caminhamos ao longo de hum rio mais cinco legoas, até hum lugar que se chamaua Bitonto, no qual nos agasalhamos aquella noite em hum bom Mosteyro de Religiosos que se chamaua **Sao Miguel**, com muyta festa & gasalhado do Prior & Sacerdotes que nelle estauão, onde nos veyo ver hum filho do Barnagais Gouernador deste imperio de Ethyopia." (*PR,* 4)

ao considerarmos as construções como um tipo geográfico, se estas aparecerem referidas por um nome próprio, teremos também um caso de EGM. Em (o) "o que se chama" é o mosteiro, a entidade é geográfica, independentemente do seu carácter hagionímico.

Do mesmo modo o teónimo *Tinaogoogoo* na concordância:

p) "E porque o embaixador adoeceo aquy de hũ inchaço nos peitos, foy acõselhado que não passasse adiãte até não ser saõ delle, pelo que assentou cõ algũs dos seus de se yr curar a hũa grande enfermaria que estaua daly doze legoas adiante em hũ pagode por nome **Tinagoogoo**, que quer dizer, deos de mil deoses, para onde partio logo, & chegou là hum sabbado ja quasi noite" (*PR,* 158)

aparece explicitamente como expressão de um edifício, um pagode. Portanto, o seu referente é uma entidade geográfica e não diretamente o deus *Tinagoogo*, como no exemplo:

q) "na qual noite se gastou infinito numero de cera nas luminarias que se fizeraõ, as quais tomauão tanto espaço de terra quanto a vista podia alcançar, o que tudo parecia então que ardia em fogo, & a razão disto era, porque

dezião que o **Tinagoogoo** deos de mil deoses era ido em busca da serpe tragadora para a matar com hũa espada que lhe viera do Ceo." (*PR,* 161)

Os casos mais ambíguos surgem quando o teónimo ou hagiónimo não é declarado explicitamente como expressão da entidade geográfica. Assim em:

r) "Do caminho que fizemos atê chegarmos ao pagode de **Tinagoogoo**." (*PR,* 158)

A expressão *Tinagoogoo* aparece agora como modificador e não núcleo do sintagma EGM. No entanto, o contexto refere o mesmo pagode que em (p). A solução que adotamos neste caso passa por um critério alheio a regras linguísticas: com o fim de obtermos mais ocorrências para o tratamento do corpus e resolução de georreferências, anotamos como EGM aqueles casos em que, tendo sido declarada a expressão de modo explícito como entidade geográfica, volte aparecer numa outra concordância acompanhada de um atributo que permite atribuir um objeto geográfico como referente.

Característica identificadora: uma expressão declarada de modo explícito no texto como entidade geográfica será anotada como tal sempre que mantiver o valor de referente de objeto geográfico, ainda quando tiver também ocorrências com o valor de entidade de uma classe não geográfica.

Assim em (o) um hagiónimo é explicitamente declarado expressão geográfica, como também o teónimo em (p) nomeia de modo inequívoco um edifício, em ambos os casos anotamos a expressão como EGM. Do mesmo modo anotamos (r), porquanto ainda estando diante de um teónimo, a expressão foi declarada em mais alguma ocorrência no corpus (p) como entidade geográfica e volta aparecer como modificador do tipo geográfico que a subclassifica (pagode). Porém, em (q), a mesma expressão tem unicamente o valor de teónimo e, em consequência, não é considerada EGM.

4.6. A entidade geográfica mencionada contém outra entidade geográfica mencionada

No exemplo:

s1)"& leuamos tambem hum Bispo Abexim, que vinha para vir a este reyno, & daquy yr a **Santiago de Galiza**, & a Roma, & dahy a Veneza, para dahy se passar a Ierusalem." (*PR,* 5)

temos *Santiago de Galiza*, entidade mencionada interpretada como a cidade com coordenadas (N 42° 52' 49", W 8° 32' 44"), mas também *Galiza*, entidade geográfica de âmbito maior, presente de feito no texto em gentílicos:

t) "duas fustas em que hião sessenta Portugueses, de hũa das quais era capitão Diogo Soarez o **Galego**" (*PR, 204*)

