TESE DE DOUTORAMENTO

A IDENTIFICAÇÃO E REFERENCIAÇÃO DE ENTIDADES GEOGRÁFICAS MENCIONADAS

O caso da *Peregrinação* de Fernão Mendes Pinto

Alfonso Javier Canosa Rodríguez

DEPARTAMENTO DE XEOGRAFÍA
FACULTADE DE XEOGRAFÍA E HISTORIA

SANTIAGO DE COMPOSTELA
2017

TESE DE DOUTORAMENTO

A IDENTIFICAÇÃO E REFERENCIAÇÃO DE ENTIDADES GEOGRÁFICAS MENCIONADAS

O caso da *Peregrinação* de Fernão Mendes Pinto

Orientadores
Rubén C. Lois González

José A. Souto Cabo

Pablo Gamallo Otero

Autor
Alfonso Javier Canosa Rodríguez

DEPARTAMENTO DE XEOGRAFÍA
FACULTADE DE XEOGRAFÍA E HISTORIA

SANTIAGO DE COMPOSTELA
2017

D. Rubén C. Lois González, Profesor Catedrático do Departamento de Xeografia da Universidade de Santiago de Compostela

D. José Antonio Souto Cabo, Profesor Doutor do Departamento de Galego da Universidade de Santiago de Compostela

D. Pablo Gamallo, Profesor Doutor do Programa de Doutoramento de Linguística da Universidade de Santiago de Compostela

Como Directores da Tese de Doutoramento titulada

« A IDENTIFICAÇÃO E REFERENCIAÇÃO DE ENTIDADES GEOGRÁFICAS MENCIONADAS. O CASO DA *PEREGRINAÇÃO* DE FERNÃO MENDES PINTO»

Presentada por
D. Alfonso Javier Canosa Rodríguez

Alumno do Programa de Doutoramento

[E5061V01] Programa de Doutoramento en Historia, Xeografía e Historia da Arte

Autorizan a presentación da tese indicada, considerando que reúne os requisitos esixidos no artigo 34 do regulamento de Estudos de Doutoramento, e que como Director da mesma non incurre nas causas de abstención establecidas na lei 40/2015.

Orientador
Rubén C. Lois González

Orientador
D. José Antonio Souto Cabo

Orientador
Pablo Gamallo Otero

Doutorando
D. Alfonso Javier Canosa Rodríguez

“Quare recta inveniendi via est ex data aliqua definitione cogitationes formare”

Baruch de Spinoza. *Tractatus de Intellectus Emendatione*.

“For example, we know that the centre of mass of the Solar System at a definite instant is some definite point, and we can affirm a number of propositions about it; but we have no immediate acquaintance with this point, which is only known to us by description. The distinction between acquaintance and knowledge about is the distinction between the things we have presentations of, and the things we only reach by means of denoting phrases. It often happens that we know that a certain phrase denotes unambiguously, although we have no acquaintance with what it denotes; this occurs in the above case of the centre of mass.”

Betrand Russel. *On denoting*.

A Fernão Mendes Pinto

Aos primeiros habitantes de uma entidade geográfica fora da Terra

AGRADECIMENTOS

À minha familia que me apoiou para dedicar-me por completo a redigir este trabalho.
A todos os meus grandes mestres, muito particularmente D. Carlos, D. José, D. Jesus, D. Camilo, D. Víctor, Chelo, D. Jesus, Xan, Rosa, Pepe, Matilde, Manolo e Rubén que sentaram as bases da minha carreira.
A Nicolás Álvarez-Aldir, Pedro Dono e Víctor López Baliño, meus companheiros de estudo e de conversas sobre filologia e para além dela.
À professora Dona Carme Hermida, professor Francisco Fernández Rei, Dona Miruca Parga e pessoal investigador nos projectos de tratamento de textos para o Atlas Dialectológico da Língua Galega.
A Xan Garrido que me orientou para os estudos de pós-graduação.
Ao doutor Henrique Rodrigues Peres, companheiro de doutoramento, de conversas on-line e ao vivo sobre linguística teórica e aplicada.
Ao centro Mercator, na Universidade de Gales, em Aberystwyth, onde começou o caminho da investigação.
Ao professor François Soler que me iniciou no processo de escritura e articulação de uma tese.
Ao doutor D. André Pena e os professores D. Luis Monteagudo e Fernando Alonso Romero cujas conversas durante anos me orientaram no trabalho sobre a toponímia, as relações marítimas e a história.
A Hugo da Nóbrega Dias que reviu e comentou o primeiro capítulo desta tese, escrita do início ao fim em galego, para comprovarmos que vá também escrita no galego internacional chamado de português.
A Rodrigo Loureiro que reviu o texto em inglês do resumo deste trabalho e das duas publicações relacionadas mais recentes.
A D. José María de la Viña Varela que me explicou o conceito de entropia na física e conversou durante horas e tardes comigo sobre os princípios (com exemplos) do método nas ciências exatas.
Ao engenheiro Pablo Domínguez Froján que contrastou comigo a convenção para as primeiras formulações lógicas de relações e funções.
Ao doutor Miro Moman, pelas conversas sobre filosofia da ciência, língua e sociedade; não poucas vezes, estímulo para novas pesquisas.
Ao professor Marosi Bela, por uns quantos cafés e chás e outras tantas conversas onde comecei a desenvolver as métricas avaliativas mais frequentemente aplicadas neste trabalho.
Ao professor Chung Lee que me recomendou aplicar a entropia como medida avaliativa.
Ao pessoal e alunado da Mongolia International University em que durante dois anos tive a oportunidade de revisar e desenvolver os fundamentos de linguística teórica e boa parte da quantitativa.
Ao professor Xosé Luis Regueira, pela sua orientação nos rascunhos do plano de investigação da tese e comentários relativamente aos modelos iniciais de aproximação fonética aos topónimos que precederam o modelo semântico desenvolvido nesta tese.
Ao professor Hugo Gonçalo Oliveira que me atendeu e comentou a introdução à base lexical difusa do CLIP 2.1.
Ao professor Rubén Lois González, já citado de modo mais abreviado anteriormente, e os professores Pablo Gamallo e José Antonio Souto, orientadores e diretores desta tese, pelos seus atentos comentários, sugestões, revisões, conversas e apoio constante.

ABSTRACT

Geographical named entities represent one of the main types of named entities. A problem arises when a geographical named entity is identified in text but there are no given coordinates to provide a location. This thesis proposes a semantic model as a solution. Entities can be divided in two groups following an epistemologic criterion: those with known coordinates and those without. *Peregrinação*, an extensive report written by a diplomat travelling through Asia in the fifteenth century, is used as a case study. A list of geographical named entities is manually extracted and commented through comparative critical analysis of descriptions in corpus, those from related bibliography, and geovisualization of relevant areas in geographical databases and programs. This list is also used to evaluate automatic solutions for annotation and geo-referencing. Annotation is examined in three stages: tagging entities by matching expressions, optimization of results with a NERC tool and, finally, full automatization from scratch. For geo-referencing, entities with known coordinates are linked to an open global database from where geographical data is extracted and added to a local relational data-base. Relative references are solved for both known and unknown entities. The problem of assigning a geographical type is related to that of creating a taxonomy. For that purpose, extraction of geographical terms is evaluated, achieving best results by combining syntactic parsing, TF-IDF metrics and validation with external resources. A machine learning approach is explored to find examples of relations among entities and geographical features, results being significant for those entities with highest frequencies. Entities are organized in an onthology to refine their relations. An index is finally extracted to provide a structured definition of each entity, its occurrences in corpus, contemporary name and coordinates when available, and relations with other entities to further develop the relative reference.

Keywords: geographical named entities, georeferencing, NERC, semantic model, Peregrinação

RESUMO

As entidades geográficas mencionadas são uma das principais classes de entidades mencionadas. Um problema ocorre quando a entidade geográfica é identificada no texto, mas não há coordenadas para localizá-la. Esta tese propõe um modelo semântico como solução. As entidades são divididas em dois grupos segundo um critério epistemológico: aquelas que têm coordenadas conhecidas e as que não. *Peregrinação*, um extenso relatório escrito por um diplomata na Ásia no século dezasseis, serve de caso de estudo. Extrai-se manualmente uma lista de entidades geográficas mencionadas e comenta-se a partir da análise crítica e comparativa das descrições encontradas no corpus, a bibliografia relacionada e a geovisualização das áreas relevantes em bases de dados e programas geográficos. Esta lista é também usada para avaliar soluções automáticas de anotação e georreferenciação. A anotação é examinada em três fases: coincidência de expressões, otimização de resultados com uma ferramenta NERC e processo de automatização completo. Para a georreferenciação, as entidades com coordenadas conhecidas são procuradas numa base de dados aberta de âmbito global de onde se extraem dados geográficos que são adicionados a uma base de dados relacional local. As referências relativas são solucionadas para todas as entidades. O problema de atribuição do tipo geográfico liga-se ao de criação de uma taxonomia. Com esta

finalidade, avalia-se a extração automática de termos: a combinatória de análise sintática, medida TF-IDF e validação com fontes externas conseguiu os melhores resultados. Explora-se o aprendizado de máquina com exemplos na procura de relações entre entidades e tipos geográficos, com resultados significativos para aquelas entidades de frequências mais altas. As entidades são instanciadas numa ontologia para organizar as relações. Finalmente, extrai-se um índice com uma definição estruturada para cada entidade, as suas ocorrências no corpus, nome contemporâneo e coordenadas quando disponíveis e relações com outras entidades para mais desenvolver a referência relativa.

Palavras-chave: entidades geográficas mencionadas, georreferenciação, reconhecimento de entidades mencionadas, modelo semântico, Peregrinação

RESUMO

As entidades xeográficas mencionadas son unha das principais clases de entidades mencionadas. Un problema ocorre cando a entidade xeográfica é identificada no texto, mais non hai coordenadas para localizala. Esta tese propón un modelo semántico como solución. As entidades son divididas en dous grupos segundo um criterio epistemolóxico: aquelas que teñen coordenadas coñecidas e as que non. *Peregrinação*, um extenso relatorio escrito por um diplomático na Asia no século dezaseis, serve de caso de estudo. Extráese manualmente unha lista de entidades xeográficas mencionadas e coméntase a partir da análise crítica e comparativa das descricións encontradas no corpus, a bibliografía relacionada e a xeovisualización das áreas relevantes en bases de datos e programas xeográficos. Esta lista é tamén usada para avaliar solucións automáticas de anotación e xeorreferenciación. A anotación é examinada en tres fases: coincidencia de expresións, optimización de resultados con unha ferramenta NERC e proceso de automatización completo. Para a xeorreferenciación, as entidades con coordenadas coñecidas son procuradas nunha base de datos aberta de ámbito global de onde se extraen datos geográficos que son adicionados a unha base de datos relacional local. As referencias relativas son solucionadas para todas as entidades. O problema de atribución do tipo xeográfico lígase ao de creación dunha taxonomia. Con esta finalidade, avalíase a extracción automática de termos: a combinatoria de análise sintáctica, medida TF-IDF e validación con recursos externos conseguiu os mellores resultados. Explórase o aprendizado de máquina con exemplos na procura de relacións entre entidades e tipos xeográficos, con resultados significativos para aquelas entidades de frecuencias mais altas. As entidades son instanciadas nunha ontoloxía para organizar as relacións. Finalmente, extráese un índice con unha definición estructurada para cada entidade, as súas ocorrencias no corpus, nome contemporáneo e coordenadas cando dispoñíbeis, e relacións con outras entidades para mais desenvolver a referencia relativa.

Palabras chave: entidades xeográficas mencionadas, georreferenciación, recoñecemento de entidades mencionadas, modelo semántico, Peregrinação

ABREVIATURAS

Alt.	altitude
cap.	capítulo
c.	*circa*
E	Este
EM	Entidade Mencionada
ex.	exemplo
fig.	figura
fól	fólio
KML	*Keyhole Markup Language*
Lat.	latitude
Long.	longitude
N	Norte
NER	*Name Entitiy Recognition*
NERC	*Name Entitiy Recognition and Classification*
PLN	Processamento da Linguagem Natural
r.	recto
s.v.	*sub voce*
REM	Reconhecimento de Entidades Mencionadas
SIG	Sistema de Informação Geográfica
tab.	tabela
TF	Frequência absoluta (*Term Frequency*)
TF-IDF	*Term Frequency - Inverse Document Frequency*
v.	verso
vid.	*vide*
vs.	*versus*
XML	(*eXtensible Markup Language*)

Bibliográficas

BDN Biblioteca Digital Nacional

BNP Biblioteca Nacional de Portugal

DHDP *Dicionário de História dos Descobrimentos Portugueses* (Albuquerque, 1994)

FMP Fernão Mendes Pinto (Pinto, 1614)

FMPP *Fernão Mendes Pinto and the Peregrinação* (Alves, 2010)

GT *Glossário Toponímico da Antiga Historiografia Portuguesa Ultramarina* (Lagoa, 1950-53)

PR *Peregrinaçam* (Pinto, 1614). Quando citada com estas siglas, a referência à obra é para o capítulo.

Índice de conteúdo

4 A elaboração do corpus

5 A identificação das entidades geográficas mencionadas

Capítulo 1

Introdução

As entidades geográficas mencionadas são expressões que referem um objeto geográfico pelo nome próprio. Duas perguntas fundamentais surgem da sua definição. A primeira, relativamente à identificação (reconhecimento e anotação no texto): "quais são as expressões que mencionam uma entidade geográfica pelo nome próprio?". A segunda diz sobre o referente, a ligação da expressão com uma entidade geográfica física (georreferenciação): "onde fica, no globo, a entidade geográfica que estamos a mencionar?"

1.1 Descrição do problema

1.1.1 A identificação da entidade geográfica mencionada no texto

Dentro de um texto, um mesmo referente, a entidade geográfica, pode ser referido de múltiplas formas: através de frases descritivas, expressões anafóricas, deíticos e o nome próprio. Nesta sua última expressão é que tem sido atendida a identificação de entidades geográficas mencionadas nos últimos 20 anos como parte do problema NERC (*Name Entitiy Recognition and Classification*; o acrónimo REM, por Reconhecimento de Entidades Mencionadas, também usado em português). Trata-se do reconhecimento e classificação de entidades mencionadas, em que as entidades são identificadas quando referidas pelo nome próprio e classificadas dentro de um tipo, os mais comuns, organização, pessoa e lugar (são também possíveis outras classificações mais amplas e hierarquias de subtipos). Foram celebradas competições para avaliar o desempenho de ferramentas dedicadas a solucionar o problema NERC, nalguns casos com provas específicas para as entidades geográficas. Estes eventos serviram também para criar corpora anotados de referência, chamados coleções douradas, com que desenvolver e avaliar futuras aplicações. Os resultados destas ferramentas mudam segundo o problema particular a solucionar e as línguas de aplicação, mas achamo-los suficientemente representativos para considerarmos o seu uso no trabalho de anotação de um corpus. No entanto, várias dificuldades surgem à hora de pôr em funcionamento uma solução num caso prático. Em primeiro lugar, a disponibilidade das ferramentas avaliadas na literatura e testes de referência, simplesmente não disponíveis, ou o facto de apresentarem pouca ou nenhuma documentação para a sua configuração. Em segundo lugar, os resultados publicados podem vir condicionados pela adaptação das ferramentas às características dos corpora dourados. Em terceiro lugar, o corpus a anotar pode apresentar dificuldades e requisitos adicionais pela variedade de

língua, usos não padronizados e muito específicos, pelo qual a capacidade de adaptação e personalização é mais um fator para conseguir um desempenho adequado.

A primeira parte do problema abordado neste trabalho passa por avaliar os métodos para anotar (identificar) as entidades geográficas mencionadas num corpus histórico, em que boa parte das referências geográficas têm uma expressão não reconhecida nos atlas e listas geográficas e o texto em que aparecem adota uma variedade não padrão que dificulta o seu processamento.

1.1.2 A georreferenciação da entidade geográfica mencionada

O problema do georreferenciação consiste em determinar uma localização para uma expressão previamente identificada como entidade geográfica mencionada. O uso de listas de topónimos associados a coordenadas geográficas permite estabelecer esta ligação. Mas, mais uma vez, vários subproblemas dificultam o trabalho. Em primeiro lugar, a granularidade e abrangência das listas. Mesmo para uma base de dados muito completa, de âmbito global, toponímia local, microtoponímia, formas históricas, variantes menos usadas ou não padrão podem ficar de fora. O uso de listas específicas contribui para a solução, porém, num mesmo texto podem coocorrer formas de distintos níveis na escala, zonas geográficas e mesmo variantes de um mesmo topónimo. Outro problema surge com as formas homónimas, entidades geográficas distintas, mas com uma mesma forma na expressão. Achamos também casos de metonímia, em que um mesmo topónimo e umas coordenadas coincidentes representam mais de um objeto geográfico, por exemplo, uma cidade, rio, e ilha que coincidem num mesmo ponto no espaço (total ou parcialmente) com um mesmo nome. Nestes casos, a simples aplicação da lista, ainda que considere todos os objetos em questão, não é suficiente. Diversas técnicas como a configuração inicial de limites geográficos numa área fora da qual não se pesquisam os georreferentes, o cálculo de um centroide a partir das entidades geográficas mencionadas coocorrentes, ou análises do léxico como mais um elemento georreferenciador, recuperam as coordenadas mais prováveis dentro do grupo de candidatos homónimos.

Existe uma dificuldade ainda maior para a georreferenciação. O tratamento daquelas entidades desconhecidas, não integradas numa lista, para as quais não há coordenadas. Em trabalhos de georreferenciamento, se não há umas coordenadas para a entidade geográfica mencionada, não se considera a entidade como referenciada. No entanto, num texto encontramos elementos descritivos que georreferenciam uma entidade em princípio desconhecida, por vezes mesmo até o ponto de permitir a elaboração de coordenadas exatas. Estas descrições são sistematizáveis: uma entidade conhecida serve de ponto de referência a partir do qual temos indicadores tais como pertença, proximidade, distância, direção, similitude, com que obtemos uma georreferência relativa da entidade desconhecida.

O desenvolvimento de um método para o tratamento das entidades cujas coordenadas não são solucionáveis mediante as técnicas convencionais de georreferenciação é o principal contributo desta tese. A nossa proposta tenciona criar um modelo semântico em que um conceito estrutura as relações georreferenciadoras da entidade. Dada a variedade de relações com valor geoespacial,

centramos o modelo em apenas duas. A primeira classifica a entidade ao lhe outorgar um tipo geográfico. A segunda delimita a sua situação como parte de outra entidade geográfica.

1.2 Objetivos

Marcamos um objetivo geral para a solução do problema. Na aplicação de um caso prático consideramos também objetivos específicos relacionados com o procedimento e materiais necessários para resolver os subproblemas com que, em conjunto, avaliarmos a nossa proposta.

Geral

O nosso objetivo principal é definir um modelo para a georreferenciação das entidades geográficas mencionadas independentemente de que as coordenadas exatas sejam ou não conhecidas numa lista prévia.

Específicos

Os objetivos específicos são:

• Criar um índice de entidades geográficas mencionadas georreferenciadas para um texto de carácter histórico, não normalizado, com entidades de âmbito global ainda que com preponderância da Ásia. Boa parte destas entidades são desconhecidas mesmo em glossários especializados. Os tipos geográficos considerados superam os limites mais comuns, abrangendo entidades menores, tais como as construções, frequentemente nem incluídas na microtoponímia. Também no tratamento gramatical dos topónimos se vai para além do convencional, integrando no mesmo índice os gentílicos, formas derivadas do topónimo mas não nome próprio, como mais uma variante dentro da categoria das entidades geográficas mencionadas.

• Elaborar um corpus padrão dourado para o estudo das entidades geográficas mencionadas.

• Integrar o corpus anotado num ambiente de pesquisa para a recuperação de concordâncias.

• Avaliar soluções disponíveis e métodos de reconhecimento de entidades geográficas mencionadas para o aumento da qualidade de um corpus anotado e na anotação de um corpus sem anotar.

• Avaliar soluções para a extração de termos do domínio geográfico (os tipos que descrevem as entidades geográficas mencionadas) e relações semânticas das entidades.

• Configurar os resultados num sistema de informação geográfica que permita a visualização e representação cartográfica dos objetos georreferenciados por coordenadas e contribua para a solução das georreferências relativas.

1.3 Hipóteses

Consideramos uma hipótese principal, objeto central da tese, e outra secundária, relacionada com o procedimento metodológico com que abordamos a primeira hipótese.

Principal

Tanto as entidades geográficas cujas coordenadas são conhecidas quanto aquelas desconhecidas são georreferenciáveis de modo relativo num modelo semântico concetual que considera o tipo geográfico da entidade e as suas relações espaciais relativamente a outras entidades.

A relação espacial considerada nesta tese é a de pertença a outra entidade.

Secundária (de procedimento)

Técnicas do Processamento da Linguagem Natural e linguística de corpus sistematizam as entidades geográficas e extraem dados relevantes para a georreferenciação, facilitando e reduzindo o tempo de indexado.

1.4 Metodologia

A validez das hipóteses é conferida através de um caso prático, um corpus padrão sobre o qual se realizam trabalhos de identificação e georreferenciação de entidades geográficas mencionadas.

Na linha do objetivo geral, com os resultados obtidos pretendemos elaborar um índice de entidades geográficas mencionadas georreferenciadas segundo o modelo semântico proposto. Este objetivo específico define as pautas para a preparação dos materiais e a ordenação do trabalho.

1.4.1 Materiais

A produção dos materiais faz-se de modo ordenado, os primeiros servem de base para a composição de novos produtos mais elaborados. Com os materiais testamos e avaliamos hipóteses e modelos de procedimento. São reutilizáveis e supõem mais um contributo da tese.

Corpus padrão

Elaboramos um corpus com as entidades geográficas mencionadas anotadas com uma marca identificativa que agrupa todas as variantes de uma mesma entidade e as classifica segundo sejam topónimo ou gentílico.

Corpus paralelo

Como trabalho prévio a esta tese, criamos um corpus paralelo que contém capítulos do corpus padrão e a sua correspondente tradução para inglês com unidade temática em redor do espaço geográfico da Tartária. O corpus paralelo foi aplicado para testes no trabalho de identificação.

Base documental

Glossários geográficos e estudos específicos sobre as entidades geográficas anotadas. Boa parte das entidades são desconhecidas, portanto, foi preciso um trabalho amplo de documentação em obras quer genéricas sobre a temática do corpus, quer específicas para uma área ou topónimo.

Base de dados relacional

Os dados obtidos do corpus por meio de métodos da linguística de corpus e da base documental pela análise crítica e geovisualização dos objetos geográficos são integrados numa base de dados relacional. Esta base é consultada num painel de pesquisa para a recuperação de concordâncias enriquecidas com dados de frequências e georreferências. Serve também como material nos experimentos sobre a georreferenciação das entidades geográficas mencionadas no corpus.

1.4.2 Procedimento

A elaboração do índice de entidades geográficas define o procedimento numa direção linear. Começamos pelo processamento de dados de corpus e análise crítica de fontes e documentos. Integramos os dados em ambientes de trabalho para melhorar a operabilidade. Definimos testes com que confrontar soluções para os problemas de identificação e georreferenciação, conferimos resultados e, finalmente, extraímos o índice a partir dos dados obtidos conforme ao modelo proposto.

Preparação do corpus padrão

A partir de uma transcrição digital do texto original, procedemos à revisão e correção de erros conferindo com uma edição digital fac-similar do original. Anotamos as entidades geográficas mencionadas neste documento digital.

Preparação de um estudo crítico e geovisualização dos objetos geográficos

Analisamos criticamente uma base documental composta por glossários de geografia histórica, estudos das navegações e descobrimentos, geografia descritiva e histórica, e estudos específicos sobre a obra do corpus. Elaboramos um documento KML (*Keyhole Markup Language*), notação XML (*eXtensible Markup Language*) para a visualização de objetos geográficos em mapas digitais e navegadores 3 dimensões.

Base de dados relacional e painel de consulta

Os dados obtidos das análises do corpus, estudo crítico da base documental e geovisualização dos objetos são integrados numa base de dados e ligados às entidades geográficas mencionadas por um índice relacional. Para cada entidade citamos variantes, referências aos capítulos e páginas em que ocorrem, descrição da georreferência segundo o corpus, e análise comparada das distintas soluções para o georreferente nos glossários e estudos específicos. Desenvolvemos também um sistema de consulta para a recuperação de concordâncias e dados da base relacional.

Trabalho de identificação das entidades geográficas mencionadas no texto

Consideramos a disponibilidade e desempenho de soluções de anotação automática para o português e o inglês. A disponibilidade de uma ferramenta facilmente adaptável para o português fez com que a tenhamos usado na pesquisa de erros de uma anotação manual. A maior variedade, disponibilidade, assistência e documentação propiciou que, para uma anotação totalmente

automática, tenhamos testado o desempenho de sistemas de base estatística e de regras com listas sobre o componente em inglês do corpus paralelo.

Trabalho de georreferenciação

A georreferenciação aplica o modelo semântico da hipótese. Tem dois componentes, por sua vez segmentáveis em subtrabalhos. Por um lado, precisa solucionar um tipo geográfico classificatório da entidade, por outro, define uma georreferência que a localiza.

A partir da base documental e geovisualização do objeto obtemos as georreferências, por seu turno de dois tipos, absolutas (de coordenadas) e relativas (obtidas por descrição a partir do corpus).

Processamos num SIG as entidades georreferenciadas por coordenadas, representadas como pontos num mapa de vetores.

Realizamos testes para conferir o desempenho da extração automática de termos de domínio para a descrição das entidades relativamente a um tipo geográfico.

Integramos na base de dados relacional as relações para uma georreferência relativa obtidas do estudo crítico, usando uma convenção para facilitar o seu tratamento automático.

Análises e métricas de resultados

Em base ao componente prático da tese, os resultados dos procedimentos para a solução dos trabalhos de identificação e georreferenciação são analisados conforme a métodos quantitativos. Definimos variáveis discretas e operamos com métricas específicas.

Na análise do corpus, frequências absolutas e relativas contabilizam a variável central da tese. A frequência relativa serve como base empírica da probabilidade. Em casos pontuais, como medida de procedimento, a entropia avalia a desordem (incerteza) do âmbito em que procuramos a variável. Nos trabalhos de extração de entidades mencionadas e termos geográficos, medimos os desempenhos em base à precisão, abrangência e medida-F. Para lhe darmos maior coerência dentro da tese, os resultados de testes de aprendizado de máquina são também apresentados em função da abrangência e precisão.

Elaboração do índice

A combinatória do modelo semântico, PLN e análise de corpus resulta na obtenção de tabelas de frequências e dados de georreferências. Ordenamos todas as entidades geográficas mencionadas conforme a um índice alfabético obtido ao aplicar regras de ordenação e selecionamos os dados mais diretamente relacionados com a georreferência.

1.5 Contributos

A evolução dos sistemas e técnicas de identificação e referenciação de entidades mencionadas propiciou o desenvolvimento de coleções douradas a partir de textos normalizados e contemporâneos. Esta tese avalia a identificação e georreferenciamento sobre textos irregulares,

numa variedade de língua para a qual, em princípio, os sistemas automáticos não têm sido configurados.

Uma das dificuldades da aplicação de listas de entidades geográficas é a sua abrangência; listas muito globais consideram de preferência as entidades geográficas maiores, listas muito locais ficam sem abrangência fora da sua área. Os topónimos considerados no corpus cobrem âmbitos geográficos muito amplos, de local para global, da escala mínima das construções com nome próprio até às entidades da macrotoponímia como países, impérios e continentes. Para além dos topónimos, introduzimos um caso pouco estudado, os gentílicos, não nome próprio, mas morfologicamente mais uma variante do topónimo, anotados também como entidade geográfica mencionada. Ampliamos a abrangência da variável para maximizarmos as descrições georreferenciadoras.

O próprio corpus apresenta a dificuldade de não existir lista nenhuma que georreferencie por coordenadas a maior parte dos seus topónimos. Dado que não existe um referente conhecido, a georreferenciação tem de ser feita em termos relativos a uma outra entidade geográfica. Esta tese desenvolve um modelo de georreferenciamento em que, uma vez esgotadas as georreferenciações por coordenadas, consideramos uma georreferência relativa como solução alternativa.

O corpus anotado, índice de entidades geográficas, e dados obtidos do estudo individualizado de cada entidade foram integrados numa base de dados relacional que pode ser utilizada em pesquisas no âmbito da geografia histórica e, particularmente, no estudo da *Peregrinação* de Fernão Mendes Pinto (1614, 1653).

1.6 Estrutura da tese

Ordenamos a tese por capítulos cujo contributo para a solução do problema se relaciona a seguir.

O capítulo 1, *Introdução*, oferece uma primeira definição da variável objeto central deste trabalho, descreve o problema e subproblemas a resolver, os materiais e linhas metodológicas aplicadas para o trabalho sobre a variável e apresenta a estrutura da tese.

O capítulo 2, *Disciplinas, métodos e modelos para a abordagem das entidades geográficas mencionadas*, enquadra a variável dentro das disciplinas e subáreas mais relevantes e introduz o caso prático.

O capítulo 3, *Definições relativamente às entidades mencionadas*, define os termos chave mais importantes e estrutura-os segundo a sua procedência disciplinar.

O capítulo 4, *A elaboração do corpus*, descreve a metodologia e procedimento para a criação do corpus padrão, elabora as variáveis quantitativas de frequência e mostra dois exemplos práticos de aplicação: primeiro para a descrição do corpus no seu conjunto, depois para a caracterização geográfica dos capítulos da obra base do corpus.

O capítulo 5, *A identificação de entidades geográficas mencionadas*, apresenta três casos de

identificação automática no trabalho de anotação, com a dificuldade adicional de considerar os gentílicos como mais uma variante da entidade mencionada. No primeiro, analisamos em detalhe o procedimento empregado na anotação do corpus, a aplicação da lista de topónimos e gentílicos da base documental sob um critério de simples coincidência dos caracteres da expressão, e consideramos os problemas que surgiram e obrigaram à sua revisão. No segundo, introduzimos as métricas para a avaliação de resultados de extração de entidades e termos do corpus, usando como exemplo uma ferramenta NERC baseada em regras e configurada com a mesma lista de entidades geográficas mencionadas da elaboração do corpus padrão. Finalmente, conferimos os resultados de dois modelos de soluções NERC, estatísticas e de regras, fazendo uso de três ferramentas sobre o componente em inglês do corpus paralelo.

O capítulo 6, *A referenciação da entidade geográfica mencionada*, formaliza o modelo concetual a partir do qual pretendemos solucionar a georreferenciação das entidades geográficas mencionadas no corpus padrão e desenvolve os problemas derivados da relação entre a expressão da entidade e o seu referente.

No capítulo 7, *A georreferenciação por conhecimento prévio*, estabelecemos uma distinção epistemológica entre as entidades geográficas mencionadas, aquelas para as quais não existem discrepâncias na base documental nem contradição com o seu contexto no corpus, chamadas de conhecimento prévio, e as restantes, chamadas de conhecimento descrito enquanto limitamos a nossa georreferência à descrição obtida da análise do corpus. Ilustramos o processo de elaboração de uma lista com estudo crítico a partir da base documental e geovisualização dos objetos geográficos e, selecionadas as entidades resolvidas como conhecidas, classificamo-las dentro dos tipos geográficos no topo da taxonomia de uma ontologia geográfica. Finalmente, elaboramos uma base de dados para a sua aplicação num SIG.

O capítulo 8, *A georreferenciação por descrição*, descreve o georreferenciamento relativo das entidades geográficas mencionadas e formaliza as relações precisas dentro do modelo concetual da hipótese em que toda entidade tem um tipo geográfico (relação de hiponímia) e estabelece uma relação de pertença (*é_Parte_de*, relação de meronímia) com outra entidade. Como caso prático de aplicação do corpus no trabalho de georreferenciação por descrição, realizamos um exercício de extração de termos do domínio geográfico validados semanticamente por listas especializadas e termos associados numa base lexical.

O capítulo 9, *A definição do georreferente*, formaliza a escolha de uma expressão como representativa de todas as variantes de uma entidade geográfica mencionada. Aplicando a forma representativa como critério de ordenamento alfabético, extraímos um índice da base de dados relacional com todas as entidades georreferenciadas, quer por coordenadas, no caso de conhecimento prévio, quer relativas, nas de conhecimento descrito. Como caso prático usamos os dados do índice e a análise do corpus para a classificação das entidades segundo um tipo geográfico, e exploramos as relações semânticas do modelo usando como exemplos as entidades de maior frequência no corpus.

O capítulo 10, *Síntese de resultados, principais contributos e trabalho futuro*, examina os principais resultados e contributos, as suas limitações e possíveis aplicações, e marca linhas de trabalho consideradas prioritárias para o desenvolvimento futuro do modelo de georreferenciação proposto.

Capítulo 2

Disciplinas, métodos e modelos para a abordagem das entidades geográficas mencionadas

Neste capítulo apresentamos as disciplinas relacionadas com a variável central deste trabalho. Duas grandes áreas enquadram o objeto e a metodologia: a geografia modela a entidade geográfica mencionada entanto que expressão de um objeto físico e a linguística, auxiliada pelas disciplinas que atuam no campo do Processamento da Linguagem Natural (PLN), opera com o objeto textual. A base documental para a análise crítica das entidades geográficas do caso prático tem um carácter humanístico, sustentado em estudos filológicos, de geografia histórica e humana e históricos.

2.1 A geografia

O estudo das entidades geográficas é objeto prioritário da geografia, a disciplina que estuda o lugar e as relações entre lugares (Gregory & Ell, 2007) e procura a causalidade na maior ou menor proximidade dentro de uma escala que condiciona a interpretação (Knowles 2007). Neste trabalho tem mais frequentemente um carácter instrumental: proporciona as ferramentas para analisarmos os objetos enquanto sujeitos da geografia física. Por outro lado, iniciamos a análise crítica para resolvermos a georreferenciação a partir de trabalhos de geografia humana e histórica.

Dois componentes fundamentais dos dados geográficos modelam a variável estudada: o atributo classifica a entidade geográfica mencionada pela sua pertença a um tipo (físico ou administrativo), e o espaço, a georreferência propriamente, indica-nos onde se situa (Gregory & Ell, 2007).

Sistemas de Informação Geográfica

Para a ordenação e estudo das entidades geográficas mencionadas como dados geográficos fazemos uso de Sistemas de Informação Geográfica (SIG) (Gregory & Ell, 2007; Steinberg & Steinberg, 2015). A aplicação de SIG estende-se a todos âmbitos das humanidades, assim os estudos históricos, literários e linguísticos, em que a análise espacial é aplicada em pesquisas com um componente geográfico (Alves & Queiroz, 2015; Gregory, Donaldson, Murrieta-Flores, & Rayson, 2016). Quando a fonte selecionada como caso de estudo é uma fonte histórica, trabalhamos com um SIG histórico (Knowles 2007; Bol 2013). Exemplo particularmente relevante para este trabalho é o

China Historical GIS (Bol & Ge, 2005) pelo seu âmbito geográfico e aproximação metodológica: as entidades geográficas, independentemente da sua natureza geométrica (pontos, linhas ou polígonos), são definidas como pontos (Bol, 2007).

Cartografia e geovisualização

Na pesquisa do referente da entidade geográfica é preciso visualizar o objeto geográfico, explorar a sua relação com outros objetos e captá-lo dentro de distintos níveis de escala, em formatos que vão da simples representação de vetores de pontos a imagens tomadas de satélite. A geovisualização é assim compreendida como a aplicação da cartografia para a investigação do objeto, visualizado com distintos objetivos, a forma mais elaborada a produção de um mapa (MacEachren, 1995; Kraak, M. J., & Ormeling, 2010; Hennig, 2016).

Glossários e recursos geográficos

Produtos geográficos específicos são usados nas secções experimentais deste trabalho. Nos testes de identificação automática das entidades geográficas mencionadas (cap. 5) servimo-nos de listas de topónimos como componente das ferramentas NERC. Nos capítulos 7 e 8, glossários e trabalhos de geografia histórica formam o recurso mais habitual da base documental com que estudamos todas as entidades geográficas mencionadas. Consultamos bases de dados geográficos de domínio público para georreferenciarmos as entidades geográficas e usamos a taxonomia de GeoNames[1], com cobertura sobre toponímia mundial, para uma classificação inicial das entidades geográficas georreferenciadas. Para criarmos listas de termos geográficos do corpus (cap. 8) aplicamos tanto a taxonomia de GeoNames quanto o glossário de termos usados no mapeamento do Brasil (IBGE, 2015).

2.2 Linguística e Processamento da Linguagem Natural

A linguística estuda a linguagem humana estruturada em níveis atendidos pelas matérias centrais: fonética e fonologia (Martins, 1988), morfologia (Haspelmath & Sims, 2010), sintaxe (Radford, 1997; Valin, 2001; Haegeman, 2007) e a semântica com a incorporação da pragmática (Leech, 1981; Lyons, 1995; Riemer, 2010).

A área e método da linguística baseados na observação e uso de textos digitalizados é chamada *linguística de corpus* (Peres Rodrigues 2000; McEnery & Hardie, 2011). Orientada pela observação mais que pela experimentação (Wallis, 2014), vem justificada inicialmente porque o volume de dados estudado não é processável por métodos manuais ou requereria demasiado tempo e recursos se fosse processado por humanos. Exemplos de análise por corpus são as coocorrências de termos, conhecidas como concordâncias (análise qualitativa) e as frequências, a contagem de ocorrências dos termos (quantitativa) (McEnery & Hardie, 2011). Da análise de frequências avaliamos fenómenos linguísticos (lexicais e gramaticais), aplicando métodos estatísticos e probabilísticos (Bybee & Hopper, 2001).

[1] http://www.geonames.org

Neste trabalho, a linguística de corpus proporciona o método para preparar e realizar operações analíticas sobre as entidades geográficas mencionadas. A semântica tem particular relevância na conformação do modelo para o estudo das entidades geográficas, motivo pelo qual, das áreas centrais da linguística, é a de maior aplicação.

2.2.1 Semântica

A semântica é a disciplina que estuda o significado (Leech 1981; Lyons 1995; Riemer 2010; Wang 2016). Da perspetiva das unidades linguísticas, distinguimos dois grandes tipos de problemas semânticos: aqueles que atendem, de preferência, ao léxico, e a semântica das orações em que analisamos o valor proposicional e a conformação do significado do todo a partir das partes.

As distintas teorias do significado propiciam aproximações também diversas ao problema semântico. Para a elaboração deste trabalho combinamos propostas que atendem ao léxico (num modelo concetual) e as suas relações (num modelo distribucional). Importante para a compreensão das aplicações do modelo proposto, ainda que sem tanta relevância nos experimentos práticos aplicados, é o modo de as unidades se relacionarem para conformar unidades de significado maiores, ao nível da oração, num esquema composicional ligado a um modelo denotacional em que o referente é avaliado a respeito de um valor de verdade lógica.

Modelo concetual de base cognitiva

Tomamos como modelo as estruturas conceituais (Jackendoff 1990; 2010) em que os significados são codificações mentais na base do pensamento humano (à semelhança com outros animais como os primates). Estas estruturas não têm sempre fronteiras bem definidas. Uma proposta de formulação consiste na sua elaboração a partir de um número finito de primitivos com categorias como coisa, evento, estado, acção, lugar, direcção, propriedade e quantidade, e princípios combinatórios que permitem compreender um número ilimitado de significados. Nesta teoria, o referente da expressão linguística forma parte do constructo mental, é parte do conceito entanto concebido na mente humana e não objeto no mundo real.

Jackendoff elabora inicialmente a sua proposta dentro do modelo generativista (Chomsky 1965) pondo as estruturas conceituais em paralelo com as sintáticas e fonológicas, relacionadas por meio de regras, formando assim uma representação lógica da faculdade de linguagem. Nesta proposta estabelecemos uma relação entre a estrutura concetual e a representação espacial que chega através do sistema visual (Jackendoff 2010, pp. 85-137) e liga os conceitos, esquemas algébricos, com a representação espacial como uma codificação geométrica do espaço.

Neste trabalho usamos um modelo simplificado de conceito para a elaboração das definições das entidades geográficas mencionadas. Tiramos do modelo de estrutura concetual a construção de um conceito em base a uns primitivos e relações básicas, não obstante, o nosso constructo responde à necessidade de definição prática da georreferência, autónoma face a qualquer consideração de esquemas cognitivos inatos ou representação mental do objeto. Ainda mais, situamos sempre o referente como objeto físico externo, entidade material no mundo real independente da sua

concetualização.

Modelo distribucional

O modelo distribucional captura as propriedades semânticas de um termo através das coocorrências num corpus. Na hipótese distribucional, palavras que ocorrem num contexto similar tendem a ser semanticamente similares (Mitchell & Lapata, 2010; Baroni, Bernardi & Zamparelli, 2014). Num modelo de distribuição semântica quantificamos as coocorrências da expressão, os termos semanticamente próximos aparecem a menor distância. Resgatamos assim o modo em que o significado é expressado, a relação entre a expressão e o conceito independente do objeto físico.

A semântica distribucional apresenta modelos especialmente concebidos para operar com corpora e captura de relações semânticas. A hipótese distribucional subjaz nos procedimentos metodológicos para a extração de termos do domínio geográfico (cap. 8) e na captura de relações tais como a hiponímia e a meronímia (cap. 9) consideradas neste trabalho.

Modelo denotacional

Quando o significado vem determinado pela relação entre a expressão e o objeto denotado no mundo real (o referente da expressão) aplicamos uma teoria do significado denotacional (Lyons, 1995). Nestes modelos, o significado é formulado em termos lógicos (Cann, Kempson, & Gregoromichelaki, 2009).

A hipótese mais relevante nesta tese é a composicional. Propõe a elaboração de proposições cujo significado emerge a partir do valor de verdade que unidades menores têm dentro de unidades maiores, isto é, o significado da oração é composto como uma função das partes.

Neste trabalho usamos formalismos e concebemos regras e mecanismos de inferência para a georreferenciação segundo modelos da semântica formal, baseados na lógica, que capturam as entidades e predicados como relações legíveis em linguagem natural (Cann, Kempson, & Gregoromichelaki, 2009).

2.2.2 Processamento da Linguagem Natural

O problema da compreensão da linguagem escrita é uma área específica da Inteligência Artificial (Rich, Knight, & Nair, 2009; Russell & Norvig, 2010) conhecida como Processamento da Linguagem Natural (PLN). A aparição da Inteligência Artificial avança paralela à linguística desenvolvida por Chomsky (1978) com uma formalização da linguagem natural sustentada na lógica que a faz programável (Russell & Norvig, 2010, pp. 15-6). Existem pontos diversos de convergência entre disciplinas mais computacionais e aquelas mais específicas da linguagem humana. Assim, um campo comum em que a linguística, psicologia e informática trabalham conjuntamente para resolver este tipo de problemas é o da ciência cognitiva (Rich, Knight, & Nair; 2009, p. 20). Zapparoli (2010) situa a linguística informática numa área que abrange a linguística de corpus, a linguística computacional e o PLN como linha de investigação científica dentro dos quadros teóricos da linguística.

2.2.3 Lógica e formalismos na descrição da linguagem natural

A preocupação por superar a ambiguidade das linguagens naturais contribuiu para o desenvolvimento da linguagem simbólica e o uso de formalismos para a sua descrição (Robins, 1969; Sampson, 1980). A formulação de uma linguagem perfeita, sem ambiguidades, capaz de representar o pensamento racional de modo similar ao expressado pela linguagem natural, ultrapassando as suas ambiguidades e limitações, é trabalho acometido pela lógica (Russell 1905; Wittgenstein, 1922). Esta forma de linguagem mais elaborada é capaz de voltar sobre a própria linguagem natural como ferramenta para estruturar os distintos níveis linguísticos e as suas unidades, do fonema à oração, permitir a formulação de gramáticas e teorias, e explicar o conjunto de um sistema linguístico (Chomsky, 1978).

O aproveitamento de formalismos lógicos para a descrição da linguagem natural é particularmente relevante nos domínios específicos da sintaxe (Chomsky, 1965) e supõe a base da semântica formal, com métodos de análise da lógica proposicional e de predicado (Cann, Kempson, & Gregoromichelaki, 2009) também usados na Inteligência Artificial (Generseth & Nilsson, 1988; Nilsson, 1991; Russell & Norvig, 2010). O estudo da lógica é também relevante para o modelado da inferência das relações semânticas (Cann, Kempson, & Gregoromichelaki, 2009, pp. 17-21).

Neste trabalho usamos formalismos para a descrição do sistema com que modelamos a linguagem natural. Para a notação seguimos convenções da proposta introdutória de Restall (2006). Conforme avançamos na tese, a expressão de relações teve de usar mais nomes, motivo pelo qual adotamos um critério de expansão das formulas, definindo as relações mediante símbolos transparentes a partir da sua expressão em linguagem natural (Cann, Kepmson, & Gregoromichelaki, 2009; Russell & Norvig, 2010).

2.3 Geografia, linguística de corpus e análise de textos

A combinatória de técnicas da linguística de corpus e SIG aparece como uma nova área com aplicações nas humanidades, uma das suas utilidades, facilitar a visualização das geografias dum texto (Speriosu, Brown, Moon, Baldridge, & Erk, 2010; Gregory, Baron, Murrieta-Flores, Hardie, & Rayson, 2013; Speriosu & Baldridge, 2013; Alves & Queiroz, 2015; Purves & Derungs, 2015; Gregory, Cooper, Hardie, & Rayson, 2015; DeLozier, Baldridge, & London, 2015; DeLozier, Wing, Baldridge, & Nesbit, 2016). Num procedimento tipo de análise geográfica preparamos um corpus e anotamos as entidades geográficas mencionadas usando um sistema de reconhecimento automático. O resultado pode ser mais enriquecido, por exemplo processando termos chave ou mais entidades com que se queiram relacionar as localidades. Dos textos georreferenciados extraem-se dados para um SIG que facilita a representação cartográfica e geovisualização de entidades e as suas relações combinando os resultados provenientes das frequências e coocorrências no corpus com a análise espacial da geografia.

Como exemplo de análise geográfica de textos históricos preparados para a extração de dados e processamento num SIG, Gregory e Hardie (2011) usam um corpus reduzido de notícias de jornais

de Londres em 1653-4, anotado semanticamente, de onde obtêm colocações das entidades geográficas mencionadas para mais pesquisarem termos reveladores de uma temática, combinando os dados de corpus com a análise espacial. No âmbito da literatura portuguesa, Alves e Queiroz (2013) georreferenciam as entidades geográficas mencionadas em um corpus de 35 obras, publicadas entre 1854 e 2009, para estudar a evolução do espaço urbano de Lisboa. Bruggmann e Fabrikant (2014) identificam as entidades geográficas mencionadas num corpus histórico da Suiça por meio de uma lista específica e processam as frequências das suas coocorrências, assumindo representam uma relação semântica, para visualizar gráfica e cartograficamente a evolução histórica da relação entre topónimos. No contexto da geolocalização de um corpus literário do Lake district na Grande Bretanha (Gregory, Baron, Cooper, Hardie, Murrieta-Flores, & Rayson, 2014), Donaldson, Bushell e Gregory (2016) revisam aplicações para a criação de mapas digitais nas humanidades e expõem as limitações dos SIG no trabalho com textos literários.

Uma ferramenta que combina o processo de anotação e representação geográfica numa só aplicação é o Edinburgh Geoparser. Tem sido empregado para a georreferenciação de coleções de textos históricos (Grover, Tobin, Byrne, Woollard, Reid, Dunn, & Ball, 2010) e permite a integração de listas de topónimos pelo usuário. Alex, Byrne, Grover e Tobin (2015) mostram três exemplos do seu uso nas humanidades, no primeiro caso para o estudo do comércio no Império Britânico no século XIX, no segundo para a georreferenciação de textos clássicos, no terceiro para textos históricos locais. Nestes casos a resolução das entidades geográficas faz-se com listas de entidades geográficas especializadas, o Geoparser permite a sua identificação no corpus e visualização cartográfica. Os resultados podem ser posteriormente reelaborados por meio de técnicas de corpus e análise espacial (Gregory, Baron, Cooper, Hardie, Murrieta-Flores, & Rayson, 2014).

2.4 O caso da *Peregrinação* de Mendes Pinto

Este trabalho usa como caso prático um corpus elaborado a partir da edição de uma crónica de viagens, principalmente pela Ásia, publicada em Lisboa em 1614 (Pinto, 1614). Adicionamos como suplemento em forma de corpus paralelo os capítulos referidos à estadia com os Tártaros da primeira tradução em inglês (Pinto, 1653). O aparato crítico procedente da história e geografia histórica, combinado com uma base filológica, permite determinar o contexto e a compreensão de feitos, fazendo-os acessíveis na expressão (no caso de distância linguística), e na apreensão do seu significado (distância nos feitos históricos, sociais, peritos). Numa gradação que vai da simples compreensão até à capacidade de valoração de feitos, o aparato crítico humanístico facilita a interpretação crítica especializada ao fazer acessível a expressão, os feitos, os referentes e fatores que condicionam a extração e análise dos dados.

Estudos históricos e de geografia histórica formam a base documental com que analisamos criticamente todas as entidades mencionadas estudadas neste trabalho. A sua aplicação permitiu-nos elaborar um estudo crítico com que acometemos a georreferenciação das entidades geográficas. Aquelas com a maior probabilidade na atribuição do referente, identificáveis com um objeto geográfico na atualidade (admitindo um certo grau de variação na geografia humana,

particularmente na sua dependência administrativa, e uma relação de proximidade para o caso de urbes que mudaram o seu centro) formam a base que chamamos de conhecimento prévio. O aparato bibliográfico usado para a georreferenciação destas entidades aparece explicado no capítulo 7.

2.5 Conclusão

Neste capítulo apresentamos as disciplinas em que enquadramos o estudo das entidades geográficas mencionadas e a base documental de apoio para o caso prático. A geografia visualiza o objeto físico, a linguística modela a expressão no corpus. No caso prático é preciso ademais um trabalho de documentação fundamentalmente humanístico, procedente da filologia, a geografia e a história.

Capítulo 3

Definições relativamente às entidades geográficas mencionadas

Neste capítulo introduzimos termos chave da tese. Continuando a compartimentação disciplinar do capítulo anterior, em primeiro lugar consideramos a variável estudada como entidade física no espaço, posteriormente como expressão linguística na dimensão gramatical e objeto textual dentro de um corpus. Finalmente, definimos termos de modelos estatísticos e semânticos com que operaremos sobre as entidades geográficas mencionadas.

3.1 Entidades geográficas

3.1.1 Lugar

O conceito de lugar dentro da geografia abrange desde as noções mais físicas, numa relação topográfica com o espaço, até às visões mais holísticas em que o sentido vem dado como contexto de interação humana (Agnew, 2011). A noção de lugar mais frequentemente usada neste trabalho responde a um conceito abstrato de espaço aplicado com a finalidade de modelar uma taxonomia geográfica (Santos & Chaves, 2006; Ballatore 2016). A sua representação é análoga a um ponto na geometria euclidiana: unidade considerada como um todo não composto de partes. Lugar compreende assim qualquer entidade geográfica independentemente da extensão e posição dentro de uma classe ou subclasse. Tem a propriedade de ser escalável (Vasardani & Winter, 2016). Assim, *Pequim* é um lugar dentro da *China*, por sua vez um lugar dentro da *Ásia*, que ocupa um lugar dentro do planeta Terra. Quando uma entidade geográfica é um lugar ponto no espaço, pertence a uma unidade maior: o espaço que a contém.

Toda entidade geográfica tem a propriedade de ser um lugar (ponto no espaço), das mais amplas (o *Oceano Índico* é um lugar do planeta Terra dentro do Sistema Solar), até às unidades menores (o lugar de *Iacur* no *Reino do Bata*, na *ilha de Samatra*).

A determinação de uma entidade geográfica como lugar (ponto no espaço) é um problema de escala (por exemplo, uma cidade dentro de um país aparece como um ponto) ou de simplificação na definição de uma área (por exemplo, a definição de uma entidade geográfica administrativa a partir

da sua metrópole sem considerar os seus limites no espaço).