Existe a possibilidade de anotar uma entidade dentro de outra entidade:

s2)"& leuamos tambem hum Bispo Abexim, que vinha para vir a este reyno, & daquy yr a <LUGAR>**Santiago de** <LUGAR>**Galiza**</LUGAR> </LUGAR>, & a <LUGAR> Roma </LUGAR>, & dahy a <LUGAR>Veneza</LUGAR>, para dahy se passar a <LUGAR>Ierusalem</LUGAR>." (*PR*, 5)

Usando um critério de simplificação optamos por processar a forma complexa como um todo e marcamos assim:

s3)"& leuamos tambem hum Bispo Abexim, que vinha para vir a este reyno, & daquy yr a <LUGAR>**Santiago de Galiza** </LUGAR>, & a <LUGAR> Roma </LUGAR>, & dahy a <LUGAR>Veneza</LUGAR>, para dahy se passar a <LUGAR>Ierusalem</LUGAR>." (*PR, 5*)

Característica identificadora: a entidade geográfica mencionada tem como referente aquele que cobre o conjunto da forma complexa independentemente de um dos seus componentes ser entidade geográfica mencionada independente.

5. Considerações finais

Nos exemplos analisados considerei as limitações surgidas ao aplicar uma lista que contém todas as entidades geográficas mencionadas no corpus num processo de anotação por coincidência de expressões. Achamos problemas que resultam da dificuldade para determinar se os *tokens* recuperados são EGMs ou não. A listagem não pretende ser exaustiva, mas simplesmente servir de mostra das ambiguidades mais comuns que aparecem ao tentar automatizar o processo de anotação. Embora o caso prático aqui estudado atende as EGM, é de esperar que problemas similares aconteçam com outro tipo de entidades, como ficou parcialmente ilustrado nos exemplos de ambiguidades relativas a nomes de pessoas.

Para cada dificuldade apontei como conclusão uma característica identificadora, sem que isto implique que seja necessariamente a única solução possível. Os critérios aplicados evidenciam a subjetividade inicial da anotação que, necessariamente, condicionará o desempenho de uma ferramenta mais avançada, particularmente quando esta não for concebida nem adaptada para os objetivos específicos do corpus.

As expectativas criadas na automatização da anotação de entidades mencionadas devem, portanto, considerar que uma parte importante do problema de automatização reside na definição prévia do que se deve ou não anotar. Uma mesma ferramenta, mesmo em supostos muito favoráveis, em que trabalha com uma lista com abrangência total das entidades contidas no texto, como é o caso aqui analisado, terá de resolver ambiguidades cuja solução será satisfatória em função dos critérios previamente definidos por quem vai operar com o produto final, o corpus anotado.

Os exemplos mostram também como as regras necessárias para a identificação de uma entidade como EGM chegam a ser de uma especificidade tal que resulta de difícil formulação em termos morfossintáticos: a solução passa mais facilmente por um critério semântico e os ativadores lexicais, de os querer utilizar, obrigam à consideração de regras muito específicas. Se se operar por treino, os casos são tão particulares que dificilmente têm relevância estatística (não há concordâncias suficientes). O trabalho da especialista é que determina o necessário equilíbrio entre a validação experta e a adequação e melhora da ferramenta de automatização, imperfeita, porém, suficientemente eficaz para fazer desnecessária uma boa parte do tedioso trabalho de anotação manual.

Agradecimentos

Os parágrafos da secção 4 deste artigo foram inicialmente redigidos como parte da tese de doutoramento defendida na Universidade de Santiago de Compostela intitulada *Entidades geográficas mencionadas. O caso da Peregrinação de Fernão Mendes Pinto*. O autor agradece a orientação e comentários dos professores Paulo Gamallo, Rubén Lois González e José António Souto durante o período de redação da tese. A ideia de dar forma de artigo surgiu pelo convite realizado pelo professor Xavier Varela, também da USC, para participar no congresso *Lingüística Histórica e Toponimia Galego-Portuguesa* celebrado em Santiago de Compostela, do 25 ao 26 de janeiro de 2018. A preparação do texto final em forma de artigo foi realizada

durante o período de trabalho no projeto de desenvolvimento de uma ferramenta para a anotação de topónimos em textos medievais no CITIUS da USC (Outubro 2017 – Fevereiro 2018) no marco da rede galega de investigação TECANDALI, ED341D R2016/011. Finalmente, o autor agradece os comentários e sugestões do professor Alberto Simões da Universidade do Minho e os pareceres recebidos no processo de avaliação para este volume da *Diacrítica* que contribuíram para uma nova versão, notavelmente acrescentada, do artigo.