Uma segunda aceção de lugar, mais restrita e específica, aparece no corpus caso prático de estudo como tipo geográfico particular, unidade de povoação de dimensão menor do que uma vila, frequentemente a mínima habitada, composta por uma casa ou mais e pertencente a uma unidade administrativa maior. Como tal tipo é recolhido em descrições pontuais das entidades mencionadas e nos testes de capturas de termos do domínio geográfico.

Na taxonomia das entidades geográficas do corpus, *lugar* será usado como genérico, na primeira aceção aqui considerada, aplicado a qualquer entidade, for da natureza que for, quando não houver indicativo outro nenhum que a adscreva a um tipo particular.

3.1.2 Atributo e tipo geográfico

Um atributo geográfico é uma propriedade de uma ou mais entidades geográficas. Usamos o termo *atributo* quando nos referimos a uma propriedade da entidade geográfica mencionada. Quando nos referimos à terminologia usada na geografia e cartografia para descrever as entidades geográficas, usamos também a denominação *tipo geográfico*.

O atributo geográfico, dentro de uma ontologia ou nas relações estudadas nesta tese, refere inicialmente um ente abstrato, definidor de classe e não designador rígido. Por exemplo, uma *ilha* denota qualquer segmento de terra rodeado pelo mar. Neste trabalho o atributo geográfico é uma categoria que classifica as entidades geográficas.

Uma entidade geográfica ativa um atributo que a faz instanciar uma classe. Por exemplo, *Goa* é uma instância de *cidade*; *Miaygimaa* é, segundo o contexto, uma instância de *ilha*, de *povoação* ou de *templo*.

3.1.3 Representação cartográfica

Num SIG (Gregory & Ell, 2007) as entidades podem vir representadas em mapas de dois tipos: matriciais (*raster*) e vetoriais. Os primeiros quadriculam o espaço em tesselas regulares (pixels). Num modelo vetorial consideramos os elementos:

Ponto. Sem dimensão e correspondente com a noção abstrata de lugar. Neste trabalho, todas as entidades geográficas mencionadas georreferenciadas são representadas como um ponto numas coordenadas de latitude e longitude, mesmo aquelas que têm uma área definida.

Linha. As linhas unem pontos e representam objetos tais como rios e grandes construções, assim a Grande Muralha da China.

Polígono. Um polígono define uma área. Por exemplo, um país ou uma divisão administrativa. A dificuldade de definir áreas em SIG históricos, caso do *China Historical GIS* (Bol & Ge, 2005; Bol, 2007), também experimentada nesta tese, faz com que as representemos por um ponto que procura localizar a sua máxima unidade administrativa ou o centro aproximado da sua área.

A noção mais importante para esta tese é a de ponto. Todas as entidades, quando georreferenciadas em termos de um único par de coordenadas, são representadas como pontos.

3.2 Entidades mencionadas e nomes próprios

3.2.1 Entidades geográficas mencionadas

Usamos o termo *entidade mencionada* mais frequentemente para expressões de nomes de pessoas, organizações e lugares (Sang & Meulder 2003; Nadeau & Sekine, 2007), enquanto que nome de entidade. O termo pode provocar certa confusão, porquanto *entidade* propriamente designa o referente, no entanto, o seu uso mais comum na literatura NERC e mesmo em definições (Amaral 2013) aponta sem ambiguidade à expressão, isto é, à menção no texto. Assim, neste trabalho, se não houver mais nenhum modificador, *entidade mencionada* será sempre o objeto textual, o nome da entidade.

Entidade geográfica mencionada é aquela expressão que refere o tipo de entidade lugar, uma classe dentro das entidades mencionadas, caracterizada por apontar para uma *entidade geográfica*, objeto físico no espaço. O termo mais frequentemente usado nesta tese é *entidade geográfica mencionada*, contudo, também achamos *entidade geográfica nomeada* na literatura em português. Quando quisermos referir o objeto geográfico usaremos *entidade geográfica*, *referente* ou *objeto geográfico*.

A forma gramatical característica das entidades mencionadas em trabalhos de PLN é o nome próprio. Particularidade desta tese é considerar os gentílicos como variante de um nome próprio geográfico e, portanto, como mais uma expressão de entidade geográfica mencionada.

3.2.2 Nomes próprios

As entidades mencionadas pertencem à categoria gramatical de nomes próprios (exceto quantidades e datas). Uma definição semântica aparece gizada já em Barros (1540, fol. 5 recto) e segue a ser válida para os objetivos deste trabalho. Nomes próprios referidos a entidades geográficas e o seu atributo geográfico, nome comum, aparecem como exemplo de contraste: “Nome próprio, é aquelle que se nam póde atribuir a mais que a hũa só cousa: como este nome. Lisboa, por ser próprio desta cidade, e nam conuem a Roma: nẽ ô e Cesar a Cipiam, peró se dissermos cidade, que é geral nome a todas, em tam será comũ”.

Características semânticas

Da definição de Barros tiramos três características dos nomes próprios que também encontramos em exposições contemporâneas (Herbelot, 2015):

Unicidade : têm um único referente “ se nam póde atribuir a mais que a hũa só cousa” (Barros, 1540, fol. 5 recto).

Assim, uma entidade geográfica mencionada alude a um único referente ainda quando a expressão

seja transparente: *Ilha de Fogo* é uma ilha particular, não qualquer ilha onde haja fogo.

Instância: instanciam uma classe. Por exemplo, *Lisboa* é uma cidade.

Individualidade: apenas referem o indivíduo e não a classe, se não "será comũ". Em Mendes Pinto (1614), *Varella* como nome próprio refere um lugar concreto na costa da *Cauchenchina*, e apenas um. Porém, o nome comum *varella*, aplicado pelo mesmo autor para um tipo de templo, é qualquer um de muitos, sem distinção de lugar.

Características morfossintáticas

Também Barros (1540, fól. 10 r.) traz a característica morfológica principal dos nomes próprios: "Todo nome próprio tẽ singulár e nã plurár: assy como, Cípiam, Lísboa. Etc. Tiranse desta regra algũus nomes próprios que se declinam pelo plurár e nã tem singulár: como, Torres uedras, Torres nóuas, As pias, Alhos uedros, alfarelos, e outros desta calidade."

Número e concordância: os nomes próprios são singulares e, no caso de terem uma forma plural, concordam em singular com o verbo. Ex. *As Pias fica em Portugal*.

Características gráficas

Os nomes próprios têm **maiúscula inicial**. Esta propriedade é a base da regra mais comum e fator determinante nas aplicações de reconhecimento automático de entidades mencionadas.

3.2.3 Topónimos e gentílicos

3.2.3.1 Topónimo

Um topónimo é um nome próprio que refere uma entidade geográfica. A toponomástica ou toponímia estuda os topónimos de modo similar ao lexicon de um idioma, preocupada pela evolução da expressão no tempo e o significado do nome de lugar (Dick, 1975). Como topónimo, a entidade geográfica mencionada é um nome próprio que forma parte de uma língua.

A toponímia é também estudada da perspetiva da geografia física e humana (Gammeltoft, 2016). Critérios geográficos permitem classificar os topónimos. Uma primeira classificação, baseada na extensão do espaço abrangido, distingue entre topónimos maiores e menores. A microtoponímia atende às unidades mínimas cujo limite estaria por cima dos objetos arquitetónicos e engenharia civil como entes individuais. Neste trabalho, a classificação dos topónimos faz-se como entidade geográfica mencionada numa taxonomia que atende principalmente à espacialidade nas características físicas e administrativas do objeto geográfico. Deste modo, entidades que normalmente ficam fora da toponímia, tais as construções e obras de engenharia civil, serão também incorporadas quando tiverem as características dos nomes próprios (§3.2.2).

3.2.3.2 Gentílicos como variante de topónimo

Gentílico é aquele nome ou adjetivo que designa a origem ou procedência de um lugar. Diferenciamo-lo do topónimo e da definição canónica das entidades mencionadas por não partilhar as propriedades definitórias principais do nome próprio. Expomos as características que nos levaram, não obstante, à sua consideração como entidade geográfica mencionada nesta tese:

Tipografia. A nível tipográfico, a *Peregrinação* não distingue entre gentílico e nome de lugar e ambos são tratados do mesmo modo, capitalizando o carácter inicial. A forma *Achem* aparece assim como topónimo "yr tambem com elle ao Achem" (*PR*, 14), e como gentílico ou nome próprio "tyranno Achem" (*PR*, 14). Tanto a *China* (*PR*, 1), a designar um território, quanto os *Chins* (*PR*, 1), habitantes da China, são grafados com maiúscula. Por vezes o topónimo ou gentílico é também nome próprio de pessoa representante de um território, assim o *Reyno Bata* (*PR*, 15), gentílico ou topónimo cujo rei é *o Bata* (*PR*, 13), outras vezes chamado *Rey dos Batas* (*PR*, 15).

Morfologia. De um ponto de vista formal, o procedimento mais comum de formação de gentílicos é morfológico, o topónimo adiciona um sufixo. O topónimo propriamente fica, portanto, na raíz. Mesmo nos casos de maior distância morfológica (supleções fortes, quando se mantém o mesmo significado mas muda o radical), é frequente que a morfologia do gentílico responda a um topónimo anterior, alternativo, do mesmo referente geográfico.

Semântica. Ainda não partilhando as características semânticas que caracterizam o nome próprio (§3.2.2), o principal traço semântico do gentílico é comum com o topónimo: ambos designam um lugar ou área geográfica.

A coincidência no principal atributo identificativo das entidades mencionadas (maiúscula), a similitude formal (até o ponto de topónimo e gentílico poderem coincidir com idêntica expressão) e o facto de terem um mesmo núcleo semântico, permitem uma interpretação do gentílico como variante de um topónimo. O motivo principal deste proceder neste trabalho é, não obstante, o rendimento nas descrições das entidades geográficas mencionadas. O facto de indexar os gentílicos permite dispor de um maior número de ocorrências, ou mesmo a única, para a descrição da entidade geográfica.

3.3 A entidade geográfica mencionada no texto

3.3.1 Texto e corpus

Um *texto escrito* é uma unidade completa com um princípio e um fim (Lyons, 1995). Características do texto são a coesão e a coerência: o que se diz numa unidade é relevante para as unidades anteriores (Lyons, 1995). Uma coleção de textos (Kornai, 2008) ou simplesmente texto processado para proceder à sua análise (Wallis, 2014) chamamos de *corpus*.

Um corpus é, essencialmente, texto preparado para ser analisado; a extensão da linguística que centra o seu trabalho sobre ele é também chamada de corpus. Corpora podem ser usados com

finalidade prática: elaboração de dicionários, extração de terminologia, treino de sistemas de tradução automática, padrões anotados para medir a eficácia de ferramentas e soluções práticas de PLN são exemplos. Neste trabalho, o corpus é usado com esta finalidade prática mais do que para a pesquisa de evidências linguísticas.

3.3.2 Expressão, token, tipo

O termo *expressão* refere uma unidade linguística sem considerarmos o significado. A expressão de uma entidade geográfica no corpus é o conjunto de caracteres (espaço em branco incluído quando for multi-palavra) que a representam. Como sinónimo associado a expressão, usamos menos frequentemente o termo *forma*.

Quando queremos computar expressões presentes num corpus ou segmento de um corpus, empregamos o termo *token*, com valor de unidade de medida (Kornai, 2008; Haspelmath & Sims, 2010). Quando nos referimos a uma mesma expressão que aparece uma ou mais vezes no corpus (com um ou mais tokens), usamos *tipo*.

Exemplo:

> "Do que me aconteceo despois que me party deste reyno de Aarû" (*PR*,18) (3a)

Em (3a) temos 12 tokens, mas só 10 tipos, pois as expressões *que* e *de* ocorrem cada uma duas vezes.

3.3.3 Lexema e lema

A unidade que agrupa expressões similares, construídas com regras gramaticais simples, chamamos de *lexema* (Lyons, 1995, pp. 47, 50-1; Haspelmath & Sims, 2010). Por convenção, em lexicologia e morfologia, um lexema representa a variação da expressão ligada a uma mudança no significado gramatical ou processos morfológicos muito comuns. No trabalho com corpora anotados é comum também o uso do termo *lema* como conceito análogo (Garside, Leech, & McEnery, 1997). Um modo prático de captar a noção de lexema é a entrada de um dicionário: representa as variantes criadas por flexão (paradigmas gramaticais) e formas muito produtivas de derivação (caso dos diminutivos e aumentativos em português). Neste trabalho empregamos esta noção mais prática e abrangente de lexema, análoga a lema, para considerarmos todas as variantes de um nome de lugar (gráficas ou devidas a processos gramaticais desconhecidos pela sua condição de exotopónimos) junto com os seus gentílicos quando partilham o mesmo radical (supletivos fortes são considerados à parte como não pertencentes ao mesmo lexema).

3.3.4 Frequências, probabilidade e entropia

3.3.4.1 Frequência absoluta e relativa

Dado um corpus, a *frequência absoluta* de um tipo $w_i \in W$ = {tipos do corpus} é o número de tokens em que o tipo w_i ocorre (Kornai, 2008).

Dado um corpus com um número de tokens N, a *frequência relativa* de um tipo w_i é o número de tokens em que o tipo *w* ocorre divido pelo número de tokens do corpus N.

3.3.4.2 Probabilidade

A *probabilidade objetiva* (Bod, 2003) de um evento vem dada pela observação da frequência com que o evento sucede dentro de um número de resultados possíveis, o espaço amostral Ω.

$$\text{P(evento X)} = \frac{|evento\ X|}{|resultados\ possíveis\ \Omega|}$$

Ex. A probabilidade de uma entidade geográfica w_i ser um topónimo vem dada pelo espaço amostral Ω ={topónimo, gentílico}

$$\text{P}(w_i \text{ topónimo}) = \frac{|w_i = topónimo|}{|\Omega|}$$

A frequência relativa de um tipo w_i, num corpus suficientemente representativo, é uma estimação da sua probabilidade (Baayen, 2008) e expressa o grau de certeza com que, dadas umas condições similares, o tipo volveria ocorrer.

Neste trabalho usamos a probabilidade sobre dois tipos de dados. O mais comum, como base do cálculo da entropia (§3.3.4.3) em que obtemos a probabilidade pela observação empírica das frequências relativas das entidades no corpus. Num segundo caso, a probabilidade expressa a certeza que temos de que um objeto geográfico seja o referente de uma entidade mencionada segundo os dados da base documental. Nos resultados mostrados nesta tese, a probabilidade da georreferência vem dada pelo grau de divergência na base documental, Ω = { georreferencias distintas para a entidade geográfica mencionada w_i na base documental}.

3.3.4.3 Entropia

A entropia (Karmeshu, 2003) quantifica a incerteza num sistema probabilístico. Uma forma de a compreendermos é como sendo indicativo do grau de desordem, a média da incerteza sobre uma variável num sistema: quanta maior incerteza, maior desordem. A formulação empregada neste trabalho é a de Shannon:

$$\text{H(X)} = -\sum_{i=1}^{n} p_i \log p_i$$

Usamos a entropia em dois exercícios relacionados com os atributos geográficos (cap. 8). Numa taxonomia, a entropia representa a desordem produzida pelo número de tipos geográficos sobre as entidades mencionadas que os instanciam (§8.2.3.4). Num segundo caso, avaliamos o efeito da frequência absoluta sobre medidas de desempenho na recuperação de candidatos a termos do domínio. A entropia mede a maior incerteza achada na captura de candidatos dentro de um grupo de termos com uma mesma frequência numa distribuição de Zipf (§8.4.3.1).

3.4 Significado, referenciação e relações semânticas

3.4.1 Significado

A noção de significado tem diversidade de interpretações (Wang, 2016), um modo de apreender a variedade de aproximações é considerar distintos tipos: Leech (1981) distingue três, cinco Lyons (1995). Os mais relevantes nesta tese estão relacionados com os modelos semânticos utilizados (§2.2.1):

1. O concetual (Leech 1981), também chamado cognitivo ou denotativo, requer a assunção de estruturas mentais e categorias que permitem subcategorizações. Uma palavra pode assim ser definida por atributos que a estruturam e contrastam.

2. O associativo (Leech 1981). Em que o contexto da palavra determina o seu significado. Incluímos aqui a proposta distribucional (Baroni, Bernardi & Zamparelli, 2014), em que os atributos definitórios de um termo são calculados a partir de coocorrências por métodos estatísticos.

3. De verdade condicional e verificável (Lyons, 1995). Associado ao valor lógico de verdade das unidades que compõem a oração e da oração mesma como proposição.

Dentro das teorias do significado, encontramos propostas que focam num ou noutro tipo. Por um lado temos as de base *cognitiva* ou *representacionais*, o significado como sendo um constructo mental de base *concetual*; por outro, as teorias em que o significado se liga ao referente como objeto no mundo real (Cann, Kempson, & Gregormichelaki, 2009, pp. 9-11). Nos modelos *distribucionalistas*, o significado de um termo aparece mais limitado ao plano da expressão e as suas concordâncias. Dentro da diversidade de teorias, há pontos em comum e mesmo é possível a assunção de que não têm de ser propostas incompatíveis (Jackendoff 1990, p. 12), de facto, as aplicadas nesta tese são desenvolvidas como complementares.

3.4.2 Denotação e referenciação

Uma distinção fundamental para o estudo da relação entre uma expressão e o objeto que representam é a denotação do referente.

Uma expressão, entidade mencionada, que aponta para entidades (o referente) no mundo real, pode ter mais de um sentido e estabelecer, consequentemente com cada sentido, relações semânticas com outras expressões. Quando consideramos apenas o seu referente, estamos a focar a atenção (e levar o sentido) para o objeto físico.

Ex. *O juiz da Almada*, *o autor da Peregrinação* e *um mercador muito rico de Malaca*, são expressões tiradas de contextos diversos cujo referente é uma mesma pessoa no mundo real.

Neste trabalho distinguimos *denotação* para apontarmos o referente baseando-nos nalgumas características como ente físico, particularmente a quantificação dentro de um grupo. Nos exemplos anteriores, o mesmo referente é denotado de modo distinto, como um indivíduo particular (e só um) incluído num grupo de juízes, de escritores ou de mercadores em âmbitos geográficos que formam

grupos também distintos (na Almada, indefinido mas no âmbito da escrita em português, e em Malaca, respetivamente).

O termo *referenciação* aponta também para o referente, mas usamo-lo preferencialmente para instanciar, afirmar a existência, assinalar o objeto físico enquanto matéria. Assim, o referente em todos os exemplos anteriores é Mendes Pinto (c.1510-1583).

3.4.3 Relações semânticas

Expressões com significado aparecem ligadas por relações semânticas (Cann, Kempson, & Gregormichelaki, 2009). São relevantes nesta tese:

3.4.3.1 Sinonímia

A sinonímia implica uma relação de equivalência no significado. A equivalência entre sinónimos raramente é total, mais comummente encontramos uma similitude de termos equivalentes em determinados contextos.

Ex. *Povoação*, *lugarejo*, *lugar* e *povoado* são sinónimos, ainda que não em todos os contextos. *Povoação* pode ter o significado de "pequeno lugar povoado", mas também o de "espaço povoado numa área determinada" independentemente do seu tamanho.

Ex. *Metrópole* e *cidade* são sinónimos, ainda que tenham conteúdos semânticos diferentes.

Esta propriedade de equivalência dos sinónimos é aproveitada para agrupar termos de um domínio cuja diferença no significado apresenta apenas um matiz semântico pouco relevante. À hora de definir um atributo ou classe numa taxonomia, termos sinónimos, quando considerados muito próximos (relativamente a outros elementos do grupo), podem ser reduzidos a um mesmo tipo (§8.2.3.4).

3.4.3.2 Hiponímia e hiperonímia

O sentido de inclusão numa classe vem dado pela relação de *hiponímia*.

Ex. *cidade*, *vila*, *povoado* são hipónimos de *povoação* no seu sentido de "espaço habitado numa área determinada". Por seu turno, *povoação* é um hiperónimo de *cidade*, *vila* e *povoado*.

Os termos membros de uma mesma classe são *cohipónimos*.

Neste trabalho, consideramos que as *instâncias* (indivíduos particulares membros de uma classe, isto é, as entidades geográficas como objetos) têm uma relação de hiponímia com o tipo geográfico. Assim, *Pequim* e *cidade* entram em relação de hiponímia (ou hiperonímia segundo a direção da relação).

3.4.3.3 Holonímia e meronímia

A relação de pertença, *parte de*, chamamos de meronímia. O seu inverso é a holonímia.

Ex. *Esteiro*, *rio*. O *esteiro* é parte de um *rio*. *Rio* é holónimo e *esteiro* merónimo.

A sua formalização neste trabalho, *é_Parte_de*(*x*,*y*), expressa uma georreferência relativa do merónimo.

Ex. *é_Parte_de*(*Pequim*,*China*). *China* é holónimo e *Pequim* merónimo. *Pequim* é parte da *China*, portanto, *Pequim* fica situado dentro da área abrangida pela *China*.

3.5 Bases lexicais e ontologia

3.5.1 Ontologia

Uma ontologia ordena os objetos por meio de relações numa taxonomia. O termo apareceu na filosofia para se referir às entidades e às suas categorias e passou para o âmbito computacional (Poli & Obrst, 2010) como atividade concernida com a criação de modelos semânticos da realidade. Concebida uma concetualização como uma abstração do mundo que queremos representar, uma ontologia é uma especificação explícita da concetualização (Gruber, 1995) que formaliza um vocabulário comum e axiomas para o troco de conhecimento (Gruber, 1995; Buitelaar, Cimiano & Magnini, 2005). O conjunto de objetos representados por uma ontologia é o universo de discurso (Gruber, 1995). A ontologia permite classificar os objetos, estabelecer hierarquias (Jones & Paton, 1999) e realizar inferências e predições sobre os objetos classificados (Russell & Norvig, 2010; Rich, Knight & Nair, 2009).

3.5.2 Base de conhecimento lexical

Uma base de dados lexical organiza vocábulos segundo relações semânticas, de um modo similar a um tesouro lexicográfico, estruturado por relações tais como a sinonímia e antonímia, ou grupos de termos de um mesmo domínio. A uma base lexical organizada conforme a uma ontologia damos o nome de base de conhecimento lexical.

Uma base lexical ordenada segundo relações semânticas é a WordNet (Fellbaum, 1998). Serve de modelo para a criação de bases de conhecimento lexical em português (Felippo & Almeida, 2010; Gonçalo Oliveira & Gomes, 2010; 2014). Numa WordNet, os termos são classificados pela sua categoria gramatical e organizados em listas chamadas *synsets*. Os termos relacionam-se entre si por relações semânticas tais como a hiperonímia e meronímia para os substantivos.

3.6 Oração e proposição

A unidade mínima de segmentação de corpus neste trabalho é a *oração*. Como elemento do corpus, vem definida pela presença de um sinal gráfico, marca de pontuação. No nosso caso prático, por particularidades quer estilísticas, quer editoriais, tem uma longitude que a aproxima ao pseudoparágrafo. Conforme usada nesta tese, a oração tem também uma definição sintática: unidade completa que contém como mínimo um verbo. A maior parte das orações no caso de estudo são

compostas e complexas (mais de um verbo).

No tocante às unidades que compõem a oração, duas são usadas nesta tese. Chamamos de *cláusula* aquela oração contida dentro de outra oração, e *frase* qualquer unidade menor, com um núcleo categoria gramatical que lhe dá nome, conjunto com independência sintática dentro da oração.

Uma ou várias orações que expressarem um mesmo feito, examinável em termos lógicos de verdadeiro ou falso, serão também expressão de uma mesma *proposição*.

Quando o significado proposicional de duas orações é o mesmo dizemos que são paráfrases:

O_1 : Pequim é a capital da China.

O_2 : A capital da China é Pequim.

O_2 é uma paráfrase de O_1.

3.7 Conclusão

Neste capítulo definimos termos chave com que trabalharemos na exposição da tese. Três grandes áreas definem a entidade geográfica mencionada: Na geografia, como objeto físico no espaço, vem caracterizada por atributos que a classificam e permitem a sua representação cartográfica. Na linguística, como tipo gramatical com características gráficas, morfológicas e sintáticas que a singularizam. Na linguística de corpus, como expressão num texto, na sua dimensão qualitativa (tipos e lexemas) e quantitativa (frequência e probabilidade).

A entidade geográfica mencionada é modelada no significado lexical como conceito, em relação com outras entidades e termos do corpus. A organização dentro de uma taxonomia com relações e capacidade de inferência forma uma ontologia. Quando as relações são de tipo semântico, como num tesouro léxico, mas ordenadas conforme a uma ontologia, temos uma base de conhecimento lexical. Finalmente, a entidade geográfica mencionada aparece como mais um elemento da oração que, enquanto afirma factos sobre as entidades geográficas, é expressão de uma proposição.

Capítulo 4

A elaboração do corpus

Para estudarmos as entidades geográficas mencionadas devemos definir um enfoque metodológico, termos o material de estudo preparado para respondermos as questões de pesquisa e dispormos de ferramentas de análise que nos permitam chegar a resoluções das hipóteses iniciais. Nesta secção consideramos o processo de preparação do corpus, explicando as distintas etapas da sua elaboração até termos obtido um produto adaptável às necessidades das nossas pesquisas.

4.1 Aplicações do corpus

A preparação do corpus tem um objetivo material e outro metodológico-analítico:

- Material: Dispor de um mecanismo com que pesquisarmos seletivamente os contextos e estudarmos os topónimos para resolver a sua georreferenciação. O resultado final é um painel de consulta com múltiplas disposições do texto: por capítulos, por orações ou todos os fragmentos referidos a um topónimo.

- Metodológico-analítico: Extrair dados para formularmos hipóteses iniciais sobre as entidades geográficas mencionadas como fenómeno de corpus. O resultado final, nesta etapa de preparação, são tabelas de frequências para um estudo estatístico do texto.

As técnicas e trabalhos de elaboração material do corpus são atendidas por Habert, Nazarenko e Salem (1997) e Manning e Schutze (1999). Considerações mais específicas sobre os objetivos e métodos de corpora anotados são considerados por Nelson, Wallis e Aarts (2010, pp. 257-283) e Wallis (2007; 2014). O uso de corpora para o estudo do léxico é atendido por Ooi (1998), a emergência da estrutura linguística pela análise frequentista e probabilística tem uma boa introdução e casos de estudo em Bybee e Hopper (2001). A base estatística da análise de corpus, particularmente nas medidas de frequências, tem introduções em Muller (1970), Peres Rodrigues (2000) e Jurafsky, Bell e Gregory (2001).

4.2 Considerações metodológicas

O processo de elaboração de um corpus compreende o uso de um material básico de partida (o texto) e a elaboração de ferramentas que permitam o seu tratamento automático para chegarmos a um texto recuperável de tal modo que satisfaça uns objetivos de pesquisa definidos a priori, susceptíveis de serem reelaborados.

No nosso caso, o objetivo principal é o estudo das entidades geográficas mencionadas seguindo uma ordem lógica de anotação, abstração e análise (Wallis 2007). É um processo cíclico, de contínua reelaboração, que pode exigir novas formas de abstração. O texto anotado atualiza corpus; a análise conduz à revisão (adicionar novas marcas ou melhorar os algoritmos para extrairmos novos dados sobre o corpus anotado).

Podemos entender o corpus como a matéria prima numa cadeia de produção. Em função dos objetivos aplicaremos um ou outro procedimento e adicionaremos uns ou outros materiais complementares. Por exemplo, para procurarmos pontos de referência com que resolver as coordenadas geográficas de uma entidade geográfica mencionada, interessará recuperar apenas aqueles trechos que oferecerem contexto suficiente para situar o objeto geográfico em questão (a oração como unidade mínima e o capítulo como unidade temática mais ampla), e só aqueles que forem pertinentes (seleção para uma rápida focalização). Uma vez definidos os objetivos de uso, o texto é enriquecido com marcas (anotação das entidades) e elaboramos algoritmos para a automatização de processos e pesquisas com que obtermos os dados que serão objeto de análise.

4.3 O procedimento de anotação

4.3.1 Transcrição e revisão do texto

Partimos de uma transcrição do conjunto da primeira edição da *Peregrinação* (Pinto, 1614) adquirida a Edições Vercial[2]. O documento que nos foi enviado apresentou algumas dificuldades. Destacamos:

- A omissão de parágrafos correspondentes com uma página completa do original. Em total foram achados 4 fólios omitidos na transcrição que nos foi enviada. Ex. no capítulo 98 desaparece aproximadamente 1/3 do texto.
- A substituição do texto omitido por parágrafos doutros capítulos dificultou a deteção do erro e a sua correção. Por exemplo, o texto omitido no capítulo 98 é substituído por um texto do capítulo 97 de extensão aproximadamente similar e com temática comum (descrição de uma cidade num rio).

Uma vez transcritos e corrigidos os capítulos pertinentes, o resultado final é um documento de texto com o conjunto do texto da *Peregrinação* do capítulo 1 ao 226. A tabela 4.1. esquematiza os materiais de partida, técnicas e procedimentos para a obtenção do texto transcrito final sobre o qual realizamos a anotação.

[2] http://alfarrabio.di.uminho.pt/vercial/evercial/index.html

Material	Formato	Uso	Procedimento	Labor de corpus realizado
Peregrinaçam. Transcrição. (Ed. Vercial)	Documento digital (.odt, .txt, .pdf)	Texto base de partida	Editores de texto e pesquisas REGEX	Revisão da transcrição
Fernão Mendes Pinto and the Peregrinação. (Alves, 2010)	Documento impresso	Instrumento de comparação da qualidade do texto base	Leitura crítica	Revisão da transcrição
Peregrinaçam. fac-similar da BNP (Pinto, 1614)	Documento digital (.pdf)	Instrumento de comparação da qualidade do texto base	Leitura crítica	Revisão da transcrição
Peregrinaçam. Transcrição revista.	Documento digital (.txt, .odt)	Texto base para a anotação e análise	Editores de texto, linguagens de programação (Python e PHP), base de dados (MySQL)	Correção de erros de transcrição. Segmentação (capítulos e orações)

Tabela 4.1: Materiais, formatos e procedimentos para a obtenção do texto base do corpus.

4.3.2 Levantamento de entidades geográficas para a sua anotação

Fichas. Inventariamos manualmente as entidades a partir de repetidas leituras do texto em:

- Fac-similar do original da Biblioteca Digital da Biblioteca Nacional de Portugal.
- Lopes da Costa (Ed. e rev.). *Fernão Mendes Pinto and the Peregrinação. Restored Portuguese Text*. In Jorge Santos Alves (Ed.), *Fernão Mendes Pinto and the Peregrinação* (Vol. 2). Lisbon: Fundação Oriente. 2010.
- *Peregrinaçam*. Documento digital. Edições Vercial.

Listagem estruturada. A partir das fichas criamos uma lista em formato digital que recolhe:

- Uma entrada principal para cada entidade geográfica (equivalente ao lexema nos nomes comuns). Ex. *Cauchenchina*.
- As variantes com que aparece a entidade geográfica no texto. Ex. *Cauchenchina, Cauchim, Cauchins*.
- A categoria da entrada principal (nome próprio por defeito, gentílico se só aparece no texto como adjetivo ou nome comum).

Índice primário de lexemas. Guardamos a listagem numa base de dados e, segundo ordem alfabética, criamos um índice para cada entrada que servirá de referência para a sua recuperação.

Índice de variantes. A partir da lista de lexemas geramos uma nova tabela que indexa todas as variantes de cada entidade geográfica, quer gráficas, quer gramaticais, seguindo a ordem de

aparição na lista de lexemas. Uma vez criada a primeira, o algoritmo permite processar também as variantes para obter uma lista de lexemas. Isto é, os lexemas são recuperáveis a partir das variantes e vice-versa.

Anotação automática. Da lista de variantes o algoritmo anota cada uma das entidades geográficas mencionadas no conjunto do texto.

4.3.3 Preparação do corpus para a análise das georreferências

O texto anotado tem de ser processado para permitir a recuperação dos segmentos relevantes na pesquisa do contexto das entidades geográficas mencionadas. Duas unidades textuais foram consideradas: os capítulos e as orações.

Segmentação por capítulos. Com o texto anotado, criamos um ficheiro-texto para cada capítulo com o objeto de otimizar pesquisas e facilitar a indexação das entidades geográficas nomeadas.

Controle da anotação de entidades geográficas. Para cada variante indexada, comprovamos que aparece no texto e realizamos correções se é preciso.

- Detetamos variantes que não aparecem por leituras erradas, gralhas ou resultado de realizar o levantamento de distintas edições (por exemplo, diferenças na transcrição da nasalidade: *Iapão* aparece também no original como *Iapaõ*, não obstante, no levantamento das fichas, só anotáramos *Iapão*).
- Alguns termos foram anotados várias vezes por aparecerem como parte de uma forma complexa (ex. *Çambilão* e *Pullo Çambilão*).

Aproveitamos o controle para recolhermos dados sobre a frequência de variantes em cada capítulo.

Segmentação por orações. Escolhemos o ponto (.), a interrogação (?) e exclamação (!) como delimitadores oracionais depois de observarmos que o ponto e vírgula (;) segmenta cláusulas dentro de orações complexas que interessa conservar no seu conjunto para captar o significado completo (que perderíamos se omitirmos o verbo principal).

A segmentação assim obtida é ainda imperfeita, porquanto o uso do ponto (.) nem sempre indica o final da oração. Criamos exceções para:

- Abreviaturas, datas e quantidades em números que levam sistematicamente um ponto no texto. Ex. (…) *& vim do reyno na armada do Marichal no anno de 1513***.** *na nao S***.** *Ioaõ* (*PR*, 176).
- Eliminar todos os cabeçalhos com o título de capítulo em número romano do original, elemento de marca estrutural mais do que de conteúdo. Ex. *CAP. II*.

Controle da anotação por orações. Da listagem de orações contamos ocorrências de cada variante de modo análogo ao controle que tínhamos feito dos capítulos, isto é, procuramos evidência factual de cada variante. Se uma variante não tem pelo menos uma ocorrência, repetimos o procedimento de preparação e revisão. Deste modo controlam-se possíveis erros e localiza-se a correção:

- Na anotação das variantes: quando a variante não foi indexada no processo de marcado automático nem foi detetada no controle por capítulos.
- No algoritmo: quando o segmento em que aparece a variante não foi segmentado corretamente. Por exemplo, detetamos que quando um número aparece no final de oração

introduz mais uma marca de final de oração, portanto, procede uma anotação especial ou modificação do algoritmo para solucionar mais uma exceção às exceções (optamos por uma solução de anotação nestes casos).

4.4 Tabelas de frequências das entidades geográficas mencionadas

Os dados de frequências são processados por *scripts* num ambiente web, recuperáveis por consultas individuais sobre uma entidade em particular ou em forma de tabelas para o conjunto, em formatos variáveis segundo for requerido (base de dados, folha de cálculo, documento texto). Os *scripts* processam as unidades anotadas quer lematizadas, quer como variantes, atendendo aos distintos modos de segmentação do corpus (conjunto, capítulos e orações).

4.4.1 Variantes

Por capítulos. Observamos cada variante através da segmentação por capítulos. Cada capítulo é uma observação. Somamos o total de ocorrências para cada variante.

$$f_{cap}(v)= \sum_{i=1}^{c} v_i \quad \text{sendo } v \text{ a variante, } c \text{ o número de capítulos.} \tag{4.1}$$

Por orações. Observamos cada variante através da segmentação por orações. Cada oração é uma observação. Somamos o total de ocorrências para cada variante.

$$f_{or}(v)= \sum_{i=1}^{o} v_i \quad \text{sendo } v \text{ a variante, } o \text{ o número de orações.} \tag{4.2}$$

Controle. Sendo o conjunto de todos capítulos $C=\{c_1, c_2, \ldots, c_n\}$ o volume total do texto, e sendo o conjunto de todas as orações $O=\{ o_1, o_2, \ldots, o_k\}$ também o volume total do texto, para qualquer variante *v* a soma do total das suas frequências tem de ser igual for esta resultado de (4.1) ou de (4.2). O único que muda é o número de observações, resultado de distinto número de segmentações $c \neq o$. Isto é,

$$f_{abs}(v)= \sum_{i=1}^{c} v_i = \sum_{i=1}^{o} v_i \quad = \text{frequência absoluta da variante v no corpus.} \tag{4.3}$$

4.4.2 Lexemas

Por soma de variantes. Agrupamos as variantes que correspondam a um mesmo lexema. Cada variante é uma observação. Somamos o total de ocorrências de cada variante para um mesmo lexema.

$$f_{var}(l)= \sum_{i=1}^{v} f_i \tag{4.4}$$

onde *v* é o número de variantes que concorrem num lexema e f_i a frequência de cada variante. Como mecanismo de controle aplicamos a fórmula tanto para as frequências obtidas em (4.1) como as de (4.2), isto é, aguardamos o mesmo resultado $f_{var}(l)$ para f_i obtido de (4.1) do que f_i obtido de (4.2).

Por lematização. Na listagem de lexemas, procuramos as variantes correspondentes a cada lexema, assim como as suas ocorrências, isto é, contaremos as ocorrências de cada lexema independentemente de qual for a variante.

Em capítulos. Cada capítulo é uma observação. Estabelecemos mais um controle comparando com a frequência absoluta obtida anteriormente na listagem de variantes em (4.1).

$$f_{cap}(l) = \sum_{i=1}^{c} l_i \quad \text{sendo } l \text{ o lexema, } c \text{ o número de capítulos.} \qquad (4.5)$$

Em orações. Cada oração é uma observação. Estabelecemos um controle comparando com a frequência absoluta obtida anteriormente da listagem de variantes em (4.2).

$$f_{or}(l) = \sum_{i=1}^{o} l_i \quad \text{sendo } l \text{ o lexema, } o \text{ o número de orações.} \qquad (4.6)$$

Controle. De modo análogo a (4.3) temos:

$$f_{abs}(l) = \sum_{i=1}^{c} l_i = \sum_{i=1}^{o} l_i = \text{frequência absoluta do lexema } l \text{ no corpus que, aliás, já}$$

fora obtida por (4.4), podendo efetuar um controle adicional:

$$f_{abs}(l) = \sum_{i=1}^{c} l_i = \sum_{i=1}^{o} l_i = \sum_{i=1}^{v} f_i \qquad (4.7)$$

4.4.3 Controle final das tabelas de frequências absolutas

Para o conjunto de todas as georreferências anotadas no corpus, a sua frequência absoluta virá dada pela soma de todas as frequências de lexemas:

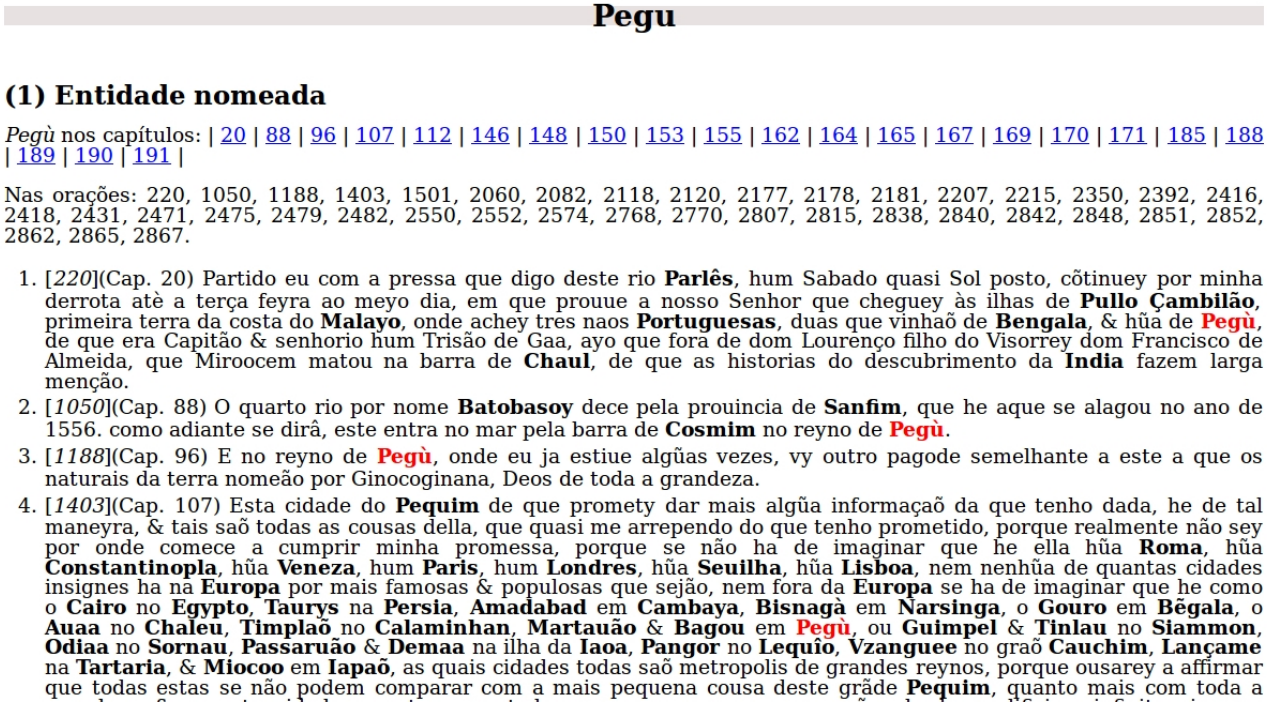
$$f_{abs}(\text{anotações}) = \sum_{i=1}^{l} f_i \quad \text{sendo } l \text{ o número de lexemas e } f_i \text{ a frequência de cada lexema.} \qquad (4.8)$$

Consequentemente, o total da soma das observações de cada tabela de frequências obtidas até o momento, for por variantes em capítulos, variantes em orações, lexemas por soma de variantes, ou lexemas por lematização em capítulos ou em orações, terá de oferecer um resultado final = f_{abs}(anotações).

4.5 Resultados

4.5.1 Concordâncias

Ao pesquisarmos uma entidade geográfica mencionada obtemos todos os seus contextos por capítulos e orações. No ambiente de consulta do corpus podemos ademais obter o texto íntegro da primeira edição da *Peregrinação* com as entidades geográficas mencionadas destacadas. A figura 4.1 mostra as primeiras concordâncias recuperadas para a entidade geográfica mencionada *Pegù.* A expressão aparece salientada no texto de modo distinto às outras entidades geográficas mencionadas, também destacadas.



Pegu

(1) Entidade nomeada

Pegù nos capítulos: | 20 | 88 | 96 | 107 | 112 | 146 | 148 | 150 | 153 | 155 | 162 | 164 | 165 | 167 | 169 | 170 | 171 | 185 | 188 | 189 | 190 | 191 |

Nas orações: 220, 1050, 1188, 1403, 1501, 2060, 2082, 2118, 2120, 2177, 2178, 2181, 2207, 2215, 2350, 2392, 2416, 2418, 2431, 2471, 2475, 2479, 2482, 2550, 2552, 2574, 2768, 2770, 2807, 2815, 2838, 2840, 2842, 2848, 2851, 2852, 2862, 2865, 2867.

1. [*220*](Cap. 20) Partido eu com a pressa que digo deste rio **Parlês**, hum Sabado quasi Sol posto, cõtinuey por minha derrota atè a terça feyra ao meyo dia, em que prouue a nosso Senhor que cheguey às ilhas de **Pullo Çambilão**, primeira terra da costa do **Malayo**, onde achey tres naos **Portuguesas**, duas que vinhaõ de **Bengala**, & hũa de **Pegù**, de que era Capitão & senhorio hum Trisão de Gaa, ayo que fora de dom Lourenço filho do Visorrey dom Francisco de Almeida, que Miroocem matou na barra de **Chaul**, de que as historias do descubrimento da **India** fazem larga menção.
2. [*1050*](Cap. 88) O quarto rio por nome **Batobasoy** dece pela prouincia de **Sanfim**, que he aque se alagou no ano de 1556. como adiante se dirâ, este entra no mar pela barra de **Cosmim** no reyno de **Pegù**.
3. [*1188*](Cap. 96) E no reyno de **Pegù**, onde eu ja estiue algũas vezes, vy outro pagode semelhante a este a que os naturais da terra nomeão por Ginocoginana, Deos de toda a grandeza.
4. [*1403*](Cap. 107) Esta cidade do **Pequim** de que promety dar mais algũa informaçaõ da que tenho dada, he de tal maneyra, & tais saõ todas as cousas della, que quasi me arrependo do que tenho prometido, porque realmente não sey por onde comece a cumprir minha promessa, porque se não ha de imaginar que he ella hũa **Roma**, hũa **Constantinopla**, hũa **Veneza**, hum **Paris**, hum **Londres**, hũa **Seuilha**, hũa **Lisboa**, nem nenhũa de quantas cidades insignes ha na **Europa** por mais famosas & populosas que sejão, nem fora da **Europa** se ha de imaginar que he como o **Cairo** no **Egypto**, **Taurys** na **Persia**, **Amadabad** em **Cambaya**, **Bisnagà** em **Narsinga**, o **Gouro** em **Bẽgala**, o **Auaa** no **Chaleu**, **Timplaõ** no **Calaminhan**, **Martauão** & **Bagou** em **Pegù**, ou **Guimpel** & **Tinlau** no **Siammon**, **Odiaa** no **Sornau**, **Passaruão** & **Demaa** na ilha da **Iaoa**, **Pangor** no **Lequîo**, **Vzanguee** no graõ **Cauchim**, **Lançame** na **Tartaria**, & **Miocoo** em **Iapaõ**, as quais cidades todas saõ metropolis de grandes reynos, porque ousarey a affirmar que todas estas se não podem comparar com a mais pequena cousa deste grãde **Pequim**, quanto mais com toda a

Figura 4.1: Consulta da entidade geográfica mencionada *Pegù* sobre o corpus anotado. Resultados das quatro primeiras concordâncias para a variante *Pegù*.

4.5.2 Tabelas de frequências

Obtidas as frequências absolutas para lexemas e variantes, por capítulos e orações, processamos os resultados numa base de dados. O cálculo de frequências absolutas tem dupla utilidade:

Deteção de erros no corpus. Ao conferirmos os resultados de frequências de entidades segundo modos de segmentação, soma de variantes e agrupamento em lexemas, dentro das combinatórias consideradas na secção anterior, os casos de discordâncias nos resultados são indício de um erro, quer no tratamento do corpus, quer na anotação. Por exemplo, no capítulo 132, a abreviatura de capítulo do cabeçalho não vai seguida de um ponto (em todos os demais 225 capítulos sim), motivo pelo qual se produziu um erro de segmentação ao omitir a primeira oração do capítulo, em que aparecem as georreferências *Huzamguee*, *Tanixumaa* e *Japão*. Foi o próprio algoritmo, ao comparar as frequências obtidas nos capítulos, que advertiu da presença de um erro até então não identificado.

Elaboração de conjeturas e hipóteses iniciais. Começamos a formular perguntas para o estudo das entidades geográficas mencionadas que queremos responder com o corpus:

1. A distribuição das frequências absolutas permite estabelecer comparações para tentar um primeiro modelo aproximativo que as descreva. Por exemplo, descreve o corpus a primeira lei de Zipf?
2. A observação das ocorrências oferece uma unidade em que procurar preferências de agrupamentos, clusters, distribuições semânticas, probabilidades condicionais. Oferecem as entidades geográficas alguma evidência de interação?

3. A ordenação dos lexemas por ordem de frequência outorga uma escala de preferência geográfica, já que as entidades com uma frequência mais alta oferecem um maior índice de georreferenciamento para as entidades geográficas desconhecidas. São os dados de frequências extraídos por análise de corpus já um indício para a georreferenciação?

4.6 Unidades e modos iniciais de análise do corpus

O objetivo principal da preparação e anotação do corpus é permitir a sua funcionalidade como ferramenta de análise. Tanto na recuperação de concordâncias quanto em trabalhos de processamento da linguagem natural operamos com segmentos textuais que servem de unidades para a análise quantitativa.

4.6.1 Unidades

Uma unidade é um segmento do corpus: da unidade menor, o carácter, até aos capítulos que conformam a unidade maior, o conjunto do corpus. As unidades que representam mais diretamente as entidades geográficas mencionadas são objeto de uma classificação mais específica: tokens e tipos (§3.3.2) e lexemas e variantes (§3.3.3).

Caracteres. Na contagem de caracteres não consideramos os signos de pontuação exceto o espaço em branco. Os dígitos são tratados como um carácter.

Tokens e tipos. A unidade principal para a análise do conjunto do corpus é o token, entendido como conjunto de caracteres delimitados por um espaço em branco ou signo de pontuação. Os tokens com a mesma forma são agrupados como pertencentes a um mesmo tipo. Dígitos entre espaços em branco ou signos de pontuação são mais um token.

Quando analisamos o corpus em termos de tokens, fazemo-lo de forma homogénea, sem considerarmos a anotação de um conjunto de tokens como uma entidade geográfica, nem a lematização das variantes das entidades geográficas. O corpus é analisado prescindindo da anotação.

Lexemas e variantes. As variantes são tokens anotados como entidades geográficas mencionadas. Quando lematizadas, recuperamos lexemas, que podem representar mais de uma variante (análogo ao token e os tipos). Se uma variante tiver uma expressão multipalavra, a anotação capturará os tokens como uma única unidade.

Orações. A oração é a unidade subsegmento do capítulo. Dentro dela computamos as unidades menores (tokens, tipos, variantes e lexemas). No caso de estudo, pelas suas características estilísticas e editoriais, fica próxima do pseudo-parágrafo como unidade textual.

Capítulos. O capítulo consta de cabeçalho e corpo. Nas análises consideradas neste capítulo, ambos são considerados do mesmo modo, segmentados em orações sem mais distinção. Os índices numéricos em romanos, precedidos da abreviação CAP (Capítulo) foram suprimidos.

Conjunto do corpus. Nas análises de conjunto observamos o texto ordenado sem distinção de

capítulos ou orações.

4.6.2 Modos do corpus

Distinguimos dois modos de análise, em função da consideração ou não das anotações.

Modo texto simples. Neste tipo de análise prescindimos das anotações. Respondemos a perguntas genéricas sobre o corpus e as suas propriedades como objeto textual. Ao não haver marcas que distingam as entidades geográficas mencionadas, capturamos as suas propriedades enquanto tokens e tipos. No modo texto simples, variantes gráficas e flexões são inicialmente processadas como tipos únicos, quer isto dizer, computam como unidade independente na observação da frequência.

Modo anotado. Consideramos as anotações para a identificação das entidades geográficas mencionadas, capturadas como lematizadas (lexemas) ou sem lematizar (variantes), dentro do conjunto do corpus, num capítulo ou numa oração.

O modo do corpus afeta a como se captura a entidade geográfica mencionada. Com o corpus anotado as variantes (e lexemas se as recuperarmos lematizadas) podem estar compostas de mais de um token quando sejam multipalavra.

Captura	Tokens	Modo	Identificado como entidade geográfica?
Santiago de Galiza	3	Texto simples	Não
Santiago_de_Galiza	1	Anotado	Sim
Pulo Çambilão	2	Texto simples	Não
Pulo_Çambilão	1	Anotado	Sim
Bengala	1	Texto simples	Não
Bengala	1	Anotado	Sim

Tabela 4.2: Exemplo de capturas de entidades geográficas mencionadas segundo o modo de análise do corpus

4.7 Exemplo de análise em modo texto simples

Uma vez elaborado o corpus interessa conhecer a sua representatividade, valorarmos o seu comportamento como objeto susceptível de ser analisado e operado com os métodos previstos (cap. 2).

4.7.1 O corpus como objeto da linguagem natural

Uma primeira questão que podemos colocar é sobre a natureza dos dados. É o corpus representativo da linguagem natural? Uma das primeiras formulações que apareceram ao estudar a língua aplicando métodos estatísticos é a conhecida como lei de Zipf.

4.7.1.1 A primeira lei de Zipf

Ao estudar a relação entre a intensidade de um fenómeno linguístico e a frequência com que este aparece, usando exemplos de base quantitativa contínua, mas enunciados como fatores discretos, Zipf (1929) achou uma relação inversamente proporcional entre a intensidade que mostra o fenómeno e a frequência com que acontece num corpus representativo da língua. Expressamos uma formulação desta relação como:

$$Y = n / X \tag{4.9}$$

onde Y é o fenómeno estudado, X a frequência com que aparece no corpus e *n* uma constante (não desenvolvida, possivelmente um coeficiente da extensão e duração). A representação gráfica proposta por Zipf para esta relação é uma hipérbole, apresentada de modo ilustrativo sem dados reais. A proposta foi posteriormente desenvolvida e chamada de primeira lei de Zipf. Descreve a relação entre os tipos (formas únicas) observados numa mostra suficientemente representativa de uma língua, ordenados de maior a menor frequência, e o valor da sua ocorrência no corpus. A relação é aceite como empírica (Tullo, & Hurford, 2003; Yan & Minnhagen, 2015), não obstante, na sua formulação inicial, é apenas uma proposta de modelo teórico que serve de aproximação a observações frequentísticas (Zipf, 1929). Assumida como representativa de um comportamento geral das línguas, com carácter descritivo mais que preditivo (Tullo, & Hurford, 2003), contestada no grau de aproximação para línguas como o chinês (Yan & Minnhagen, 2015), teve aplicações em campos diversos e foi objeto de revisões e ampliações para além do âmbito linguístico. Encontramos exemplos de desenvolvimento da sua formulação matemática em Rapoport (1983), Köhler (2002), Hilberg (2002) e Li (2002;) e criticismo em Kornai (2008), Benguigui e Blumenfeld-Lieberthal (2011) e Piantadosi (2014). Um exemplo de aplicação no âmbito da geografia é Lois-González e López-González (2013).

4.7.1.2 Aplicação sobre o corpus

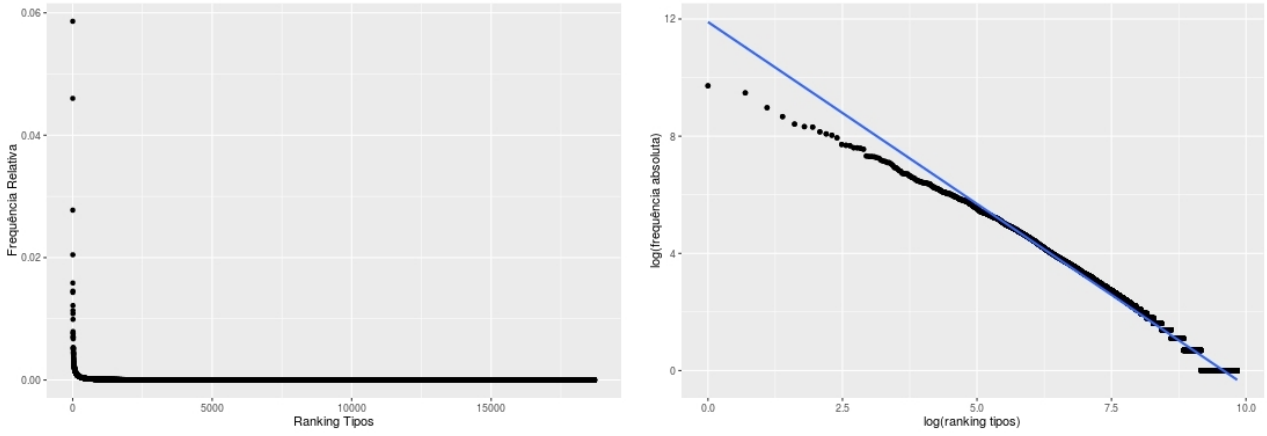


Figura 4.2: Frequências dos tipos do corpus em ordem decrescente (esquerda) e transformação logarítmica com uma linha de regressão (direita), mostrando uma relação aproximada à expressada na primeira lei de Zipf.

Para a aplicação sobre o nosso corpus tomamos a formulação expressada em Li (1991):

$$P(r)= C / r^{\alpha} \quad (4.10)$$

em que P_r representa a frequência, *r* vem dado pela ordenação decrescente das frequências, C ≈ 0.1 e $\alpha \approx 1$. A sua representação gráfica tem a forma de uma curva monotonamente decrescente e uma linha recta com declive -1 ao realizarmos a sua transformação logarítmica.

O gráfico da figura 4.2 esquerda mostra a distribuição do corpus com os termos de maior frequência decrescendo em intervalos progressivamente menores até formar a linha dos dislegomena e a maior dos hápax legomena, dominante no eixo das ordenadas. As frequências aparecem com valores relativos para facilitar a sua comparação. Observamos uma distribuição similar também sobre subsegmentos do vocabulário, por exemplo quando restrito segundo um critério de classificação semântica (entidades geográficas mencionadas) ou por fatores extralinguísticos como o tipo de conhecimento (§7.5), o qual parece indicar o seu carácter estatístico, não revelador de uma propriedade intrinsecamente linguística.

A transformação das frequências absolutas representadas em função do logaritmo dos tipos ordenados mostra uma relação linear (fig. 4.2 direita), o desvio dos dados do corpus da linha de regressão aumenta nas frequências mais altas, confirmando que apenas supõe uma aproximação (Piantadosi, 2014), não obstante, o conjunto da distribuição representa um comportamento observado repetidamente na análise de corpora (Kornai, 2008; Bian, Lin, Zhang, Ma, & Ivanov, 2016). Serve-nos, portanto, para validar o nosso corpus como objeto textual representativo da linguagem natural segundo a lei de Zipf.

4.8 Exemplo de análise em modo anotado

A anotação do corpus permite-nos analisar tokens e tipos a respeito das entidades geográficas mencionadas. Uma primeira questão, continuação da análise feita em §4.7, é qual seria a distribuição das frequências se as limitarmos apenas aos tipos das entidades geográficas mencionadas. Em §7.5 amostramos exemplos em que subsegmentos de uma distribuição aproximadamente zipfiana (as entidades geográficas mencionadas como subconjunto do total de tipos do corpus) têm, de novo, uma distribuição aproximadamente zipfiana. Mas a anotação das entidades geográficas permite fazer análises mais elaboradas desde o início. Ao incluirmos um fator que relaciona o objeto textual com um fator geográfico, podemos perguntar sobre a relação do texto com a geografia.

4.8.1 Caracterização do espaço geográfico por capítulos

Para analisarmos geograficamente o corpus, um primeiro elemento que surge como limitador é o conhecimento geográfico. Se distinguirmos que a entidade *A* é um lugar distinto de *B*, poderemos mais facilmente observar uma distribuição geográfica no corpus, se a houver. De preferência, devemos operar com entidades geográficas cuja distribuição espacial, ainda que aproximada, seja conhecida. Por exemplo, de *Goa* sabemos, sem necessidade de maior explicação, que é uma cidade

costeira da Índia, porém, de *Aapessumhee*, um lugar do reino de Aarù, por sua parte antigo reino na Samatra, contamos apenas com a descrição obtida do corpus.

Em segundo lugar, é preciso que as entidades geográfica observadas tenham uma frequência suficientemente alta para observar a sua variação (uma entidade geográfica mencionada que apareça uma vez no corpus é indicativo geográfico, mas não podemos observar como varia).

Para aplicarmos o primeiro critério, não temos nenhum elemento identificativo (será desenvolvido no capítulo 7) mais que a inspeção direta das entidades. Optamos por começar pelo segundo, que já foi solucionado: o conjunto das entidades geográficas mencionadas ordenadas pelo ranking de frequência é um subconjunto dos tipos do corpus ordenados segundo uma distribuição para o teste da lei de Zipf (§4.7).

4.8.2 Resultados

Ao selecionarmos as 20 entidades geográficas mencionadas com maior frequência no corpus em função da sua distribuição por capítulos, obtemos o gráfico da figura 4.3.

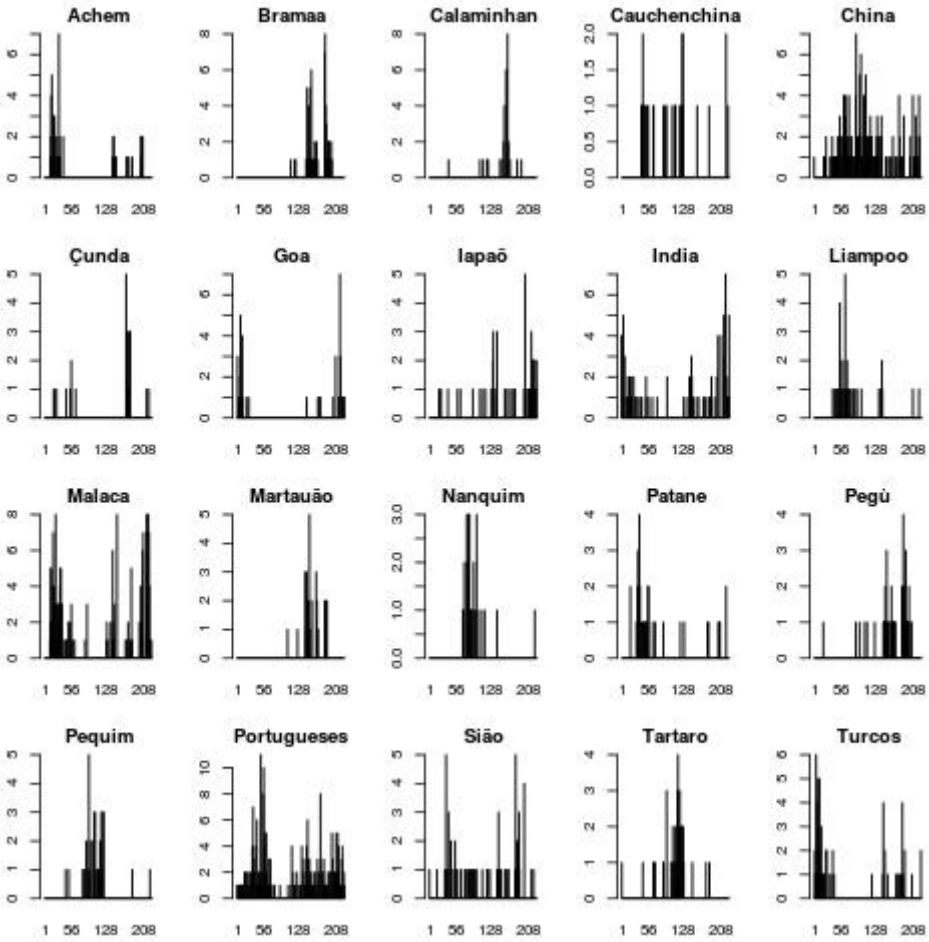


Figura 4.3: Distribuição das frequências (Y) por capítulos (X) das 20 primeiras entidades geográficas mencionadas ordenadas segundo a distribuição da primeira lei de Zipf.

Observamos:

- Os tipos que obtemos quando selecionamos as 20 primeiras frequências anotadas ordenadas segundo a lei de Zipf são entidades geográficas conhecidas ou facilmente consultáveis para uma georreferenciação aproximada. Podemos, portanto, distingui-las entre si e estabelecer relações espaciais e de atributos geográficos.
- Uma mesma entidade geográfica mencionada tende a aparecer em capítulos contínuos, há uma tendência clara a criar clusters de capítulos.
- Há uma complementaridade na distribuição entre entidades geográficas ao observarmos os máximos das frequências. Os capítulos com maior frequência do *Achem*, distam daqueles outros em que predomina a *China* ou o *Bramaa* ou o *Calaminham*. Há uma tendência a criar clusters de entidades.
- Há coincidência na distribuição entre entidades geográficas relacionadas. *Goa* aparece em faixas em que também a *India* tem as frequências mais altas. *Nanquim* e *Pequim*, entidades geográficas distintas, mas de um nível similar (grandes cidades) aparecem com certa complementaridade: ocupam capítulos diferentes nos seus máximos de frequência, mas próximos, e ambos coincidem num espaço em que também a *China* tem as suas frequências mais altas.

4.9 Conclusão

À hora de elaborarmos um corpus temos de ter uns objetivos iniciais que permitam orientar o seu desenho. As possibilidades de elaboração e os procedimentos a seguir são variados: ter umas expetativas de resultados oferece um critério de escolha e avaliação de alternativas. O processo de elaboração segue um processo inicialmente linear, que vai do texto original a um corpus anotado consultável segundo as nossas preferências analíticas. No entanto, a prática mostra que o processo vira cíclico, a superação de uma fase apresenta novas questões ou (no pior dos casos) mostra deficiências nos estádios anteriores que obrigam à sua reconsideração. Ao chegarmos ao objetivo final, contar com um corpus estável, obtemos também uma série de produtos auxiliares que complementam e mostram as possibilidades de desenvolvimento futuro: motor de pesquisa, ferramentas de leitura mais ágil e seletiva, índices e tabelas de dados com que começarmos a formular as conjeturas e hipóteses que serão o objeto das nossas pesquisas. O corpus pode ser consultado em modos distintos. Partindo de conjeturas iniciais realizamos as primeiras análises exploratórias. Os resultados obtidos de uma de análise podem ser reutilizados para novas consultas.

A conclusão mais importante é que a análise de frequências combinada com a anotação das entidades geográficas mencionadas revela uma estrutura geográfica. Uma mesma entidade tende a oferecer os seus máximos em clusters, entidades geográficas distintas e mais afastadas ocupam posições distintas no corpus, entidades geográficas relacionadas coincidem no mesmo cluster que a entidade geográfica maior que as contém.

Capítulo 5

A identificação das entidades geográficas mencionadas

No capítulo anterior descrevemos a preparação do corpus para o estudo das entidades geográficas mencionadas. Consideramos agora a primeira parte do problema da georreferenciação, a identificação, entendida como o reconhecimento de aquelas expressões que, dentro de um texto, são entidades geográficas mencionadas.

A modo de introdução, enquadramos a identificação das entidades mencionadas, de modo genérico, dentro do Processamento da Linguagem Natural (PLN) e disciplinas que atendem ao mesmo problema e descrevemos os procedimentos relacionados com a sua aplicação no caso particular das entidades geográficas assim como algumas dificuldades que, sem serem centrais, condicionam o desempenho do processo de identificação.

O resto do capítulo estuda três casos práticos de identificação automática. O primeiro analisa a anotação do corpus descrito no capítulo anterior e o modo como se solucionaram as situações criadas pela aplicação automática da lista de topónimos e gentílicos. No segundo, usamos a mesma lista, integrada numa ferramenta de PLN para a anotação morfossintática, e aproveitamos os resultados para introduzirmos as métricas de precisão, abrangência e medida-F que utilizaremos no resto da tese. O último caso estuda uma anotação totalmente automática sobre um corpus menor: a versão em inglês (Pinto, 1653) dos capítulos da Tartária, conferindo os resultados das configurações sobre três sistemas de uso livre: um de base estatística, o segundo de regras orientado para trabalhos de PLN, e um terceiro de regras orientado para o georreferenciamento (com capacidades de representação cartográfica). Como conclusão dos ensaios oferecemos uma possível solução prática para a anotação de textos históricos auxiliada por métodos de anotação automática.

5.1 Enquadramento da identificação de entidades geográficas mencionadas num corpus

5.1.1 Áreas relacionadas

A identificação de entidades (do tipo que for, não só as geográficas) é referida também como reconhecimento de entidades geográficas mencionadas (REM) (*NER*, de *Named Entity Recognition*, em inglês). Quando a identificação usa algum tipo de marcado padrão, o processo é conhecido como anotação. O problema tem grande aplicação e é abordado em campos diversos: no Processamento da Linguagem Natural (PLN), na Recuperação de Informação (Leidner 2007; Dias,

Anastácio & Martins, 2012; Janowicz, Raubal, & Kuhn 2015) e na Extração de Informação (Leidner 2007: 56-7; Anastácio, Martins, & Calado, 2011; Ji & Grishman, 2011). Da combinatória de aplicações e metodologias surgem propostas de subdisciplinas mistas que combinam SIG e linguística computacional, assim a análise geográfica de textos (Gregory, Cooper, Hardie, & Rayson, 2015).

A variedade de aproximações dá lugar também a certa confusão terminológica. Identificação de entidades mencionadas (Santos & Cardoso, 2007b, p.3), reconhecimento de topónimos (Leidner 2007, p. 29) e geo-parsing (Leidner 2007, p. 3; Jones & Purves, 2008) são termos usados como um caso especial ou extensão do problema NERC, geoparsing abrangendo tanto a identificação como a referenciação da entidade (Rupp, Rayson, Baron, Donaldson, Gregory, Hardie, & Murrieta-Flores, 2013; Gregory, Baron, Murrieta-Flores, Hardie, & Rayson, 2013). Se no PLN uma entidade geográfica mencionada está convenientemente anotada quando se marca como tal, sem necessidade de adicionar um referente ou coordenadas, numa aplicação com uma finalidade mais geográfica a identificação é apenas uma parte do problema da georreferenciação, aquela que precede a resolução do topónimo (Leidner 2007, p. 3; Tobin, Grover, Byrne, Reid, & Walsh, 2010). Esta distinção é precisa para compreender a variedade na aplicação e extensão de termos como anotação (Stokes, Li, Moffat, & Rong, 2008) ou geotagging (Pasley, Clough, & Sanderson, 2007), para se referir à resolução do topónimo associando-o a uma única localização geográfica, face a um uso mais restrito no âmbito PLN em que apenas especificam a atividade NERC prévia à resolução do referente (Tobin, Grover, Byrne, Reid, & Walsh, 2010; Grover, Tobin, Byrne, Woollard, Reid, Dunn, & Ball, 2010).

Neste trabalho usaremos o termo *identificação* para nos referirmos às atividades de reconhecimento da entidade mencionada no texto, trabalho que, também neste caso particular de estudo, se resolve por meio da *anotação* (marca sobre o texto).

5.1.2 A identificação de entidades geográficas como problema do PLN

No PLN, um labor frequente é o da classificação das unidades dentro de categorias gramaticais, entre as quais achamos os nomes e, dentro deles, os nomes próprios que, pelas suas características especiais (não aparecem no lexicon como tais), supõem um problema específico a resolver, referido como reconhecimento de entidades mencionadas. O problema REM (NER pelas siglas em inglês) é resolvido ao identificar como entidade mencionada uma unidade formada por uma palavra (ex. *Pequim*) ou grupo de palavras que formem uma unidade (ex. *Pullo Çambilão*) para classificá-la como correspondente a um nome de lugar, pessoa, organização ou aquelas outras classes que determinássemos, as mais comuns nomes próprios, ainda que também se podem incluir nesta categoria tipos especiais tais como datas e quantidades. Conhecemos a combinatória de ambos os problemas, reconhecimento e classificação, como NERC, pelas suas siglas em inglês *Named Entities Recognition and Classification*.

5.2 Processos na identificação de entidades geográficas mencionadas como atividade prévia à georreferenciação

5.2.1 Reconhecimento da entidade

O primeiro passo para a identificação de uma expressão como entidade geográfica é reconhecer as entidades mencionadas como tais, quais são os segmentos (palavras ou expressões multipalavra) que representam uma entidade mencionada no texto.

Resolvemos o problema do reconhecimento com a atribuição de uma marca (frequentemente seguindo uma série de normas padrão) que designa uma unidade como representativa de uma entidade geográfica.

> "& chegando a hum lugar que se dizia <EM>**Aapessumhee**</EM>, quatro legoas do rio de <EM>**Puneticão**</EM>, soube por algũs pescadores que ahy tomou, tudo o que na fortaleza, & no reyno era passado, & como <EM>**Laque Xemena**</EM> estaua apoderado, assi da terra como do mar esperando por elle, com a qual noua dizem que o <EM>**Heredim Mafamede**</EM> ficou muyto embaraçado, porque na verdade nunca lhe pareceo que os inimigos fizessem tanto em tão pouco tempo." (*PR*, 32) ($5a_1$)

Em ($5a_1$) quatro entidades foram identificadas adicionando-lhe uma marca (anotação): <EM>. Os métodos usados para a identificação podem ir da simples aplicação de listas e combinatória de regras até ao treino a partir de corpora com um grande número de variáveis como preditores. Neste exemplo particular ($5a_1$), um sistema pode realizar a identificação aplicando duas regras:

1. Se uma expressão aparece em maiúscula no meio do texto, é uma entidade mencionada.

2. Se dois termos com maiúscula inicial aparecem seguidos um do outro, sem nenhuma marca de pausa, são parte de uma única entidade.

5.2.2 Classificação

Seguidamente é preciso distinguir a que classe pertencem as entidades marcadas, no nosso caso, identificar quais são as geográficas (em oposição a pessoas, organizações, ou outro tipo de EM). Nas aplicações NERC orientadas para o PLN, a classificação de uma entidade como sendo geográfica, é mais uma parte do processo de identificação. Se uma unidade é reconhecida como entidade mencionada, procede a classificação dentro de uma categoria (Gamallo & Garcia, 2011). Porém, num sistema orientado especificamente à resolução de termos geográficos pode não haver processo de classificação, pois apenas se identificam as entidades geográficas (só há uma classe).

No mesmo exemplo, as entidades são classificadas mediante uma anotação:

> "& chegando a hum lugar que se dizia **<LUGAR>Aapessumhee</LUGAR>** quatro legoas do rio de **</LUGAR>Puneticão</LUGAR>,** soube por algũs pescadores que ahy tomou, tudo o que na fortaleza, & no reyno era passado, & como **<PESSOA>Laque**

> **Xemena</PESSOA>** estaua apoderado, assi da terra como do mar esperando por elle, com a qual noua dizem que o **<PESSOA>Heredim Mafamede</PESSOA>** ficou muyto embaraçado, porque na verdade nunca lhe pareceo que os inimigos fizessem tanto em tão pouco tempo." (*PR*, 32) ($5a_2$)

Os métodos aplicados para a classificação prolongam a identificação (listas específicas, regras, treino). Neste exemplo particular observamos como as entidades geográficas são, numa mesma unidade sintática (sem marca de pausa), precedidas de termos geográficos (*lugar*, *rio*) que não aparecem no caso das pessoas. Também achamos relevante o facto de que as entidades pessoais sejam formadas por mais de uma expressão com maiúscula inicial. Um sistema NERC aproveita estas características para outorgar-lhe maior ou menor probabilidade a uma classe, a maiores da aplicação opcional de uma lista de entidades que contribua para a classificação (listas de topónimos e nomes de pessoa).

5.2.3 A desambiguação geo / não-geo

Um problema específico derivado da classificação é o das formas ambíguas, aquelas reconhecidas como pertencentes a mais de uma classe (Anastácio, Martins, Calado, 2011). A resolução de este tipo de problemas, quando limitados à atribuição da entidade mencionada dentro da categoria de nome geográfico, chamamos de desambiguação geo / não-geo (Anastácio, Martins, Calado, 2009; Zhang, Jin, Lin, & Yue, 2012).

Por exemplo, *Xipatom* pode ser nome de cidade (entidade geográfica) ou usado para designar uma pessoa (entidade não-geográfica, não anotada no corpus):

> "os quais o Rey Tartaro aly trouxera de hum grande templo chamado Angicamoy que tomara na cidade **<LUGAR>Xipatom</LUGAR>** na capella dos jazigos dos Reys da China para triumphar delles quando se embora tornasse para sua terra, por que se soubesse por todo o mundo que a pesar do Rey da China lhe catiuara os seus deoses." (*PR*, 122) (5b)

> "E apos estas & outras muytas palauras dignas de serem notadas, que por regimento da casa lhe diz hum Sacerdote, o **Xipatom**, que he, como disse, o principal sobre todos os outros que gouernão este grande labarinto" (*PR*, 106) (5c)

A resolução da desambiguação geo / não-geo é uma extensão do problema de classificação, de maior relevância quando considerarmos vários tipos de entidades (não unicamente as geográficas). No corpus estudado nesta tese apenas anotamos as entidades geográficas e só em testes pontuais consideramos a anotação de entidades não geográficas.

5.2.4 A desambiguação geo / geo

É importante notar que uma vez resolvido o problema NERC, isto é, identificada a entidade como geográfica, temos de novo processos classificatórios ligados à referenciação, por exemplo, quando

distintos lugares são referidos por um mesmo topónimo ou numa mesma localidade convergem vários tipos ou acidentes geográficos com um mesmo nome. A desambiguação nestes casos é representada mediante um identificativo único para cada objeto geográfico (entidades geográficas distintas têm um número de referência que as identifica e distingue).

> “E com isto se determinou com parecer & conselho de todos, de yr inuernar os tres meses que lhe faltauão para poder fazer sua viagem, a hũa ilha deserta que estaua ao mar de <LUGAR id='329'>**Liampoo**</LUGAR> quinze legoas, que se chamaua <LUGAR id='518'>**Pullo Hinhor**</LUGAR>, de boa agoada, & bom surgidouro” (*PR*, 66) (5d)

> “Partidos nos, como ja disse atras, desta ilha de <LUGAR id='519'>**Pullo Hinhor**</LUGAR> continuamos por nossa derrota na via do porto de <LUGAR id='610'>**Tanauçarim**</LUGAR>, ao negocio que já atras disse algũas vezes, & como foy noite, receoso o Piloto dos muytos baixos que tinha por proa, se fez no bordo do mar com tenção de tanto que fosse menham tornar a demandar a terra cos ventos Oestes, que ja neste tempo cursauão da <LUGAR id='296'>**India**</LUGAR> por moução tendente.” (*PR*, 147) (5e)

No primeiro exemplo (5d) temos uma ilha a quinze léguas de *Liampoo* (Ningbo, N 29° 52' 41" E 121° 32' 58') no mar Leste da China (Lagoa 1953-59; Alves 2010), no entanto, no segundo (5e) refere-se uma viagem de Malaca a Martabão, pela costa oeste do Malayo; o porto de refêrencia mencionado é *Tanauçarim* no extremo sul da Baixa Birmânia (Tenasserim, 11º 59’ N, 99º E) (Thomaz, 2002).

Os métodos para a desambiguação geo / geo (Santos, Anastácio, & Martins, 2015) baseados em técnicas PLN usam grandes volumes de corpora georreferenciados (ex. bases de conhecimento abertas) ou corpora anotados manualmente (DeLozier, Wing, & Baldridge, 2016) para obter atributos (quando a sua natureza é mais estatística) e regras sobre a base de variáveis tais como a maior ou menor frequência da entidade geográfica (as menos comuns têm menos probabilidades de ser a referência procurada) ou elementos do vocabulário (por exemplo termos do campo semântico do mar aparecem com maior frequência junto com localidades costeiras) que outorgam valores de probabilidade para os distintos candidatos. Métodos mais próprios da análise espacial aplicam atributos próprios da geografia física e humana (assim, extensão e povoação). Também é comum a reelaboração dos dados das coordenadas para obter a distância a respeito do centroide das entidades geográficas agrupadas num cluster (por coocorrerem numa mesma unidade textual): a entidade geográfica cujas coordenadas forem mais próximas será escolhida como referente.

Nos exemplos (5d) e (5e) não temos coordenadas que resolvam definitivamente a entidade mencionada *Pullo Hinhor*. Não obstante, as entidades geográficas com que coocorre apontam para áreas geográficas distintas, um ponto perto de *Liampoo* (porto da China) em (5d), face às referências ao porto de *Tanauçarim*, na costa do *Malayo* em (5e). A desambiguação requer um estudo crítico prévio, em todos os casos muito especializado por serem entidades com coordenadas não resolvidas ou de uso pouco frequente. A solução adotada para operar com entidades ambíguas

no corpus foi anotar cada uma com um identificativo único, a desambiguação resolvida manualmente. Assim, um mesmo objeto geográfico tem um único identificativo independentemente de ser mencionado por mais de uma expressão, e expressões idênticas que referem objetos geográficos distintos são marcadas com identificativos também diferentes.

5.2.5 Problemas relacionados com o reconhecimento das entidades

Referimos aqui dois problemas independentes do reconhecimento e classificação cuja solução incide diretamente no desempenho de ferramentas NERC.

5.2.5.1 A normalização

Um nome de lugar pode ter variantes de expressão. O processamento desta variação (Rupp, Rayson, Baron, Donaldson, Gregory, Hardie, & Murrieta-Flores, 2013) consiste em normalizar ou ligar todas as variantes sob uma mesma forma: a mais representativa ou canónica (Leidner 2007: 32). Distinguimos vários tipos de variação:

- Ortográfica. Quando um mesmo som se representa com várias grafias. Ex. *Sincaapura* e *Cincaapura* têm grafias distintas, mas supomos uma mesma realização fonética.

- Fonética. Na mesma língua pode haver pronúncias distintas de um topónimo por variações dialetais ou de registo. No caso da *Peregrinação*, achamos grafias distintas que podem (mas não temos confirmação) responder a distintas percepções e interpretações da língua de origem. Em qualquer caso, representam uma realização fonética distinta na língua de chegada: *Angegumaa* e *Iangumaa* referem-se a um rio (possivelmente o Salween).

- Gralhas. Variações atribuíveis ao processo de edição. *Lequios*, *Elequios* (isto é, conjunção copulativa mais gentílico: "*e Lequios*").

- Abreviações. Ex. *Apefingau*, com aférese *Fingau* (ilhéu na Samatra).

5.2.5.2 Lematização

A lematização é o processo de ligar as variantes gramaticais de uma palavra a uma forma representativa, a que normalmente aparece nos dicionários (Roak & Sproat, 2007, p. 24). Em trabalhos NERC determinar os lemas de cada unidade é particularmente relevante quando usamos o contexto (os termos que rodeiam a entidade geográfica mencionada ou mesmo o conjunto do documento) para reconhecer e classificar uma entidade mencionada (Gamallo & Garcia, 2011; Grover, Tobin, Byrne, Woollard, Reid, Dunn, & Ball, 2010). Nestes casos procede uma lematização de todos os elementos do vocabulário.

Para a georreferenciação interessa particularmente a lematização dos topónimos. Em línguas com casos, a flexão pode afetar os nomes de lugar, por exemplo, latim e mongol. Naquelas em que os nomes próprios são considerados formas típicas sem flexão, como o português, temos o caso particular dos gentílicos que são, frequentemente, formados ao adicionar um sufixo derivativo para o topónimo. Neste trabalho, um gentílico é mais uma variante da forma representativa da entidade

geográfica mencionada e, portanto, resolvível por um processo de lematização. O termo que usamos para nos referir à expressão que representa todas as variantes de uma entidade é o de lexema (§3.3.3).

5.3 A identificação de entidades a partir de três casos práticos

Usando o corpus como material de trabalho e padrão dourado analisamos três casos de automatização no processo de identificação.

- Caso de identificação semiautomática com uma lista de entidades geográficas. Começamos com a análise da aplicação de uma lista previamente definida de entidades mencionadas extraídas do próprio corpus. Neste processo temos uma lista construída a partir de um trabalho especializado e manual que usamos posteriormente para anotar de modo automático o corpus. Analisamos em base a exemplos que ilustram os problemas originados pela anotação automática a partir de uma lista, mesmo quando esta for específica do corpus (§5.4).

- Caso de estudo de desempenho do labor de identificação de uma ferramenta NERC. Repetimos o processo de identificação automática do corpus com a mesma lista do caso anterior, mas desta vez usando uma ferramenta de reconhecimento e classificação de entidades mencionadas. Aproveitamos os resultados para introduzirmos as métricas de desempenho, precisão e abrangência que usaremos no resto desta tese (§5.5).

- Caso de estudo comparativo da configuração de sistemas NERC para a anotação automática de um corpus sem anotar. Configuramos três ferramentas NERC para identificarmos as entidades geográficas mencionadas num corpus reduzido e analisamos os resultados obtidos como exemplo de anotação automática (§5.6).

5.4 Caso de identificação por meio de uma lista de entidades geográficas *ad hoc* para a anotação semiautomática do corpus

Um recurso comum no processo identificação, particularmente nos sistemas mais orientados à georreferenciação, é o uso de listas de entidades geográficas (*gazetteers*) que incluem, para além do topónimo, informação complementar para a desambiguação e georreferenciação (Leidner 2007: 51), particularmente as coordenadas geográficas em termos de latitude e longitude. Quando um termo no texto coincide com uma entrada na lista, outorgamos o atributo de entidade geográfica e recuperamos a informação relevante disponível segundo os objetivos e o problema a resolver: a lista pode servir tanto para os labores iniciais de reconhecimento e classificação das entidades mencionadas como para a resolução geográfica (Gregory, Baron, Murrieta-Flores, Hardie, & Rayson, 2013). A qualidade e abrangência das listas resulta determinante na efetividade do sistema empregado (Pasley, Clough, & Sanderson, 2007; Gregory, Cooper, Hardie, & Rayson, 2015).

Em textos históricos e obras comentadas com aparato crítico, podemos achar índices temáticos e glossários de entidades geográficas. No caso da *Peregrinação* temos listas e dicionários (Lagoa,

1950-1953; Albuquerque 1994; Flores, Gomes, & Sousa, 1983; Alves, 2010), mas nem o mais exaustivo e específico (Alves 2010) recolhe todas as variantes, motivo pelo qual elaboramos uma lista própria (§4.3.2) com a qual anotamos o corpus. Esta lista foi extraída manualmente, em sucessivas leituras que acompanharam o estudo crítico do texto. A anotação automática a seguir consiste na aplicação de um algoritmo que recupera um termo da lista e pesquisa no corpus as expressões equivalentes.

5.4.1 Análise da aplicação de um sistema de identificação de entidades geográficas baseado numa lista específica

Nesta secção desenvolvemos um sistema simples que recolhe uma entrada da lista e pesquisa expressões coincidentes no corpus.

Seja:

L= {lista de nomes de lugar} = $\{l_1, l_2, \ldots, l_n\}$ onde l_i é um nome de lugar contido na lista

e

T={texto do corpus} = $\{w_1, w_2, \ldots w_n\}$ em que w_i é uma palavra do corpus e o seu índice expressa a ordem de aparição no texto.

Para simplificarmos a função de identificação, *l* e *w* representam quer tipos de um só token (*Goa*), quer formas complexas (*Montemor_o_Velho*).

Um processo de identificação automática, dada uma lista L, consiste na aplicação da lista ao corpus T que podemos definir como uma função:

$$I(l)=w \tag{5.1}$$

em que obtemos um resultado positivo se se cumpre a relação:

Coincide_na_expressão(l,w) = "*l* tem a mesma expressão que *w*" (5.2)

deste modo *I* é uma função que verifica a coincidência das cadeias de caracteres de um elemento $l \in$ L com as de $w \in$ T.

O procedimento de anotação descrito no capítulo anterior em §4.3 é um exemplo da aplicação deste modelo. As funções para o cálculo de frequências definidas em §4.4 são também resultado de aplicar a função *I*, sendo L={lista de variantes} e cada lexema X um subconjunto da lista aplicada sobre o corpus.

Outro cenário prático para o mesmo modelo é uma ferramenta de assistência à anotação. Cada vez que marcamos um novo item $w_i \in$ T, adicionamos também um novo elemento a L, onde $w_i=l_j$, $l_j \in$ L, e pesquisamos o corpus procurando todos os resultados positivos da relação *Coincide_na_expressão*(*l*,*w*) definida em (5.2).

Acorde com o objetivo de analisarmos as dificuldades que apresentam textos não normalizados

segundo o padrão contemporâneo, estudamos agora casos particulares de identificação que achamos em um cenário prático: a anotação do corpus caso de estudo deste trabalho. Cada vez que identificamos uma entidade geográfica *l*, esta passa a formar parte de L e revemos o corpus para anotarmos todas as formas com a mesma expressão, isto é, marcamos os resultados positivos de I(*l*)=*w*.

Analisamos a seguir as situações mais problemáticas achadas no caso prático de preparação do corpus (cap. 4), aquelas coincidências em que um dos termos não é uma entidade geográfica mencionada ou a expressão aparece de algum modo alterada ou tem difícil recuperação. Usamos os exemplos para definirmos características que nos permitem selecionar um token face a outros que, ainda coincidindo na mesma expressão, não são entidade geográfica mencionada. As características servem de regras que, em conjunto, provêm a definição de entidade geográfica aplicada para o corpus.

5.4.1.1 A entidade geográfica mencionada coincide com uma forma do vocabulário

Sejam as concordâncias:

> "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) (5f)

> "& no mesmo dia se coroou por Rey de Péguu **na varella grande**" (*PR*, 193) (5g)

> "Passados os dez dias deste encerramento, **as varellas** & pagodes, & brallas, que saõ os seus templos, amanheceraõ todos ornados de insignias de alegria" (*PR*, 184) (5h)

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 (5f) 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, entidade geográfica mencionada.

Característica identificadora: a entidade geográfica mencionada começa por maiúscula. Assim (5f) é entidade geográfica mencionada, (5g) e (5h), sendo a mesma expressão, não são entidades geográficas mencionadas.

5.4.1.2 A entidade geográfica mencionada coincide plenamente com uma forma do vocabulário

Uma dificuldade maior aparece na seguinte concordância:

> “sendo tanto auante como o rio a que os naturaes da terra chamão Tinacoreu, & os nossos **a varella**” (*PR*, 41) (5i)

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 (5f). 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 (5i) temos uma entidade geográfica mencionada, mas a sua codificação supõe uma exceção à primeira regra de identificação (§5.4.1.1): não começa por maiúscula. No nosso caso anotamo-la igualmente como entidade geográfica mencionada e a lista regista-a como mais uma variante com a particularidade de aparecer em minúscula.

Seguindo o mesmo critério anotamos como entidade geográfica mencionada:

> “A Quarta feira seguinte nos saimos logo deste **rio da varella** por nome Tinaçoreu” (*PR*, 42) (5j)

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

> “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) (5k)

> “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) (5l)

> “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) (5m)

Mesmo se todas as frases destacadas em (5i), (5j), (5k), (5l) e (5m) 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 (5i) e (5j) *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, portanto, do nome próprio (§3.2.2).

No entanto em (5k), (5l), (5m), a mesma expressão refere tanto o indivíduo quanto a classe (*varela*). Precisa, ademais, de modificadores para denotar uma individualidade (“*do Comquiay*” (5l), (5m)),

ou aparece simplesmente como modificador duma individualidade (“*do baluarte do caez da varella*” (5k)).

<u>*Característica identificadora*</u>*: a entidade geográfica mencionada aparece em contexto de nome próprio.* Assim em (5i) e (5j) temos nomes próprios que grafamos com maiúscula segundo as convenções atuais, no entanto em (5k), (5l) e (5m) estamos perante nomes comuns.

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

Seja a concordância:

> “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) (5n)

Em (5n) temos um exemplo similar a (5j), um nome comum forma parte dum topónimo, todos os termos aparecem como elementos do vocabulário, no entanto em:

> “& com este concerto jurado & assinado por todos, se vieraõ surgir a hũa **ilha que se dezia dos ladroẽs”** (*PR*, 53) ($5o_1$)

o mesmo topónimo aparece com uma cláusula intercalada. Pelo critério usado em §5.4.1.2 para (5j), (5n) é também entidade geográfica mencionada, mas em (5o) *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 onde surgiria mais uma dificuldade, ao aplicarmos a lista sobre o corpus obtemos também a concordância:

> “mãdou tambem hum Naique com vinte Abexins que nos veyo guardando **dos ladroẽs**, & prouendonos de mãtimẽto & caualgaduras ate o porto de Arquico” (*PR*, 4) (5p)

Uma possível solução à ambiguidade criada por concordâncias como (5p) 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 entidade geográfica mencionada.

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

> “& com este concerto jurado & assinado por todos, se vieraõ surgir a hũa **<LUGAR>ilha que se dezia dos ladroẽs</LUGAR>**” (*PR*, 53) ($5o_2$)

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.

> "& 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) ($5o_3$)

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, criando uma regra para 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.

5.4.1.4 A entidade mencionada é ambígua na expressão de referente geográfico

No exemplo:

> "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) ($5q_1$)

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 do nome próprio (§3.2.2) estamos perante um referente de pessoa, isto é, uma solução por exemplo do tipo:

> "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) ($5q_2$)

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*, para além do 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 entidade geográfica mencionada com a marca <LUGAR> em ($5q_2$).

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 entidades geográficas mencionadas 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 entidade geográfica mencionada, pois o primeiro termo da entidade pessoa aponta para uma entidade geográfica de forma não ambígua. Aliás, usamos o critério de frequência, portanto a mesma expressão tem uma ocorrência (5r) em que aparece como entidade geográfica mencionada independente:

> “hum Portuguez que andaua com elles, por nome Christouão Sarmento natural de **Bargança**” (*PR*, 195) (5r)

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 toponímico, são expressões de grande ambiguidade geo / geo (muito comuns), quando se quiser dar um lugar de origem entidade geográfica para o 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. O referente de *Bargança* é, portanto, uma entidade geográfica.

5.4.1.5 A expressão da entidade geográfica mencionada é uma entidade não geográfica

No exemplo:

> "& 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) (5s)

temos uma situação inversa a §5.4.1.4, estamos perante uma entidade geográfica mencionada 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:

> "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) (5t)

ao considerarmos as construções como um tipo geográfico, se estas aparecerem referidas por um nome próprio, teremos também um caso de entidade geográfica mencionada. Em (5t) "o que se chama" é o mosteiro, a entidade é geográfica, independentemente do seu carácter hagionímico.

Do mesmo modo o teónimo *Tinaogoogoo* na concordância:

> "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) (5u)

aparece explicitamente como expressão de um edifício, um pagode, portanto, o seu referente é uma entidade geográfica e não diretamente o deus *Tinagoogo*, caso do exemplo:

> “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) (5v)

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

> “Do caminho que fizemos atê chegarmos ao pagode de **Tinagoogoo**.” (*PR*, 158) (5w)

A expressão *Tinagoogoo* aparece agora como modificador e não núcleo do sintagma entidade geográfica mencionada. No entanto, o contexto refere o mesmo pagode que em (5u). A solução que adotamos neste caso passa por um critério alheio a regras linguísticas, similar ao que nos levou a considerar os gentílicos como mais uma variante dos topónimos: com o fim de obtermos mais ocorrências para o tratamento do corpus e resolução de georreferencias, anotamos como entidade geográfica mencionada aqueles casos em que, tendo sido declarada a expressão de modo explícito como entidade geográfica, volva aparecer numa outra concordância acompanhada de um atributo geográfico, ainda quando estiver também a apontar um referente não geográfico.

Característica identificadora: uma expressão declarada de modo explícito como entidade geográfica em quando menos uma ocorrência será anotada como entidade geográfica mencionada quando aparecer como modificador do tipo geográfico em que foi declarada, ainda quando tiver também o valor de entidade de uma classe não geográfica. Assim em (5t) um hagiónimo é explicitamente declarado expressão geográfica, como também o teónimo em (5u) nomeia de modo inequívoco um edifício, em ambos os casos anotamos a expressão como entidade geográfica mencionada. Do mesmo modo anotamos (5w), porquanto ainda estando diante de um teónimo, a expressão foi declarada em mais alguma ocorrência no corpus (5u) como entidade geográfica e volve aparecer como modificador do tipo geográfico que a subclassifica (pagode), no entanto em (5v), a mesma expressão não é considerada entidade geográfica, porquanto tem unicamente o valor de teónimo.

5.4.1.6 A entidade geográfica mencionada contém outra entidade geográfica mencionada

No exemplo:

> “& 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) ($5y_1$)

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:

> “duas fustas em que hião sessenta Portugueses, de hũa das quais era capitão Diogo Soarez

o **Galego**" (*PR, 204*) (5z)

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

> "& 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) ($5y_2$)

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

> "& 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*) ($5y_3$)

<u>Característica identificadora</u>: 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.

5.4.2 Considerações sobre a aplicação de uma lista de entidades geográficas

Nos exemplos analisados consideramos as limitações surgidas ao aplicarmos uma lista que contém todas as entidades geográficas mencionadas no corpus num processo de anotação automática. Achamos problemas que resultam da ambiguidade de expressões que podem ser geográficas ou não. Para cada caso apontamos como conclusão uma característica identificadora, sem que isto implique que seja necessariamente a única. A escolha de critérios de seleção evidencia a dificuldade da anotação em dois níveis. Primeiramente, pela recuperação interessada de expressões como entidades geográficas no caso de usarmos um critério extralinguístico, por exemplo quando consideramos os gentílicos para obtermos o máximo número de concordâncias e facilitarmos a georreferência das entidades geográficas mencionadas. Em segundo lugar, mesmo quando uma característica for aceite, a sua formulação em forma de regras suporá outro critério de seleção. Por exemplo: "ter contexto de nome próprio" requer definir as características de um nome próprio, abordáveis de um ponto de vista gráfico, morfológico, sintático, semântico, cada um e o conjunto com múltiplas possibilidades de formalização e resolução algorítmica.

A escolha dos critérios sobre que operar com as listas é um exemplo de regras para a identificação e classificação de entidades geográficas. As dificuldades não são, porém, exclusivas do corpus. Situação similar teríamos em outro trabalho NERC se aplicássemos listas com todas as entidades identificadas a priori, por exemplo se estivéssemos a anotar só os planetas e satélites do sistema solar, ou as freguesias de Portugal, ou as vilas e cidades da Galiza. Do confronto com o corpus emergem ambiguidades próprias das operações com a linguagem natural.

Na alínea que se segue veremos um nível mais avançado de automatização, mediante o uso de um

sistema de reconhecimento e classificação de entidades mencionadas (que chamamos de NERC pelas suas siglas em inglês) de regras para solucionar os problemas de identificação fazendo uso da mesma lista prévia que vimos de analisar.

5.5 Caso de aplicação da lista de entidades geográficas por meio de uma ferramenta NERC para a melhoria da anotação do corpus

Um modo de melhorar os resultados obtidos pela aplicação de uma lista é o uso de sistemas que implementem critérios de reconhecimento e classificação para a resolução de situações de ambiguidade. De facto, os primeiros sistemas de reconhecimento de entidades mencionadas aplicavam regras para otimizar listas (Nadeau & Sakine, 2007). Hoje em dia são comuns soluções de aprendizado de máquina e sistemas híbridos que combinam regras e métodos estatísticos (Leidner 2007; Nadeau & Sakine, 2007; Ratinov & Roth, 2009; Gamallo & Garcia, 2011; Garcia, 2014). Dentro da variedade de soluções, descrevemos uma aplicação para o *Reconhecimento de Entidades Mencionadas* (usaremos as siglas NER em inglês para nos referirmos a ferramentas com esta função) como um sistema que avalia os atributos de uma unidade do texto com vistas a reconhecê-la ou rejeitá-la como entidade mencionada. Se também seleciona o tipo de entidade entre vários possíveis (por exemplo, PESSOA, LUGAR, ORGANIZAÇÃO) usamos o termo NERC (reconhece e classifica).

Nesta secção analisamos o caso prático de aplicar uma ferramenta NERC de regras para anotar o corpus com a mesma lista de entidades geográficas da secção anterior (§5.4). Considerando as características particulares dos requisitos para selecionar o que é ou não uma entidade geográfica (necessidade de compreender os gentílicos) e o facto de o texto estar não só sem normalizar relativamente ao padrão contemporâneo, mas sem corrigir as gralhas editoriais (por exemplo, topónimos que começam com minúscula e variantes do tipo *Elequios* por *e Lequios* ou *Gucos* por *Gueos*), é importante insistir que não estamos a medir o desempenho do sistema como ferramenta NERC genérica. Pelo contrátrio, realizamos uma análise comparativa para observarmos o grau de maior ou menor coincidência de uma anotação semiautomática, muito revista (cap. 4), com o resultado da aplicação de uma configuração particular de uma ferramenta NERC inicialmente não desenhada para trabalhos de identificação tão específicos. Tomamos o corpus já anotado como padrão porque o consideramos mais próximo dos nossos propósitos, mas, propriamente, os resultados da ferramenta testada não são necessariamente incorretos (por exemplo, quando não identifica um gentílico), nem os do padrão totalmente corretos (a ferramenta, de facto, achou casos de entidades geográficas que ficaram sem anotar no corpus).

5.5.1 Métricas

Uma descrição de medidas e critérios avaliativos específica para o desempenho de sistemas NERC aparece em Santos, Cardoso e Seco (2007) e uma introdução em contexto de PLN em Manning e Schütze (1999, pp. 267-271). Como exemplo para introduzirmos as medidas usadas, escolhemos o

capítulo 120 do corpus da *Peregrinação* e o módulo NERC de Linguakit[3] (Garcia & Gamallo, 2015), um pacote de PLN (§5.6.2.2) que adaptamos para a busca de possíveis erros de anotação adicionando-lhe a nossa própria lista de entidades mencionadas (§4.3.2). Referimos o corpus previamente anotado como padrão e a anotação obtida pelo Linguakit como resultados do sistema.

Seja T um vetor indexado que contém os tokens do texto e os mostra por ordem de ocorrência :

T = {ocorrências ordenadas de todos os tokens do texto}

T = {w_1, w_2, …. w_n} em que *w* é o token e o índice expressa a ordem de aparição no texto.

Para o capítulo 120 contamos 1044 tokens:

T = {w_1="Do", w_2="caminho", w_3="que", w_4="o", w_5="Mitaquer", w_6="fez", w_7="deste", w_8="castello", w_9="de", w_{10}="Nixiamcoo", w_{11}="atê", w_{12}="chegar", w_{13}="ao", w_{14}="arrayal", w_{15}="que", w_{16}="el", w_{17}="Rey", w_{18}="dos", w_{19}="Tartaros", w_{20}="tinha", w_{21}="sobre", w_{22}="a", w_{23}="cidade", w_{24}="do", w_{25}="Pequim", …, w_{1044}}

Seja P um vetor indexado com todas as ocorrências de entidades geográficas anotadas no segmento analisado do corpus, isto é, as anotações do padrão no capítulo 120, na mesma ordem e frequência com que aparecem no texto.

P = {entidades geográficas anotadas no padrão}

P = {Nixiamcoo, Tartaros, Pequim, Pequim, Lautimey, Bumxay, Quansy, Nixiancoo, Pommitay, Pequim, Palemxitau, Tartaro, Lautir, Persia, Quansy}

Seja S o vetor indexado das ocorrências das entidades geográficas anotadas pelo sistema NERC:

S = {entidades geográficas anotadas pelo sistema NERC}

S = {Nixiamcoo, Pequim, Pequim, Lautimey, Bumxay, Quansy, Nixiancoo, Pommitay, Pequim, Palemxitau, Tartaro, Lautir, Quansy, Mitaquer}.

Queremos conferir os resultados do sistema a respeito do padrão. Uma solução consiste no alinhamento de P e S em T. Isto é, tencionamos achar a intersecção da anotação no padrão com a do corpus. Resolvemos as ocorrências anotadas como entidade geográfica pela sua expressão e o resto dos tokens como NULO, de modo que obtemos a seguinte lista para as anotações do padrão:

P_{emT} ={NULO, NULO, NULO, NULO, NULO, NULO, NULO, NULO, NULO, Nixiamcoo, NULO, NULO, NULO, NULO, NULO, NULO, NULO, NULO, Tartaros, NULO, NULO, NULO, NULO, NULO, Pequim, … p_{1044}}

Procedemos do mesmo modo para o resultado do sistema NERC:

S_{emT} ={NULO, NULO, NULO, NULO, NULO, NULO, NULO, NULO, NULO, Nixiamcoo, NULO, NULO, NULO, NULO, NULO, NULO, NULO, NULO, NULO, NULO, NULO, NULO, NULO, NULO, Pequim, … s_{1044}}

[3] Disponível em https://citius.usc.es/transferencia/software/linguakit.

Já agora, para avaliarmos os resultados do sistema com o padrão, simplesmente temos que selecionar o mesmo índice. Podemos também simplificar as séries para convertê-las em vetores onde 0 é NULO e 1 representa a anotação de uma entidade geográfica mencionada (que tem a mesma expressão nas duas listas por estarem alinhadas em T). Podemos assim aplicar as métricas que explicamos a continuação.

Verdadeiros positivos

As entidades geográficas que o sistema anota como tais e estão também anotadas no padrão. Isto é, quando $s_i = p_i \wedge s_i \neq 0$.