Referências

Amaral, D. O., Fonseca, E. B., Lopes, L. & Vieira, R. (2014). Comparative Analysis of Portuguese Named Entities Recognition Tools. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik (pp. 2554–2558). European Language Resources Association (ELRA). Disponível em: <http://www.lrecconf.org/proceedings/lrec2014/pdf/513_Paper.pdf>.

Canosa, A. X. (2017). Algumas interseções disciplinares na recuperação da geografia da *Peregrinação* de Fernão Mendes Pinto. *Fluxos e Riscos*, 2(1).

Canosa, A. X., Varela, X., Lema, P., Gamallo, P., Taboada, J. A. & Garcia, M. (2018). Uma utilidade para o reconhecimento de topónimos em documentos medievais. *Linguamática*, *11*(1).

Gregory, I. N., Baron, A., Murrieta-Flores, P., Hardie, A. & Rayson, P. (2013). Geographical Text Analysis Mapping and spatially analysing corpora. In A. Hardie, & R. Love (Eds.), *Corpus Linguistics 2013 Abstracts* (pp. 105–108). UCREL. Disponível em: <http://ucrel.lancs.ac.uk/cl2013/doc/CL2013-ABSTRACT-BOOK.pdf>.

Gregory, I. N., Cooper, D. C., Hardie, A. & Rayson, P. (2015). Spatializing and Analyzing Digital Texts: Corpora, GIS, and Places. In D. J. Bodenhamer, J. Corrigan, and T. M. Harris (Eds.), *Deep Maps and Spatial Narratives*. Bloomington: Indiana University Press. Disponível em: <http://e-space.mmu.ac.uk/579357/2/Spatializing%20and%20Analyzing%20Digital%20Texts.pdf>.

Leidner, J. L. (2007). *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names* (Tese de PhD, University of Edinburgh,). Disponível em: <https://www.era.lib.ed.ac.uk/handle/1842/1849>.

Nadeau, D. & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes, 30*(1), 3–26. Disponível em: <http://nlp.cs.nyu.edu/sekine/papers/li07.pdf>.

Santos, D. & Cardoso, N. (Eds.). (2007). *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM a primeira avaliação conjunta na área*. Linguateca 2007. Disponível em: <http://comum.rcaap.pt/bitstream/10400.26/380/1/LivroSantosCardoso2007.pdf>.

Southall, H., Mostern, R. & Berman, M. L. (2011). On historical gazetteers. *International Journal of Humanities and Arts Computing*, *5*(2), 127–145.

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

Fontes e estudos para a lista de topónimos e corpus

Albuquerque, L. (Dir.). (1994). *Dicionário de História dos Descobrimentos Portugueses.* 2 vols. Lisboa: Caminho.

Alves, J. S. (Dir.). (2010). *Fernão Mendes Pinto and the Peregrinação.* 4 vols. Lisboa: Fundação Oriente.

Bluteau, R. C. R. (1712–28). *Vocabulario portuguez e latino, aulico, anatomico, architectonico, bellico, botanico, brasilico, comico, critico, chimico, dogmatico, dialectico, dendrologico, ecclesiastico, etymologico, economico, florifero, forense, fructifero...* Coimbra, Portugal: Collegio das Artes da Companhia de Jesus. Edição digital fac-similar: Biblioteca Nacional de Portugal. Disponível em: <http://purl.pt/13969>.

Flores, A. M., Gomes, R. V. & R. H. Pereira de Sousa. (1983). *Fernão Mendes Pinto. Subsídios para a sua Bio-Bibliografia.* [Almada]: Câmara Municipal da Almada.

Lagoa, V. (1950–53). *Glossário Toponímico da Antiga Historiografia Portuguesa Ultramarina.* 4 vols. Lisboa: Junta de Investigações Coloniais.

Pereira, B. (1647). *Thesouro da Lingoa Portugueza*. Lisboa: Paulo Craesbecck. Edição digital facsimilar: Biblioteca Nacional de Portugal. Disponível em: <http://purl.pt/29129>.

Pinto, F. M. (1614). *Peregrinaçam*. Lisboa: Pedro Crasbeek. Edição digital fac-similar: Biblioteca Nacional de Portugal. Disponível em: <http://purl.pt/82>.

[recebido em 7 de março de 2018 e aceite para publicação em 31 de outubro de 2018]