Exemplo: s_{10} (“Nixiamcoo”) = p_{10} (“Nixiamcoo”).

Se o vetor é transformado em valores numéricos s_{10} (1) = p_{10} (1).

Lista de verdadeiros positivos = {Nixiamcoo, Pequim, Pequim, Lautimey, Bumxay, Quansy, Nixiancoo, Pommitay, Pequim, Palemxitau, Tartaro, Lautir, Quansy, Mitaquer}

Total de verdadeiros positivos = 14

Falsos positivos

As anotações que o sistema resolve como entidades geográficas mas não são tais no padrão. Isto é, quando $s_i \neq 0 \wedge s_i \neq p_i$.

Exemplo: s_{822} (“Mitaquer”) ≠ p_{822} (NULO).

Se os vetores são transformados em valores numéricos s_{822} (1) ≠ p_{822} (0).

Ao operarmos sobre todo o vetor obtemos:

Lista de falsos positivos = {Mitaquer}

Total de falsos positivos = 1

É de notar que a forma *Mitaquer* tem 6 ocorrências no texto analisado e apenas uma (s_{822}) aparece como falso positivo, na concordância: “O *Mitaquer* prostrado por terra, cõ as mãos aleuantadas...”. Quer isto dizer, o sistema NERC aplica corretamente uma regra que distingue a expressão como não entidade geográfica na maior parte dos casos.

Verdadeiros negativos

Todos os tokens que o sistema deixa sem anotar e também não estão anotados no padrão. Isto é, quando $s_i = 0 \wedge p_i = 0$.

Exemplo 1: s_1 (NULO) = p_1 (NULO). Se os vetores forem transformados a valores numéricos, então s_1 (0) = p_1 (0).

Para recuperarmos a expressão correspondente temos de ir a T, onde w_1= “Do”.

Exemplo 2: s_5 (NULO) = p_5 (NULO). Achamos o valor do token correspondente em T, com w_5=

“Mitaquer”. Nesta ocorrência a forma *Mitaquer* foi resolvida como no padrão, isto é, não anotada como entidade geográfica.

Lista de verdadeiros negativos = {Do, caminho, que, o, Mitaquer fez, deste, castello, de, atê, chegar, ao, arrayal, que, el, Rey, dos, tinha, sobre, a, cidade, do, Tanto, que, ao, outro, dia, foy, ... w_{1030} }

Total de verdadeiros negativos= 1030.

Falsos negativos

As entidades geográficas anotadas no padrão que o sistema não anota como tais. Isto é, quando $s_i = 0 \wedge p_i \neq 0$.

Exemplo: s_{19} (NULO) ≠ p_{19} (Tartaros). Com os elementos dos vetores em expressão numérica, s_{19} (0) ≠ p_{19} (1). Neste caso o sistema NERC não anotou o gentílico *Tartaros* que foi anotado como entidade geográfica no padrão. Enfatizamos, mais uma vez, que usamos aqui o termo falso negativo relativamente à anotação do corpus, numa avaliação convencional de desempenho NERC, o sistema procede corretamente (*Tartaros* não é um nome próprio).

Lista de falsos negativos = {Tartaros}

Total de falsos negativos= 1

	Anotado no padrão	**Não anotado no padrão**
Anotado pelo sistema NERC	verdadeiros positivos (14)	falsos positivos (1)
Não anotado pelo sistema NERC	falsos negativos (1)	verdadeiros negativos (1038)

Tabela 5.1: Tabela de contingência dos resultados obtidos pelo sistema NERC e anotações no padrão para o capítulo 120 da *Peregrinação*.

A análise dos resultados para o capítulo 120 (tab. 5.1) mostra um alto grau de coincidência entre a anotação do corpus e a da ferramenta NERC. Consideramos, portanto, que pode ser de utilidade para acharmos possíveis erros no corpus.

Se queremos avaliar a eficácia do sistema para obtermos os mesmos resultados que no padrão, temos várias medidas de avaliação ao relacionar os componentes da tabela 5.1. A medida dos verdadeiros negativos tem pouco interesse nos problemas que atendemos neste capítulo, porquanto é muito grande e representa os tokens que não são entidades geográficas no corpus. Da relação das outras três obtemos as medidas que vamos usar para avaliar o desempenho de trabalhos de PLN no resto da tese.

Para a exemplificação destas medidas usamos os resultados de mais um caso prático: o sistema NERC com a lista de todas as variantes do corpus (isto é, em princípio cobre todas as entidades geográficas presentes no texto) para pesquisarmos possíveis omissões no corpus e qual o desempenho global do sistema nas condições ideais (lista *ad hoc*). A tabela 5.2 mostra os resultados obtidos. Nas alíneas a seguir exemplificamos as métricas com estes dados.

Total entidades anotadas no corpus[4]	4488
Total entidades anotadas pelo sistema	3993
Verdadeiros positivos	3778
Falsos positivos	215
Falsos negativos	710

Tabela 5.2: Resultados de aplicação de um sistema NERC para a *Peregrinação (1614)* com a lista de todas as variantes de entidades geográficas anotadas no padrão.

Precisão

Medimos a qualidade dos resultados dos sistema. Que percentagem das anotações produzidas pelo sistema coincide com a anotação do padrão?

$$\text{Precisão} = \frac{verdadeiros\ positivos}{verdadeiros\ positivos + falsos\ positivos} \tag{5.3}$$

$$\text{Precisão} = \frac{3778}{3778+215} = 0.9462 \quad (94.62\%)$$

Abrangência

Medimos a capacidade de recuperação do sistema. Qual é a percentagem de anotações que obtivemos do total daquelas que temos no padrão?

$$\text{Abrangência} = \frac{verdadeiros\ positivos}{verdadeiros\ positivos + falsos\ negativos} \tag{5.4}$$

[4] No momento de elaboração do teste. Como resultado mesmo de aplicação do NERC descobrimos mais variantes posteriormente anotadas e adicionadas à lista. Vid. §5.5.2.

$$\text{Abrangência} = \frac{3778}{3778+710} = 0.8418 \quad (84.18\%)$$

Medida-F

Para medirmos o desempenho do sistema relativamente às duas medidas de precisão e abrangência usamos a medida-F. Considerando ambas de igual importância, aplicamos a fórmula:

$$F = \frac{2\,x\,precisão\,x\,abrangência}{precisão + abrangência} \tag{5.5}$$

$$F = \frac{2\,x\,0.9462\,x\,0.8418}{0.9462+0.8418} = 0.891 \quad (89.1\%)$$

5.5.2 Avaliação dos resultados para a melhoria da anotação do corpus

Nos exemplos anteriores, observamos que, mesmo com uma lista *ad hoc*, os resultados de aplicação de uma ferramenta NERC, ainda sendo muito altos, não representam a mesma solução que do corpus em 100%. A análise desta divergência permite-nos achar erros e, portanto, possíveis melhorias no corpus. Particularmente, o que nos interessa são os casos encontrados como falsos positivos, isto é, expressões que o NERC considerou entidades geográficas mencionadas, mas não aparecem anotadas como tais no corpus.

Ao termos avaliado os resultados do sistema em função do padrão, apenas temos de focar na precisão. Com um desempenho de 95%, fica 5% para a inspeção de respostas contraditórias. Da sua análise achamos as seguintes expressões não anotadas no corpus padrão por não terem sido levantadas no processo de elaboração manual da lista, porém, são recuperadas agora como resultado de aplicação do sistema NERC: *Ansesedaa* (rio, possivelmente gralha por Ansedaa), *Cantaõ* (cidade, variante de *Cantão*), *Ghatigaõ* (porto, gralha por *Chatigaõ*), *Ginocoginana* (pagode, inicialmente teónimo, dentro dos casos de maior ambiguidade, finalmente considerado entidade geográfica mencionada), *Lãpacau* (ilha, variante de *Lampacau*), *Latinas* (gentílico), *Minapau* (cidade), *Nacapirau* (edifício), *Pèguu*, *Pêguu* (reino, variantes de *Pegù*), *Sanchaõ* (ilha, variante de *Sanchão*), *Tobasoy* (ilha, variante de *Taubasoy*) e *Xinamguibaleu* (prisão).

A importância da análise destas divergências, ainda quando o seu número seja reduzido, vem dada por ser o único modo que tivemos para a validação duma lista que, doutra maneira, ficaria unicamente sujeita a um processo manual sem validação externa nenhuma.

5.6 Caso de configuração de sistemas NERC para o ciclo completo de anotação

Nos casos anteriores, partimos em primeiro lugar da lista elaborada manualmente para ensaiarmos um método de anotação semiautomática primeiro (por coincidência na expressão §5.4) e posterior melhoria (aplicação da mesma lista com um sistema NERC, §5.5). No caso a seguir estudamos a anotação de um corpus, relacionado com a *Peregrinação (1614)*, mas distinto em quanto é uma tradução de apenas 15 capítulos (mais facilmente inspecionável para a verificação dos resultados) com que ensaiarmos um ciclo de anotação automática completo.

5.6.1 Desempenho de trabalhos NERC

Para promover o desenvolvimento e facilitar a comparação de distintas propostas na resolução do problema NERC, celebraram-se competições entre sistemas sobre um mesmo corpus (Nadeau & Sekine, 2007; Santos & Cardoso, 2007b; Santos, Freitas, Gonçalo Oliveira, Carvalho, & Mota, 2008; Mandl, Carvalho, Di Nunzio, Gey, Larson, Santos, & Womser-Hacker, 2009; Santos, Cardoso, & Cabral, 2010). Os corpora elaborados para estes eventos servem também para testar sistemas a posteriori. Especificamente para o português, celebrou-se o HAREM num primeiro (Santos & Cardoso, 2007a) e segundo evento (Mota & Santos, 2008; Freitas, Mota, Santos, Gonçalo Oliveira, & Carvalho, 2010) que produziu corpora e métodos de referência para treinar e avaliar sistemas NERC (Amaral, Fonseca, Lopes, & Vieira, 2014; Santos e Guimarães, 2015). Observa-se uma diferença notável nos resultados citados na literatura para o inglês e o português. Finkel, Grenager e Manning (2005), em testes sobre o corpus CoNLL 2003 (Tjong Kim Sang & De Meulder, 2003), referem resultados na medida-F de 88.51% para as entidades da categoria LOC (geográficas). Para este mesmo corpus, Suzuki e Isozaki (2008), sem distinção de categorias, apresentam medidas-F por cima de 90%. Os melhores resultados para a identificação da categoria LOCAL no primeiro evento do HAREM foram de 68.03% de precisão e 73.91% de abrangência (Santos & Cardoso, 2007a, p. 331). No segundo (Mota & Santos, 2008) há uma maior variedade na avaliação de tarefas, tomamos como referência 72.12% de precisão e 80.17% de abrangência na identificação para a categoria de local (Chaves, 2008a, p.240). Amaral (2013, pp. 40-46) faz uma descrição de sistemas NERC para o português a que deve ser adicionada a sua própria proposta baseada em aprendizado de máquina: a comparação em base ao corpus do HAREM (Amaral, 2013, pp. 70-79) oferece resultados no reconhecimento de todas as categorias com uma medida-F por baixo de 60%. Uma avaliação mais recente de quatro ferramentas NERC para o português (Amaral, Fonseca, Lopes, & Vieira, 2014) mostra rendimentos da medida-F entre 50% e 60%, com uma diferença máxima de 6 pontos para todos os sistemas testados e percentagens de 62% de precisão e 66% de abrangência para os melhores resultados na categoria de lugar.

5.6.2 Seleção de aplicações para um ensaio de identificação de entidades geográficas a partir de um sistema NERC

Dado que os melhores resultados obtidos nas avaliações do HAREM (Santos & Cardoso, 2007a, p.

331) superam os obtidos no mais recente estudo crítico de Amaral e outros (2014) em que os sistemas avaliados oscilam próximos a 60% de medida-F, consideramos não só o desempenho, também a atualização, assim como o facto de os sistemas terem sido aplicados à solução de textos com características similares à *Peregrinação*, condicionantes para realizarmos o ensaio. Optamos por sistemas representativos de distintos métodos. Assim, escolhemos três ferramentas disponíveis para o uso público com versões atualizadas. São: o Stanford NER (§5.6.2.1) como típico dos sistemas de base estatística, Linguakit (§5.6.2.2) como sistema baseado em regras e listas orientado à PLN e o Edinburgh Geoparser (§5.6.2.3) como sistema baseado em regras e listas orientado para os SIG. Pela maior disponibilidade das ferramentas, usamos o componente em inglês de um subcorpus paralelo criado do alinhamento dos capítulos referidos à Tartária da primeira edição em português (Pinto, 1614) com a do inglês (Pinto, 1653).

5.6.2.1 Sistema estatístico de treino orientado a PLN

Dentro do pacote de ferramentas de PLN Stanford CoreNLP (Manning, Surdeanu, Bauer, Finkel, Bethard, & McClosky, 2014), o Stanford NER é um sistema de reconhecimento de entidades mencionadas baseado em aprendizado de máquina em que as regras (também chamadas atributos e deste modo salientando a diferença com sistemas em que as regras são inferidas por métodos não necessariamente estatísticos) são resolvidas por análise estatística a partir do treino em corpora anotados. O Stanford NER traz três modelos já configurados em que variam o número de tipos de entidades mencionadas a identificar e os corpora em que foram treinados os classificadores. Para esta prova usamos a versão mais atualizada no momento de realizar os testes[5], com os modelos e classificadores da configuração padrão mais um pacote de classificadores que prescinde do atributo de uso de maiúsculas, disponível como extra no web do Stanford NER[6].

5.6.2.2 Sistema de regras com listas orientado a PLN

Linguakit (Garcia & Gamallo, 2015) é um sistema de processamento da linguagem natural que contém dois módulos para entidades mencionadas: um para o reconhecimento e outro para a classificação. Descrito como um Transdutor de Estados Finitos, combina o uso de regras (capitalização e palavras chave) e listas para identificar entidades que posteriormente são classificadas em quatro tipos, um dos quais corresponde à categoria LUGAR (*location*). Tem versões para português e inglês e inclui listas básicas de topónimos para cada idioma, facilmente modificáveis. Neste ensaio testamos cinco configurações com cada uma sua variação nas listas usadas. Uma primeira com a lista da configuração padrão de Linguakit, outra com a lista acrescentada com todas as entidades geográficas anotadas no corpus da *Peregrinação* (1614), outra só com a lista de todas as entidades extraídas do corpus, mais uma com a lista da configuração padrão acrescentada com uma lista só de topónimos do corpus da *Peregrinação* (1614) e finalmente uma lista só de topónimos do corpus.

[5] Stanford NER - v3.6.0 – 2015-12-09
[6] http://nlp.stanford.edu/software/CRF-NER.shtml

5.6.2.3 Sistema de regras com listas orientado a georreferenciamento

Descrito como um sistema de georreferenciação (Tobin, Grover, Byrne, Reid, & Walsh, 2010), o Edinburgh Geoparser distingue duas áreas: NER (geotagger) e a resolução (georesolver) das entidades geográficas baseando-se em regras e listas. Foi adaptado para trabalhar com coleções de documentos humanísticos (Rupp, Rayson, Baron, Donaldson, Gregory, Hardie, & Murrieta-Flores, 2013; Alex, Byrne, Grover & Tobin, 2015). Grover e outros (2010) descrevem um processo típico de georreferenciação de textos históricos: partem do corpus preparado por uma suite de ferramentas de PLN que, com um texto típico em formato XML, tokeniza, segmenta orações, identifica língua, lematiza e anota categorias gramaticais. O módulo NER pesquisa expressões multipalavra, adiciona atributos (nome comum, nome de pessoa, lista), variantes e nomes alternativos das entidades, e aplica regras genéricas de identificação de entidades geográficas, de contexto (coordenação, parênteses), léxico geográfico e mesmo tipográficas (por exemplo se a itálica for usada sistematicamente num tipo de texto). O Edinburgh Geoparser usa dois tipos de listas: umas para a identificação no texto e outras para a resolução geográfica. Neste ensaio apenas avaliamos os resultados de identificação (NER), consequentemente, não consideramos a georreferenciação na análise. Para a identificação usamos o mesmo esquema de listas que no caso de Linguakit, com a lista da configuraçao padrão própria do Edinburgh Geoparser..

5.6.3 Particularidades da anotação requerida

Pretendemos uma anotação que recupere as entidades geográficas mencionadas com as mesmas características que no corpus anotado (Pinto, 1614). Nesse sentido, os ensaios não representam tanto o desempenho dos sistemas, concebidos em princípio para usos e critérios de identificação de identidades distintos dos nossos, quanto a capacidade de adaptação resultante da nossa configuração. Todos os sistemas usados são abertos e permitem configurações muito mais avançadas do que as aqui aplicadas. Os resultados por nós obtidos são, em todos os casos, melhoráveis com novas adaptações. Considerando que empregamos as configurações padrão do sistema, é importante insistir duas particularidades dos nossos requisitos de anotação que divergem das definições tipo de entidades mencionadas e afetam, portanto, o desempenho:

• O nosso corpus requer a anotação de gentílicos, não compreendidos dentro da categoria de entidade mencionada nas avaliações ao uso, em que a identificação se limita aos nomes próprios.

• No caso dos nomes próprios, uma entidade geográfica mencionada terá que ser reconhecida como tal ainda quando não aparecer em maiúscula, o qual supõe uma alteração das convenções gramaticais contemporâneas e das definições ao uso de entidades mencionadas (Santos, Freitas, Gonçalo Oliveira, Carvalho, & Mota, 2008, pp. 133-4).

5.6.4 O texto a anotar

Escolhemos uma transcrição dos capítulos referidos à Tartária da primeira edição em inglês da obra de Mendes Pinto (1653). Estes capítulos foram previamente usados para criar um corpus paralelo com o texto em português de 1614 (Pinto, 1614) (fig. 5.1 e tab. 5.3).

1	Como hum Capitão Tartaro entrou com gente nesta cidade de Quansy, & do que nella fez.	A Tartar Commander enters with his Army into the Town of Quincay, and that which followed thereupon; with the Nauticors besieging the Castle of Nixiamcoo, and the taking of it by the means of some of us Portugals.
2	Avẽdo ja oito meses & meyo que estauamos neste catiueyro em que passamos assaz de trabalhos & necessidades, porque não tinhamos de que nos sustentassemos, se não de algũas fracas esmollas que tirauamos pela cidade.	WE had been now eight months and an half in this captivity, wherein we endured much misery, and many incommodities, for that we had nothing to live upon but what we got by begging up and down the Town,
3	Hũa quarta feira treze dias do mez de Iulho do anno de 1544, sendo passada mais de meya noite se leuantou em todo o pouo hũa tamanha reuolta & vnião de repiques & gritas, que parecia que se fundia a terra, & acudindo nós todos a casa de Vasco Caluo lhe pregũtamos pela causa daquelle tumulto, & elle cõ assaz de lagrimas, nos disse, que auia noua certa de estar el Rey da Tartaria sobre a cidade do Pequim, co mais grosso poder de gente que nenhum outro Rey nũca ajuntara no mundo, desde o tempo de Adão até aquella hora,	when as one Wednesday, the third of July, in the year 1544, a little after midnight there was such a hurly burly amongst the people, that to hear the noise and cries which was made in every part, one would have thought the earth would have come over and over, which caused us to go in haste to Vasco Calvo his house, of whom we demanded the occasion of so great a tumult, whereunto with tears in his eyes he answered us, that certain news were come how the King of Tartary was fallen upon the City of Pequin with so great an Army, as the like had never been seen since Adams time;

Figura 5.1: Os três primeiros alinhamentos do corpus da Tartária correspondentes à edição de 1614 (esquerda) e 1653 (direita)

O corpus foi anotado para recuperar capítulos, títulos e páginas segundo as edições originais (fig. 5.2).

```
-<text>
 -<div type="chapter" n="38">
    <head>CHAP. XXXVIII</head>
   -<head>
    -<div2 type="page" n="149">
       A Tartar Commander enters with his Army into the Town of Quincay,
       the taking of it by the means of some of us Portugals.
     </div2>
   </head>
   -<div1>
    -<div2 type="page" n="149">
       WE had been now eight months and an half in this captivity, where
```

Figura 5.2: Excerto inicial do texto de 1653 com a marcação usada.

Para o ensaio de anotação automática de entidades geográficas nomeadas escolhemos o documento em formato texto resultado da segmentação por orações da edição em inglês de 1653. O texto a anotar aparece sem marca nenhuma, cada oração conclui com final de linha.

Tipo	PT 1614	EN 1653	Total
Capítulos	15	**5**	20
Títulos	30	**10**	40
Páginas	36	**20**	56
Orações	222	**353**	575
Alinhamentos	240	**230**	470
Palavras	3401	**2772**	6173
Tokens	18039	**18901**	36940

Tabela 5.3: Características do corpus da Tartária. Destacado o componente escolhido para a prova de identificação de entidades geográficas.

5.6.5 Elaboração do padrão

Realizamos provas iniciais com todos os sistemas e escolhemos o StanfordNER como base (fig. 5.3).

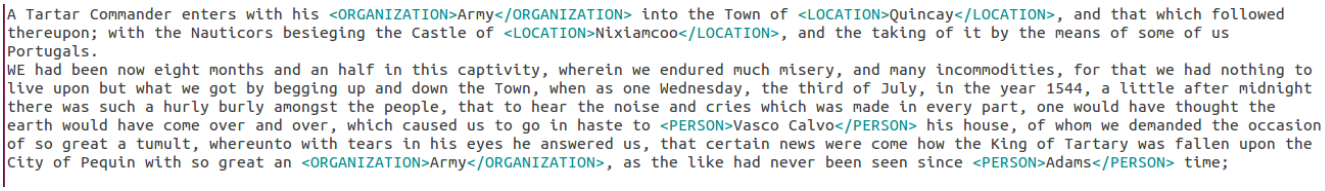


Figura 5.3: Captura do início do documento anotado pela ferramenta NERC.

Salvamos uma cópia do documento anotado pelo NERC como documento HTML para proceder à sua revisão (fig. 5.4).

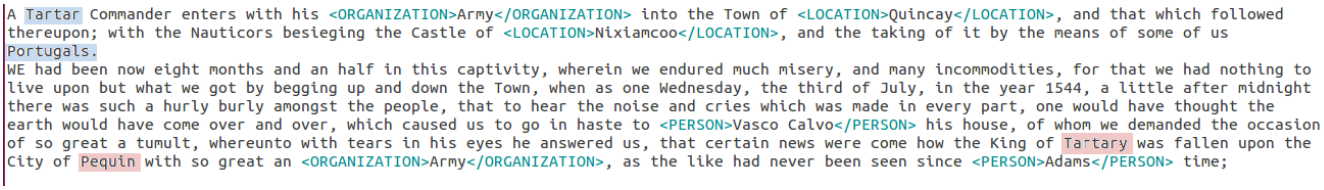


Figura 5.4: Início de revisão da anotação base. Identificação de falsos negativos (*Pequin* e *Tartary*) e elementos que queremos anotar como referências geográficas (*Tartar* e *Portugals*)

Adicionamos a marca <LOCATION></LOCATION> para as entidades geográficas mencionadas não anotadas pelo NER (falsos negativos) (vid. Fig.5.6). Anotamos também o atributo type="g" para os gentílicos: <LOCATION type="g"></LOCATION> (vid. Fig. 5.6) e o tag <LOCATION></LOCATION> nos tokens que não são entidade geográfica (falsos positivos).

```
In this room was the King of <LOCATION>Tartaria</LOCATION>, accompanied with many Princes, Lords, and Captains, amongst whom were the Kings of
<LOCATION>Pafua</LOCATION>, <LOCATION>Mecuy</LOCATION>, <LOCATION>Capinper</LOCATION>, <PERSON>Raina Benan</PERSON>, <LOCATION>Anchesacotay</
LOCATION>, and others to the number of fourteen, who in rich attire were all seated some three or four paces from the foot of the
<ORGANIZATION>Tribunal</ORGANIZATION>.
```

Figura 5.5: Exemplo de falso positivo.

Eliminamos as marcas de entidades mencionadas não geograficas (<ORGANIZATION> e <PERSON>) e falsos positivos. (Ver figuras 5.3 e 5.5 para exemplos). Obtemos o corpus final que usaremos como padrão. Salvamos uma segunda cópia adicionando a marca <B></B> para facilitarmos a visualização das entidades geográficas num navegador HTML.

```
A <LOCATION type="g">Tartar</LOCATION> Commander enters with his Army into the Town of <LOCATION>Quincay</LOCATION>, and that which followed
thereupon; with the Nauticors besieging the Castle of <LOCATION>Nixiamcoo</LOCATION>, and the taking of it by the means of some of us <LOCATION
type="g">Portugals</LOCATION>.
WE had been now eight months and an half in this captivity, wherein we endured much misery, and many incommodities, for that we had nothing to
live upon but what we got by begging up and down the Town, when as one Wednesday, the third of July, in the year 1544, a little after midnight
there was such a hurly burly amongst the people, that to hear the noise and cries which was made in every part, one would have thought the
earth would have come over and over, which caused us to go in haste to Vasco Calvo his house, of whom we demanded the occasion of so great a
tumult, whereunto with tears in his eyes he answered us, that certain news were come how the King of <LOCATION>Tartary</LOCATION> was fallen
upon the City of <LOCATION>Pequin</LOCATION> with so great an Army, as the like had never been seen since Adams time;
```

Figura 5.6: Início do documento com a anotação padrão.

5.6.6 Configuração dos ensaios

O apêndice (I) mostra todas as combinatórias de sistemas e parâmetros aplicados nos ensaios, assim como uma tabela com os resultados. Tanto os sistemas quanto o padrão apresentam distintas configurações.

Configuração dos sistemas

Usamos configurações padrão dos sistemas mais variantes em que modificamos dois parâmetros centrais da definição das entidades mencionadas. Para o modelo estatístico de treino, alteramos a sensibilidade de maiúsculas e minúsculas. Para os modelos de regras, modificamos as listas predefinidas adicionando a lista de entidades obtidas do corpus padrão da *Peregrinação* (1614) em duas modalidades: completa (topónimos e gentílicos) e só topónimos (sem gentílicos).

Configuração do padrão

O motivo para alternarmos as listas com e sem gentílicos é pesquisar possíveis efeitos da otimização dos sistemas pela consideração das entidades mencionadas como uma classe dentro dos nomes próprios. Para observarmos a incidência na anotação automática, o padrão oferece dois modos comparativos: com a anotação de todas as referências geográficas (topónimos e gentílicos) e uma outra só com topónimos.

5.6.7 Análise de resultados

5.6.7.1 Avaliação de resultados e análise de erros

Das anotações obtidas pelos sistemas, recuperamos apenas aquelas com a categoria de entidade geográfica e comparamos cada unidade anotada com as anotações do padrão. Avaliamos os

resultados conforme aos critérios em baixo.

• Se uma forma anotada pelo sistema coincidir plenamente com a anotação do padrão, ela somará como verdadeiro positivo (exemplos na tabela 5.4).

Forma anotada pelo sistema	Marca	Forma anotada no padrão	Ensaio	Positivo
Cauchenchina	LOCATION	Cauchenchina	Stanford_3_M	SIM
China	LOCATION	China	Stanford_3_M	SIM
Japan	NP00G00	Japan	Linguakit_D	SIM
Lantimay	NP00G00	Lantimay	Linguakit_D	SIM
Lingator	loc	Lingator	Geoparser_D	SIM
Luançama	loc	Luançama	Geoparser_D	SIM

Tabela 5.4: Resultados de coincidências totais.

• Se a expressão de uma entidade coincidir parcialmente com uma forma anotada no padrão, ela contará igualmente como verdadeiro positivo (exemplos na tabela 5.5).

Forma anotada pelo sistema	Marca	Forma anotada no padrão	Ensaio	Positivo
Kingdom of Pegu	LOCATION	Pegu	Stanford_3_M	SIM
Xinarau of Tartaria	LOCATION	Tartaria	Stanford_3_M	SIM
City_of_Pequin	NP00G00	Pequin	Linguakit_D	SIM
Country_Cunebetea	NP00G00	Cunebetea	Linguakit_D	SIM
city of Lingator	loc	Lingator	Geoparser_D	SIM
Cities of Luançama	loc	Luançama	Geoparser_D	SIM

Tabela 5.5: Resultados de coincidências parciais.

• Se uma entidade geográfica aparecer dentro de outro tipo de anotação sem marca de entidade geográfica ou sem anotar em absoluto, não contará como positivo ainda quando estiver anotada como entidade geográfica noutras ocorrências do corpus (exemplos na tabela 5.6).

Forma anotada pelo sistema	Marca	Forma anotada no padrão	Ensaio	Positivo
City of Pequin	ORGANIZATION	Pequin	Stanford_3_M	NÂO
King of Cauchenchina	ORGANIZATION	Cauchenchina	Stanford_3_M	NÂO
Siege_of_Pequin	NP00SP0	Pequin	Linguakit_D	NÂO
Isle_of_Ainan	NP00SP0	Ainan	Linguakit_D	NÂO
Panquinor	-	Panquinor	Geoparser_D	NÂO
Psipator	-	Psipator	Geoparser_D	NÂO

Tabela 5.6: Resultados de formas sem anotar ou anotadas dentro de outra categoria.

• O apêndice (I) mostra todos os resultados dos ensaios e um exemplo concreto de recuperação de entidades com verdadeiros e falsos positivos. O fator comum nos falsos positivos é que se trata de nomes cuja grafia contemporânea seria minúscula, mas têm maiúscula inicial no corpus. Assim temos nomes comuns como *Spades*, *Bavins*, *Kingdom*, *Champhire*, *Arras*, *Pagode*, *Falconets*, *Chappels*, *Pirot, Lake, Laulees*; expressões que, ainda tendo um valor de referente geográfico, não são propriamente uma entidade, tais como *Vantguard* e *East*; outras que aparecem perto de entidades geográficas mas são entidade mencionada de pessoa, como é o caso de *Anchesacotay* e alguns teónimos: *Pontimaqueu*, *Vanguenarau*, *Migama;* expressões usadas para se referir a pessoas e organizações que aparecem num rol semântico de lugar ou direção e precedidas por advérbios e preposições indicativos destes roles, assim *Tileymay*, *Councel*, *Mitaquer*, *Christendom* e *Broquem*. Finalmente temos um termo que é intercalação da língua Mongol, inserida no corpus sem marca nenhuma para a distinguir da língua alvo dos sistemas (*Pucau*). É salientável, como no caso da aplicação da lista de entidades por meio do sistema NERC para o corpus padrão, que uma entidade anotada como falso positivo, *Earth*, é mais uma vez um exemplo da disparidade de critérios na consideração do que é ou não uma entidade geográfica mencionada.

5.6.7.2 Valoração dos resultados dos ensaios

Para a anotação de um texto não normalizado em inglês obtivemos resultados por cima de 60% na medida-F, um patamar que se situa entre os melhores resultados em avaliações mesmo para o português em padrão contemporâneo (Amaral, Fonseca, Lopes, & Vieira, 2014). Ambos os dois grandes tipos de sistemas, de treino e de regras e listas, conseguiram superar a barreira de 60%, no entanto o sistema de treino tem o seu desempenho mais alto numa configuração predefinida, sem necessidade de melhoria nenhuma. No caso prático da anotação do corpus *Tartaria 1653*, o ponto de partida foi um texto sem anotar. O sistema de treino realizou uma anotação com resultados mais que suficientes para servir de base na elaboração do padrão (§5.6.5). No entanto, os sistemas de regras e listas aumentam o desempenho conforme a lista cobre o âmbito com que se relaciona o texto-alvo.

Ainda que não usamos uma lista *ad hoc*, é muito similar. O desempenho dos sistemas de listas aparece então condicionado pela disponibilidade de glossários ou listas específicas próximas às entidades mencionadas no corpus.

Os ensaios para a anotação só de topónimos oferecem resultados por cima dos da anotação de todas as variantes, mas em ambos os dois casos os melhores resultados se dão na faixa >60% e <70%. Dado que nenhuma das três ferramentas NERC usadas considera os gentílicos como entidades geográficas, surpreende um resultado tão próximo. Uma explicação é que, no corpus usado, os gentílicos são grafados com maiúscula inicial como os nomes próprios. No casos dos sistemas de regras e listas, ademais, temos que considerar que na lista acrescentada de todas as variantes introduzimos os gentílicos juntamente com os topónimos.

5.7 Conclusão da identificação de entidades geográficas

Enquadramos o problema da identificação de entidades geográficas como comum à Recuperação de Informação Geográfica, Sistemas de Informação Geográfica e Processamento da Linguagem Natural. A variedade de aproximações faz com que haja também divergências terminológicas, e subproblemas como a classificação e desambiguação têm distintas formulações segundo os objetivos últimos da identificação: a classificação em PLN consiste comummente em determinar um tipo de entidade mencionada (ex. local, pessoa, organização), mas em sistemas mais orientados à georreferenciação, em que se vista principalmente resolver as georreferências, pode chegar à atribuição um tipo de entidade geográfica (ex. ilha, rio, país, cidade).

Definimos duas atividades no processo de identificação: o reconhecimento e a classificação, ainda apontamos algumas dificuldades que complicam a resolução automática das entidades geográficas. Usamos o corpus da *Peregrinação* (Pinto, 1614) para analisarmos as dificuldades de aplicarmos uma lista de entidades geográficas, mesmo sendo esta *ad hoc*, concluindo na necessidade de estabelecer critérios que permitam resolver se a coincidência de um elemento da lista com um termo do corpus (reconhecimento) é ou não uma entidade geográfica (classificação de tipo geo / não geo).

Apresentamos soluções de anotação automática e medidas usadas para a sua avaliação. Na introdução das métricas de precisão, abrangência e medida-F, empregamos um sistema NERC de regras e listas: mesmo com a mesma lista de entidades extraídas do corpus, os resultados diferem do padrão. A análise das divergências permitiu recuperar entidades geográficas que não foram identificadas no processo de elaboração do corpus.

Considerando a disponibilidade das ferramentas, realizamos um teste com a tradução em inglês da *Peregrinação* (1653), selecionando os capítulos agrupados sob a unidade temática da Tartária. Conferimos os resultados da anotação de dois tipos de sistemas: estatístico duma parte, e de regras e listas da outra. As configurações dos sistemas foram avaliadas a respeito de duas anotações: todas as entidades geográficas, compreendendo gentílicos e topónimos, e só topónimos. Obtivemos resultados por cima de 60% para cada um dos cenários. A configuração do sistema de treino não requereu nenhum tipo de melhoria, os melhores resultados serviram de base à anotação de um

corpus. Nos sistemas de listas e regras o seu maior rendimento veio condicionado pelo tipo de listas usadas.

Consideramos que os sistemas de anotação automática são de aplicabilidade prática para a melhoria de um corpus já anotado, na pesquisa de entidades sem anotar, mediante o simples acrescentamento da lista do sistema com as entidades já anotadas no corpus.

Os ensaios com o corpus da Tartária mostram que a configuração de um sistema estatístico, mesmo sendo treinado em textos não relacionados, oferece rendimentos sastisfatórios para automatizar o processo inicial de anotação. Partindo das configurações predefinidas, o sistema de treino ofereceu um produto com que preparar o corpus padrão. Quando contamos com um glossário relacionado disponível, um sistema de regras obtém desempenhos que também permitem automatizar o processo inicial de anotação.

À vista das análises e resultados obtidos neste capítulo e no anterior (cap. 4) a nossa proposta para a anotação completa do corpus é:

1) início da anotação com sistema de treino ou de regras e listas no caso de contarmos com glossário relacionados (§5.6),

2) revisão (§5.6.5),

3) elaboração ou correção de listas específicas (§4.3.2),

4) melhoria e afinamento dos critérios de anotação por pesquisas sobre a lista (§5.4)

5) teste com sistema de regras e lista *ad hoc* (§5.5).

5.8 Sumário de objetivos

Os objetivos desta secção foram:

- Definir os processos aplicados para a identificação de entidades geográficas como parte do problema da georreferenciação.
- Analisar problemas e dificuldades encontrados na aplicação de uma lista de entidades geográficas para a anotação de um corpus e na melhoria de um corpus já anotado.
- Avaliar soluções para a identificação automática de entidades geográficas na anotação de um texto não normalizado.
- Propor possíveis melhorias para a anotação de textos não normalizados.

Capítulo 6

A referenciação da entidade geográfica mencionada

No capítulo anterior descrevemos a identificação das entidades geográficas mencionadas no corpus. Estudamos as limitações da aplicação de uma lista de entidades geográficas como simples coincidência de uma expressão e ensaiamos modos de identificação automática que facilitam o processo de anotação. Uma vez resolvida a identificação no texto, neste capítulo introduzimos um modelo semântico para o georeferenciamento das entidades geográficas mencionadas. Apresentamos uma proposta com o referente como objeto físico e o conceito um esquema em que inserimos um número finito de atributos e regras que denotam o objeto. O problema da georreferencição fica resolvido pela ligação da expressão da entidade geográfica mencionada com o referente, o objeto geográfico. Desenvolvemos duas possíveis soluções. Na primeira, o objeto é referido diretamente através de umas coordenadas que apontam o espaço físico. A segunda, quando não houver coordenadas prévias, resulta da elaboração do conceito, expressado como uma definição composta por um tipo geográfico e uma relação com uma outra entidade geográfica.

6.1 A entidade geográfica mencionada num modelo semântico

Uma caracterização semântica da entidade geográfica mencionada em termos de signo distingue um composto formado por:

1) um conceito (a ideia, pensamento, componente mental com que compreendemos a entidade, mais especificamente neste trabalho, o seu constructo),

2) uma expressão (mais comummente, o som, conjunto de caracteres ou símbolo),

3) o referente (o objeto físico no caso das entidades), e finalmente

4) o significado (o modo em que este é concebido, neste capítulo, de tipo referencial e verificável, dado pelo feito da existência do objeto (§3.4.1)).

Assumido um modelo referencial do significado, o referente determinará o valor dos predicados em função do grau de certeza que tivermos sobre o facto da sua existência. Nas secções a seguir desenvolvemos a aplicação de um modelo para o caso específico das entidades geográficas.

6.2 Proposta de modelo semântico para as entidades geográficas mencionadas

Considerando um modelo de significado referencial em que o foco é o referente, o objeto geográfico no mundo real, analisamos a ligação entre os distintos elementos implicados na construção do significado: a expressão, o conceito e o referente.

6.2.1 Os componentes da referenciação

6.2.1.1 Expressão: *Entidade geográfica mencionada*

A expressão é a materialização ou forma física de um elemento da linguagem sem considerar o seu significado. Assim, a expressão pode ser uma sequência de caracteres (no nosso corpus, caracteres UNICODE, por sua vez codificados como bits; sinais de tinta sobre um papel no original de 1614) ou as ondas sonoras emitidas ao pronunciar essa mesma sequência.

Ex. *Pequim* é uma expressão usada na *Peregrinação* para se referir a uma entidade geográfica.

A entidade geográfica mencionada é expressão enquanto que cadeia de caracteres num texto.

Seja assim:

W= {expressões das entidades geográficas mencionadas}

A lista com todas as expressões resultado de extrair as anotações das entidades geográficas mencionadas no corpus.

6.2.1.2 Referente: *Entidade geográfica*

O referente é a entidade geográfica como objeto físico.

> Ex. *Pequim* como a extensão de terreo no planeta Terra com centro aproximado em 39º 54' 20" N, 116º 23' 29" E.

Neste capítulo, para assinalar o referente, apontamos para a sua situação por meio de uma descrição que nos leve à entidade geográfica como objeto real no planeta Terra.

Uma forma simples de indicar o referente é mediante as suas coordenadas geográficas. Se estas não estiverem disponíveis, procuraremos uma descrição que permita localizá-lo dentro de um raio ou a uma distância aproximada de outro referente conhecido.

Chamemos de:

G = {referentes para as entidades geográficas mencionadas de W}

à lista que contém a descrição dos referentes das entidades geográficas mencionadas.

A relação que liga uma expressão com um referente é:

Tem_georreferente(w,g)= “*w* tem o georreferente *g*” (6.1)

em que *w* é o nome de uma entidade geográfica mencionada e *g* a sua situação expressada em termos o mais geograficamente exatos possíveis.

Para obtermos uma expressão associada a um referente *g* é necessária uma função:

$$\text{Topónimo}(g) = w \tag{6.2}$$

em que para um referente $g \in G$ obteremos uma expressão $w \in W$ quando se cumprir a relação Tem_georreferente(w,g) definida em (6.1).

A função inversa:

$$\text{Georreferente}(w) = g \tag{6.3}$$

recupera o referente a partir da expressão.

6.2.1.3 Conceito: *Definição da entidade geográfica*

Num modelo semântico o conceito costuma ser a representação mental, estrutura e atributos composicionais (Jakendoff 2010; Leech 1981, pp. 89-90) que categorizam uma entidade. Um modelo de conceito, como esquema formulado em termos de um número finito de atributos e regras, é desenvolvido por Jakendoff (2010, pp. 85-134) dentro de uma arquitetura que processa todo o lexicon e os seus componentes fonológicos e sintáticos. O referente é percebido através do sentido da vista como uma representação espacial geométrica parte do próprio conceito e não objeto físico externo. Nesta tese elaboramos o conceito também a partir de um esquema, composto de uns atributos (os tipos geográficos) e relações a modo de regras (*é_Parte_de*), mas o referente é sempre um objeto físico, a entidade geográfica, objeto na superfície do planeta Terra.

Uma forma simples de representar o conceito é mediante uma definição.

Ex. “Capital da China” é uma definição para a entidade geográfica mencionada *Pequim*.

Neste trabalho, a definição mais simples de uma entidade geográfica é um predicado diádico em que o duplo <c, a> interpreta a relação:

$$\text{é\_Parte\_de}(c,a) = \text{“}c \text{ é parte de } a\text{”} = \text{“}c \text{ de } a\text{”} = \text{“}c \text{ pertencente a } a\text{”} \tag{6.4}$$

em que

$$c \in C,\ C = \{\text{Lista de atributos geográficos}\}$$

e

$$a \in A,\ A = \{\text{Lista de entidades geográficas nomeadas}\}$$

Isto é, *c* confere um atributo e *a* estabelece uma relação da entidade geográfica relativamente a outra entidade geográfica.

Denominemos de:

$$D = \{\text{lista com as definições para as entidades geográficas mencionadas de W}\}$$

à lista com as definições para as entidades geográficas mencionadas de W.

Uma definição $d \in$ D virá dada por uma função *Define(w)* que, para uma expressão *w*, obtenha a relação entre um atributo e uma entidade geográfica com referente conhecido segundo (6.1)

$$d = \{ \text{Define(w)} \mid <c,a> \} \quad (6.5)$$

Ex. d_i = { Define(*Bardees*) | <*Barra*, *Goa*>} = "Barra de Goa"

6.2.2 A ligação entre a entidade geográfica mencionada e o objeto geográfico

O objetivo da georrefenciação é achar o referente da expressão. A esta relação entre a expressão e o objeto físico chamamos de *referência*. Segundo §6.2.1.2 a resolução da referência de uma entidade geográfica mencionada tem, no seu modo mais simples, forma de coordenadas. Na sua ausência, o conceito aplica atributos e regras para ligar a expressão com o objeto, representado aqui por meio de uma definição que descreva a entidade e represente algum modo de relação espacial com outro objeto geográfico. No trabalho de georreferenciação achamos situações de maior dificuldade, que requerem um conceito mais elaborado e com mais relações das consideradas nesta tese, ou cenários em que operamos com objetos geográficos para os quais não há ainda uma expressão. A seguir apresentamos as direções da ligação da expressão e o referente exemplificadas no caso prático da *Peregrinação*.

6.2.2.1 Da expressão diretamente ao referente

Por um lado, a ligação pode vir de modo direto, ostensivo, ao declararmos a expressão apontando para o objeto. É o equivalente ao que realizamos quando, perguntados pelo que é, seja por caso, um carvalho, em vez de o explicar, usamos uma vara para apontarmos uma árvore da espécie *Quercus robur* que temos defronte.

No caso das entidades geográficas mencionadas, um modo de ligarmos a expressão diretamente ao referente é apresentarmos as suas coordenadas. Volvendo ao exemplo do carvalho, as coordenadas seriam o posicionamento numa determinada direção da vara. A vara não é o objeto referido (aliás pode ser que a tivermos com objetivos muito distintos, por exemplo, realizarmos cálculos trigonométricos ou simplesmente para nos ajudar na caminhada), mas um instrumento que o assinala diretamente. Assim, as coordenadas geográficas permitem a geovisualização num SIG, sensores remotos ou mesmo a orientação no deslocamento para a observação *in situ*.

Chamamos de ligação da expressão diretamente ao referente ao modo de referenciar as entidades mencionadas que foram identificadas quando aplicamos uma lista de coordenadas.

Como exemplo, a referência para a entidade geográfica mencionada *Pequim* é resolvida pela função (6.3) que recupera as coordenadas a partir da expressão:

Georreferente(Pequim) = "Lat. 39.9075, long. 116.39723"

Isto é, a georreferência da expressão *Pequim*, é resolvida com g_{Pequim} = "Lat. 39.9075, long.

116.39723”.

6.2.2.2 Da expressão ao referente pelo conceito

Há uma outra relação entre a expressão e o objeto geográfico (referente) que passa pela sua concetualização. Neste trabalho usamos um esquema concetual onde o referente fica como objeto no mundo real. Assim a definição da entidade geográfica (6.5) a partir de um número reduzido de atributos geográficos permite ligar o conceito a uma representação geométrica mantendo o referente como objeto geográfico físico, representado em termos de coordenadas geográficas ou uma definição que descreva a sua espacialidade, ainda que esta fosse relativa. A relação entre o conceito e o referente como entidade no mundo real chamamos de **denotação**.

A aplicação do conceito também permite recuperar uma georreferência quando não temos uma entrada na lista de entidades ligada a umas coordenadas, no entanto, sim uma definição que denote o objeto. A georreferência nestes casos é a própria definição.

Georreferente(Queitor) = “Rio de Auaa”

Isto é, a georreferência da expressão *Queitor* é resolvida por $g_{Queitor}$ = “Rio de Auaa”.

Para o caso prático deste trabalho, é importante, primeiro, que o conceito usa apenas uma relação com valor espacial (*é_Parte_de*) para introduzir um modelo em que relações mais precisas (ex. Distância_a) podem ser adicionadas (mas não aparecem desenvolvidas nesta tese). E, segundo, mesmo dentro do conceito aplicado, há distintas possibilidades na resolução da definição. Assim:

Georreferente(Queitor) = “Rio de Ásia”

resolve uma definição válida para a entidade geográfica mencionada *Queitor*.

Não obstante, a definição que procuramos é aquela cujo valor denotativo se aproxime mais à unidade, preferencialmente, aquela que aponta um único referente. “Rio de Auaa” permite distinguir um único objeto geográfico, no entanto, “Rio de Ásia” denota o conjunto de rios do continente asiático. Recuperando o modelo referencial como determinador do nível de certeza sobre o feito da existência da entidade, a função *Georreferente* deve devolver a definição cuja denotação seja 1 ou otimizar a solução mais próxima quando o seu valor for >1.

6.2.2.3 Do referente ao conceito e a expressão

Um termo que descreve um ponto de partida do referente é *reverse geo-coding* (georreferenciação reversa) usado no âmbito tecnológico para a pesquisa do objeto e a sua área (§7.3.4.2). Esta situação temo-la na *Peregrinação* quando pesquisamos um topónimo (cuja forma atual desconhecemos) a partir de uma área ou itinerário com referentes identificáveis pela elaboração da sua descrição no corpus (distâncias, características físicas da paisagem).

Por exemplo, ao geovisualizarmos a área do objeto geográfico (§7.3.4) com expressão *Auaa* e coordenadas lat. 21.85479, long. 95.97635, achamos um rio. Ao pesquisarmos no mesmo recurso em que geovisualizamos o objeto qual é a sua expressão, obtivemos: *Irrawaddy*.

Topónimo(Lat. 21.86741, long. 95.98247) = Irrawaddy

O corpus mostra também objetos cuja expressão é recuperável através do conceito.

> "Quatorze dias auia ja que estas cousas erão passadas, nos quais o tyranno se occupou sempre em fortificar a cidade cõ grande presteza & cuydado, quando lhe chegou noua certa pelas espias que nisso trazia, que da cidade do ***Auaa*** era partida pelo rio de **Queitor** abaixo hũa armada de quatrocentas vellas de remo" (*PR*, 156) (6a)

Em (6a) a função resolve:

Topónimo("Rio de Auaa") = Queitor

Assumindo o referente como centro do modelo, é importante notar que tanto "Rio de Auaa" como "Lat. 21.86741, long. 95.98247" georreferenciam uma mesma entidade, assim:

Topónimo("Rio de Auaa") = Irrawaddy

é uma solução válida (o referente aparece ligado a mais de uma expressão). Como também vimos para *Georreferente*, a aplicação prática da função obriga a considerar critérios de escolha nos casos de mais de uma possível solução. Nas alíneas a seguir consideramos as situações de ambiguidade derivadas destas situações.

6.3 A desambiguação da expressão e do referente

Dada a expressão de uma entidade geográfica mencionada, usamos o termo *georreferenciação* para designar o processo de resolução de um referente geográfico seja pela referência ostensiva (§6.2.2.1) ou pela denotação do referente através de um conceito (§6.2.2.2). A relação entre a expressão e o referente pode apresentar situações de ambiguidade: as entidades geográficas mencionadas são designadores rígidos, isto é, denotam um único referente, o qual facilita a sua referenciação comparativamente com os nomes comuns (§3.2.2), não obstante, no trabalho com um corpus achamos uma série de dificuldades produto quer das múltiplas ocorrências de uma mesma expressão, quer das múltiplas referenciações de um mesmo objeto. A tabela 6.1 introduz as situações derivadas do valor múltiplo de um dos termos da relação que desenvolvemos nas subsecções em baixo.

Expressão	Referente	A resolver
1	>1	Homonímia
1	>1	Polissemia
1	>1	Metonímia
>1	1	Anáfora (correferência)
>1	1	Lexema
>1	1	Sinonímia

Tabela 6.1: Valores da relação expressão – referente e problemas a resolver.

Procedendo segundo um método cíclico (§4.2), um mesmo problema pode ter várias soluções, aparecer em distintas fases do trabalho com o corpus, ou ser requerido para um fim definido a posteriori. Nesta secção apresentamos os problemas e as soluções adotadas na criação de uma primeira lista de referentes que interpretam a relação *Tem_georreferente(w,g)* = "*w* tem o georreferente *g*" (6.1) para o corpus *Peregrinação 1614*.

Como ponto de partida, assumimos que todas as entidades geográficas anotadas são recuperáveis pela função (6.2), isto é, são georreferenciáveis, quer por coordenadas geográficas, quer por elaboração de um conceito que denote a georreferência.

Na exposição dos casos, introduzimos em primeiro lugar os exemplos e ilustramos a dificuldade que apresentam, a seguir, e derivado da exposição, definimos o problema em termos dos componentes do modelo semântico, a intuição usada para a sua abordagem e a solução adotada para a anotação do corpus.

6.3.1 Uma expressão tem mais de um referente geográfico

Em (§5.2.4) introduzimos a desambiguação geo / geo como problema a resolver uma vez classificada a expressão como entidade geográfica mencionada. Encontramos este tipo de ambiguidade quando a expressão tem mais de um referente. A tipologia que propomos usa termos procedentes do estudo do léxico a partir de casos práticos extraídos do corpus da *Peregrinação 1614*.

6.3.1.1 Homonímia e homografia

Um caso típico é o processo de classificação de expressões identificadas como geográficas nos processos de georreferenciação automática. Tomemos como exemplo a expressão *Goa*.

Expressão	País	Latitude	Longitude
Goa	Índia	N 15° 20' 0"	E 74° 5' 0"
Goa	Filipinas	N 13° 41' 52"	E 123° 29' 21"
Goa	Botswana	S 18° 17' 0"	E 21° 50' 0"
Goa	Burkina Faso	N 12° 36' 0"	W 2° 54' 0"
Goa	Chade	N 9° 47' 53"	E 15° 56' 17"
Goa	Noruega	N 58° 59' 0"	E 5° 38' 0"
Goa	Rússia	N 41° 45' 12"	E 47° 42' 3"

Tabela 6.2: Entidades geográficas com expressão *Goa.*

Uma pesquisa na base de dados geográfica GeoNames captura, entre outros, os referentes associados à expressão *Goa* mostrados na tabela 6.2.

Para o caso da *Peregrinação 1614* a expressão *Goa* tem como referente uma cidade na Índia com

coordenadas 15° 20' 0" N, 74° 5' E em todas as ocorrências. Denominamos de desambiguação geo / geo (§5.3.4) ao problema de determinar qual dos referentes da tabela 6.2 se corresponde com a entidade mencionada no corpus.

Dizemos que estas expressões coincidentes têm uma relação de **homonímia** quando as escrevemos e pronunciamos igual, são **homógrafas** se apenas coincidem na escrita.

Quando listarmos formas homónimas, operamos com entidades geográficas distintas e, assim sendo, entradas não relacionadas.

Problema: Como selecionar o referente quando temos uma lista com expressões homónimas?

Intuição: Num texto unitário, os casos de homonímia são exceção.

Solução adotada na anotação do corpus: Partimos de um princípio de coerência, assumimos que uma expressão tem um mesmo referente em um texto unitário. Reduzimos assim o problema à definição de um único objeto geográfico comum para todas as ocorrências da mesma expressão no corpus. Num corpus como o nosso caso de estudo, os exemplos de homonímia registados sendo mínimos e com frequências baixas, a revisão manual do resultado da anotação automática baixo o princípio de coerência resultou mais efetiva que a aplicação de novas regras de procedimento para as exceções.

6.3.1.2 Polissemia

Achamos um caso mais problemático, assimilável à homonímia, nas seguintes concordâncias para *Babylonia* extraídas do corpus da *Peregrinação 1614:*

> “No cabo dos tres meses prouue a nosso Senhor que receoso elle que por ser insofriuel perdesse o que dera por mim, como algũs seus vizinhos lhe tinhão ja dito, me vendeo a troco de tamaras por preço de doze mil reis a hum Iudeu por nome Abrão Muça, natural da cidade do Toro, duas legoas & meya do monte Sinay, o qual em hũa Cafila de mercadores que partio de **Babylonia** para Cayxem me leuou a Ormuz” (*PR*, 6) (6b)

> “E mandandoo sayr para fora da tenda se praticou sobre a resolução deste feito, em o qual por peccados nossos se não tomou nenhũa, por auer nesta junta tantas diuersidades de opinioẽs & de pareceres, que **Babylonia** em seu tempo não lançou de sy mais variedades de lingoas” (*PR*, 148) (6c)

Em (6b) temos uma cidade em Médio Oriente com atividade comercial, segundo DHDP s.v. *Babilónia* é a forma usada nas fontes dos descobrimentos para o sultanato do Cairo e referir-se-ia originalmente à cidade de Heliopolis em 30° 07' 46.3" N, 31° 17' 20" E. No entanto, em (6c) a mesma expressão tem como referente a cidade do mito de Babel, com referente segundo o glossário GT (Lagoa, 1950-1953) em 32º 28' N, 44º 48' E.

Neste caso temos dois referentes distintos, duas cidades situadas em pontos distantes que existiram

em distintos períodos da história. Do ponto de vista do significado referencial são tão diferentes como os casos estudados para *Goa* na tabela 6.1, um caso de homonímia que invalida a solução inicial proposta por intuição em §6.3.1.1. No entanto, da perspectiva da expressão, não temos duas formas coincidentes de origens distintas, mas uma mesma expressão que por uma similitude concetual ("capital de um império de Médio Oriente") passa a designar um novo objeto (Lakoff & Johnson, 1980). No campo do léxico, quando temos uma similitude no significado e uma mesma origem da expressão, dizemos que estamos perante uma relação de **polissemia** (Lyons, 1995, p. 58). Se por significado se entende apenas o referencial, em (6b) e (6c) temos referentes totalmente distintos, este é mais um caso de homonímia. Mas, se considerarmos o facto de que em ambos os casos há uma origem comum na expressão e uma única etimologia, e aceitarmos uma definição intensiva do referente como significado ("capital de um império de Médio Oriente") procederá aplicar a noção de polissemia.

Problema: Quando listamos formas polissémicas, operamos com uma mesma expressão e consideramos distintos referentes. Como indicar que uma mesma expressão tem mais de um sentido?

Intuição: a solução final requer uma listagem de conceitos que relacione referentes distintos em base à sua concetualização.

Solução adotada na anotação do corpus: Se aplicarmos a mesma solução que para a homonímia em §6.3.1.1., uma só expressão sem formas homónimas, teremos que aceitar um nível de erro proporcional às ocorrências com um referente distinto ao selecionado como mais provável. Não obstante, dado o carácter excecional de esta situação, a revisão da anotação automática sob o princípio de coerência também aplicado para a homonímia foi mais efetiva que a aplicação de novas regras para o tratamento da exceção. A efeitos práticos do corpus, o caso de polissemia tem a mesma solução que a homonímia.

6.3.1.3 Metonímia

Sejam as concordâncias da expressão *Aarù***:**

> "E partidos deste porto de Panaajù, chegamos cõ duas horas de noite a hũ ilheo, que se dizia Apefingau, obra de hũa legoa & meya da barra, pouoado de gẽte pobre, que viue pela pescaria dos saueis, de que, por falta de sal, não aproueitão mais que sòs as ouas das femeas, como nos rios de **Aarù**, & Siaca, nestoutra costa do mar mediterraneo." (*PR*, 18) (6d)

> "E embarcandome hũa terça feyra pela menham cinco dias de Outubro do anno de 1539. continuey meu caminho até o Domingo seguinte que cheguey ao rio de Puneticão, onde està situada a cidade de **Aarù**." (*PR*, 21) (6e)

> "E tornando outra vez a auer conselho sobre a determinaçaõ deste negocio, se

> assentou que por rodas as vias lhe fizesse guerra como a inimigo capital, & se entendesse logo primeyro que tudo em se tomar o reyno de **Aarù**, & a fortaleza de Puneticão, antes que o Achem o fortificasse mais." (*PR*, 32) (6f)

Em (6d) temos como referente um rio, em (6e) uma cidade e em (6f) um reino, todos três ligados à expressão *Aarù*. Em todos os casos supomos uma mesma expressão com uma mesma origem, portanto, segundo os critérios apontados para a polissemia em §6.3.1.2, temos formas polissémicas. Já agora, os significados referenciais, ainda tendo coordenadas geográficas distintas, apresentam uma relação particular. O reino (6f) cobre uma área que abrange as coordenadas geográficas dos referentes do rio (6d) e a cidade (6f). Neste caso particular falamos de uma relação de **metonímia** em que uma parte aparece pelo todo, ou o todo por uma parte, ou há uma relação de continuidade entre as partes.

Uma forma de sistematizar esta relação, aplicada aos nomes comuns, é a **meronímia** (§3.4.3.3), em que um termo entra em relação com um todo. À parte denominamos de **merónimo** e ao todo de **holónimo**. Assumimos que as expressões neste caso são complexas e compreendem também a frase preposicional que precede o topónimo. Assim, *cidade de Aarù* é um merónimo de *reyno de Aarù* e, na outra direcção, *reyno de Aarù* é um holónimo de *cidade de Aarù*.

Problema: Quando listamos formas polissémicas numa relação parte-todo operamos com uma mesma expressão e referentes relacionados. A expressão de distintos referentes aparece condicionada a expressões não toponímicas e contextos muitas vezes ambíguos. Convém indicar todos os referentes possíveis?

Intuição: A especificação de um referente parte de um todo é um problema de escala. A solução final depende da concetualização e do nível de resolução (granularidade) que se quiser aplicar no georreferenciamento.

Solução adotada na anotação do corpus: consideramos unicamente o todo como referente.

6.3.2 Um mesmo referente tem mais de uma expressão

Também o referente pode ser ambíguo, no sentido de ter distintas expressões para designá-lo. Surge assim o problema de escolher qual das expressões é mais operativa (§6.3.2.1) e a necessidade de agruparmos as expressões que se referem a um mesmo referente (§6.3.2.4). Os problemas derivados da relação entre o referente e a expressão afetam, por um lado, a anotação, determinando a escolha das unidades léxicas e como têm de ser anotadas e, por outro, ao indexado das entidades geográficas (cap. 9). A designação deste tipo de dificuldades como desambiguação do referente procede da aplicação do nosso modelo (§6.2), mas são também abordadas dentro do problema NERC (§5.2.4).

6.3.2.1 Anáforas (correferências)

Uma entidade geográfica mencionada, enquanto que designador rígido, denota um só referente, no entanto um mesmo referente pode ser referido no mesmo texto com distintas expressões, não necessariamente nomes próprios.

Na concordância (6g) para *Pequim* destacamos as expressões que têm um mesmo referente, a cidade capital da China, com coordenadas 9° 54′ N, 116° 23 E.

> "(…) me pareceo conueniente dar algũa pequena informação desta cidade do **Pequim**, que com verdade se pode chamar **metropoli da Monarchia do mundo**, & de algũas cousas que n**ella** notey, assi da abastança, policia, & grandeza d**ella**, como do regimento & grande gouerno da sua justiça" (*PR*, 105) (6g)

De quatro referências, apenas uma é topónimo, nos outros três casos temos um nome comum com um modificador frase preposicional e nas outras duas uma mesma expressão, um pronome pessoal (deítico). Às expressões que aparecem salientadas depois do nome próprio em (6g), substituíveis pelo topónimo da entidade geográfica mencionada, denominamos de **referências anafóricas** ou **correferências**.

Nesta tese, a concordância de um nome próprio é um segmento com unidade de sentido que abrange a oração segundo a segmentação aplicada em §4.3.3, portanto, as referências anafóricas não recuperadas ao pesquisar no corpus pelo topónimo serão aquelas que estiverem afastadas mais de uma oração.

Problema: as referências anafóricas podem conter elementos que permitam a referenciação quando a entidade geográfica mencionada é desconhecida na lista.

Intuição: O segmento aplicado para a recuperação de concordâncias pretende abranger como mínimo uma unidade que permita a compreensão do contexto mais direto da entidade, por isso, mais que recuperar a anáfora (que pode não existir), é mais fácil alargar o contexto dentro do capítulo.

Solução adotada na anotação do corpus: a análise das anáforas forma parte do contexto da entidade geográfica mencionada. Intuímos o segmento recuperado pela concordância, a oração, suficiente para os objetivos da georreferenciação no caso prático. Quando pretendemos ampliar o contexto, processamos a unidade maior, neste caso o capítulo, também considerado unidade operativa no corpus (§4.6.1). Como unidade léxica operamos unicamente com entidades geográficas mencionadas com forma de topónimos e gentílicos.

6.3.2.2 Lexemas e variantes

As formas agrupadas sob um mesmo lexema (§3.3.3), além de partilharem significado lexical, têm em comum boa parte da expressão. Um lexema representa variantes de um mesmo tipo de expressão (quando houver diferença morfológica muito forte ou total falaremos de formas irregulares e supleção).

No caso dos nomes próprios de lugar, nas listas de topónimos e índices geográficos, uma forma padrão aparece associada a umas coordenadas geográficas e pode ir acompanhada de variantes adicionais, formas abreviadas ou longas, traduções noutras línguas de uso comum, históricas ou gráficas quando houver vários sistemas de escrita. Os gentílicos são, porém, formas indexadas num

dicionário junto com o resto do vocabulário de um idioma.

Seja o exemplo:

> "o **Sornau** de **Odiaa**, que se intitula Rey de **Sião**, cujo senhorio cõfina por distãcia de setecentas legoas de costa, como he de **Tanauçarim** a **Champaa** cos **Malayos** & **Berdios** & **Patanes**" (*PR*, 124) (6h)

Em (6h), os gentílicos *Malayos*, *Berdios* e *Patanes* contribuem para situar o reino de *Sião*: o valor georreferenciador, a sua frequência, a aparição em clusters junto com topónimos, aconselha o seu processamento junto com as entidades geográficas mencionadas por um nome próprio. A similitude entre o nome próprio e o gentílico, e o facto de apenas se precisarem operações morfológicas facilmente sistematizáveis para chegar ao topónimo quando este é regular (reduzir o afixo por sufixação e flexões de número e género se as houver) apresenta o lexema como unidade válida para definir a forma que representa tanto o gentílico quanto o topónimo como variantes de um mesmo tipo.

Na lista de expressões temos ainda outro tipo de variação que convém agrupar, são os exemplos referidos em §5.2.5.1 como variantes ortográficas, gralhas e abreviações. Os sinónimos constituem um caso diferente que analisamos em §6.3.2.3.

Fica assim um conjunto de expressões, variantes de um mesmo tipo, com um único referente geográfico, cuja variação é devida a processos morfológicos sistematizáveis (gentílicos) ou a causas mais aleatórias (gralhas, inconsistências editoriais) ou desconhecidas nesta altura, produto da transcrição limitada pelos recursos gráficos e fonéticos do português para representar fenómenos das línguas asiáticas de onde chega o topónimo (tais como tonalidade, quantidade e qualidade vocálica, pontos e modos de articulação consonântica não recuperáveis no nosso estado de conhecimento atual da codificação no texto).

Como exemplo as concordâncias:

> "a que os escritores Chins, Siames, Gueos, **Elequios** nomeão nas suas geografias por pestana do mũdo" (*PR*, 1) (6i)

Em (6i) temos uma gralha editorial. A edição correta seria "a que os escritores Chins, Siames, Gueos **e Lequios** nomeão...". No entanto, se queremos respeitar o texto da edição original, e por ter ficado, embora erro evidente, não corregido nas edições modernas, marcamos *Elequios* como variante.

> "ou por seu respeito virmos nos a perder toda a banda do Sul, como he Malaca, Banda, Maluco, Çunda, Borneo, & Timor, a fora no Norte, a China, Iapaõ, **Lequios**, & outras muytas terras & portos" (*PR*, 26) (6j)

Em (6j) temos uma expressão, *Lequios*, pelo contexto um nome próprio. A sua função como entidade geográfica mencionada é clara, ao formar um cluster de topónimos. No entanto:

> "hia de veniaga como mercador que era para a ilha dos **Lequios** a fazer sua fazenda"

(*PR*, 41) (6k)

Em (6k) temos a mesma expressão, agora parece um gentílico, a sua função de georreferência fica, porém, igualmente óbvia.

> “& todas as mais senhoras **Lequias** nos vierão logo ver” (*PR*, 143) (6l)

Em (6l) temos o gentílico com género feminino, o qual contribui para interpretar as formas em (6j) e particularmente (6k) como possíveis gentílicos.

> “na ilha da Iaoa, Pangor no **Lequîo**, Vzanguee no graõ Cauchim, Lançame na Tartaria, & Miocoo em Iapaõ, as quais cidades todas saõ metropolis de grandes reynos” (*PR*, 107) (6m)

Em (6m) temos um masculino singular, uma vogal com acento circunflexo cujo valor ou causa desconhecemos (pode ser gralha ou representar um valor fonético distinto).

> “Esta ilha **Lequia** jaz situada em vinte & noue graos, tem duzentas legoas em roda, sessenta de cõprido, & trinta de largo.” (*PR*, 143) (6n)

Em (6n) temos uma forma que concorda em género com o nome comum que a precede, e se comporta como um gentílico. O valor de georreferência é o mesmo que um nome próprio.

> “Tem mais toda esta terra do **Lequio** muyto ferro, aço, chumbo, estanho, pedrahume, salitre, enxofre, mel, cera, açucar, & grande quantidade de gengiure muyto milhor & mais perfeito que o da India.” (*PR*, 143) (6o)

Em (6o) temos uma forma similar a (6m) mas sem o acento. Desconhecemos a causa da variação a respeito de (6m) e se esta representa algum valor fonético.

Todos os exemplos (6i) a (6o), *Elequios*, *Lequios* (em contexto de nome próprio), *Lequios* (em contexto de gentílico), *Lequias*, *Lequîo*, *Lequia*, *Lequio*, mostram expressões relacionadas.

No caso de estudo deste trabalho chamamos de lexema a aquela forma que representa estas expressões de um mesmo referente cuja similitude formal é evidente, seja qual seja a causa da variação**.** A sua noção recolhe o uso assimilável no trabalho com corpus ao lema (agrupa as variantes numa anotação), e o da tradição lexicográfica (termo que representa formas relacionadas num índice ou dicionário), ainda quando as expressões agrupadas no caso do corpus não se ajusta ao tratamento convencional dos gentílicos em português (representados por um lexema independente do nome próprio). As expressões agrupadas baixo um mesmo lexema chamamos de **variantes**.

Problema: Qual a expressão para representar o lexema?

Intuição: Dada a diversidade de línguas a considerar, de aglutinantes (mongol) a analíticas (chinês) e flexivas (português), em muitos casos não identificáveis, *aplicar uma regra de derivação para o gentílico não garante a recuperação do topónimo*, por não haver elementos evidentes que delimitem a afixação. O gentílico fica assim reduzido a uma variante com a mesma consideração

que as variantes do nome próprio do topónimo.

Solução adotada na anotação do corpus: O valor inicial do lexema (a sua expressão) será a variante com a frequência mais alta no corpus. Isto é assim ao objeto de aplicar um mesmo critério para todas as entidades geográficas mencionadas dado que, em princípio, não é possível determinar critérios morfológicos para a recuperação da raiz nas expressões de línguas desconhecidas.

A lista de variantes e lexemas vem dada no processo de elaboração do corpus como comentado no capítulo 4 deste trabalho.

6.3.2.3 Sinonímia

Um mesmo referente pode ter duas expressões com lexemas distintos que o mencionem. Seja primeiro um exemplo onde o referente se explicita no texto:

> "o Nautaquim principe desta ilha Tanixumaa, & senhor de nossas cabeças manda & quer que todos vós outros, & assi os mais que habitão a terra dantre ambos mares honrem & venerem este **Chenchicogim** do cabo do mundo" (*PR*, 105) (6p)

> "os dias passados me certificarão homens que vierão dessa terra que tinheis nessa vossa cidade hũs tres **Chenchicogins** do cabo do mũdo" (*PR*, 135) (6q)

Em (6p) e (6r) temos o gentílico *Chenchicogim* para se referir a aquele ou aqueles que vêm "do cabo do mundo", isto é, *Portugal*.

A equivalência de duas expressões para um mesmo referente não tem por que ser explícita:

> "a esta minha cidade Fucheo veyo a mim de seu mandado Fernão Mendez Pinto com hũa carta de sua real senhoria, & hum presente de armas & de outras peças muyto agradaueis a minha tenção, que muyto estimey por serem da terra do cabo do mũdo por nome **Chenchicogim**, onde por poderio de armadas muyto grossas, & exercitos de gentes de diuersas naçoẽs reyna o lião coroado do grande **Portugal**" (*PR*, 225) (6r)

Em (6q) temos *Chenchicogim* e *Portugal*, duas expressões para se referir ao mesmo referente geográfico. *Chenchicogins*, um gentílico equivalente a *Portugueses* aparece em (6r) sem nenhum elemento que permita estabelecer a equivalência.

A relação entre duas expressões não variantes de um mesmo lexema, mas com um mesmo referente geográfico chamamos de **sinonímia** (§3.4.3.1).

Problema: sendo os sinónimos expressões distintas, não recuperáveis uma pela outra, isto é, com lexemas diferentes, devemos incluí-los como entradas independentes num índice de entidades geográficas mencionadas, não obstante, se a recuperação da lista vier do referente, ambas expressões serão equivalentes de um modo similar às formas relacionadas baixo um mesmo lexema.

Solução adotada na anotação do corpus: parelha à categoria de lexema, que considera a relação de expressões similares (variantes) para um mesmo referente, procede operar com outra categoria para

todas as expressões de um mesmo referente, independentemente de existir entre elas similitude formal na expressão ou não (isto é, podem ser lexemas distintos sem nenhum parecido na forma). O representante de um referente que agrupa todas as variantes abrangidas pelos seu lexemas (se houver mais de um) chamaremos de **expressão representativa** quando a expressão usada estiver presente no corpus. À forma normalizada correspondente com um uso contemporâneo, chamamos de **nome padrão**. Para o caso do corpus, a escolha da expressão representativa segue critérios baseados na frequência (§9.4).

Do ponto de vista do referente, as formas sinónimas são equivalentes e, portanto, agrupáveis sob um mesmo nome padrão. Assim, numa lista de objetos geográficos, todas as formas, sejam variantes ou sinónimos, aparecem num mesmo grupo (ex. base de dados geográfico tipo GeoNames). Não obstante, quando foi criado um índice (§9.4) em que a consulta se faz através da expressão (ex. índice alfabético num atlas), as formas não relacionadas baixo um mesmo lexema são recolhidas como entradas independentes para possibilitar a sua pesquisa.

6.4 Conclusão

Chamamos de referência à relação entre a expressão e o objeto da entidade geográfica. O objeto geográfico, ente físico, é o referente da expressão, esta última mais frequentemente chamada (em rigor, de modo ambíguo, porém comum na literatura NERC) de entidade geográfica mencionada.

Tanto na expressão quanto no referente achamos situações de ambiguidade que dificultam a georreferenciação. Assim, uma mesma expressão pode ser usada para referir objetos geográficos distintos, o caso mais comum a homonímia. Objetos geográficos distintos podem convergir num mesmo espaço partilhando um mesmo topónimo, caso da metonímia. Nestas situações é preciso desfazer a ambiguidade, indicando qual é o referente físico da expressão. Para o caso do corpus objeto de estudo nesta tese, dada a sua excecionalidade, resolvemos mediante a anotação manual, assumindo para o conjunto um princípio de coerência.

Outro caso de ambiguidade, neste caso na escolha de uma expressão para o referente, aparece quando o mesmo objeto geográfico tem distintas expressões para o referir. Quando há similitude formal dizemos que são variantes e agrupamo-las sob uma mesma forma comum, o lexema. Se as expressões não tiverem nenhuma relação formal, mas partilharem um mesmo referente, consideraremos uma situação de sinonímia. O nome padrão de uma entidade geográfica é aquela expressão escolhida como representativa de todos os lexemas sinónimos (e as suas variantes) que a mencionam, se os houver. Aparece assim como a entrada principal num índice, atlas ou lista de entidades geográficas. Não obstante, os índices podem incluir formas alternativas (particularmente sinónimos) como entradas independentes para permitir a sua recuperação alfabética.

6.5 Sumário de objetivos

Objetivos desta secção foram:

- Introduzir um modelo semântico para georreferenciar as entidades geográficas mencionadas do corpus.
- Distinguir dois tipos de georreferência possíveis dentro do modelo. Um, direto da expressão ao referente, aponta o referente por coordenadas. Outro, elaborado através do conceito, denota o referente mediante uma definição.
- Considerar o modelo como uma primeira aproximação à georreferenciação com que observamos as dificuldades e problemas derivados da ligação entre a expressão e o referente.

Capítulo 7

A georreferenciação por conhecimento prévio

Uma vez identificada a entidade geográfica mencionada e desambiguada para um único referente, desenvolvemos o georreferenciamento. Com esta finalidade, apresentamos uma distinção epistemológica que classifica, por uma parte, aquelas entidades mencionadas cuja expressão conhecemos com anterioridade e cujas concordâncias no corpus são suficientes para lhe acharmos um referente no globo sem ambiguidade, de maneira que obtemos coordenadas de longitude e latitude. Pela outra parte, introduzimos neste capítulo e mais desenvolvemos nos seguintes, aquelas entidades cujo conhecimento se limita à descrição recuperada no corpus, ou cujo conhecimento prévio entra em contradição com o referente tal e como é descrito no corpus até ao ponto de não podermos definir coordenadas com toda probabilidade sem realizarmos uma análise mais exaustiva. Uma vez estabelecida esta distinção, o resto do capítulo resolve as entidades previamente conhecidas a partir do caso prático.

7.1 Conhecimento prévio e conhecimento adquirido

É importante notar que, no nosso modelo (§6.2), ligamos a expressão a um referente objeto físico, entidade no mundo real. Esta ligação pode produzir-se diretamente da expressão ao referente (§6.2.2.1) ou através de um conceito (§6.2.2.2). Neste último caso, segundo a definição em (§6.2.1.3), um conceito requer uma outra entidade geográfica nomeada com a que estabelece uma relação de pertença (meronímia) que em última instância, dado o seu carácter hierárquico, permitirá resolver um holónimo definido em termos de coordenadas geográficas a que a entidade geográfica inicialmente desconhecida pertence.

Não conhecermos, portanto, as entidades geográficas mencionadas do mesmo modo. Há umas entidades que nos servem para situarmos o referente de outras, apontando para umas coordenadas conhecidas com anterioridade. Consideramos assim dois tipos de conhecimento. O **prévio**, ou que vem dado, georreferencia aquelas entidades conhecidas antes de serem mencionadas no texto. O **adquirido por descrição** (Russel, 1905, p. 479), elabora uma georreferência a partir do contexto em que a entidade é mencionada.

7.1.1 Entidades geográficas referenciadas por conhecimento prévio

Deste modo, quando encontramos a expressão *Pequim* por primeira vez no texto, temos um conhecimento prévio, independente da sua ocorrência no texto. Tem um referente que damos como certo e podemos localizar em termos coordenadas geográficas. Num atlas convencional obtemos

segundo (6.3):

Georreferente(Pequim) = “39º 54' 20" N, 116º 23' 29" E”

Temos uma relação direta que liga expressão e referente, cumpre (6.1):

Tem_georreferente(Pequim, 39º 54' 20" N, 116º 23' 29" E) = “Pequim tem georreferente 39º 54' 20" N, 116º 23' 29" E”

Este conjunto de entidades que assumimos como georreferenciadas *a priori* chamamos entidades geográficas referenciadas por conhecimento prévio e designaremos como G'.

G' é um subconjunto da lista de referentes definida em §6.2.1.2.

G' = {lista de entidades geográficas referenciadas por conhecimento prévio} $\subset$ G

O valor de $g \in$ G' são coordenadas geográficas definidas em valores de latitude e longitude.

Regra de georreferenciação 1*: O valor inicial do referente* g *para uma expressão* w *que denotamos por conhecimento prévio são as suas coordenadas geográficas conhecidas*.

7.1.2 Entidades geográficas referenciadas por descrição

Por outro lado, temos aquelas outras expressões que chegam a nós pela primeira vez ou cujo conhecimento é limitado.

Seja o exemplo:

> “E embarcandome hũa terça feyra pela menham cinco dias de Outubro do anno de 1539. continuey meu caminho até o Domingo seguinte que cheguey ao rio de **Puneticão**, onde està situada a cidade de **Aarù**.” (*PR*, 21) (7a)

Puneticão não é um nome que achemos num atlas ou SIG, portanto não temos um referente com umas coordenadas geográficas que possamos recuperar com a função *Georreferente* (6.3). No entanto, podemos operar com um conceito.

Pela descrição em (7a) elaboramos uma definição a partir do duplo <rio, Aarú>.

Segundo (6.5):

$d_{\text{Puneticão}}$ ={Define(Puneticão) | <rio, Aarú>} = “rio de Aarù”

Isto é, a pesar de não termos um referente interpretado em termos de coordenadas geográficas exatas, é possível inferir um conceito que denote um objeto geográfico mencionado pela expressão *Puneticão.*

O subconjunto de referentes conhecidos por descrição receberá o nome de G".

G" = {lista de entidades geográficas referenciadas por descrição} $\subset$ G

O valor de $g \in$ G" é a sua definição como conceito (§6.2.1.3).

Regra de georreferenciação 2*: O valor inicial do referente* g *para uma expressão* w *que denotamos por conhecimento descrito é a definição do seu conceito*.

7.2 A georreferenciação por conhecimento prévio

O uso de conhecimento prévio é recurso comum nos sistemas NERC. Nos sistemas de listas e regras, vimos a eficácia do sistema vir muito condicionada pela lista (uma forma de conhecimento prévio também). No trabalho de georreferenciação, é preciso, ademais de a identificar, relacionar a entidade geográfica mencionada com coordenadas geográficas. Um exemplo de aplicação que, para além da identificação, georreferencia e até projeta cartograficamente as entidades georreferenciadas, é o Edinburgh Geoparser (§5.6.2.3; Grover, Tobin, Byrne, Woollard, Reid, Dunn, & Ball, 2010). Podemos, portanto, considerar a aplicação de um sistema NERC para obter as coordenadas geográficas de um topónimo a partir de uma lista como sendo uma forma de georreferenciação por conhecimento prévio.

No caso da *Peregrinação 1614* temos a dificuldade de não dispormos de grandes corpora digitalizados com que treinar um sistema, nem listas especializadas que cubram as particularidades linguísticas e geográficas requeridas pelo corpus. Contamos, de qualquer modo, com glossários e estudos impressos, porém restritos no seu uso, quando não parciais e dispersos se fossem acessíveis. A georreferência para uma entidade pode aparecer de modo contraditório num ou noutro recurso. O trabalho de elaboração de uma lista georreferenciada *ad hoc* requer, portanto, uma aproximação humanística clássica, capaz de selecionar os estudos sobre a fonte e, posteriormente, resolver por análise crítica. Enquadramos a sua aplicação no âmbito SIG como uma modalidade de SIG histórico (§2.1).

O resultado final também não é suficiente para abranger todas as entidades geográficas mencionadas da *Peregrinação*. Trata-se de um ponto de partida, a seleção de pontos de referência que foram já estudados ou são de conhecimento geral. A seguir desenvolvemos o processo de elaboração de uma lista específica para operar com as entidades mencionadas cujo referente damos com o máximo valor de probabilidade.

7.2.1 Base documental

O conjunto dos recursos documentais usados no processo de georreferenciação receberá o nome de *base documental*. A partir dos dados de coordenadas, cartografia, nomes contemporâneos e referências históricas obtidas da documentação, geovisualizamos as áreas e objetos geográficos para anotarmos referências mais prováveis, avaliarmos contradições e ambiguidades e contrastarmos os dados com as descrições do corpus.

O desenvolvimento dos estudos da história das navegações transoceânicas e a sua cartografia, junto com a importância de Mendes Pinto como referência histórica e figura literária de primeira magnitude, produziram uma bibliografia que permite seguir o seu percurso com o apoio de um aparato crítico em que tem especial relevância a geografia (Marques, 1991).

7.2.1.1 Glossários e dicionários específicos

• **[GT]** *Glossário Toponímico da Antiga Historiografia Portuguesa Ultramarina* (Lagoa, 1950-53). A obra mais notável para a toponímia das navegações na Ásia tem na *Peregrinação* uma das suas fontes principais. Uma versão on-line disponível no web do Centro de História de Além Mar (CHAM)[7] foi de particular utilidade com anterioridade à consulta da obra impressa (Lagoa, 1950-53). Com uma orientação prática, o GT fornece um índice alfabético explicativo dos topónimos dos descobrimentos e junta as coordenadas geográficas em termos de latitude e longitude segundo a cartografia britânica do século XX. Não tivemos disponível outro trabalho original de 1949 deste mesmo autor, baseado especificamente na obra de Pinto, mas sim o mapa interpretativo dos itinerários editado como extra-texto a uma edição atualizada da *Peregrinação* (Pinto, 1989, vol. 1).

• **[DHDP]** *Dicionário de História dos Descobrimentos Portugueses* (Albuquerque, 1994). Uma ferramenta de grande utilidade para se introduzir e revisar as grandes áreas e culturas assim como os topónimos mais importantes dos descobrimentos portugueses. Permite entender as regiões às que pertencem outros topónimos de entidades menores de que achamos múltiplas referências indiretas. Se a obra do Visconde da Lagoa (1950-53) constitui um intento de localizar o máximo de topónimos possíveis, Albuquerque (1994) oferece uma panorâmica de conjunto, mais seletiva, que estuda a fundo as entradas, comentadas por um conjunto variado de especialistas. Aliás, as vozes referem distintos campos do conhecimento relevantes para os descobrimentos, dos lugares e os personagens aos produtos do comércio à navegação.

• *Roteiro Geográfico através da Peregrinação* (Gomes, 1983). Apoiado principalmente no trabalho do Visconde de Lagoa e num estudo de Le Gentil de 1947 (que nós não tivemos acessível), Reinaldo Varela Gomes (1983) apresenta um glossário breve, mas o primeiro que achamos específico para a obra de Mendes Pinto.

• **[FMPP]** *Fernão Mendes Pinto and the Peregrinação* (Alves, 2010). Finalmente, um volume organizado por Zoltán Biedermann (2010), complemento da reimpressão do texto da primeira edição da *Peregrinação*, contém uma praticamente definitiva lista de topónimos (Alves, 2010, vol. 4, pp.11-37) e grupos étnicos e sociais (pp. 72-77) que, junto com as notas explicativas ao texto (Alves 2010, vol. 3), constitui o mais completo e jeitoso índice para seguir as viagens de Mendes Pinto que encontramos até à data.

7.2.1.2 Estudos genéricos

Com carácter mais complementar, a modo de aproximações de conjunto, foram especialmente úteis os trabalhos dirigidos por Albuquerque (1989a; 1989b) sobre a geografia dos descobrimentos e a sua evolução histórica; os estudos de Jaime Cortesão (1981) para os itinerários comerciais, Albuquerque (1983, 1987) para as navegações e Boxer (1977) para uma visão de conjunto do Estado da Índia.

[7] http://cham.fcsh.unl.pt/pages/glossario_visconde_lagoa.htm

7.2.1.3 Estudos de área

O trabalho divulgador de Barreto (2000) sobre a Ásia, pela sua simplicidade e concisão, proveu uma primeira definição das grandes áreas do espaço em que com frequência se desenvolvem os itinerários de Mendes Pinto. Redefinimos a proposta deste autor para uma classificação geográfica dos estudos mais citados na análise crítica dos topónimos e gentílicos do corpus:

• **Índico Ocidental.** Costa do Nordeste da África e Mediterrâneo Orientais e o litoral asiático do Índico Ocidental (fig. 7.1). A sua fronteira oriental estaria nos mares de Ceilão, em termos de geografia física, e budismo, da humana. Particularmente relevante para Mendes Pinto são Etiópia (Aubin, 1996; Graça, 1989), Estreitos e Malabar (Thomaz 1998; Subrahmanyam, 1998). Alguns trabalhos sobre cidades e portos específicos serviram para uma melhor compreensão e contextualização do espaço: Cochim (Aubin 1996; John 1998; Tavim, 2002), Diu (C. A. Pinto, 2007), Goa (Chandeigne, 1996; Rodrigues, 2007) e Meliapor (Subrahmanyam, 1990).

• **Índico Oriental**: da fronteira definida para o Índico Ocidental até aos mares do norte da Insulíndia e pelo interior até ao Nanyang da China e Champá (Barreto, 2000, pp. 20, 32-42) (fig. 7.2). Dada a grande densidade de topónimos mencionados no corpus para esta área, a região a leste do Champaa ficou classificada como área à parte com a denominação de Indochina no índice das entidades geográficas mencionadas do corpus (§9.2). O Sudeste Asiático continental é caracterizado linguisticamente por Vittrant (2010). Para Sião toda a obra de Flores (1991) foi da maior utilidade na contextualização histórica e identificação geográfica com menções diretas e aclarações de passagens inteiras da *Peregrinação*; mais especificamente, o seu glossário geográfico oferece as coordenadas geográficas dos topónimos do Sião (Flores, 1991, p.155). Carvalho (2006) estuda a metrópoli de Ayuttahaya citando Pinto como uma das principais fontes históricas. Leider (2010) analisa o budismo principalmente em Burma também a partir da *Peregrinação*, pelo que serviu também para a contextualização geográfica humana da zona.

• **Ásia Oriental**: do limite do Índico Oriental para Oriente (Barreto, 2000) (fig. 7.3). Consultamos Wheeler (2010) para a costa do Vietname. Na China (Cruz, 1570; Costa 1995; 76- 147), Macau foi objeto de estudo mais detido (Boxer 1989; Loureiro 1996). Para os povos adjacentes e interior da Ásia, salientamos as notas da edição e estudo crítico coordenado por Alves (2010). Consultamos possíveis localizações e a costa de Coreia em atlas (NGII 2009a; 2009b). Japão tem sido mais detidamente estudado da ótica das navegações e a própria *Peregrinação* (Moraes 1920; Marques 1996; Costa 1995). Pela relevância histórica de Mendes Pinto, são especialmente úteis as secções específicas do trabalho de Costa (1995; 119-147) que, sem se limitarem exclusivamente a Pinto como fonte, mas sim com atenção particular, servem para contextualizar os capítulos da obra e documentação externa de Pinto referidos ao Japão.

• **Insulíndia.** Distinguimos a Insulíndia (fig. 7.4) como área que compreende a Malásia e as ilhas da Samatra até às Filipinas e Papua. Com carácter geral a obra de Ferrand (1922) permite ver a evolução da descrição de Samatra e Java nas fontes orientais. O estudo de Thomaz (2002) serviu de introdução com atenção particular ao Sul da Samatra e Java. Específico para o Achem, consultamos

o estudo introdutório e glossário toponímico de Alves e Manguin (1997).

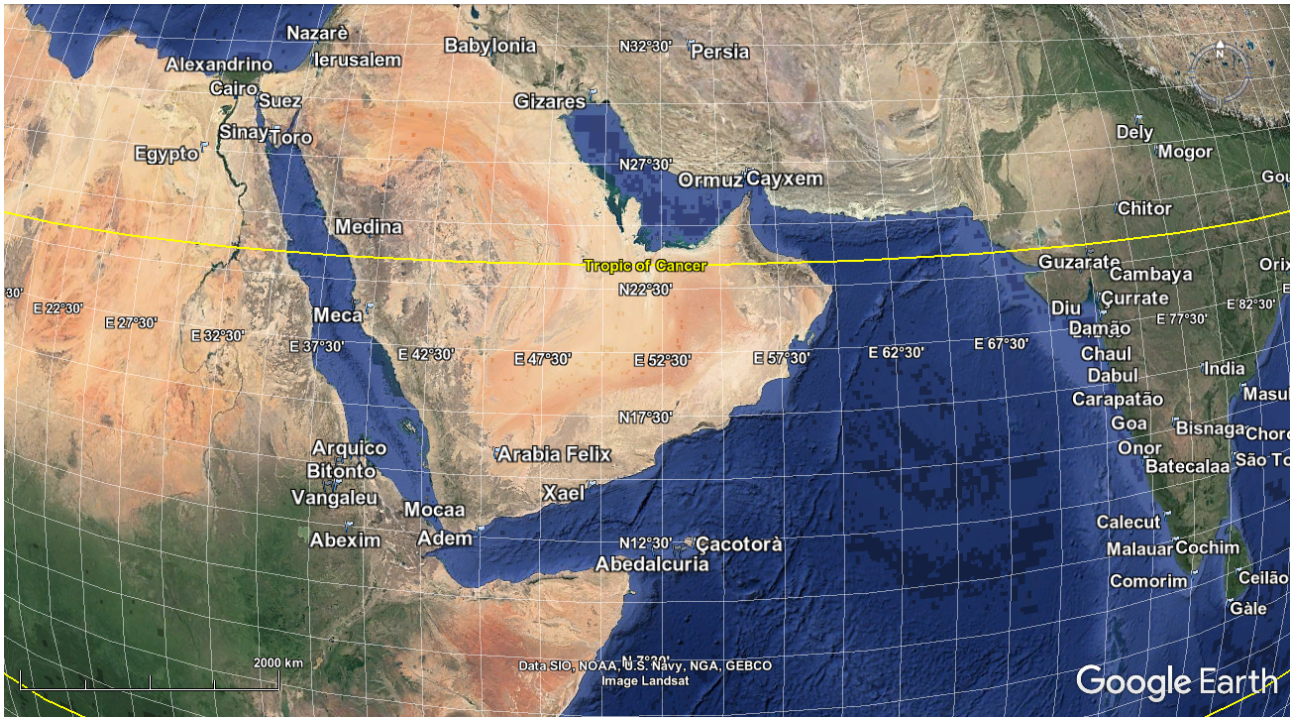


Figura 7.1: Imagem do Google Earth para a área do Índico Ocidental na fase de geovisualização dos objetos pesquisados na base documental.

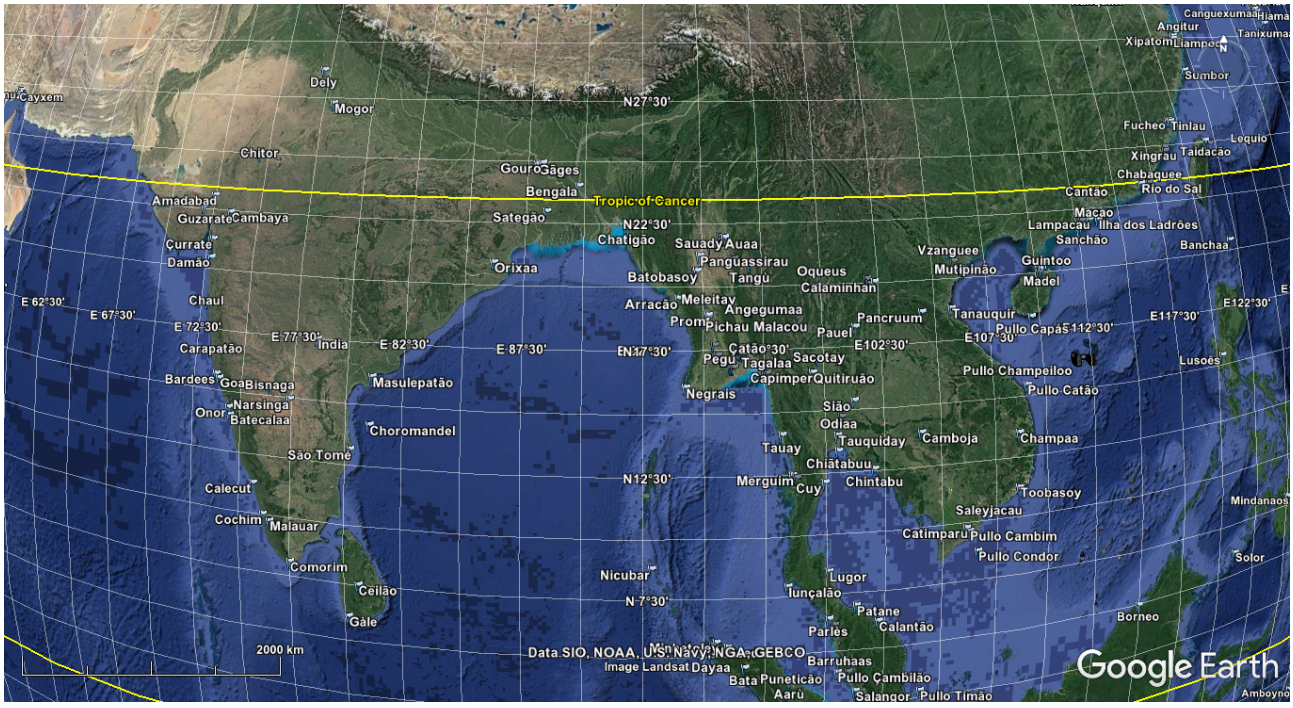


Figura 7.2: Imagem do Google Earth para a área do Índico Oriental na fase de geovisualização dos objetos pesquisados na base documental.

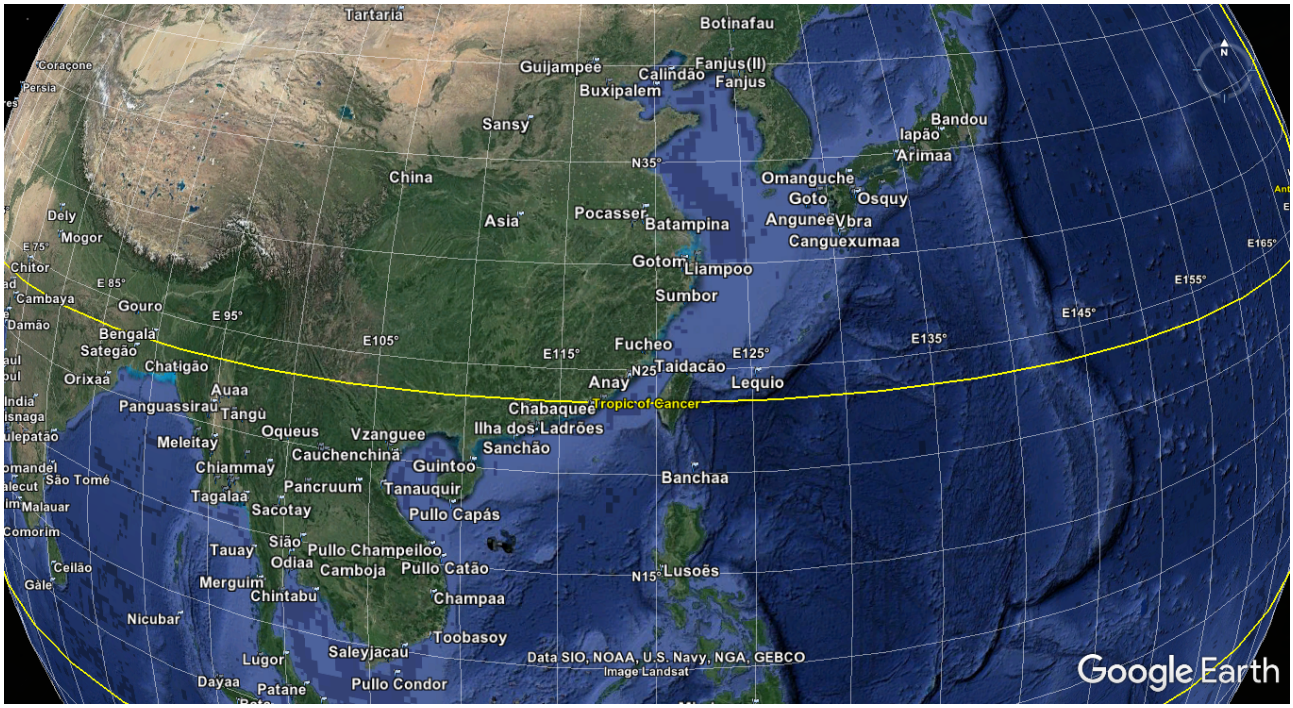


Figura 7.3: Imagem do Google Earth para a área da Ásia Oriental na fase de geovisualização dos objetos pesquisados na base documental.

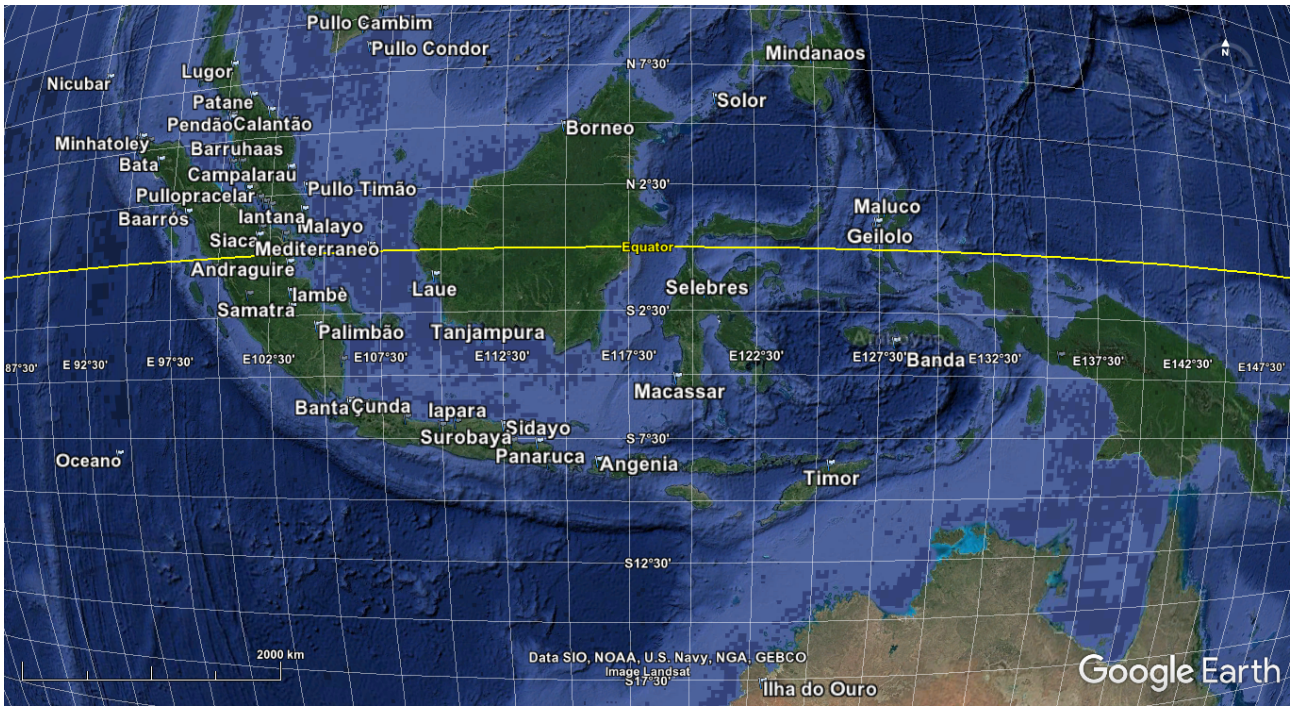


Figura 7.4: Imagem do Google Earth para a área da Insulíndia na fase de geovisualização dos objetos pesquisados na base documental.

Figura 7.5: Imagem do Google Earth para a área da Europa na fase de geovisualização dos objetos pesquisados na base documental.

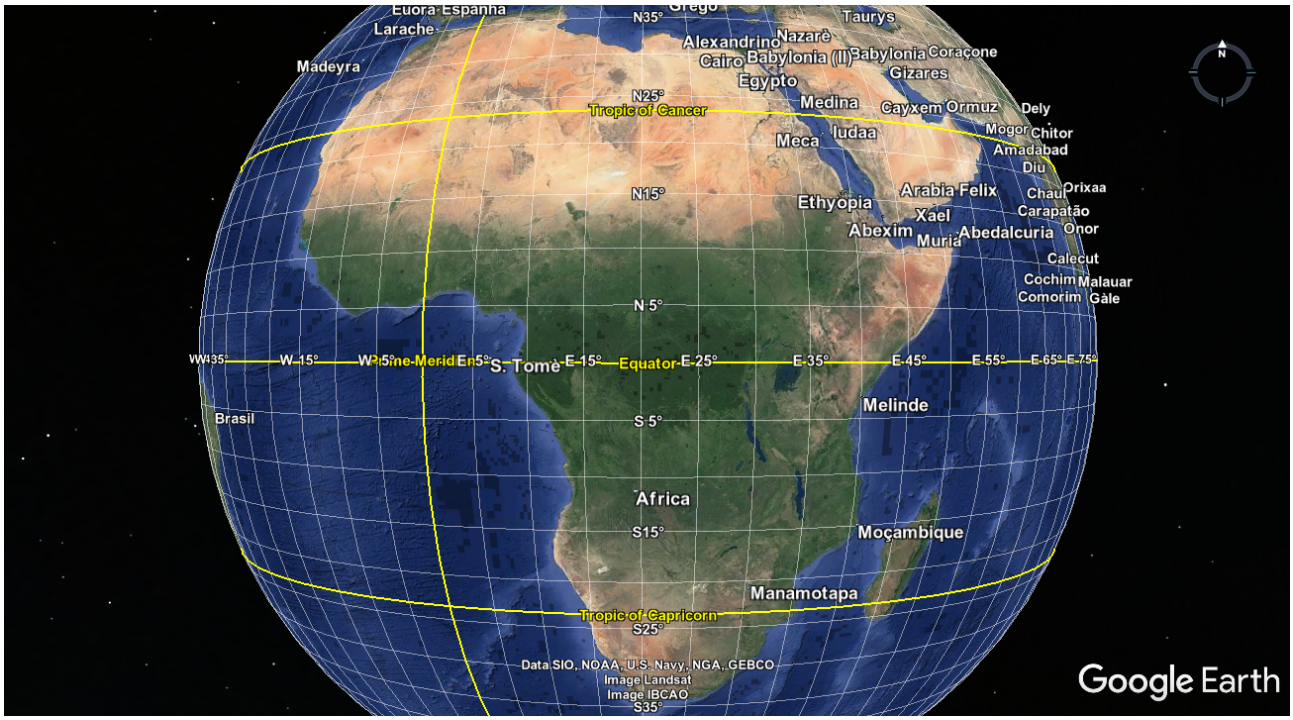


Figura 7.6: Imagem do Google Earth para a área de África na fase de geovisualização dos objetos pesquisados na base documental

• **Europa**. A geografia europeia (fig. 7.5) não requereu o uso de um aparato bibliográfico específico, ainda que alguns topónimos precisam notas aclaratórias por causa de modificações históricas ou evoluções culturais (caso por exemplo da menção que se faz de Sardenha em termos de 'malhorqui'). Os guias Mercator (Thomas, 1993-1999), em cuja elaboração participamos, têm um introdução aos grupos linguísticos menos conhecidos ou mais difíceis de referenciar da Europa.

• **América e África Ocidental**. Apenas há três topónimos georreferenciados na América. Caso particular foi a localização de Nova Espanha (Meyer, Sherman e Deeds 2007), topónimo que até o momento não acháramos referido em nenhum glossário consultado para a obra de Pinto. Para pesquisas particulares sobre o Atlântico, relações e exploração da África (fig. 7.6), consultamos principalmente os artigos específicos das obras dirigidas por Albuquerque (1989b; 1994).

7.2.2 Lista com estudo crítico

A partir da análise das obras da base documental, junto com a geovisualização dos objetos e o contraste crítico com o corpus, elaboramos uma lista com todas as entidades geográficas mencionadas que encontramos estudadas na bibliografia (§7.2.1). A lista inclui todas as referências com maior probabilidade na atribuição do referente e uma análise crítica das prováveis e possíveis comparando a evidência documental.

Como ferramenta de geovisualização, usamos GoogleMaps[8] e GoogleEarth[9] para uma localização aproximada das entidades geográficas estudadas, contraste de coordenadas e inspeção mediante mapas e imagens de satélite dos objetos. Cada entrada constitui uma única entidade geográfica, o termo mais usual serve de lema para a ordenação alfabética e vai seguido das variantes observadas no texto da *Peregrinação.*

Dentro do estudo crítico, incluímos referências à edição fac-similar da Biblioteca Digital Nacional (BDN) da Biblioteca Nacional de Portugal (BNP)[10]. Indicamos o número de página segundo correspondia à numeração tal e como era apresentada nos URLs da BDN em 2011, isto é, fazíamos coincidir com o endereço web e não a referência de fólio, porque naquela altura consideramos que facilitava a consulta direta on-line. Este sistema ficou obsoleto com a mudança de URLs e nova disposição do fac-similar na BDN. Nas novas versões da lista, as referências são feitas de preferência ao capítulo para facilitar a compatibilidade entre edições e ao fólio da edição impressa quando queremos destacar uma passagem da primeira edição.

Para o conjunto da entrada usamos como convenção:

Entrada do topónimo [*Variante(s)*] (página segundo a BDN[número de coluna;]). Tipo geográfico. [Explicação do topónimo] [Bibliografia e dados procedentes de bibliografia]

Ex. **Adem**. *Adẽ* (14ii; 48ii). Porto. DHDP, GT s.v. Adam, Adém, Aden: 12º 45' N 45º 4' E

[8] https://maps.google.pt
[9] http://www.google.com/intl/pt-PT/earth/index.html
[10] http://purl.pt

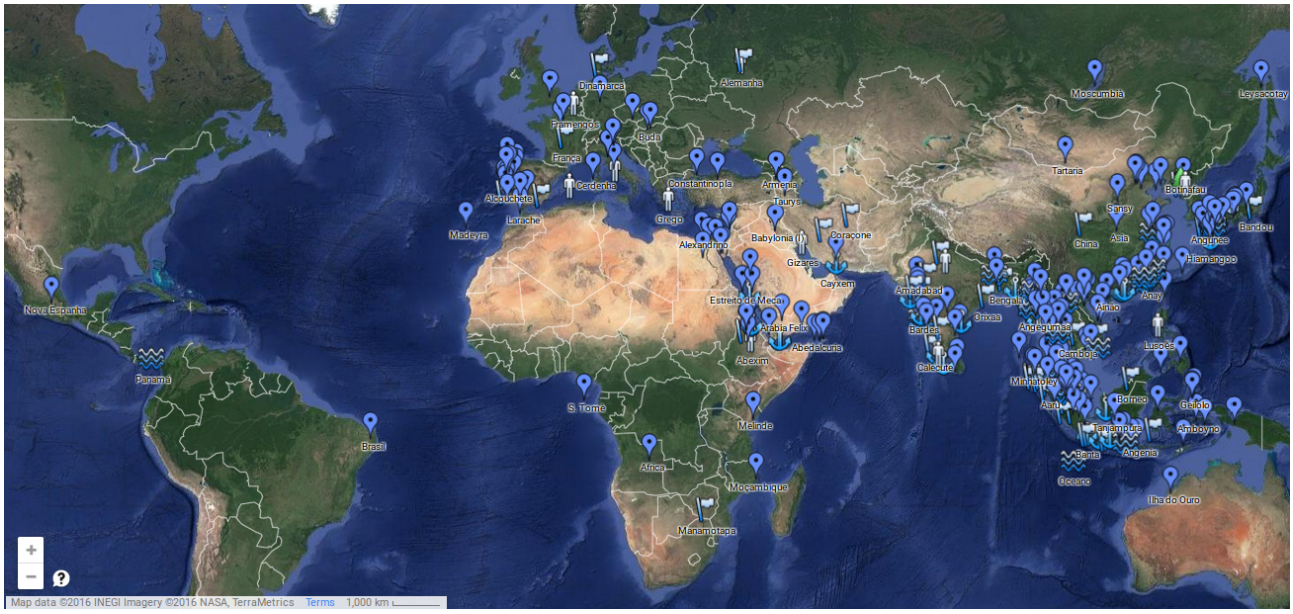


Figura 7.7: Vista de conjunto do mapa de entidades da lista com estudo crítico no GoogleMaps.

Para o estudo crítico trabalhamos com visualizações a escala regional (fig. 7.8) e local (fig. 7.9). Na (fig. 7.7) mostramos uma imagem de conjunto com todas as entidades geográficas mencionadas de que se obteve uma georreferência considerada como provável na definição do conceito e cujas coordenadas damos de aproximadas a exatas.

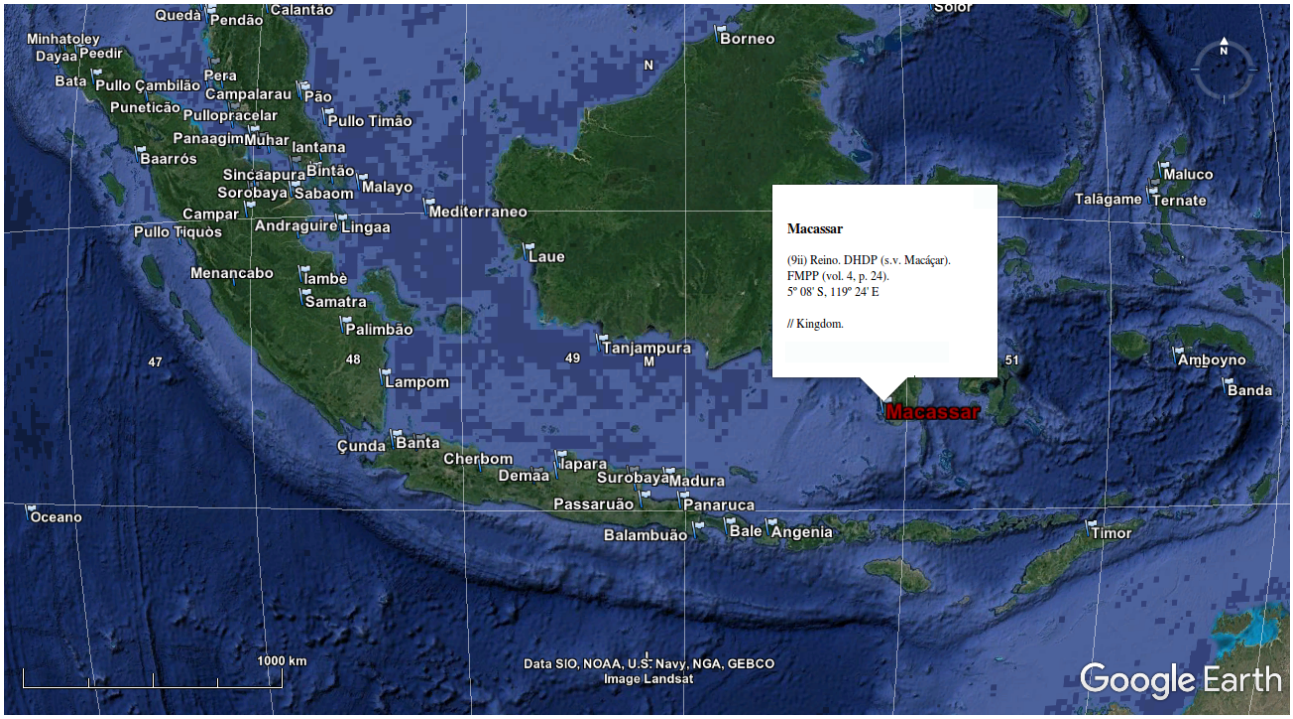


Figura 7.8: Vista regional para a entrada *Macassar* (Google Earth)

Figura 7.9: Vista local para a entrada *Dalaa* (Google Earth).

7.2.3 Integração do estudo crítico no corpus

Pullo Çambilão

(1) Entidade nomeada

Pullo Çambilão nos capítulos: | 20 | 144 | 205 |

Nas orações: 220, 2002, 3133.

1. [*220*](Cap. 20) Partido eu com a pressa que digo deste rio **Parlês**, hum Sabado quasi Sol posto, cõtinuey por minha derrota atè a terça feyra ao meyo dia, em que prouue a nosso Senhor que cheguey às ilhas de **Pullo Çambilão**, primeira terra da costa do **Malayo**, onde achey tres naos **Portuguesas**, duas que vinhaõ de **Bengala**, & hũa de **Pegù**, de que era Capitão & senhorio hum Trisão de Gaa, ayo que fora de dom Lourenço filho do Visorrey dom Francisco de Almeida, que Miroocem matou na barra de **Chaul**, de que as historias do descubrimento da **India** fazem larga menção.
2. [*2002*](Cap. 144) Eu lhe aceitey a viagem de boa vontade, & me party hũa quarta feyra noue dias do mez de Ianeyro do anno de 1545 desta fortaleza de **Malaca**, & seguy minha derrota com vẽtos bonanças ate **Pullopracelar**, onde o piloto se deteue por respeito dos baixos que atrauessauão todo este canal da terra firme â ilha **Çamatra**, & despois de sermos fora delles inda que com trabalho, vellejamos por nossa derrota as ilhas de **Pullo Çambilão**, onde me mety nũa manchua bem esquipada que leuaua, & nauegando sempre nella por espaço de mais de doze dias, cõforme ao regimento que leuaua de Pero de Faria, espiey toda a costa deste **Malayo**, que saõ cento & trinta legoas atè **Iunçalaõ**, entrando em todos os rios de **Barruhaas**, **Salangor**, **Panaagim**, **Quedaa**, **Parlés**, **Pendão**, & **Sambilão Sião**, sem em nenhum delles achar noua certa destes inimigos.
3. [*3133*](Cap. 205) Esta armada se partio do porto de **Malaca** hũa sesta feira vinte & cinco de Oitubro do anno de 1547. & vellejando todos por sua derrota aos quatro dias chegarão a **Pullo Çambilão** sessenta legoas dõde tinhaõ partido.

(2) Objecto geográfico interpretado

Conhecimento: P.

100.535805,4.006868,0.000000

(47i; 355i; 363i; 542i) Arquipélago. Thomaz (2002, p. 445): Pulau Sembilam, “as nove ilhas” em malaio, junto à foz de Perak em 4º N, 100º 32’ E. FMPP (vol. 3, p. 61): "Nine islands" na foz do rio Perak, 20 km a Sul de Pulau Pangkor. GT Pulo Sambilão (I): No estreito de Malaca, junto à foz do Perak em 4º N, 100º 32' E.

Figura 7.10: Recuperação de concordâncias (1) e estudo crítico da base documental (2) para *Pullo Çambilão*.

A avaliação do referente requer contrastar todas as ocorrências das expressões anotadas que o mencionam. Para facilitar a análise crítica, integramos o estudo crítico no painel de recuperação das concordâncias (fig. 7.10).

7.3 Criação de uma lista inicial de referentes com conhecimento prévio

A lista com estudo crítico (§7.2.2) inclui as entidades geográficas analisadas, mas nem todas têm o mesmo grau de precisão nas coordenadas (de aproximadas a exatas), nem certeza no conceito (de possível a muito provável e assumido como certo), nem em todas há unanimidade na atribuição de um referente a partir dos estudos que conformam a base documental. Necessitamos uma lista com referentes precisos para operarmos com conceitos (§6.2.1.3) e abordarmos uma nova análise crítica em que consideremos tanto as entidades geográficas mencionadas georreferenciadas por conhecimento prévio como aquelas outras que não encontramos referenciadas em estudo nenhum e, portanto, conhecemos apenas por descrição (§7.1.2). Esta lista recebe o nome de **lista de entidades geográficas referenciadas por conhecimento prévio** e foi definida em (§7.1.1). A seguir, descrevemos o critério de seleção e casos práticos da sua aplicação.

7.3.1 Critério de seleção de referentes

Selecionamos os referentes a partir da lista com estudo crítico, excluindo aquelas entradas que precisam uma maior análise crítica, isto é, aquelas cujas coordenadas são só possíveis (não exatas) ou cuja definição do conceito (tipo geográfico *c* pertencente a entidade geográfica *a*) (6.2) situamos por debaixo de muito provável (numa escala obtida a partir da análise da base documental). Estabelecemos um critério para considerar um referente geográfico *g* como sendo parte da lista inicial de referentes com conhecimento prévio: quando uma entidade geográfica mencionada aparecer referenciada como certa ou muito provável na base documental e o contexto das concordâncias extraídas não oferecer contradições com o estudo crítico (fig. 7.4), aceitaremos o referente como entrada na lista G' (§7.1.1). Uma expressão *w* entrará na lista de conhecimento prévio se tiver um referente geográfico com coordenadas geográficas *g*, e apenas um único referente *g* (qualquer outro valor *y* que se achar, terá de ser equivalente a *g* a efeitos da descrição do objeto geográfico).

***Regra de georreferenciação 3**: Uma expressão referenciada por conhecimento prévio tem um único referente.*

$$\forall\, g \in G' \,(\exists\, w \in W, \text{Tem\_georreferente}(w, g)\ \Lambda\ (\forall\, y\ \text{Tem\_georreferente}\,(w,y)\ \rightarrow y = g) \qquad (7.1)$$

A probabilidade de uma expressão *w* ter um referente *g* segundo a função Georreferente(w) = g (6.3) para um objeto geográfico incluído na lista de conhecimento prévio que cumpre a relação Tem_georreferente(*w, g*) = "*w* fica situada em *g*" (6.1) é:

$$P(w \mid g \in G') = 1 \qquad (7.2)$$

***Regra de georreferenciação 4**: A probabilidade outorgada por um valor inicial $g \in G'$ para uma expressão w que denotamos por conhecimento prévio ao cumprir a relação "w fica situado em g" é* $P(w \mid g \in G') = 1$.

$$\forall\, g \in G' \,(\exists\, w \in W, \text{Tem\_georreferente}(w, g) \rightarrow P(w \mid g \in G') = 1) \qquad (7.3)$$

7.3.2 Processo de seleção de um referente para a lista de conhecimento prévio

Para resolver um referente na lista de conhecimento prévio temos que considerar, por um lado, a entidade geográfica mencionada dentro do seu contexto imediato (concordância) no corpus, por outro, a interpretação produzida no estudo crítico. Há um passo prévio de agrupamento de todas as variantes sob um mesmo lexema (§6.3.2.2) no caso do corpus, e de seleção de apenas um referente para casos de homonímia (§6.3.1.1) na base documental. A avaliação do objeto descrito no estudo crítico como coincidente com o mencionado no corpus tem, nesta altura, valor inicial (atualizável), mesmo quando o referente fosse seguro na base documental. Procuramos apenas que não haja contradição dentro de um contexto limitado pelo exame das concordâncias (fig. 7.10).

Seguidamente descrevemos os passos a partir de um exemplo, *Pullo Çambilão*, arquipélago recuperado com três ocorrências na *Peregrinação 1614* (concordâncias completas na fig. 7.10).

7.3.2.1 Extração das concordâncias no corpus

O primeiro passo é a recuperação de todas as ocorrências para uma mesma entidade geográfica mencionada.

> "Partido eu com a pressa que digo deste rio *Parlês*, hum Sabado quasi Sol posto, cõtinuey por minha derrota atè a terça feyra ao meyo dia, em que prouue a nosso Senhor que cheguey às ilhas de **Pullo Çambilão**, primeira terra da costa do *Malayo*" (*PR*, 20) (7a)

> "& me party hũa quarta feyra noue dias do mez de Ianeyro do anno de 1545 desta fortaleza de *Malaca*, & seguy minha derrota com vẽtos bonanças ate *Pullopracelar*, onde o piloto se deteue por respeito dos baixos que atrauessauão todo este canal da terra firme â ilha *Çamatra*, & despois de sermos fora delles inda que com trabalho, vellejamos por nossa derrota as ilhas de **Pullo Çambilão**, onde me mety nũa manchua bem esquipada que leuaua, & nauegando sempre nella por espaço de mais de doze dias, cõforme ao regimento que leuaua de Pero de Faria, espiey toda a costa deste *Malayo*, que saõ cento & trinta legoas atè *Iunçalaõ*, entrando em todos os rios de *Barruhaas*, *Salangor*, *Panaagim*, *Quedaa*, *Parlés*, *Pendão*, & *Sambilão Sião*" (*PR*, 144) (7b)

> "Esta armada se partio do porto de *Malaca* hũa sesta feira vinte & cinco de Oitubro do anno de 1547. & vellejando todos por sua derrota aos quatro dias chegarão a **Pullo Çambilão** sessenta legoas dõde tinhaõ partido." (*PR*, 205) (7c)

7.3.2.2 Integração do estudo crítico

Da base documental (§7.2.1) obtemos três descrições para o estudo crítico (§7.2.2):

> Thomaz (2002, p. 445) Pulo Sambilão: Pulau Sembilam, “as nove ilhas” em malaio, junto à foz do Perak em 4º N, 100º 32’ E.
>
> FMPP (vol. 3, p. 61) Pullo Çambilão: "Nine islands". Na foz do rio Perak, 20 km a Sul de Pulau Pangkor.
>
> GT Pulo Sambilão (I): No estreito de Malaca, junto à foz do Perak. 4º N, 100º 32' E.

FMPP dá *Pullo Sambilan* como correspondente a *Pullo Çambilão. A* forma anotada por Thomaz (2002) e GT, *Pulo Sambilão*, coincide na pronúncia com o corpus, variando apenas a forma gráfica.

A georreferenciação aparece como segura no estudo crítico. Thomaz (2002, p. 445), FMPP (vol. 3, p. 61) e GT, nenhum oferece dúvidas para a localização. Não obstante, é de notar que na base documental (§7.2.1) GT recolhe uma segunda entrada, *Sembilam*, ilha na Samatra com coordenadas 4º 09' N , 98º 15' E. Convém logo considerar que, ainda quando na lista do estudo crítico o referente seja seguro, a sua interpretação no corpus tem apenas valor de probabilidade. Neste ponto do estudo, aplicamos o princípio de coerência declarado em (§6.3.1.1) pelo qual, se não houver evidências de contradição ao respeito, consideraremos que uma mesma expressão tem o mesmo referente em todo o corpus.

7.3.2.3 As descrições do estudo crítico permitem apontar para um mesmo referente

Figura 7.11: Pulau Sembilan (esquerda) na foz do rio Perak (direita). Imagem recuperada em GeoNames para a pesquisa *Pulau Sembilan*, em Perak (Malaysia), 4° 2' 12" N, 100° 32' 59" E.

Continuando com o mesmo exemplo, todas as descrições obtidas da base documental (Thomaz, 2002, p. 445; Alves, 2010, vol. 3, p. 61; Lagoa, 1950-53, s. v. Pulo Sambilão I), embora possam variar na precisão, apontam para uma mesma entidade geográfica, referenciam o mesmo arquipélago visualizável na figura 7.11 acima.

Em termos de coordenadas geográficas, qualquer ponto que as assinale interpreta o mesmo objeto geográfico. De facto, ao representar o objeto em GoogleMaps, as coordenadas obtidas são 100.535805 (Long.), 4.006868 (Lat.), 0.000000 (Alt.), o mesmo objeto recuperado em GeoNames tem coordenadas de referência 4° 2' 12" N, 100° 32' 59" E. A escolha de um valor ou outro é nesta altura simples convenção. Com a finalidade de representarem *Pullo Çambilão* como $g \in G'$, são coordenadas equivalentes.

7.3.2.4 O contexto das concordâncias da entidade geográfica mencionada coincide com o referente do estudo crítico

As descrições que achamos nas concordâncias para a entidade geográfica mencionada são coerentes com o referente segundo descrito na base documental. O exame das ocorrências permite interpretar como provável o objeto geográfico apontado no estudo crítico.

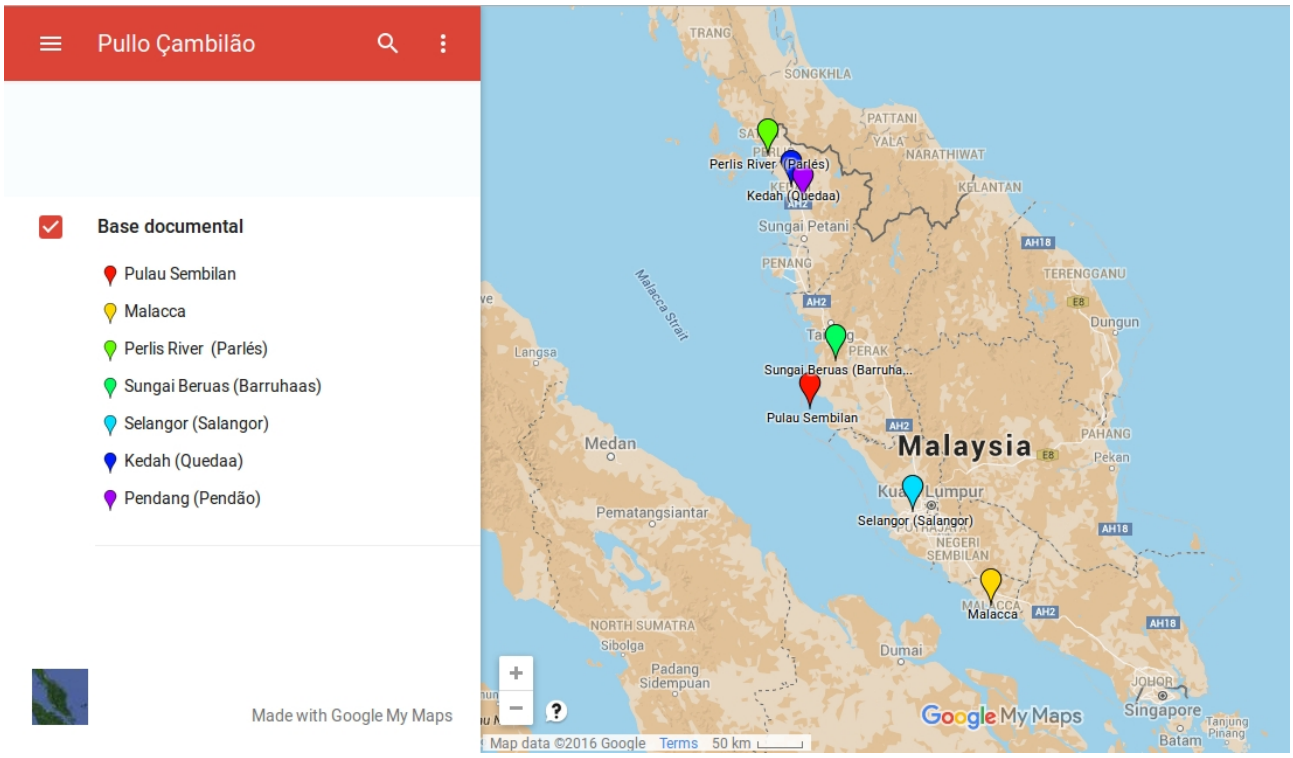


Figura 7.12: *Pullo Çambilão* e entidades mencionadas com que coocorre no corpus georreferenciadas no estudo crítico da base documental (GoogleMaps).

A análise da entidade faz-se segundo o modelo concetual proposto, com um tipo geográfico e uma relação de pertença a outra entidade.

Tipo de objeto geográfico. Em todas as ocorrências (7a-c), *Pullo Çambilão* aparece num conjunto de ilhas, favorecendo a interpretação de arquipélago confirmada no estudo crítico (Thomaz 2002; FMPP; GT).

Holónimo (é_Parte_de). Em (7a) as ilhas *Pullo Çambilão* são situadas na costa malaia ocidental, ponto comum em todas as descrições selecionadas da base documental (Thomaz 2002; FMPP; GT). Em (7b) concorrem uma série de entidades geográficas que coincidem com o espaço geográfico da georreferência considerada e, finalmente, (7c) descreve um itinerário também coerente com a georreferenciação dada pelo estudo crítico (fig. 7.12).

Confirmada a coincidência das descrições do corpus com as do estudo crítico, sem nenhum elemento de contradição evidente, o arquipélago com coordenadas 4º N, 100º 32’ E entra na lista de referentes por conhecimento prévio para a expressão *Pullo Çambilão.*

7.3.3 Referentes com conhecimento prévio excluídos da lista de referentes por conhecimento prévio

Apenas as entidades geográficas mencionadas que cumprem o critério de ter sido desambiguadas para um único referente (7.1) entram na lista de conhecimento prévio. No entanto, a lista do estudo crítico contém entradas de entidades geográficas mencionadas para as quais também temos conhecimento, mas não há certeza na atribuição de coordenadas, isto é, existe ambiguidade (um mesmo estudo oferece mais de um referente, ou vários estudos não coincidem no referente sem que haja critérios evidentes para escolher um como válido) ou aparece uma contradição evidente no exame das concordâncias. Consideramos estes casos a seguir.

7.3.3.1 Divergência nas soluções do referente na base documental

Os estudos da base documental (§7.2.1) discrepam na georreferenciação sem termos evidência suficiente para escolhermos um referente como sendo seguro.

Exemplo: na figura 7.13 abaixo mostramos uma captura de ecrã com o estudo crítico para a entidade geográfica mencionada *Calaminham*, no Tibete segundo GT, em Laos para FMPP. Temos, portanto, mais de um candidato para representar *g*, com o qual não se cumpre a condição definida em (7.1): ter um único referente. A entidade geográfica mencionada *Calaminham* fica assim fora da lista inicial de referentes por conhecimento prévio.

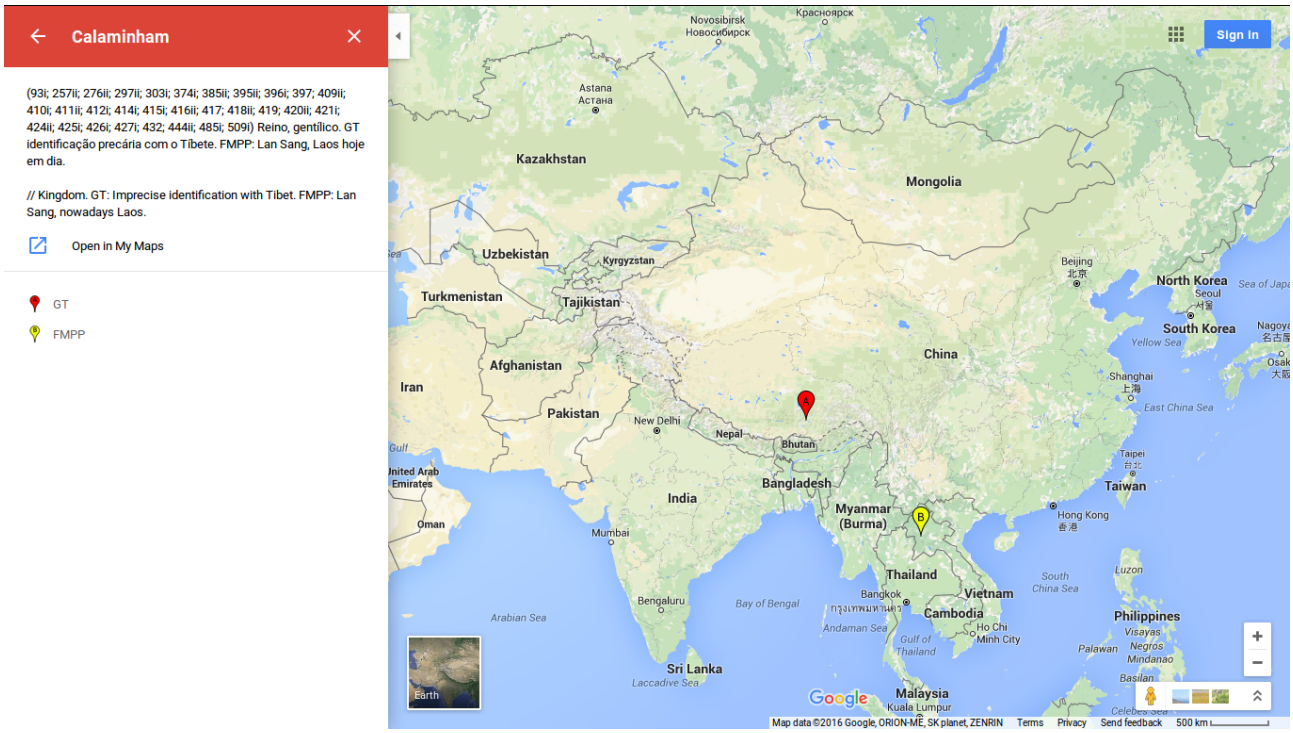


Figura 7.13: Divergência de soluções para *Calaminham*. (A) GT no Tibete, (B) FMPP em Lan Sang (Laos) (GoogleMaps).

7.3.3.2 A solução do referente é apenas possível

Os estudos da base documental dão uma solução, mas oferecem dúvidas sobre a sua precisão.

Exemplo: **Paneticão** é um rio mencionado no corpus. Para a sua georreferenciação, GT não dá nenhum referente seguro, mas sugere o *Tungkam* em 4º 05’ N, 98º 08’ E. Anthony Reid (Alves, 2010, vol. 3, p. 65) estuda distintas possibilidades a partir do reino de *Aru* e propõe o *Sungai Petani* ou ainda mais ao sul o *Panai* ou o *Barumun*.

Temos, portanto, não um, mas vários referentes e, dentro destes, todos possíveis. Não se cumpre a condição definida em (7.1), portanto, *Paneticão* também não entra dentro da lista de referentes iniciais.

7.3.3.3 A solução do referente entra em contradição evidente com o corpus

Há um referente na base documental, mas entra em contradição com o contexto do corpus.

Exemplo: Em GT, DHD e FMPP, *Curia Muria* é um único topónimo para se referir a um grupo de ilhas perto da costa arábiga. *Abedalcuria* é identificada como ilha distinta, a *Abd al-Kûrî*, parte do arquipélago de Socotorá.

No corpus temos a concordância:

> “ouuemos vista das ilhas de **Curia**, **Muria**, & **Abedalcuria**, nas quais estiuemos de todo perdidos, sem nenhũa esperança de vida; & tornandonos, por não auer outro remedio, na volta do sudueste, prouue a nosso Senhor que ferramos a ponta da ilha **Çacotorà**, hũa legoa abaixo donde esteue a nossa fortaleza que dom Francisco d'Almeida, primeyro Visorrey da **India** fez” (*PR*, 143) (7d)

Porém, na base documental, por um lado aparece *Curia Muria* como uma única entidade, e *Abedalcuria* por outro, completamente distintas e afastadas uma da outra (fig. 7.14).

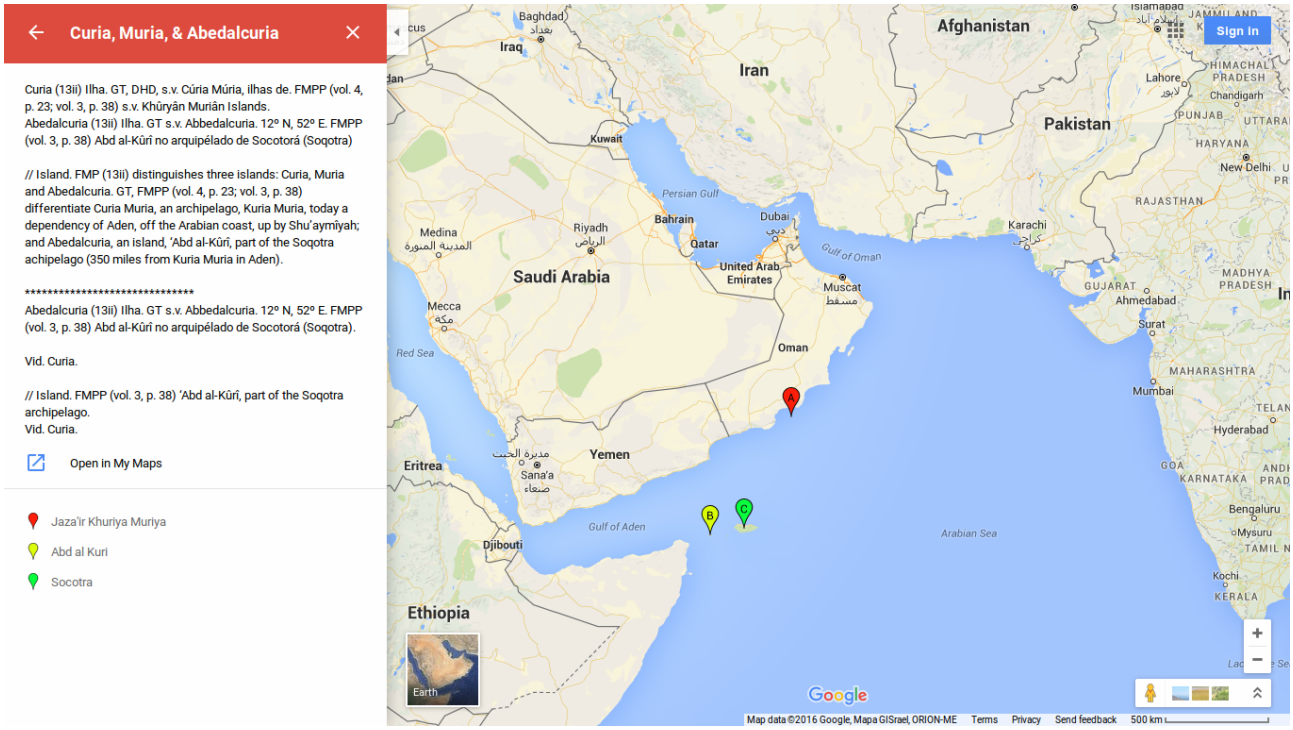


Figura 7.14: As ilhas (A) *Curia Muria* GT, DHD, FMPP (vol. 4, p. 23; vol. 3, p. 38) Khûryân Muriân Islands, aparecem muito distanciadas de (B) *Abedalcuria*, GT, FMPP (vol. 4, p. 23; vol. 3, p. 38) Abd al-Kûrî da sua vez perto de (C) *Socotorá* (Soqotra). (GoogleMaps).

No entanto, em (7d) temos três ilhas mencionadas como se formassem um único arquipélago. Há uma contradição que requer um estudo da descrição mais pormenorizado. Não damos nenhum referente como seguro, portanto, não resolvemos a relação que liga a expressão com o referente (6.1), consequentemente, também não o critério de seleção (7.1).

Excluímos, assim sendo, *Curia*, *Muria* e *Abedalcuria* da lista de referentes iniciais.

7.3.4 O objeto geográfico

Descrevemos o referente, segundo §6.2, mediante umas coordenadas geográficas e um conceito, os elementos que nos permitem a ligação com a expressão. A lista de referentes por conhecimento prévio tem de ter, portanto, coordenadas geográficas de referência, um tipo de entidade geográfica e

uma relação de pertença.

Ao operarmos com o objeto geográfico contrastamos os resultados obtidos do processamento do corpus e a base documental com dados procedentes da análise geoespacial. Na elaboração do estudo crítico e seleção de referentes começamos a operar com ferramentas para visualizar a entidade geográfica e criar mapas que a representem. Para a lista de referentes de conhecimento prévio interessa fixar um objeto no espaço com um identificativo único que o represente e integre em recursos SIG e obter assim dados geográficos que contribuam para a georreferenciação do corpus. GeoNames é uma base de dados global que associa um objeto geográfico a um identificativo único e estável, outorgando-lhe umas coordenadas geográficas de referência. Oferece também uma ontologia de atributos geográficos e liga a entidade identificada numa relação de pertença com outro objeto geográfico. É, aliás, uma das listas usadas pelo NERC Edinburgh Geoparser, aplicável, portanto, numa solução que, além de georreferenciar, identifica e classifica as entidades mencionadas (§5.6.2.3).

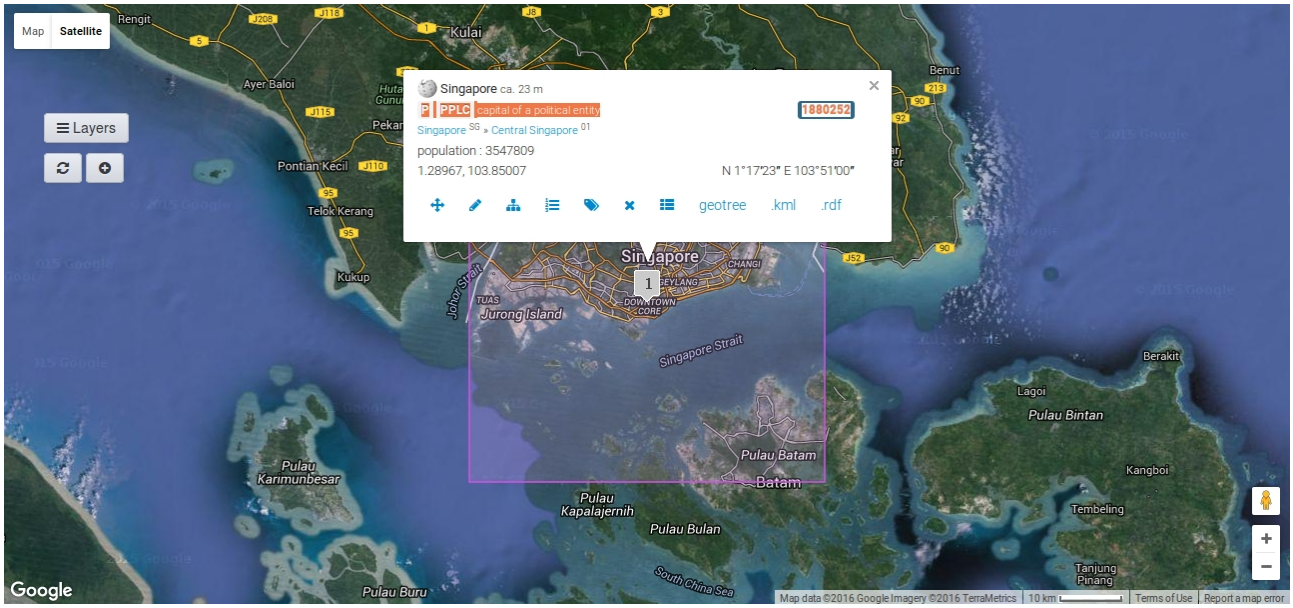


Figura 7.15: Singapore (*Cincaapura*, *Cincapura*, *Sincaapura* no corpus *Peregrinação* 1614) recuperado em GeoNames.

7.3.4.1 Pesquisa a partir da expressão

Para a pesquisa em GeoNames usamos a forma contemporânea tal e como foi solucionada no estudo crítico. GeoNames gera uma lista que, pela sua abrangência, produz frequentes homonímias (tab. 6.2). Como critério primeiro de escolha, selecionaremos o referente que tiver as coordenadas geográficas mais aproximadas às obtidas da base documental. Quando houver vários objetos com as mesmas coordenadas coincidentes com o estudo crítico, preferiremos o objeto cujo tipo geográfico melhor representar o descrito no corpus. Finalmente, quando a entidade geográfica mencionada for metonímica (§6.3.1.3), recuperaremos o tipo geográfico correspondente ao holónimo de entre os objetos que coincidam em coordenadas e tipo geográfico.

7.3.4.2 Pesquisa a partir das coordenadas geográficas

Quando o topónimo obtido do estudo crítico não ofereça resultados (frequentemente devido a variações na transcrição não incluídas dentro dos topónimos alternativos em GeoNames), usamos uma pesquisa por coordenadas geográficas (*reverse geocoding*) para acharmos a forma padrão aceite na base de dados global.

Em muito menor medida, quando a base documental não tiver oferecido um topónimo contemporâneo (porque entendemos que a entidade geográfica desapareceu como tal), pesquisamos, em primeiro lugar a partir das coordenadas, um objeto próximo que represente o objeto geográfico descrito no corpus. Ainda que GeoNames recupera topónimos, a escolha vem condicionada pela inspeção visual do espaço. Procuramos, na imagem de satélite, um objeto que descreva concetualmente a entidade mencionada no corpus (ex. um porto). Apenas se aceita um topónimo se, dentro das coordenadas aproximadas dadas pelo estudo crítico, representa também a entidade mencionada no seu tipo geográfico.

7.3.5 Resultado da seleção: lista de entidades georreferenciadas por conhecimento prévio

Excluídas as entidades geográficas que, mesmo tendo parte de conhecimento prévio, têm que ser também denotadas por descrição (§7.1.2), fica uma lista de referentes relacionados com as expressões do corpus com valor de probabilidade $P(w \mid g \in G')=1$ na solução da georreferência (enquanto não aparecerem novos dados indício do contrário).

A integração dos dados obtidos de uma base de dados toponímica fornece um identificativo único para cada objeto geográfico e permite elaborar conceitos que ligam um tipo geográfico e um holónimo com o objeto referenciado pela entidade geográfica mencionada, obtendo assim georreferências por definição (regra de georreferenciação n. 2) que complementam as coordenadas geográficas (regra de georreferenciação n.1). É importante considerar que a definição que usa os dados obtidos em GeoNames denota um objeto no presente segundo uma base de dados global particular. O procedimento consistiu em pesquisar o topónimo contemporâneo quando partimos da expressão (§7.3.4.1) e geovisualizar a área quando tivermos as coordenadas geográficas (§7.3.4.2). Se houver metonímia, o objeto selecionado será o holónimo (§6.3.1.3).

7.3.5.1 Tipos e distribuição das entidades georreferenciadas por conhecimento prévio

A integração da lista de referentes com GeoNames permite obter uma primeira taxonomia para classificarmos as entidades e observar a sua distribuição geográfica. Na recuperação do identificativo para cada entidade obtivemos a classe e tipo (subclasse) de objeto geográfico, o topónimo contemporâneo, o continente e país ou países a que pertence (holónimos) e as coordenadas de latitude e longitude de referência. A figura 7.16 mostra a disposição das classes segundo continentes e número de objetos.

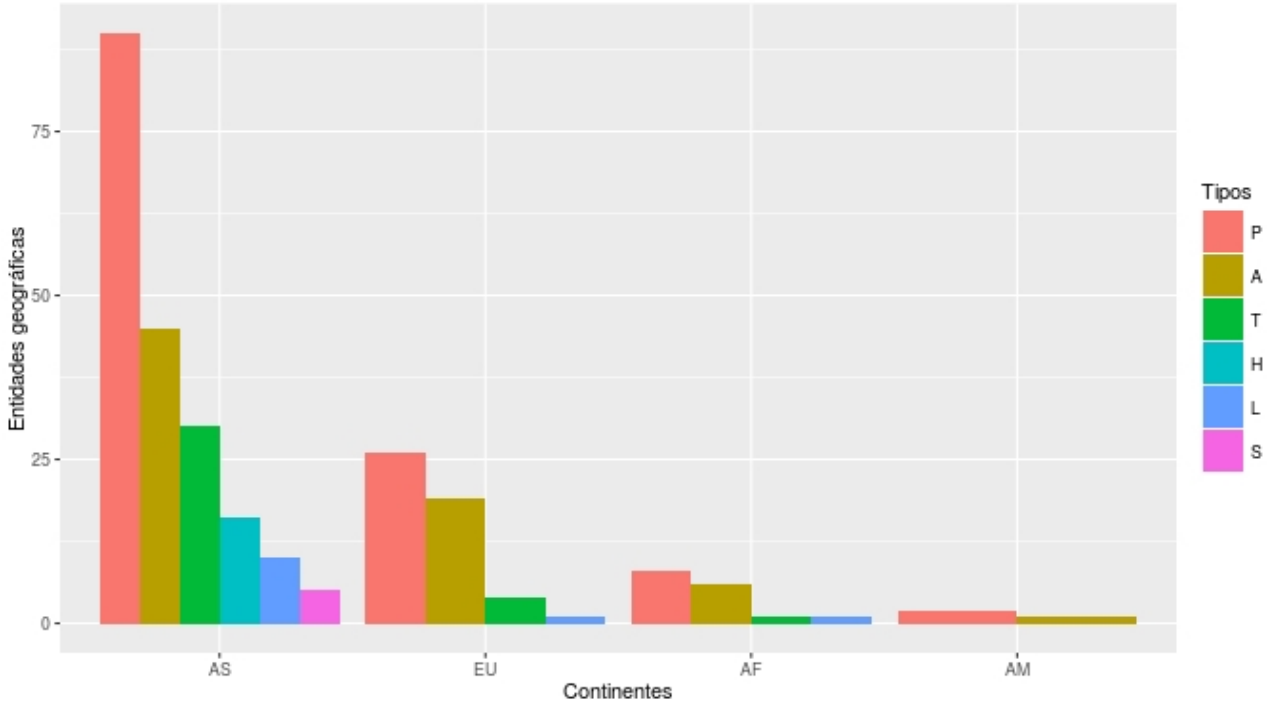


Figura 7.16: Objetos referenciados (entidades geográficas) por conhecimento prévio segundo continentes: AS=Ásia, EU=Europa, AF=África, AM=América; e tipos geográficos: P=Cidades e povoações; A=Países e divisões administrativas; T=Ilhas, montanhas e acidentes físicos de terra; H=Acidentes hidrológicos, praias e portos; L=Grandes áreas e regiões; S= Construções.

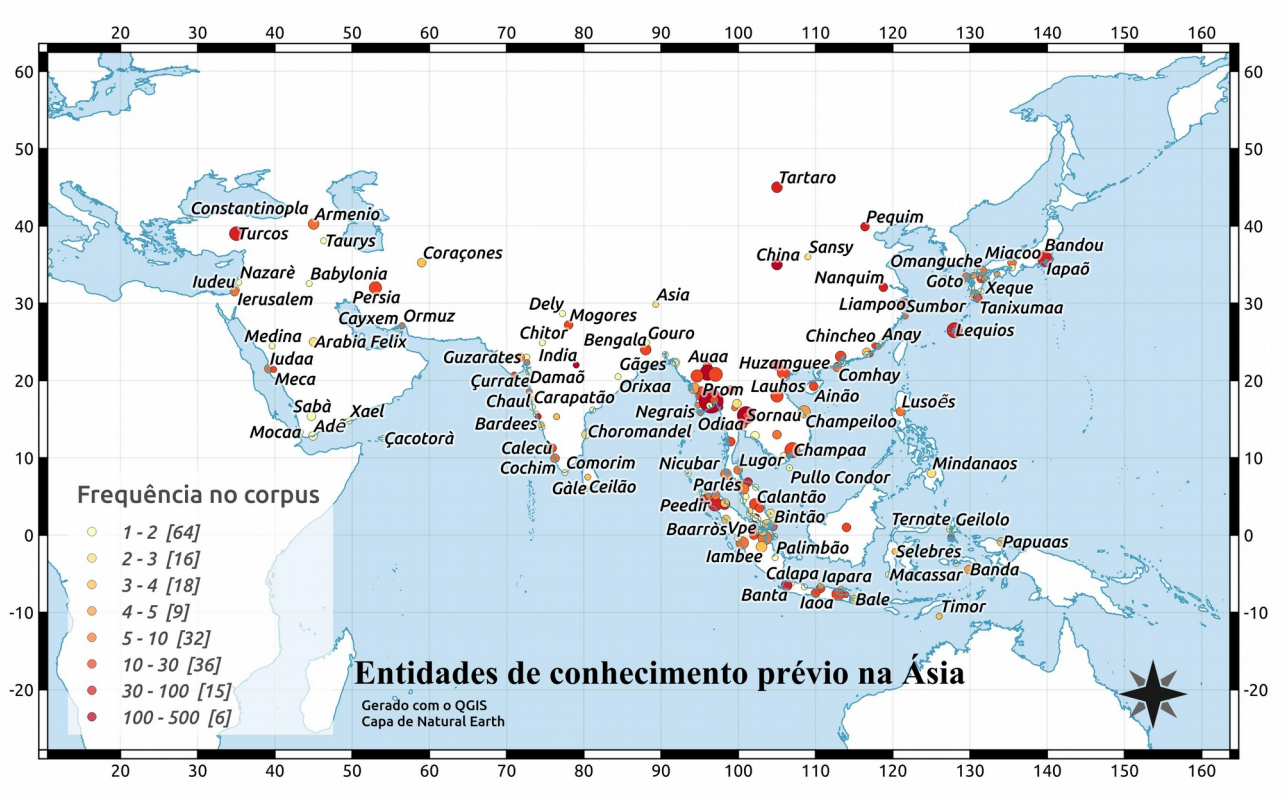


Figura 7.17: Entidades geográficas georreferenciadas por conhecimento prévio na Ásia.

Os do primeiro tipo são cidades (P) seguido de países e as suas principais divisões administrativas

(A); ilhas e geografia física terrestre (T); rios, praias e portos (H); grandes áreas e regiões (L) e casos pontuais de obras de engenharia e construções (S). Isto é, temos os primeiros elementos para a lista C definida em §6.2.1.3 que designamos pelos códigos:

C = {Lista de atributos geográficos} = {“A”, “P”, “T”, “H”, “L”, “S”}

O maior número de objetos georreferenciados em Ásia corresponde também com o facto de, tirado o caso de *Portugal*, serem também as entidades geográficas mencionadas com maior frequência no corpus. A distribuição das entidades num mapa (fig. 7.17) permite observar que, mesmo se há um maior número de entidades do tipo cidades e povoações, a geografia física aponta para zonas costeiras. Uma mesma entidade pode, portanto, pertencer a mais de uma categoria (por exemplo, ser uma cidade e um porto), porém, conforme foi apontado para os casos de metonímia (§6.3.1.3), em que mais de um tipo geográfico implica a consideração de objetos diferentes, selecionamos apenas o objeto mais abrangente para a classificação por tipos.

7.3.5.2 Abrangência da lista por conhecimento prévio sobre as entidades mencionadas no corpus

A figura 7.18 abaixo mostra como a lista de conhecimento prévio obtém uma abrangência de 80% nas ocorrências das entidades geográficas mencionadas no corpus *Peregrinação 1614* ficando apenas 20% por georreferenciar.

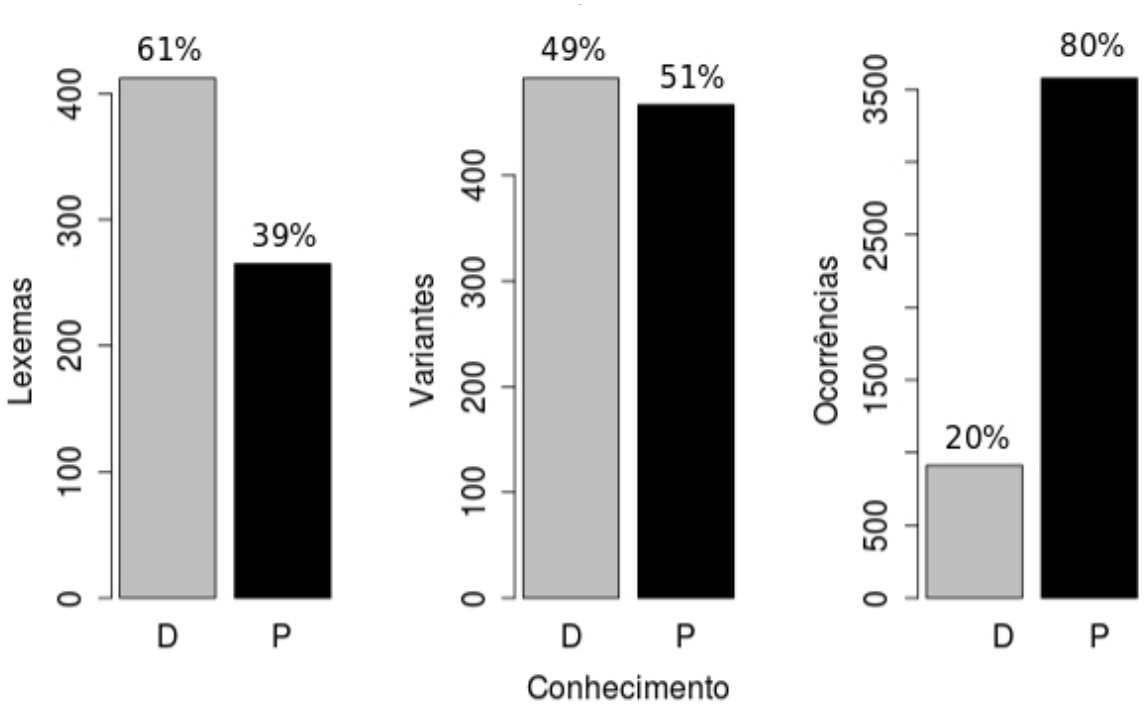


Figura 7.18: Entidades geográficas mencionadas no corpus da *Peregrinação 1614* abrangidas pela lista por conhecimento prévio (P) vs. conhecimento adquirido pela descrição (D) segundo lexemas, variantes e ocorrências totais.

Quanto a lexemas distintos, só 39% têm um referente associado, ficando 61% sem georreferenciar.

Isto deve-se a que as entidades geográficas mencionadas com maiores frequências são também aquelas de que temos um melhor conhecimento prévio: na fig. 7.19 com apenas a soma das frequências três primeiros lexemas obtemos um valor > 1000 ocorrências.

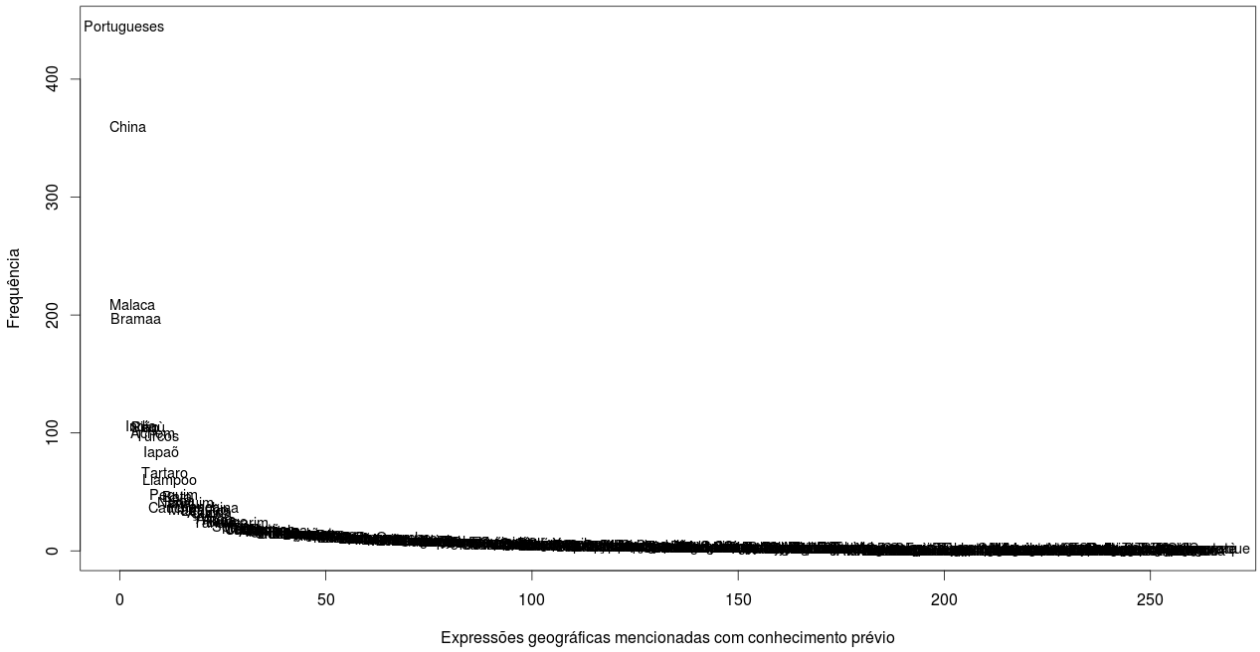


Figura 7.19: Frequências dos lexemas abrangidos pela lista por conhecimento prévio.

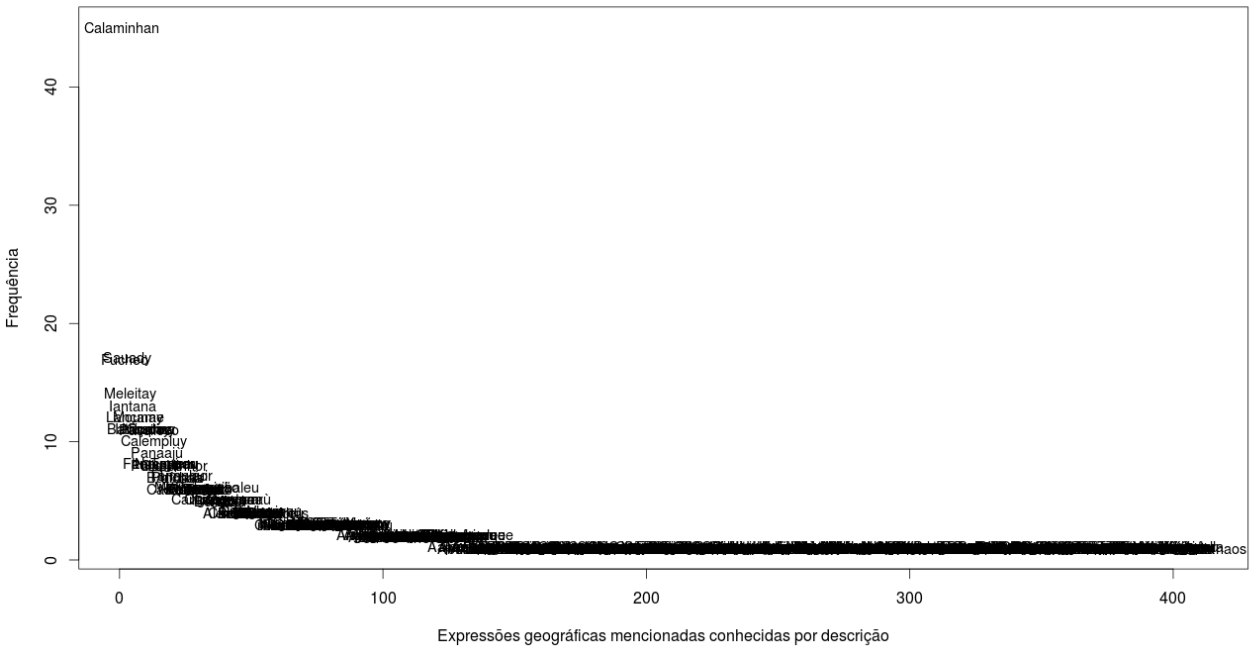


Figura 7.20: Frequências dos lexemas não abrangidos pela lista por conhecimento prévio.

No entanto, das entidades geográficas mencionadas não abrangidas pela lista de conhecimento prévio (fig. 7.20), a escala é um fator 10^{-1} menor, a entidade geográfica mencionada com maior

frequência, *Calaminhan*, ficaria na parte baixa da curva que leva à linha de hápax no conhecimento prévio. Observamos também que a lista se prolonga no eixo horizontal (linha contínua de hápax legomena) para além dos 400 lexemas (fig. 7.20) (61% na figura 7.18), no entanto a do conhecimento prévio (fig. 7.19) não chega aos 300 (39%, na figura 7.18), isto é, a lista de conhecimento prévio abrange menos lexemas que os que ficam por georreferenciar na lista de entidades conhecidas por descrição.

Quanto a tipos de objetos geográfico (§7.3.5.1), são as entidades com maior volume de povoação (países e cidades) as que ocupam as posições mais altas na lista de conhecimento prévio, uma observação comum (§4.7.1; Lois-González & López-González, 2013) para a análise das entidades mencionadas no corpus e o objeto geográfico que representam.

7.3.5.3 SIG de entidades de conhecimento prévio

Extraímos o índice de entidades georreferenciadas para um mapa vetorial num formato padrão, conforme a um SIG histórico (Gregory & Ell, 2007; Bol 2013). Todas as entidades, independentemente de se corresponderem com pontos, linhas ou áreas, são representadas como pontos. Esta solução também tem sido adotada no SIG histórico da China (Bol & Ge, 2005) pela dificuldade de definir áreas para as entidades de tipo administrativo.

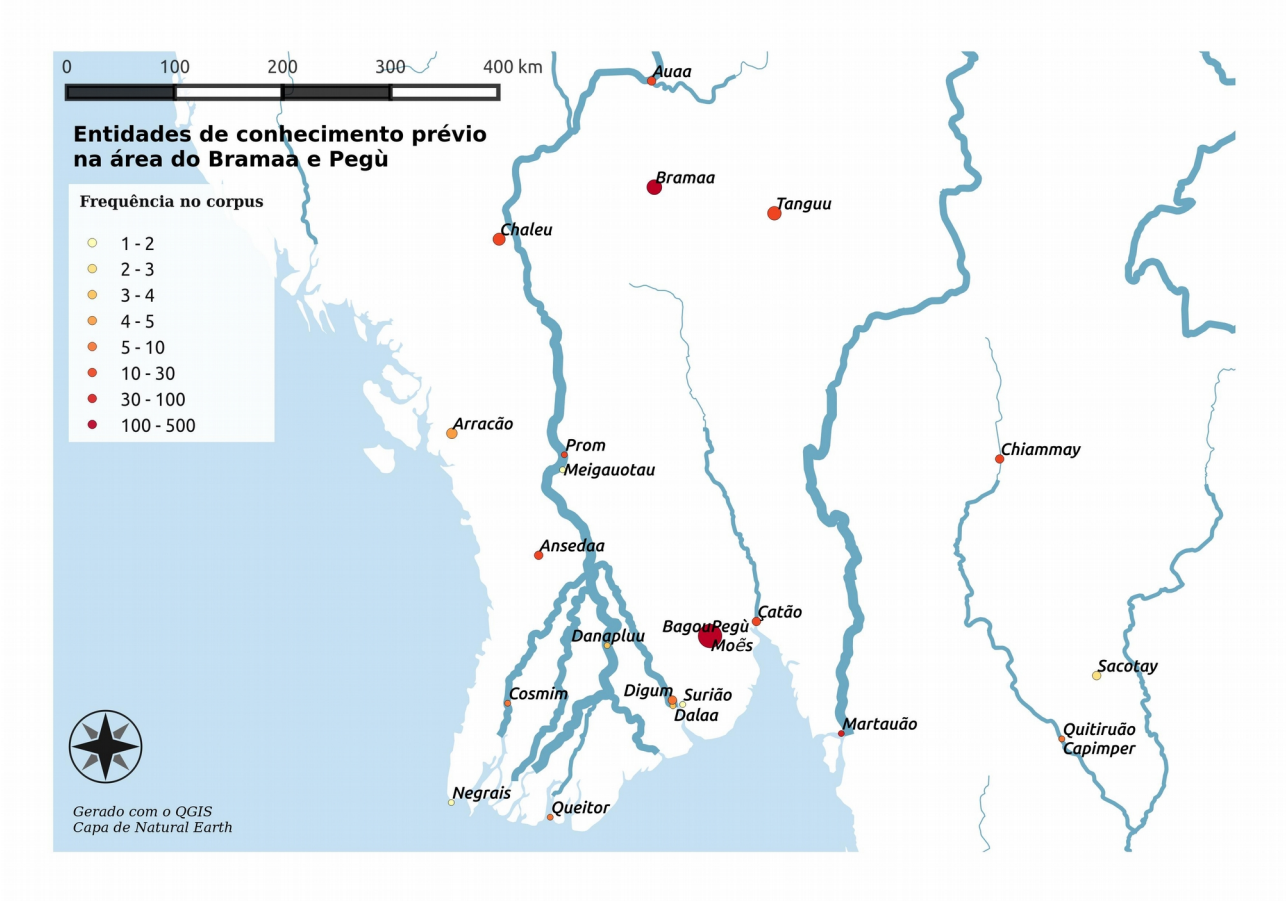


Figura 7.21: Entidades georreferencidas por conhecimento prévio na área do *Bramaa* e *Pegù*.

Os dados da nossa base podem ser combinados e visualizados sobre novas capas para facilitar a

compreensão e pesquisar relações com outros dados geográficos. Na figura 7.21 mostramos uma representação cartográfica das entidades da área do *Bramaa* e *Pegù* sobre uma capa vetorial de rios e oceanos facilitada por Natural Earth[11]. As entidades que se correspondem com linhas (caso do rio *Queitor*) e polígonos (reinos como o *Bramaa* e *Pegù*) são formalizadas como pontos pelas coordenadas recuperadas nas pesquisas sobre o objeto geográfico (§7.3.4). O objetivo principal da representação cartográfica é auxiliar o georreferenciamento das entidades que têm de ser resolvidas por descrição.

7.4 Conclusão da georreferenciação com a lista de conhecimento prévio

Neste capítulo aplicamos um modelo referencial para operarmos com os objetos geográficos. Invertemos a orientação da expressão (entidade geográfica mencionada) para o objeto geográfico do mundo real sendo assim que criamos uma lista de referentes (entidades geográficas físicas) ligados às expressões das entidades geográficas mencionadas no corpus. Mesmo quando introduzimos mais um elemento na relação, o conceito, o referente é sempre um objeto físico. No nosso modelo, o conceito opera com uma categoria (tipo de objeto geográfico) e uma relação (pertença do objeto a outro objeto geográfico que o inclui). As relações entre expressão, conceito e referente permitem-nos definir uma tipologia e sistematizar as dificuldades nos trabalhos de desambiguação que foram introduzidas ao atender à identificação automática de entidades geográficas mencionadas (cap. 5). Com o objetivo de facilitar a solução do referente, estabelecemos uma primeira classificação de todas as entidades geográficas mencionadas em dois tipos: aquelas cujo conhecimento prévio permite a georreferenciação diretamente e aquelas outras que temos de georreferenciar ampliando a descrição oferecida pelo corpus.

Para a criação de uma lista de referentes de conhecimento prévio, em primeiro lugar criamos uma base documental ampla, a partir da qual elaboramos uma lista com estudo crítico e, finalmente, estabelecemos uma regra (critério) para selecionarmos apenas aqueles referentes cujo conhecimento seja dado como certo e não só possível, não apresentem contradições dentro da base documental nem contradição evidente com o contexto mais imediato da sua expressão no corpus (as suas concordâncias). Incorporamos o estudo crítico no painel de consulta do corpus, elaboramos mapas de apoio e ligamos os resultados a uma base toponímica de âmbito global. A integração dos dados da base toponímica com o corpus permite geovisualizar e descrever os referentes das entidades geográficas mencionadas em termos de tipos geográficos, estabelecer relações de holonímia e recuperar coordenadas geográficas associadas.

O resultado obtido é uma lista de referentes com uma probabilidade, segundo a base documental e o estudo crítico, P(georreferência da expressão) = 1, e uma abrangência de 80% do total de ocorrências de entidades geográficas mencionadas no corpus. Isto é, 80% das ocorrências estariam já georreferenciadas. Esta proporção muda a respeito da abrangência por tipos (lexemas), apenas 39% das entidades geográficas mencionadas são conhecidas como certas por conhecimento prévio,

[11] http://www.naturalearthdata.com/

a explicação é as entidades mencionadas por conhecimento prévio ocuparem as posições mais altas na organização das expressões por categorias de frequência (distribuição de Zipf).

Processamos os resultados obtidos da elaboração da lista de conhecimento prévio num SIG para obtermos uma cartografia de referência da *Peregrinação*. Um problema é a representação das entidades cuja forma vetorial são linhas e polígonos, que resolvemos pela atribuição de umas coordenadas geográficas a partir de um ponto representativo obtido do seu referente contemporâneo. Toda entidade por conhecimento prévio tem, portanto, uma representação por coordenadas.

7.5 Sumário de objetivos

Os objetivos principais desta secção foram:

- Criar uma base documental de conhecimento prévio e um estudo crítico para as entidades geográficas mencionadas do corpus da *Peregrinação 1614.*
- Integrar o estudo crítico extraído da base documental no corpus da *Peregrinação 1614.*
- Criar uma lista de entidades geográficas de conhecimento prévio, aplicá-la e avaliar os seus resultados ao referenciar o corpus da *Peregrinação 1614.*
- Processarmos os resultados da lista de conhecimento prévio num SIG.

Capítulo 8

Atributos geográficos para a georreferenciação por descrição

A georreferenciação por conhecimento prévio fornece uma lista de entidades geográficas cuja localização damos, em princípio, em termos de coordenadas com a maior probabilidade. O procedimento para o georreferenciamento consistiu na elaboração de um estudo crítico que seleciona as entidades cujo referente é resolvido sem achar contradições evidentes nem na base documental nem na descrição do corpus. Ficaram por georreferenciar todas aquelas cuja localização apresenta dificuldades, quer por não formarem parte ou serem duvidosas na base documental, quer por entrar a georreferência prévia em contradição evidente com a análise das concordâncias.

No modelo semântico introduzido no capítulo 6, indicamos também dois modos de resolver o georreferente: 1) pelas suas coordenadas e 2) por uma definição, esta última elaborada a partir de um tipo geográfico e uma relação com uma outra entidade. Consequentemente, procede a abordagem das entidades geográficas mencionadas na medida em que são descritas no corpus a fim de obtermos uma definição para cada uma, seja a entidade conhecida na base documental ou não. Neste capítulo trabalhamos sobre o primeiro elemento da definição, o atributo geográfico, para elaborarmos uma lista de tipos que descrevam todas as entidades mencionadas. O objetivo é criarmos uma taxonomia como primeiro elemento descritivo das entidades, combinando termos geográficos extraídos do corpus com as classes aplicadas na descrição das entidades de conhecimento prévio.

Iniciamos o trabalho expondo a necessidade de selecionarmos quais são os segmentos do corpus a processar (aqueles que expressem feitos sobre as entidades geográficas mencionadas) para, a seguir, formalizarmos a relação entre a entidade e os atributos geográficos. Como exemplo prático partimos da lista de atributos obtidos das entidades de conhecimento prévio que agora queremos alargar para todas as entidades presentes no corpus e assim considerarmos as dificuldades que surgem da aplicação de um taxonomia dada. O resto do capítulo aborda o caso da extração automática dos termos do domínio geográfico. Começamos expondo o modo em que processamos o corpus para obter um subcorpus do domínio geográfico e introduzimos as métricas e métodos usados para a recuperação de termos. Aplicamos três blocos de testes. O primeiro, de carácter mais teórico, explora a incidência da frequência absoluta como único critério na recuperação de candidatos a termos do domínio numa distribuição zipfiana do corpus (fenómeno característico da linguagem humana e alheio ao desempenho da métrica). O segundo bloco analisa e extrai conclusões sobre o efeito do processamento do corpus, considerando distintos níveis de anotação e segmentação das

orações, junto com a aplicação de métodos comuns para a filtragem e otimização das métricas na seleção de candidatos. Finalmente, um terceiro bloco aproveita os resultados do melhor desempenho do bloco anterior para proceder a uma melhoria de resultados por meio da validação semântica das listas de candidatos mediante a aplicação de uma base lexical.

8.1 A elaboração do conceito por descrição (conhecimento adquirido)

Em §7.1 distinguimos dois tipos de conhecimento aplicáveis à referenciação de uma entidade geográfica mencionada, o prévio e o descrito. O primeiro permite referenciar diretamente a expressão mediante coordenadas geográficas precisas. No modelo de georreferenciação que deixáramos introduzido em §6.2, também elaboramos um conceito para a entidade geográfica mencionada (§6.2.1.3) exprimível por meio de uma definição em que o primeiro termo é um atributo geográfico que caracteriza a entidade. Nesta secção formalizamos os elementos precisos para operarmos com o conhecimento descrito na recuperação dos tipos geográficos das entidades mencionadas anotadas no corpus.

8.1.1 Proposições e orações

O conhecimento descrito é aquele obtido da análise do texto. Dada uma entidade geográfica *w* mencionada no corpus, interessa-nos selecionar apenas aqueles contextos que exprimam proposições que contribuam para a georreferência. Usaremos a noção de proposição (§3.6) para nos referirmos a uma forma elaborada da oração com valor de verdade sobre um feito. Uma mesma proposição pode ser expressada por orações distintas. A proposição permite-nos reduzir o trabalho de georreferenciação limitando o número de orações sobre as quais operar: uma vez elaborada a proposição de uma oração, qualquer outra oração que expresse a mesma proposição não precisa ser mais processada.

No nosso corpus, a unidade oração é assimilável de modo aproximado à de concordância enquanto unidade textual de um corpus recuperada a partir de uma unidade lexical, por exemplo, quando pesquisamos uma expressão no painel de consulta. A elaboração de proposições para a georreferenciação de uma entidade geográfica mencionada *w* começa, portanto, com a extração de concordâncias.

8.1.2 Da oração ao conceito

No exemplo introduzido em (7a) elaboramos um conceito a partir das premissas extraídas por uma simples concordância.

> “(…) chegey ao rio de **Puneticão**, onde està situada a cidade de **Aarù**.”

O primeiro fenómeno que notamos é a multiplicidade de proposições com valor geográfico extraíveis de uma simples concordância:

Puneticão é um rio

Aarù é uma cidade.

Puneticão passa por Aarù.

Aarú está situada no rio Puneticão.

O rio Puneticão está situado em Aarù.

Também, para a georreferenciação de uma entidade geográfica, elaboramos a noção de conceito (§6.2.1.3).

Ex. Puneticão: <rio, Aarù>

Uma forma de formalizar a expressão do conceito é por meio de uma definição (§6.2.1.3). Se analisamos as definições como orações obtemos proposições tais como:

Puneticão é um rio de Aarú.

Assim, as proposições que nos interessa extrair do corpus são aquelas que afirmam feitos sobre o conceito, isto é, aquelas que consideram:

1) o atributo geográfico da entidade geográfica mencionada

2) uma relação da entidade mencionada com outra previamente conhecida ou relacionável com mais outra entidade previamente conhecida.

Neste capítulo atendemos ao primeiro elemento da elaboração do conceito, a predicação de um atributo geográfico. A relação de uma entidade geográfica mencionada com outra será abordada no capítulo 9.

8.1.3 A atribuição do tipo geográfico

O atributo é uma qualidade do objeto geográfico que o classifica como instância de uma classe particular de tipos geográficos. A condição prévia para determinarmos o tipo é afirmarmos a existência da entidade, isto é, termos um referente para a expressão mencionada no corpus.

8.1.3.1 Existência de um referente

Dada a expressão de uma entidade geográfica mencionada $w_i \in W$ estabelecemos o enunciado existencial do referente como uma relação binária entre uma expressão e o objeto no mundo real através do seu georreferente:

$$\exists\, g \in \mathrm{G},\ \mathrm{Tem\_Georreferente}(w,g)$$

Isto é, ao declararmos que uma entidade mencionada tem um georreferente, estamos a afirmar também a sua existência. Exemplos de orações que expressam esta proposição:

Aarù é.

Existe Aarú.

Há uma cidade chamada Aarú.

A regra de georreferenciação inicial para qualquer entidade geográfica mencionada cujo referente é conhecido apenas por descrição para o caso de estudo desta tese é:

$$\forall\, w \in \mathrm{W}\ (\exists\, g \in \mathrm{G},\ \mathrm{Tem\_Georreferente}(w,g)\ \Lambda\ \mathrm{g} = g_0) \qquad (8.1)$$

Sendo g_0 um ponto inicialmente indefinido no Planeta Terra.

O problema da determinação da existência da entidade geográfica é resolvido pela anotação da entidade geográfica mencionada (cap. 5). Quando uma expressão é identificada e anotada como entidade geográfica mencionada, estamos a afirmar a existência de um objeto geográfico que, pela regra (8.1), vem limitado a um lugar no planeta Terra no presente ou num momento histórico recuperável (isto é, não em Marte ou num mundo fictício ou possível mas não real).

A georreferência da entidade mencionada declara a existência do referente, mas o conceito tal e como resolvido por (8.1) denota o conjunto de entidades no planeta Terra. O resultado da georreferenciação será tanto mais preciso quanto *g* mais próximo estiver de denotar apenas uma unidade (isto é, a apontar para um único objeto geográfico).

8.1.3.2 Predicado monádico

Uma vez que temos declarado que o objeto geográfico existe, procuramos as suas propriedades. Uma descrição monádica da entidade geográfica com expressão w_i é aquela relação composta unicamente de um predicado e un nome, o da entidade geográfica mencionada w_i.

Ex. Ilha(w_i)	Ex. Reino(w_j)
w_i = Çamatra	w_j = Sião
Çamatra é uma ilha	Sião é um reino

Estes predicados descrevem e classificam as entidades ao lhe atribuírem uma propriedade. Para a definição do conceito, procuramos os predicados da lista de atributos geográficos (§6.2.1.3).

Se uma entidade mencionada não tiver uma descrição do tipo geográfico, ser-lhe-á outorgado um tipo genérico. Toda entidade geográfica mencionada é predicada, por falta de um tipo específico, pelo atributo *Lugar*.

$$\forall\, w_i \in \mathrm{W},\ \mathrm{Lugar}(w_i) \qquad (8.3)$$

Definido o lugar como um ponto (§3.1.1), o tipo *Lugar* para qualquer entidade geográfica é independente da extensão ou característica física da entidade geográfica individual. Verifica-se em termos de escala para o conjunto, uma entidade tende a converter-se num ponto a medida que aumenta o espaço representado.

8.1.3.3 O atributo como tipo da classe

A consideração dos tipos geográficos como predicados numa relação monádica com a entidade

geográfica obriga a considerar uma relação distinta para cada um dos tipos. Isto é, para a elaboração da definição do conceito necessitamos, para além da relação *Parte_de*, tantas relações como tipos geográficos distintos. Uma solução que simplifica o problema é considerar uma única relação que recupere os tipos da lista de atributos geográficos que deixáramos definida em (§6.2.1.3):

c = {Lista de atributos geográficos} = {lugar, ilha, rio, …}

A relação que liga uma expressão com um tipo geográfico é:

Tem_tipoGeográfico(w,c)= "*w* tem o tipo geográfico *c*" (8.4)

Assim, nos exemplos da secção anterior, resolvemos o tipo com uma única relação:

Ex. Tem_tipoGeográfico(w_i,c_i)	Ex. Tem_tipoGeográfico(w_j,c_j)
w_i = Çamatra	w_j = Sião
c_i = Ilha	c_j = Reino
Çamatra é uma ilha	Sião é um reino

8.2 Os atributos geográficos no corpus

A descrição das entidades geográficas mencionadas em termos de um tipo foi já aludida nos capítulos anteriores. No estudo crítico (§7.2.2) classificamos manualmente as entidades geográficas mencionadas por meio da comparação das concordâncias no corpus, os dados da base documental e a geovisualização do objeto geográfico. No caso das entidades de conhecimento prévio usamos uma base de dados geográficos global com uma taxonomia que lhe outorga um tipo geográfico único ao objeto (§7.3.5.1). Nas subsecções a seguir consideramos primeiro as limitações e dificuldades de aplicar uma taxonomia prévia e da aplicação só dos tipos que aparecem no corpus. Posteriormente analisamos mais uma solução, híbrida, que combina uma taxonomia prévia para estruturar os tipos em classes com a elaboração de um vocabulário para a definição dos tipos a partir das descrições das entidades no corpus.

8.2.1 Recuperação de atributos numa taxonomia prévia

A elaboração dos atributos geográficos pode partir de uma lista preexistente, por exemplo, o glossário geográfico ou ontologia na qual selecionamos aqueles termos relevantes para descrever as entidades mencionadas do corpus. No caso das entidades geográficas por conhecimento prévio, ao aplicarmos um georreferente contemporâneo segundo a base de dados de GeoNames, obtemos uma taxonomia genérica (§7.3.5.1). O principal problema a solucionar com a hierarquia predefinida consiste em selecionar os tipos instanciados pelos objetos do corpus, isto é, escolher só aquele atributo que descreva a entidade mencionada georreferenciada. Surgem várias dificuldades da aplicação:

• O objeto na atualidade não refere a unidade geográfica histórica mencionada no corpus. Por exemplo, *Lequios* no corpus representa um conjunto de ilhas, *Taiwan* e as *Ryu-kyu*, hoje em dia não aparecem como unidade e, portanto, não serão assimiláveis a um único tipo geográfico administrativo se quisermos selecionar o tipo mais próximo àquele do corpus (*reino*) dentro dos disponíveis na taxonomia.

• O objeto na atualidade refere um tipo geográfico distinto ao mencionado no corpus. Por exemplo, *Goa* aparece descrita no corpus como um porto, hoje em dia a classificação mais comum é: unidade administrativa da *Índia*.

• A entidade mencionada tem relações metonímicas com múltiplos objetos que dificultam a escolha de um tipo. Por exemplo, *Malaca* é uma unidade administrativa, mas também uma cidade.

• O tipo geográfico no corpus não aparece na taxonomia. Por exemplo, o atributo *império* descreve várias entidades no corpus, não obstante, não aparece como disponível na taxonomia aplicada para as entidades de conhecimento prévio.

Pelo exposto, a aplicação de termos geográficos independentes do corpus ficou limitada aos casos do estudo crítico quando usamos os mesmo termos propostos pelos glossários históricos (Albuquerque, 1994; Lagoa, 1950-1953; Alves, 2010) para a descrição da entidade no estudo crítico, e à taxonomia de GeoNames para a georreferenciação de entidades com conhecimento prévio.

8.2.2 Recuperação de atributos no corpus

Para além das dificuldades de adequação, temos que considerar a expansão da taxonomia a todas as entidades do corpus, isto é, deve abranger aquelas não consideradas como conhecidas previamente, cuja caracterização não é contrastável nas instâncias associadas à taxonomia prévia ou aparecem muito superficialmente (ou não aparecem associadas a tipo nenhum) na base documental. Neste caso, ao não haver um referente recuperável, a principal, se não única, descrição da entidade é a do corpus.

Uma taxonomia *ad hoc*, específica para o corpus, resolve parte dos problemas citados anteriormente. Consideramos primeiro um levantamento manual, soluções de recuperação automática de termos do corpus serão analisadas em §8.2.4. No caso da extração manual dos termos a partir das concordâncias das entidades mencionadas achamos também dificuldades:

• Quando a entidade mencionada refere vários tipos geográficos. Por exemplo: *Meleitay* é uma *cidade*, um *rio* e uma *fortaleza*. Neste caso temos de aplicar uma regra ou regras que escolham o tipo preferente e recolhermos os alternativos. No caso prático selecionamos o tipo mais abrangente (o de maior extensão), e o tipo administrativo sobre o físico.

• Quando a entidade mencionada não tem um tipo específico que a descreva de modo explícito. Neste caso, se não for possível inferir um tipo geográfico pelo contexto, aplicamos a regra (8.3), o tipo outorgado é *lugar* como representante do valor mais neutro na taxonomia e a entidade

geográfica não pertence a classe geográfica específica nenhuma.

8.2.3 Método híbrido

A combinatória da taxonomia prévia e o aproveitamento de dados anotados no corpus recolhe as vantagens das duas taxonomias. O apêndice II mostra os tipos geográficos obtidos primeiro da análise de concordâncias das entidades (recuperação no corpus) e a sua posterior melhoria com os critérios considerados no resto desta secção. Não obstante, surgem contradições para a escolha dos tipos de uma ou outra taxonomia.

8.2.3.1 O tipo da taxonomia prévia define a entidade cujo tipo não é explicitado no corpus

No caso de uma entidade ser conhecida previamente, quando não acharmos o seu tipo descrito no corpus, aplicamos aquele recuperado por conhecimento prévio.

8.2.3.2 A categoria da entidade geográfica define o tipo

Ex. *Gentílico → nação*. Neste caso usamos uma propriedade da entidade anotada do corpus, o facto de ser gentílico, para, em caso de não ter nenhum tipo geográfico que a descreva, outorgar-lhe o tipo *nação* (preferida aqui a soluções mais neutras como *povo* por evitar a ambiguidade de povo como tipo geográfico unidade de povoação menor que *cidade*).

8.2.3.3 A ambiguidade determina a seleção do termo

Se o tipo da taxonomia prévia produz ambiguidade no corpus, escolhemos a forma menos ambígua dentro de expressões sinónimas ou quase-sinónimas. O caso anterior serve de novo de exemplo. Um gentílico como etnónimo pode ser descrito na taxonomia como *povo* ou *nação*, apenas é preciso um dos dois termos. Não obstante, a forma *povo* entra em ambiguidade com outro tipo, *povo* como sendo entidade de povoação menor do que *cidade*. O critério de ambiguidade condiciona a seleção. Escolhemos *nação* como tipo para a taxonomia por ser menos ambíguo que *povo*.

Caso excecional é o tipo *lugar*, cujo valor é tanto o de espaço indefinido como pequena entidade de povoação (§3.1.1). Dada a sua relevância na taxonomia, apenas o tipo genérico é considerado para desfazer a ambiguidade. No caso de que uma entidade mencionada tenha como tipo *lugar* com o valor de pequena povoação, é classificada como *povoação*.

Ex. “& chegando a hum lugar que se dizia ***Aapessumhee***, quatro legoas do rio de **Puneticão**” (*PR*, 32)

Aapessumhe aparece como *lugar* no corpus, é classificado como *povoação*.

O termo *lugar* na taxonomia fica limitado para aquelas entidades cujo tipo não se conhece ou é muito ambíguo, seja da classe que for, e independentemente da sua maior ou menor superfície.

Ex. “ao qual este Principe mandaua resgatar por hum mercador **Iudeu** natural de ***Azebibe***”

Não achamos elementos no corpus que determinem qual é o tipo da entidade *Azebibe*, se uma

pequena *povoação*, uma *cidade*, uma *região* ou qualquer outro atributo físico ou administrativo. Portanto, é classificado como *lugar*, isto é, dado que é uma entidade geográfica, ocupa um espaço e tem um tipo geográfico, mas desconhecemos a sua extensão e características.

8.2.3.4 Redução da entropia da classe

Um problema surde ao considerar aqueles tipos que classificam apenas uma ou um número muito reduzido de entidades, particularmente quando há outro tipo semanticamente muito próximo. A eliminação de tipos dificulta a recuperação das entidades no corpus quando forem pesquisadas pelo seu atributo geográfico, no entanto, uma taxonomia com muitas classes para poucas instâncias é pouco operativa nos trabalhos analíticos e de inferência. Calculamos a entropia das classes da taxonomia e reduzimos o número de tipos agrupando formas similares dentro de uma mesma subclasse. A tabela 8.1 mostra o valor entrópico da classe antes e depois da redução.

Código da classe	Termos do domínio aceites como subclasse	Entidades mencionadas na classe	Entropia
P	(5) 5	196	(1.542653) 1.542653
A	(8) 4	109	(1.50368) 1.144638
T	(15) 9	102	(2.217072) 1.695278
H	(15) 13	153	(2.439175) 2.3622
L	(3) 3	52	(1.13154) 1.13154
S	(22) 8	50	(3.941148) 2.33779

Tabela 8.1: Termos geográficos, entidades distintas e entropia das classes da taxonomia obtida pela combinatória de termos de corpus e taxonomia prévia antes (entre parêntese) e depois da redução de número de tipos.

Consideramos dois tipos de agrupamentos:

• Similitude morfológica. Ainda oferecendo um matiz semântico para além do valor gramatical, os plurais (*edifício*, *edifícios*; *ilha*, *ilhas*) têm uma relevância menor na classificação e reduzem os tipos da classe.

• Similitude semântica. Formas agrupáveis num mesmo synset numa base lexical semântica a respeito dos demais termos do grupo são reduzidas a um mesmo tipo nas classes com maior entropia. Assim, para a classe *Construções*, temos os termos: *abadia* (2 entidades), *ermida* (2), *mosteiro* (1), *pagode* (7), *templo* (6), todos agrupáveis baixo um tipo comum *templo*. Do mesmo modo *antesala* (1), *aposento* (1), *casa* (2), *hospedaria* (1), *paços* (1), instanciados por uma única entidade exceto *casa* que tem 2, podem ser todos agrupados baixo um mesmo termo, neste caso obtido da taxonomia prévia, *residência*.

8.2.3.5 Resultados

Como resultado, obtemos uma taxonomia que descreve todas as entidades geográficas do corpus em dois níveis (fig. 8.1). A relação entre os tipos geográficos e a classe é hiponímica (§3.4.3.2), assim quando nos referimos aos tipos como membros da classe, caso de *fortaleza*, dizemos que é um hipónimo de *construções*. Pela sua parte, se nos referimos ao atributo que define a classe, a relação é de hiperonímia: *construções* é um hiperónimo de *casa*, *castelo*, *fazenda* e *fortaleza*. Finalmente, os membros de uma mesma classe são cohipónimos. Assim, *costa*, *enseada*, *estreito*, *lago*, *mar*, *ponta*, *porto* e *rio* são cohipónimos agrupados na mesma classe como sendo tipos hidrológicos.

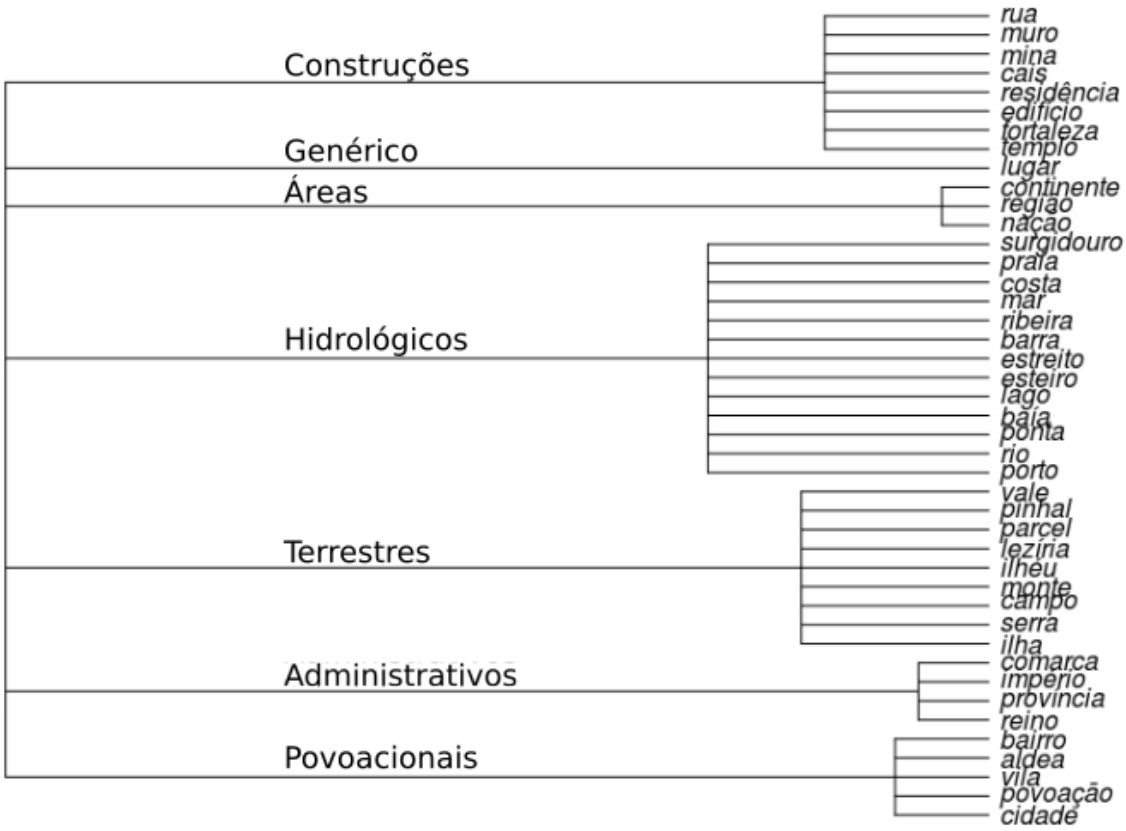


Figura 8.1: Taxonomia para a classificação das entidades geográficas do corpus.

8.3 Abordagens automáticas

Os termos de um domínio são aqueles característicos de uma temática e conformam a terminologia específica de uma área. Neste sentido, a recuperação de termos pode ser entendida como uma atividade lexicológica que distingue nos vocábulos um valor diferencial relativamente aos termos comuns e aqueles não específicos do domínio característico do corpus. A frequência dos termos (TF, pelo inglês *term frequency*) tem sido aplicada para análise da variedade lexical como representativa de uma variedade do português (Peres Rodrigues, 2000) e para a extração de léxicos preferenciais (Zapparoli & Camlog, 2002; Zapparoli, 2010), a assunção principal, as frequências reflectem a

variedade ou temática de um corpus. Um método de análise comum na linguística de corpus consiste na comparação das frequências normalizadas do léxico contido em vários documentos aplicando uma medida para capturar os vocábulos que oferecem o maior contraste. Como específico para a extração de termos, Lopes, Fernades e Vieira (2016) comparam o efeito das métricas de base estatística partindo deste mesmo princípio de contraste de corpora de distintos domínios.

A extração de termos aparece também no início do processo de aprendizagem de ontologias (Buitelaar, Cimiano & Magnini, 2005), resultante de operar sobre corpora e recursos textuais estruturados (Buitelaar, Cimiano & Magnini, 2005; Wilks & Brewster, 2006; Gonçalo Oliveira & Gomes, 2010). Zahra, Malucelli, Freddo e Tacla (2014) revisam software disponível para a aprendizagem de ontologias a partir do texto com atenção às soluções para português e Guimarães (2015) descreve trabalhos relacionados.

Como parte de uma proposta de aprendizado de máquina, Conrado (2014) oferece uma visão de conjunto e compara abordagens e métricas para a extração de terminologia baseada em corpora de domínio para o português.

A identificação de termos específicos do domínio geográfico é atendida em trabalhos dirigidos à criação de corpora com o objetivo de capturar entidades, terminologia, expressões e estruturas sintáticas reveladoras de conteúdo espacial. Stock, Pasley, Gardner, Brindley, Morley e Cialone (2013) revisam métodos e descrevem a elaboração de um corpus geoespacial de modo semiautomático.

As abordagens aplicadas no reconhecimento automático de termos baseiam-se em métodos estatísticos e atributos linguísticos, a soma de ambos propicia modelos híbridos (Maynard, Li, & Peters, 2008; Conrado, Pardo, & Rezende, 2015).

O problema da caracterização dos tipos geográficos das entidades mencionadas é também atendido como parte do labor NERC, como mais uma subclassificação dentro da classe entidade mencionada lugar; uma solução, aceitar uma taxonomia predefinida (Martins & Silva, 2007). Como problema independente das entidades mencionadas e do domínio geográfico, a extração de termos recupera listas de candidatos para um domínio e aplica critérios de seleção para reduzir o número de candidatos (Conrado, 2014). Neste sentido, automatiza o processo de elaboração do vocabulário, que finalmente tem de ser validado por especialistas do domínio (Almeida, Aluísio, & Oliveira, 2001) ou contrastado com listas especializadas e recursos externos (Wendt, Lopes, Martins, Vieira, & Lima, 2010).

No resto do capítulo exploramos o problema da elaboração da taxonomia como parte do trabalho de extração de termos de um domínio, usando a entidade mencionada anotada no corpus para a recuperação de termos.

8.4 Caso prático de extração automática de tipos geográficos do corpus

No caso do corpus estudado neste trabalho, empregamos uma combinatória de métodos estatísticos e linguísticos para concluirmos com um método totalmente automático, fazendo uso de um único corpus para a aplicação das métricas.

8.4.1 Procedimento

Começamos pela classificação das orações segundo contiverem como mínimo uma anotação de uma entidade geográfica mencionada. O pressuposto básico para a obtenção da mostra do corpus é que os termos geográficos aparecem perto das entidades geográficas. Partindo desta premissa, usamos procedimentos do PLN para chegarmos a um número reduzido de termos candidatos sobre que aplicamos a distribuição de Zipf. Os candidatos serão validados com o objetivo de os inserir numa ontologia em que instanciaremos as entidades geográficas mencionadas do corpus. Usamos o software estatístico R (R Core Team, 2016) para o processamento estatístico segundo métodos de linguística de corpus (Baayen, 2008; Gries, 2009) e análise de dados (Peng, 2015; Peng & Matsui, 2015).

Os pressupostos para o procedimento metodológico são:

1) A probabilidade (considerada pela observação empírica da frequência) de ocorrência dos termos pode ser utilizada para capturar os termos representativos do domínio ou temática do corpus.

2) Os termos de um domínio podem ser organizados em relações semânticas.

8.4.2 Processamento do corpus

Usaremos a oração delimitada pela marca de pontuação como unidade de análise e consideraremos as entidades geográficas anotadas como elemento que nos permite discriminar uma oração pertencente ao domínio específico da geografia. Isto é, operamos com um subconjunto e não com o total de orações do corpus.

8.4.2.1 Seleção das orações para o subcorpus

Selecionamos todas as orações que contenham como mínimo uma entidade geográfica mencionada como topónimo. O motivo para não considerarmos os gentílicos é duplo, por um lado, foca-se a análise linguística num objeto homogéneo (nome próprio), por outro, limita-se o número de orações a processar, em previsão de revisões para testar a qualidade das operações de PLN.

8.4.2.2 Semi-normalização

Com a finalidade de facilitar o trabalho das ferramentas de análise morfossintática e aumentar a eficácia de listas filtro para a deteção de termos mais comuns sem valor lexical, normalizamos formas gráficas cuja frequência e regularidade permite a aplicação de uma regra de substituição (grafias que têm sempre a mesma equivalência no padrão contemporâneo e formas pronominais e

terminações verbais regulares). Deste modo otimiza-se o uso de vocabulários contemporâneos e o processamento sintático automático da oração.

8.4.2.3 Redução dos topónimos e gentílicos a um tipo único

Uma das dificuldades para a extração dos termos geográficos é a dispersão dos tipos em função da diversidade dos topónimos. Assim, aguardamos que *Pequim* coocorra com o atributo *cidade*, mas não (ou muito menos frequentemente) com o de *ilha*. Reduzindo todos os topónimos a um tipo único capturamos os atributos coocorrentes com as entidades mencionadas numa única operação.

A tabela 8.2 mostra como os topónimos, identificados como pertencentes à classe *t* na anotação, são substituídos pelo representante genérico LOCAL000, com o qual duas entidades geográficas mencionadas, *Aarû* e *Paneticão*, são reduzidas a uma mesma expressão. A mesma operação é aplicada aos gentílicos, marca *g* na anotação, expressados como GENTILICO000 .

Texto anotado no corpus	Da armada que o <place id='5'><placeName class='g' id='12'>Achem </placeName></place> mandou contra el Rey de <place id='2'><placeName class='t' id='3'>Aarû</placeName></place>, & do que lhe socedeo chegando ao rio de <place id='459'><placeName class='t' id='656'>Paneticão</placeName></place>.
Texto com anotações substituídas	Da armada que o GENTILICO000 mandou contra el Rey de LOCAL000, & do que lhe socedeo chegando ao rio de LOCAL000.

Tabela 8.2: Exemplo de substituição das entidades geográficas anotadas por um tipo único que as classifica.

8.4.2.4 Lematização e marcado morfossintático

Consideramos duas operações de PLN. Uma para alargarmos a frequência dos tipos alvo, outra para reduzirmos a frequência dos tipos não-alvo.

• A lematização permite maximizar as frequências de um tipo geográfico no corpus. Assim:

freq(CIDADE)= freq(“cidade”) + freq(“cidades”)

• A análise sintática para a obtenção de categorias gramaticais elimina termos não alvo. No domínio dos tipos geográficos aplicado ao corpus, assumimos os termos serem nomes comuns formados por expressões simples (uma única palavra). A anotação morfossintáctica permite selecionar apenas os nomes comuns do subcorpus.

Na tabela 8.3 abaixo mostramos um segmento obtido pelo conjunto de ferramentas PLN do Linguakit (Garcia & Gamallo, 2015). Os tipos geográficos *barra* e *povoação* aparecem classificados como NC (nome comum). O topónimo com que coocorrem foi previamente processado para o reduzir ao tipo único dos topónimos LOCAL000.

Daqui de+aqui SPS00+RG desta de+esta SPS00+* paragem paragem NCFS000 nos o PP3MPA00 fomos ser VMIS1P0 demandar demandar VMN03S0 a o PP3FSA00 barra barra NCFS000 de de SPS00 LOCAL000 LOCAL000 Z , , Fc	onde onde RG chegamos chegar VMIP1P0 quasi quasi NC00000 à a+a SPS00+* meia meio AQ0FS0 noite noite NCFS000 que que PR0CN000 surgimos surgir VMIS1P0 na em+a SPS00+DA boca boca NCFS000 da de+a SPS00+* barra barra NCFS000 defronte defronte RG de de SPS00 uma um DI0FS0 povoação povoação NC00000 pequena pequeno NCFS000 que que PR0CN000 se se PP3CN000 dizia dizer VMII3S0 LOCAL000 LOCAL000 Z , , Fc

Tabela 8.3: Lematização, marcado morfossintático e subdivisão da oração com o Linguakit.

8.4.2.5 Subsegmentação da oração

A oração, tal e como segmentada no corpus (§4.3.3), é complexa e frequentemente assimilável ao pseudoparágrafo. Para limitarmos ainda mais os resultados ao âmbito de dependências da entidade geográfica mencionada, selecionamos o segmento mínimo em que a expressão da entidade ocorre.

Dois procedimentos foram aplicados. Um, sintático, resultado da subsegmentação do anotador morfossintático. Outro, numérico, pela definição de uma janela com longitude *n*, em que *n* representa o número de termos capturados tendo a entidade geográfica como centro. Ambos os dois critérios foram combinados para criar subsegmentações ainda mais seletivas que selecionam apenas os lemas dos nomes comuns mais próximos à entidade geográfica. Criamos assim representações reduzidas da oração com a entidade geográfica como centro.

A tabela 8.4 abaixo mostra o processamento de uma oração no texto original até à segmentação em unidades menores de cláusulas e frases. Os segmentos menores podem ser ainda reduzidos em n-gramas, usando critérios numéricos (janelas), ou aproveitando o marcado do anotador morfossintático para selecionar padrões baseados em categorias gramaticais, sobre os quais se podem volver aplicar mais critérios numéricos. O exemplo extremo seleciona apenas os núcleos

nominais situados a n=-1 distância da entidade geográfica mencionada.

Texto original		
Do que passey atè chegar ao reyno de Quedâ na costa da terra firme de Malaca do que ahy me aconteceo		
Normalização e substituição		
Do que passei até chegar ao reino de LOCAL000, na costa da terra firme de LOCAL000, e do que aí me aconteceu.		
Subsegmentação		
Sintática	**Janela**	**Sintática e janela**
Do que passei até chegar ao reino de LOCAL000,	Do que passei até chegar ao reino de LOCAL000,	reino LOCAL000, costa terra LOCAL000
na costa da terra firme de LOCAL000	na costa da terra firme de LOCAL000	
	e do que aí me aconteceu.	

Tabela 8.4: Processamento e segmentação da oração.

8.4.2.6 Resultados do processamento e modos do corpus

Para a extração dos termos consideramos os modos de processamento do corpus:

CTT. Conjunto textual formado pelas orações normalizadas que contêm como mínimo uma entidade geográfica mencionada com categoria de topónimo. O texto é processado como uma só unidade documental.

DATA2. CTT segmentado em orações, cada oração é uma unidade de processo independente.

DATA3. CTT processado de modo que cada oração contém unicamente os nomes comuns selecionados segundo o anotador morfossintático.

DATA4. CTT subsegmentado em cláusulas, apenas os nomes comuns que coocorrem com um topónimo são selecionados.

DATA5. Vetor com a lista dos termos mais próximos à entidade geográfica mencionada extraídos em DATA4. Corresponde-se com a subsegmentação *morfossintática e janela* na tabela 8.3.

8.4.3 Extração de candidatos a termos geográficos

Uma vez preparado o subcorpus, ensaiamos métodos para a extração de candidatos a termos do domínio geográfico. A figura 8.2 mostra o processo para a elaboração de uma lista. Partindo do texto do subcorpus, processado segundo uma das modalidades referidas em §8.4.2.6, aplicamos métricas seletivas de forma que obtemos uma lista que posteriormente avaliamos relativamente à precisão (número de termos que pertencem ao domínio).

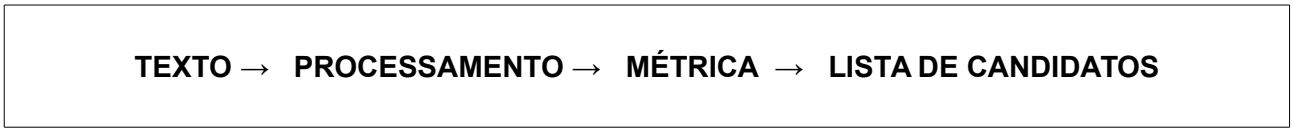


Figura 8.2: Processo de extração de candidatos a termos do domínio.

As subsecções a seguir avaliam os efeitos de distintas métricas sobre os distintos modos do corpus. Isto é, os testes conferem o desempenho de uma configuração do corpus (que representa o maior ou menor nível de PLN) combinado com uma métrica (que compara a efetividade das distintas fórmulas). A medida de avaliação mais importante é a precisão. A abrangência é considerada só quando definimos uns resultados aguardados a partir de um glossário ou lista de termos geográficos predefinida, exceto na primeira subsecção (§8.4.3.1) em que analisamos de um modo mais teórico o erro inerente à aplicação da frequência como métrica numa distribuição de Zipf.

A métrica fundamental de partida é a frequência absoluta, a partir da qual se elaboram métricas mais complexas. Como preâmbulo, um primeiro ensaio (§8.4.3.1) mostra o efeito da frequência absoluta sobre um corpus para o qual assumimos de antemão um número dado de verdadeiros positivos. Este teste tem uma base teórica e serve para mostrar a limitação da aplicação da frequência absoluta, que recupera necessariamente falsos positivos conforme diminui o valor de frequência dos candidatos. O resto das subsecções são de natureza empírica: consideram métodos para a redução dos falsos positivos com o objetivo de incrementar a precisão sobre o caso prático do corpus.

8.4.3.1 Desempenho da recuperação de candidatos baseada na frequência

Consequentemente com o pressuposto inicial (§8.4.1), dentro de um corpus de domínio, os termos podem ser recuperados pela sua frequência. Se o corpus é suficientemente representativo, os termos de domínio tendem a ter uma maior relevância e aparecem por cima da linha de hápax numa distribuição de Zipf do vocabulário (§4.7.1.1). Portanto, apenas os termos que superem uma frequência > *n*, onde *n* é o valor considerado relevante para o corpus, é que são considerados candidatos. Termos abaixo desta linha são menos relevantes para o domínio. Intui-se conveniente estabelecer um limite na frequência mínima necessária para recuperar um vocábulo. No entanto, mesmo com o limite definido, há um certo nível de ruído: a percentagem de termos não de domínio situados na mesma frequência que os verdadeiros positivos. O raciocínio é simples e mostramos

com um caso hipotético: se no corpus tivermos 50 termos do domínio e fixamos a frequência 4 como limite por abaixo do qual não se aceitam candidatos, mas há 200 termos com frequência 4 no corpus, a recuperação de termos baseada unicamente na frequência recuperará os 200 termos, quando em realidade apenas são necessários 50. Para analisarmos este efeito no corpus elaboramos uma série de simulações em que, dado um número de candidatos a termos *N* e um número de termos *T* para o total de termos representativos do domínio presentes no corpus, obtemos a frequência mínima precisa para recuperarmos *T* em função de um nível de desempenho na precisão predefinido segundo os seguintes parâmetros:

Distribuição de Zipf: resultado de ordenar as frequências segundo uma distribuição de Zipf para o corpus em modo CTT.

Precisão: A precisão face à que considerar os efeitos da distribuição das frequências. Os melhores resultados que encontramos (Conrado, Pardo & Rezende, 2015) superam 30% de medida-F em trabalhos de extração de termos para o português, com uma precisão em torno de 25%. Usamos esta percentagem como o nível de precisão suposto no desempenho da métrica sobre o corpus.

Número de termos a capturar: realizamos a simulação sobre 640 possíveis cenários, que se correspondem com o suposto de o corpus conter 6 termos geográficos (o mínimo usado no topo da classificação das entidades georreferenciadas por conhecimento prévio) até 645 (o número de tipos usados na aplicação da taxonomia de GeoNames para classificar as entidades geográficas na altura de elaboração dos testes). A efeitos comparativos realizamos um segundo teste que considera a presença de 10 até 2000 termos do domínio no corpus.

	precisão	medida_f	ranking_zipf	recuperados	termos_dom	frequência_mínima
1	0.25	0.4	24	24	6	706
2	0.25	0.4	28	28	7	615
3	0.25	0.4	32	32	8	532
4	0.25	0.4	36	36	9	465
5	0.25	0.4	40	40	10	414
636	0.2340270	0.3792899	234	2739	641	4
637	0.2343921	0.3797693	234	2739	642	4
638	0.2347572	0.3802484	234	2739	643	4
639	0.2351223	0.3807272	234	2739	644	4
640	0.2354874	0.3812057	234	2739	645	4

Tabela 8.5: Resultados para os 5 primeiros e 5 últimos supostos da simulação aplicada sobre o corpus recuperado em CTT assumindo uma precisão de 25% e uma abrangência de 100%.

A tabela 8.5 mostra os cinco primeiros e cinco últimos resultados do teste considerando de 6 a 645 termos do domínio geográfico presentes no corpus (os verdadeiros positivos). Numa distribuição de Zipf, os termos com uma frequência maior aparecem no topo, *ranking_zipf* é a variável que ordena os termos. A variável *recuperados* representa o número de candidatos obtidos com o ruído aceite,

termos_dom são os termos do domínio presentes no corpus, e *frequência_mínima* o valor da frequência do último candidato a termo recuperado.

A modo de exemplo, considerando que a métrica recuperasse todos os termos exigidos ao assumir uma precisão de 25%, temos:

Com 10 termos de domínio no corpus (linha 5), o número de verdadeiros positivos é 10 (todos são recuperados), e o número de falsos positivos 30 (o total recuperado, 40 - 10, os verdadeiros positivos).

$$\text{Precisão} = \frac{10}{30+10} = 0.25 \quad (25\%)$$

Como se pode observar na tabela 8.5 acima, os postos mais altos não apresentam variação entre o número de termos recuperados e o ranking, pois, como aguardado na distribuição, apenas há um tipo em cada frequência, não obstante, nos postos inferiores, conforme observado numa distribuição de Zipf, o número de tipos numa mesma frequência tende a aumentar exponencialmente, portanto, também o número de falsos positivos aceites. A figura 8.3 mostra como a medida-F não decresce linearmente conforme aumenta o número de termos, mas mantém uma distribuição irregular, em que os máximos se correspondem com um salto no ranking das frequências.

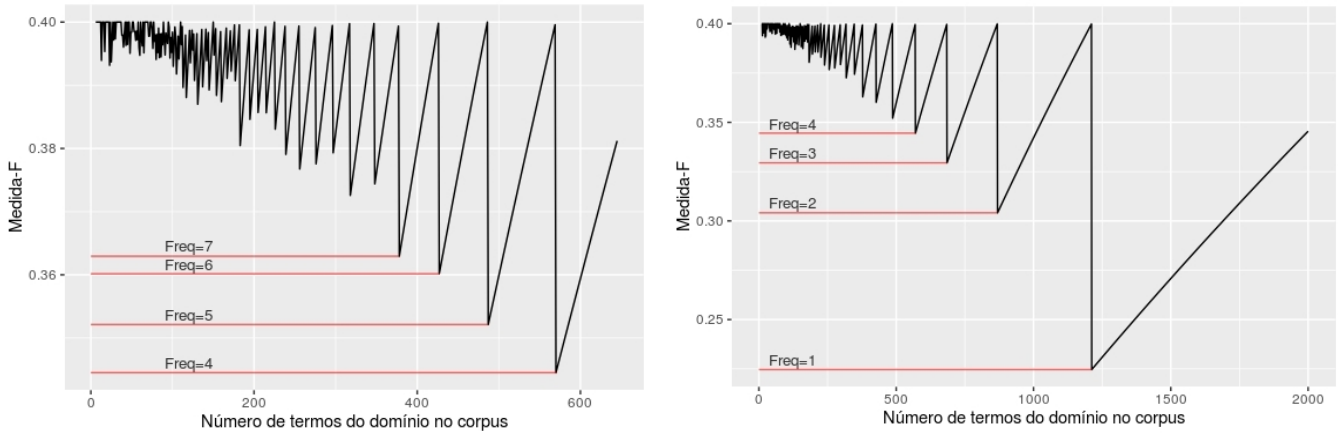


Figura 8.3: Resultados da medida-F supostos sobre a distribuição de frequências extraída do corpus CTT considerando os falsos positivos criados por um nível de precisão de 25% e 100% de abrangência. Efeitos ao considerar de 6 a 640 termos verdadeiros positivos (esquerda) e de 10 a 2000 (direita).

A frequência mínima necessária para recuperar os termos do domínio, considerando que a métrica captura todos os termos dentro do nível de ruído aceite (precisão) fica por cima da linha de frequência 4 na simulação que supõe de 6 a 640 verdadeiros positivos (fig. 8.3 esquerda). Na simulação representada na figura 8.3 direita, menos realista no número de termos de domínio no corpus, haveria que chegar até aos 1212 termos de domínio verdadeiros positivos para deixar o limite de captura na frequência mínima (1). Ambas as simulações mostram claramente que mesmo para trabalhos de recuperação de termos com grande número de candidatos (ex. 500) a consideração

de um limite na frequência fica justificada (assumindo, como fora estabelecido em princípio, que o ranking das frequências captura os termos do domínio).

A relação entre a oscilação do desempenho e a distribuição de frequências aparece indicada na figura 8.3 pelas linhas vermelhas, que assinalam também o ponto em que o primeiro termo de uma nova frequência é recuperado. Uma vez superada a barreira de termos presentes numa frequência, o salto à frequência inferior supõe recuperar todos os termos com esse mesmo valor no ranking (o troco de um máximo a um mínimo dá-se numa distância de apenas um termo). A magnitude da oscilação é também maior conforme se descende no valor da frequência. O número de termos a recuperar e a distribuição do corpus condicionam, portanto, os resultados da medida-F pela própria natureza da distribuição, independentemente do desempenho da métrica (que em termo médio tem uma precisão de 25%), pela presença de maior número de tipos quanto menor seja a frequência (§4.7.1.2). A figura 8.4 esquerda mostra esta situação. Conforme se descende na frequência mínima, maior número de termos são recuperados. O número de falsos positivos será também maior no ponto de transição, contribuindo com um salto da entropia (figura 8.4 direita), maior quanto menor for a frequência.

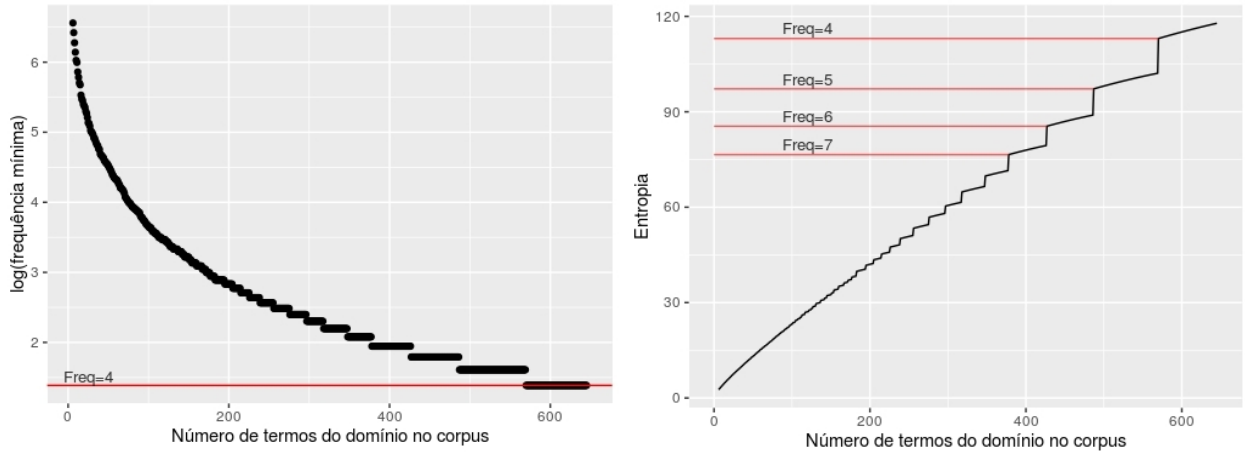


Figura 8.4: Frequência mínima atingida para recuperar os termos de domínio de um corpus (esquerda) e evolução da entropia (direita) assumindo uma abrangência de 100% e precisão média de 25%.

O teste mostra como, por um lado, o método usado para a recuperação de candidatos, a ordenação dos termos segundo a sua frequência, condiciona os níveis de desempenho na precisão recuperando mais candidatos dos necessários e, de outra parte, a redução do número de candidatos contribui para a melhoria dos resultados. Dirigimos o procedimento de extração, então, nesta direção: implementação de métodos para reduzir o número de candidatos.

8.4.3.2 Métodos para a redução do número de candidatos

Na secção anterior consideramos a frequência mínima como único critério de seleção dos candidatos. Concluímos que a própria distribuição zipfiana do corpus introduz um nível de erro que se incrementa conforme diminui a frequência mínima a partir da qual se considerem termos

candidatos. O nosso objetivo é superar o desempenho na precisão reduzindo o número de candidatos, eliminando o maior número de falsos positivos condicionados pelo erro inerente à ordenação baseada na frequência. Ensaiamos uma série de combinatórias de modos do corpus e métodos (especificações descritas no apêndice III) baseadas no filtro de candidatos e na otimização da métrica para compararmos os seus efeitos na extração de termos do domínio geográfico. Nesta secção descrevemos os filtros e métricas aplicadas.

Longitude dos tokens: em domínios técnicos, a longitude dos tokens é fator discriminante, porquanto os termos do domínio são maioritariamente pouco comuns (termos menos comuns tendem a ser mais longos). O caso que nos ocupa atende a um domínio com termos que formam também parte do léxico patrimonial e comum, pelo qual a longitude dos tokens não é um fator tão discriminante como nos âmbitos especializados em que o recurso a neologismos e derivação morfológica são norma para a criação de novos termos. Formas como *mar*, *rio*, *rua*, com apenas três caracteres, entram dentro do domínio e são, de facto, tipos básicos na taxonomia das entidades geográficas mencionadas. O critério de longitude dos tokens mostra-se assim limitado a formas por baixo dos três caracteres para o domínio alvo. Aplicamos o filtro longitude dos tokens com valor 3 sobre todos os modos do corpus.

Lista filtro. Dado que os tipos mais comuns num idioma são aqueles que aparecem independentemente da temática, a criação de listas filtro é simples, obtém-se pela análise das posições mais altas numa tabela de frequências. Listas de termos comuns foram elaboradas para o português (Biderman, 1998; Peres Rodrigues, 2000) e as aplicações de PLN acostumam incorporar as suas próprias. No nosso caso elaboramos duas listas a partir do corpus (Apêndice III). Uma filtra determinantes, pronomes, advérbios, preposições e verbos auxiliares. Uma segunda melhora o desempenho do *script* de semi-normalização (§8.4.2.2), incluindo formas incorretamente padronizadas que, não sendo nomes comuns no corpus, foram anotadas como tais (ex. *chamamos*, *dizer*). As duas são fusionadas numa só, especificada no apêndice III. Filtramos também as expressões LOCAL000 e GENTILICO000 que representam as entidades geográficas mencionadas.

Anotação morfossintática. Um anotador permite selecionar candidatos usando critérios mais complexos, assim um tipo frasal ou clausal conforme a uma anotação do tipo:

Nome + preposição + entidade

Ex. *cidade* + *de* + *Pequim*

V + Preposição + Determinante + Nome + Preposição + Entidade

Ex. *Partimos para a ilha de Ainão*

O uso de regras e métodos estatísticos para aprender e detetar de modo automático patrões morfossintáticos é comum nos sistemas NERC usados no capítulo 5. Nos testes de extração de termos usamos o anotador morfossintático para filtrarmos os tipos de maneira que considerarmos unicamente os nomes comuns como candidatos. O modo do corpus DATA3 avalia o efeito da anotação; DATA4, a anotação mais o efeito da restrição da concordância da entidade geográfica ao

seu subsegmento clausal ou frasal dentro da oração (§8.4.2.6). Adicionamos também à lista filtro de palavras comuns os casos pontuais de anotações anómalas (algumas formas verbais e maiormente determinantes numerais anotados como nomes comuns) (Apêndice III).

Proximidade. Filtramos os candidatos considerando um critério sintático e de janela, apenas o tipo mais próximo dentro do segmento sintático mínimo anotado é considerado. O modo do corpus DATA5 (§8.4.2.6) é o que mais diretamente representa uma restrição por proximidade.

Métricas. Em §8.4.3.1 usamos a frequência absoluta do termo (TF) para observarmos os efeitos do limite de frequência e a redução de tipos. Outro modo de extrair os termos é a aplicação de medidas que outorgam um maior peso às formas com uma maior relevância semântica no corpus. As métricas mais comuns (estatísticas, linguísticas e híbridas) aplicadas na extração de termos são estudadas por Conrado, Felippo, Pardo e Rezende (2014) e avaliadas por Conrado, Pardo e Rezende (2015). A métrica que obtém uns melhores para a medida-F, com 36%, é a TF-IDF. Será também a usada para os nossos testes pelos resultados referidos e por ter sido amplamente estudada e de comum aplicação na minaria de textos.

TF-IDF (Salton & Buckley, 1988) combina a frequência do termo com a sua distribuição em unidades menores do corpus (documentos). Numa distribuição de Zipf, os termos mais comuns aparecem com frequências altas e os menos comuns ocupam posições de hápax. Num texto com uma temática particular, os termos que representam o tópico apareceriam com uma frequência por cima da linha de hápax, mas sem ocuparem as posições do topo. No entanto, num documento fora do tópico, estes mesmos termos têm uma probabilidade mais baixa, própria de posições de hápax. Exemplificado no corpus objeto de estudo: nos capítulos com cenário uma cidade, o suposto inicial é termos geográficos tais como *rua*, *edifício*, *cidade*, aparecerem com uma maior frequência que naqueles outros capítulos com cenário no mar, onde aguardamos sejam mais frequentes termos como *mar*, *ilha*, *porto*, *costa*.

A formulação de TD-IDF contempla esquemas em que as variáveis consideram as frequências do termo e dos documentos (Singhal, Salton & Buckley, 1996; Mannig & Schütze, 1999). O esquema inicial de referência para os testes é:

$$tf-idf(i,j)=\begin{cases} 1+\log(tf_{i,j})\log\dfrac{N}{df_i} & \text{quando } tf_{i,j}\geq 1 \\ 0 & \text{quando } tf_{i,j}=0 \end{cases} \qquad (8.3)$$

em que $tf_{i,j}$ é a frequência do termo t_i no documento d_j, N o número de segmentos do corpus e df_i o número de segmentos do corpus em que t_i ocorre. O apêndice (III) mostra as variações para aplicar combinatórias exploratórias da frequência (absoluta, relativa e logarítmica) e normalização (frequência relativa ao documento para a frequência do termo e cosseno para o peso final) sobre os modos do corpus obtidos em DATA2, DATA3 e DATA4. No modo DATA5, o de maior elaboração PLN, apenas se aplica a métrica da frequência absoluta (TF) por se não considerar o fator número de documentos.

Ranking. Conrado (2014) obtém os melhores resultados de precisão limitando os candidatos aos 50 primeiros postos no ranking. Conrado, Pardo e Rezende (2015) apresentam os melhores resultados de medida-F ao considerar entre 200 e 400 candidatos. Dado que o que queremos é criar uma taxonomia e caracterizar relações dentro de uma ontologia para o nosso corpus, interessa sobretudo a efetividade do sistema classificatório, capaz de reduzir sinónimos e classes pouco discriminatórias dentro de um mesmo tipo, mais do que a recuperação de todos os termos do domínio. Consideramos, portanto, apenas os 50 primeiros candidatos.

8.4.4 Análise dos resultados

O apêndice III descreve os testes e mostra os resultados quantificados em número de verdadeiros positivos. A figura 8.5 mostra os resultados de precisão para todos os testes por modos do corpus, métrica e tipo de filtro. Analisamos a seguir cada um dos fatores considerados.

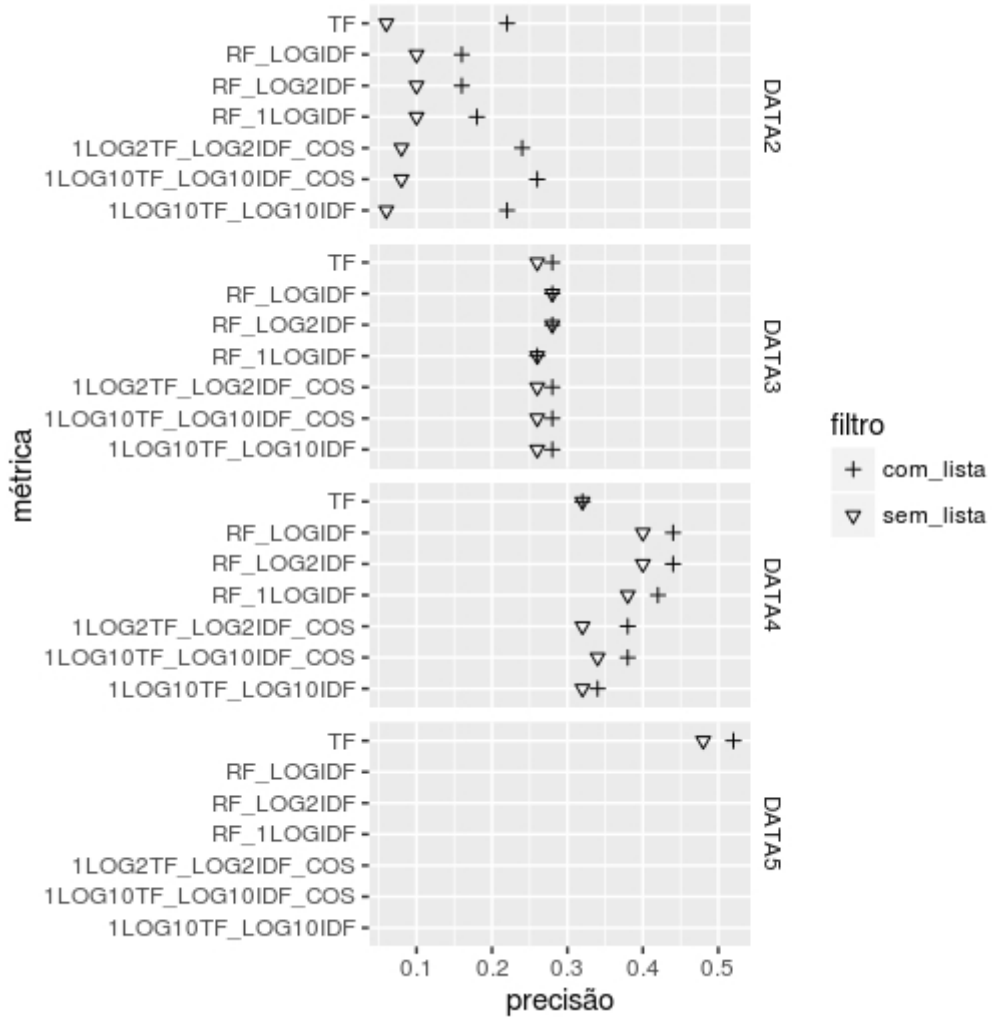


Figura 8.5: Precisão obtida na recuperação de termos do domínio geográfico segundo modo de corpus, métrica e uso ou não de lista filtro.

Como padrão de avaliação, elaboramos uma lista de verdadeiros positivos totais, obtidos do conjunto dos testes, que validamos manualmente. Realizamos a implementação das métricas com o software estatístico de R (R Core Team, 2016) e os nossos próprios *scripts*. Para a obtenção de matrizes de coocorrências, aplicação das listas filtros e comparação de resultados das métricas TF-IDF, usamos o pacote TM (Feinerer, 2008; Feinerer, Hornik, & Meyer, 2008). Nesta secção descrevemos o desempenho da precisão nas combinatórias das métricas e processamento do corpus.

8.4.4.1 Fatores no desempenho da precisão

Lista filtro. A aplicação de lista filtro tem uma maior incidência quando não há anotação morfossintática, chegando a incrementar o desempenho por cima de 100%. Nos modos do corpus anotados pelo parser, a aplicação da lista melhora os resultados, porém, de modo menos significativo. Em algum caso (DATA5), o filtro apenas é relevante pela inclusão de um termo incorretamente processado que não deveria estar incluído no modo do corpus, isto é, o desempenho melhoraria com a simples correção dos níveis de PLN no pré-processamento, independentemente da lista filtro.

Métrica. A métrica TF-IDF contribui para uma melhoria do desempenho relativamente à frequência absoluta (TF) quando se não usa outro método de redução de candidatos. Em caso de aplicarmos uma lista filtro, o modo de normalização mostra-se relevante. O melhor resultado, quando o nível de processamento do corpus (DATA2) é menor, foi para o esquema do logaritmo da frequência e normalização pelo cosseno. No modo de maior restrição por PLN em que TF-IDF é aplicável (DATA4), obteve-se o melhor desempenho com o esquema da frequência relativa.

Processamento. Os melhores resultados observam-se quanto maior é o nível PLN. O desempenho das métricas TF-IDF melhora conforme aumenta o processamento. Aliás, o maior nível de processamento PLN (DATA5), em que só se aplica a frequência absoluta, consegue os resultados mais altos.

8.4.4.2 Melhor resultado absoluto

Observamos os melhores resultados pelo incremento do PLN. Um processamento simples, de anotação morfossintática e subsegmentação da oração para obtermos o termo mais próximo à entidade geográfica mencionada, consegue uma precisão de 48% sem aplicarmos mais métrica do que a frequência absoluta (TF sem lista sobre DATA5 na fig. 8.5; R7 no apêndice III) e 52% com uma lista filtro (TF com lista sobre DATA5 na fig. 8.5; R8 no apêndice III).

Como conclusão, o fator mais relevante para a extração de termos nos nossos ensaios é o nível de processamento do corpus (modos do corpus, de DATA2 a DATA5), isto é, obtemos os melhores resultados pelo incremento do PLN. Não obstante, mantendo o nível de processamento estável, a aplicação de métricas e filtros é relevante para a melhoria da precisão.

8.4.5 Validação semântica

Como resultado dos testes anteriores obtivemos uma lista de termos de domínio com uma precisão

similar aos melhores resultados de referência para a língua portuguesa que encontramos para testes similares (Conrado, Pardo, & Rezende, 2015), mas a lista obtida fica por baixo do nível requerido para a considerar representativa de domínio, com maioria de termos falsos positivos.

Aplicamos finalmente um método de melhoria dos resultados por meio da validação semântica. Consiste no emprego de glossários específicos do domínio geográfico e recursos obtidos de uma base de conhecimento lexical (§3.5.2) com que filtramos os candidatos a termos de domínio. O apêndice IV mostra as configurações dos glossários, um teste de exemplo e os resultados totais dos ensaios. A seguir descrevemos o procedimento metodológico, analisamos os resultados e mostramos um exemplo da sua aplicação dentro da taxonomia.

8.4.5.1 Validação com glossários de termos geográficos

Aplicamos dois glossários, o IBGE (IBGE, 2015) com 126 termos geográficos usados no mapeamento contemporâneo do Brasil e a lista de códigos de GeoNames[12], originalmente em inglês e traduzida para o português mediante o sistema de tradução automática de Google[13], na altura destes ensaios, com 667 termos recuperados. A primeira lista inclui o fator da concisão, os termos orientados pela precisão, limitada na abrangência pela não consideração de sinónimos e termos geográficos não representados no mapeamento contemporâneo. A segunda maximiza a abrangência, mas inclui ruído (falsos positivos) pelas limitações da tradução automática. Da união obtemos um terceiro glossário, IBGE ∪ GeoNames_trad, que contém 725 termos, os comuns e não comuns de IBGE e a tradução de GeoNames (vid. apêndice IV para mais pormenores).

Como lista a validar, usamos o melhor resultado de extração de termos (§8.4.4.2). A avaliação do desempenho dos glossários faz-se em relação à abrangência, isto é, computamos quantos termos geográficos são verdadeiros positivos como efeito da aplicação de cada glossário. A tabela 8.6 mostra o resultado da validação com os glossários sobre a lista TR8 (melhor resultado nos testes de extração de termos do domínio) considerando o total de verdadeiros positivos obtido do conjunto dos testes de extração de termos como representativo de 100% da abrangência.

Glossário	Precisão	Abrangência	Medida-F
IBGE	**100%**	54%	70%
GeoNames_trad	**100%**	42%	59%
IBGE ∪ GeoNames_trad	**100%**	65%	**79%**

Tabela 8.6: Comparativa do efeito da validação da lista de candidatos a termos geográficos TR8 por meio de listas específicas do domínio geográfico.

Com a aplicação dos glossários geográficos obtemos uma precisão de 100% (0 falsos positivos), isto é, da lista de 50 candidatos obtidos no teste R8 (Apêndice III), apenas ficamos com aqueles que a lista de termos geográficos valida como pertencentes ao domínio. Porém, há uma diminuição na

[12] http://www.GeoNames.org/export/codes.html
[13] https://translate.google.com/?hl=pt-PT

abrangência que, da sua parte, melhora com o glossário mais completo (IBGE ∪ GeoNames_trad). Como no caso da anotação das entidades geográficas (cap. 5), a especificidade da lista aparece como fator relevante no desempenho. O IBGE usa apenas termos usados para a cartografia, o que a faz uma lista reduzida: a modo de exemplo, termos como *país*, *povoação*, *esteiro* não estão incluídos no glossário, como também não outros que são relevantes no corpus (*reino*, *império*) mas não nos usos cartográficos comuns. No caso de GeoNames temos de considerar que a tradução automática insere formas espúrias condicionadas pelo desempenho do tradutor.

8.4.5.2 Validação com listas geradas por uma base lexical difusa

Como terceiro recurso consideramos a aplicação de uma base lexical, CLIP 2.1 (Gonçalo Oliveira & Gomes, 2014; 2016), organizada em synsets como numa WordNet (Fellbaum, 1998; Gonçalo Oliveira, Paiva, Freitas, Rademaker, Real, & Simões, 2015) mas os termos agrupados de modo quantitativo difuso, numa associação determinada pela coocorrência (a partir de ligações em bases lexicais de relações semânticas), medida como frequência absoluta normalizada pelo número de termos do cluster e número de documentos considerados (Santos & Gonçalo Oliveira, 2015). O CLIP 2.1 permite-nos pesquisar os termos pelo grau de associação e estabelecermos uma medida de corte mínima para considerar um candidato como sendo relacionado com um termo dado pertencente ao domínio. Para selecionarmos os synsets do CLIP 2.1 usamos os glossários de termos do domínio geográfico (§8.4.5.1) e a lista de verdadeiros positivos do melhor resultado dos testes de validação semântica por glossário geográfico (Apêndice IV), validada pela lista de termos geográficos com melhor desempenho (IBGE ∪ GeoNames_trad). Esta nova lista tem a particularidade de limitar os termos do domínio geográficos aos positivos a respeito do corpus.

Os ensaios realizados com o CLIP 2.1 aparecem no Apêndice V. Aplicamos um total de 400 esquemas para avaliar uma lista de candidatos, a obtida no melhor resultado dos testes de extração de termos (§8.4.4.2). Cada esquema inicializa um glossário que seleciona synsets a partir de um valor mínimo na base lexical. Os termos dos synsets selecionados são de novo filtrados segundo um segundo valor de corte na medida da associação. Obtemos assim uma nova lista de termos associados ao domínio geográfico expandida pela base lexical com que validamos a lista de candidatos (repetindo o procedimento usado no apêndice IV).

A figura 8.6 (abaixo) mostra os resultados em medida-F e verdadeiros positivos do processo automático de extração de termos ao validar a lista que chamamos de TR8, obtida anteriormente como melhor resultado nos testes de extração de termos (Apêndice IV). Todos os glossários conseguem recuperar synsets capazes de conseguir, nas suas melhores configurações, 100% de precisão ou 100% de abrangência. Obtemos o melhor resultado na medida-F, de 90%, ao aplicar o glossário com o menor número de termos (17). No entanto, o glossário com maior número de termos (725) também consegue desempenhos similares, com uma medida-F de até 88%.

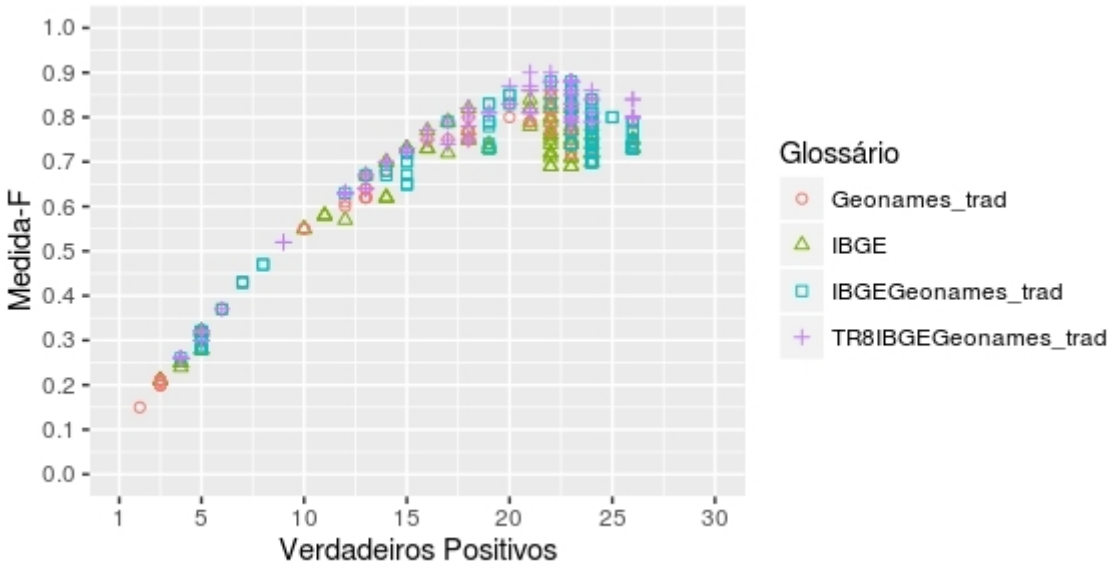


Figura 8.6: Verdadeiros positivos e medida-F na validação do melhor resultado de extração de termos (TR8) pelas listas de domínio obtidas do CLIP 2.1.

A tabela 8.7 mostra os melhores resultados para as três medidas consideradas. A abrangência aumenta pela redução do valor de corte. Achamos o melhor resultado da medida-F (84%), com um 100% de abrangência, nos valores de corte 0.05 na seleção e 0.01 para a limitação de pesquisa dentro do *synset*. Uma precisão do 100% mantém uma abrangência de 81% com o valor de seleção de *synset* em 0.15 e recuperação de termos dentro do *synset* em 0.25. A aplicação da base lexical consegue, portanto, melhorar a precisão, é possível configurá-la para abranger todos os positivos da lista de partida e otimiza a medida-F.

Precisão	**Abrangência**	**Medida-F**	**Seleção de *synset***	**Seleção de termos no *synset***
100%	81%	90%	0.15	0.25
72%	100%	84%	0.05	0.01

Tabela 8.7: Melhores resultados de precisão e abrangência a respeito da medida-F e valores de corte na medida de associação semântica estabelecida pelo CLIP 2.1.

8.5 Integração dos resultados na taxonomia prévia de tipos geográficos

No capítulo 7 definimos uma lista de atributos geográficos para classificarmos as entidades geográficas por conhecimento prévio. A tabela 8.8 considera o melhor resultado na precisão dos testes de extração automática (TR8_TR8IBGEGeoNames_0.15_0.25, vid. Apêndice V), com 21 termos recuperados, para formalizar subclasses a respeito da taxonomia usada para as entidades de conhecimento prévio. Conforme a §8.3.4.2, os atributos geográficos mais genéricos definem uma classe, dentro da qual um tipo geográfico representa uma subclasse.

Tipo geográfico extraído	⊂ classe	Código classe
cidade, lugar, povoação	Cidades_e_povoações	P
	Países_e_divisões_administrativas	A
campo, ilha, monte, serra, terra	Ilhas_e_geografia_física_terrestre	T
barra, costa, enseada, estreito, lago, mar, ponta, porto, rio	Rios_praias_portos	H
	Grandes_áreas_e_regiões	L
casa, castelo, fazenda, fortaleza	Construções	S

Tabela 8.8: Classificação dos resultados do teste TR8_TR8IBGEGeoNames_0.15_0.25 dentro da taxonomia usada para as entidades geográficas de conhecimento prévio.

Conseguimos extrair termos do domínio geográfico por procedimentos totalmente automáticos com uma precisão de 100%, mas recuperamos termos para apenas 4/6 das classes no topo da hierarquia dos tipos geográficos. Relativamente à taxonomia elaborada a partir das concordâncias do corpus, só um termo dos validados pela base lexical, *fazenda*, não foi considerado como relevante para a descrição das entidades, todos os demais são relevantes e foram recolhidos quer antes da redução de termos (inicialmente considerados, posteriormente representados por um cohipónimo, por exemplo, *fortaleza* e *castelo* foram reduzidos a uma mesma classe, *castelo*), quer como representantes finais da classe (ex. *cidade*, *lugar*, *povoação*).

8.6 Conclusões

Para georreferenciarmos uma entidade mencionada é preciso um atributo geográfico que relacione a entidade com outra numa relação de pertença. Neste capítulo atendemos à primeira parte do problema: a extração de termos para uma lista de atributos geográficos que classifique as entidades do corpus.

Introduzimos a noção de proposição como sendo uma forma elaborada da oração que exprime um facto. Os feitos que nos interessam são aqueles relativos às entidades geográficas. Múltiplas orações expressam uma mesma proposição. A identificação de orações expressão de proposições com valor geográfico vem dada pela anotação da entidade geográfica mencionada. Desta maneira, a anotação permite selecionar as orações relevantes dentro do corpus. As proposições que procuramos expressam factos e mostram relações com valor georreferenciador a respeito da entidade geográfica mencionada. No nosso modelo, duas relações são relevantes para descrever a entidade: uma monádica (de propriedade) que confere um atributo geográfico, e outra diádica de pertença a respeito de uma outra entidade. Um conceito formaliza esta descrição, expressada através de uma definição com que georreferenciamos de modo relativo uma entidade geográfica mencionada. Neste capítulo centramo-nos num aspecto da primeira relação: o problema da criação de uma lista de tipos geográficos para descrevermos as entidades do corpus.

Atendemos assim ao problema de criação de uma taxonomia para classificarmos as entidades geográficas e exploramos as possibilidades de extração de termos do domínio para a criação de uma lista de atributos classificatórios. Em primeiro lugar, analisamos os limites e possibilidades de usarmos uma taxonomia externa ao corpus, predefinida. Posteriormente vimos as dificuldades de elaborarmos uma outra manualmente e, finalmente, expusemos o procedimento para chegar a uma classificação que combina ambas possibilidades.

Como caso prático de trabalho sobre o corpus, exploramos a extração de termos do domínio geográfico para criar o vocabulário de uma taxonomia de modo automático. Selecionamos apenas orações que tenham valor proposicional georreferenciador para criar um subcorpus sobre o qual aplicamos testes para avaliar o rendimento do PLN e as métricas mais comuns na extração de termos. Exploramos as propriedades do léxico num corpus, através de uma distribuição de Zipf, como condicionantes na oscilação do resultado do desempenho independentemente do próprio rendimento do método quando se selecionam os candidatos por um critério de frequência absoluta. Posteriormente, aplicando uma métrica bem conhecida, TF-IDF, conferimos esquemas e modos de elaboração do corpus, obtendo os melhores resultados quanto maior for a elaboração por PLN. Aproveitamos o melhor resultado destes testes para, avaliada manualmente uma lista de verdadeiros positivos do corpus, comparar com a validação automática por meio de glossários geográficos e uma base lexical difusa. Conseguimos um resultado de precisão de 100% para uma lista de 21 termos de domínio (no capítulo dedicado ao conhecimento prévio apenas usamos 6 tipos) e obtivemos esquemas para otimizar o desempenho de todo o processo automático na precisão, abrangência ou medida-F. Finalmente, integramos o melhor resultado de medida-F e 100% precisão dentro da taxonomia usada para classificar as entidades geográficas de conhecimento prévio.

Como conclusão final mais relevante do capítulo, a validação com a base lexical difusa permitiu incrementar os resultados de extração de termos, duplicando a abrangência, mesmo nos casos de melhores resultados de partida (similares aos melhores que achamos para labores similares na literatura especializada), sendo possível também obter a maior precisão sem por isso diminuir consideravelmente a abrangência, chegando a obter, nos melhores resultados, medidas-F > 90%.

8.7 Sumário de objetivos

Os objetivos desta secção foram:

- Selecionar as orações que expressem proposições com valor georreferenciador para a criação de um subcorpus geoespacial.
- Definir e simplificar a descrição da entidade mencionada, a começar pela atribuição de um tipo geográfico como primeiro elemento da sua definição.
- Elaborar uma taxonomia para a descrição de todas as entidades geográficas mencionadas no corpus do caso prático.
- Automatizar o processo de extração de termos geográficos num corpus.

Capítulo 9

A definição do georreferente

Nos capítulos anteriores desenvolvemos um modelo de georreferência com duas possíveis soluções para o referente: por um lado, o resolvido por coordenadas, por outro, a georreferência relativa em que descrevemos a entidade a partir de uma definição. Neste capítulo aplicamos finalmente o modelo para criarmos um índice com georreferentes para todas as entidades do corpus. Em primeiro lugar, revisamos os componentes e a forma das definições. Posteriormente revemos as relações necessárias para ordenarmos as entidades geográficas mencionadas numa ontologia. Aproveitamos os resultados da análise de corpus para realizarmos uma série de ensaios sobre a captura automática das relações semânticas usadas na definição da georreferência relativa. Dada a limitação da representatividade estatística do conjunto das entidades (frequências em linha de hapax ou muito próximas), selecionamos os casos de frequências mais amplas numa aproximação exploratória.

Como atividade final e produto mais elaborado da tese, usamos a estrutura taxonómica da ontologia para criarmos um índice com todas as entidades geográficas mencionadas, caracterizadas por um tipo e um georreferente, a que lhe adicionamos dados selecionados do estudo crítico prévio e resultados da análise de corpus.

9.1 A definição do conceito

A descrição das entidades geográficas mencionadas no corpus contribui para a georreferenciação quando não há umas coordenadas por conhecimento prévio. Segundo o modelo de conceito (cap. 6), é preciso um tipo geográfico e outra entidade geográfica para referenciar uma entidade de modo relativo. Estabelecido um tipo geográfico e uma relação de meronímia, obtemos uma definição para a georreferencia relativa de uma entidade previamente desconhecida.

9.1.1 A definição da entidade como georreferência relativa

Seja a concordância:

> “Este tyrão Rey **Achem** foy aconselhado pelos seus, que se queria tomar **Malaca**, por nenhũa maneyra o poderia fazer cometendoa de mar em fora, como ja por seis vezes tinha tentado no tẽpo de dom Esteuão da Gama, & de outros Capitaẽs atras passados, senão com se fazer primeyro senhor deste reyno de **Aarù**, & se fortificar no rio de ***Panetição***, (...)” (*PR*, 26) (9a)

Temos dois modos de georreferenciar as entidades mencionadas. As resolvidas por conhecimento prévio têm coordenadas:

Achem: Lat. 4, long. 97

Malaca: Lat. 2.196, long. 102.2405

Aarù: Lat. 4, long. 98.2102

No caso de *Paneticão*, não há coordenadas resolvidas. Porém, a partir das suas concordâncias no corpus, a base de dados relacional regista um tipo geográfico e uma relação de meronímia como parte de *Aarù*. A definição da sua georreferência fica resolvida assim:

d = {Define(w) | <c,a>} (6.5)

$d_{Paneticão}$ = {d(Paneticão) | <rio, Aarù>}

Paneticão = rio de Aarù.

Achem, *Malaca* e *Aarû* têm uma georreferência exata; *Paneticão*, uma referência relativa que mostra o seu tipo geográfico (*rio*) e aponta para a sua localização dentro de *Aarù*. Esta georreferência é denominada relativa porquanto depende do conhecimento de outra entidade (*Aarù*) e da descrição do tipo (um rio) para localizarmos o objeto geográfico referido.

9.1.2 Relações semânticas na definição

Para a elaboração da definição estabelecemos duas relações: entre a entidade e o tipo geográfico, e entre a entidade e uma outra entidade. Ambas podem ser desenvolvidas como relações semânticas.

9.1.2.1 Hiponímia

A relação de hiponímia (§3.4.3.2) ordena as entidades mencionadas como instâncias de um tipo geográfico (ex. ilha, cidade, rio, …), por sua vez membros de uma classe maior (terrestres, hidrológicos, construções, ...). Particularmente importante para o objetivo final de estruturar as entidades numa ontologia é considerarmos o tipo geográfico como classe e a entidade geográfica mencionada como sendo instância de uma classe. Assim, o objeto *Çamatra* é uma instância do tipo geográfico *ilha*. Os tipos formam cada um sua classe e agrupam-se em classes maiores numa relação semântica que tem a propriedade transitiva e foi apresentada como hiponímia (do membro à classe) e hiperonímia (da classe ao membro) (§3.4.3.2)

Ex. *Çamatra* é uma instância da classe *ilha* que é um hipónimo (membro da clase) de acidentes_geográficos_terrestres (o seu hiperónimo).

9.1.2.2 Meronímia

Em §6.2.1.3 introduzimos a relação de pertença *é_Parte_de* (6.4) para a elaboração do conceito.

Nesta relação assumimos que certo tipo de entidades geográficas são parte de outras (administrativas com a sua área de influência, físicas no espaço que cobrem). Uma tal relação equivale a uma relação semântica de meronímia em que o continente é o holónimo e o conteúdo um merónimo. Formulamos a sua expressão do modo:

é_Parte_de(x,y)

em que *x* é o merónimo e *y* o holónimo.

Exemplos de oração que expressa uma proposição de situação ou pertença:

> "E chegando nós ao porto de **Chatigaõ** no reyno de **Bengala**, onde naquelle tẽpo auia muytos Portugueses, me embarquey eu logo nũa fusta de hum Fernão Caldeyra que hia para Goa, onde prouue a nosso Senhor que cheguey a saluamento." (*PR*, 171) (9b)

De onde tiramos a relação:

é_Parte_de(Chatigaõ, Bengala)

que tem valor semântico: *Chatigaõ* merónimo, *Bengala* holónimo.

A relação de meronímia também é transitiva. Se Chatigaõ é parte de *Bengala* e *Bengala* parte da *Índia*, *Chatigaõ* é parte da *Índia*.

9.2 Caracterização de uma ontologia para entidades geográficas do corpus

No trabalho sobre o corpus partimos da consideração da entidade geográfica como um objeto físico. Descrevemos uma ontologia (cap. 3) como uma representação das entidades geográficas numa taxonomia dentro da qual se estabelecem relações. Nesta secção procuramos povoar uma ontologia com as entidades geográficas do corpus.

No âmbito das entidades geográficas mencionadas, Zhu, Hu, Janowicz e McKenzie (2016) comparam ontologias específicas do domínio. A criação de ontologias foi particularmente atendida por Chaves (2008b, 2009) a partir de uma base de conhecimento geográfico (Chaves, Silva & Martins, 2005) para criar um vocabulário que descreve os tipos, propriedades e relações das entidades geográficas (Martins, Silva & Chaves, 2007; Chaves, Rodrigues & Silva, 2007; Lopez-Pellicer, Silva & Chaves, 2010). GeoNames, recurso geográfico consultado para a lista de topónimos de conhecimento prévio, oferece uma ontologia de domínio geográfico cujos níveis mais altos da taxonomia aplicamos no capítulo 7. Nas alíneas a seguir mostramos exemplos ilustrativos das relações usadas na ontologia, as requeridas para a elaboração da definição. Formalizamos a expressão das relações seguindo um modelo genérico (Russell & Norvig, 2010).

9.2.1 Instâncias e classes

Os tipos geográficos definem uma classe, as entidades geográficas mencionadas são instâncias que pertencem à classe.

Ainão $\in$ Ilha

Pequim $\in$ Cidade

Singrachirau ∈ Muro

9.2.2 Subclasses

Um tipo geográfico define uma subclasse dentro de outra classe.

Ilha ⊂ Ilhas_montanhas_e_acidentes_físicos_de_terra

Cidade ⊂ Cidades_e_povoações

Muro ⊂ Construções

9.2.3 Relações

A relação principal das entidades geográficas na ontologia é a de meronímia, em que um objeto é parte de outro:

Ex. é_Parte_de(Aapessumhee, Aarù).

A meronímia tem a propriedade transitiva:

Ex. é_Parte_de(Aapessumhee, Aarù) Λ é_Parte_de(Aarù, Çamatra) → é_Parte_de(Aapessumhee, Çamatra).

A relação de holonímia é definida como inversa à meronímia e formalizamos:

Ex. é_Parte_de(Aapessumhee, Aarù) → Contém(Aarù, Aapessumhee).

9.2.4 Inferência

Para derivarmos relações que não foram declaradas de modo explícito usamos a inferência. Assim, instanciamos uma entidade geográfica como pertencente às categorias no topo da taxonomia.

Por exemplo, se uma entidade geográfica é um muro, um muro é uma construção. *Singrachirau* é um muro, portanto, pertence à classe de *Construções*.

x ∈ Muro → x ∈ Construções

Singrachirau ∈ Muro ⊢ Singrachirau ∈ Construções

A inferência mais usada procede de considerar a relação é_Parte_de(x,y)

Por exemplo, inferência pela propriedade transitiva:

é_Parte_de(Pequim,China) Λ é_Parte_de(China,Ásia) ⊢ é_Parte_de(Pequim,Ásia)

Pela relação inversa:

é_Parte_de(Pequim,China) ⊢ Contém(China,Pequim)

9.2.5 Implementação da ontologia para as entidades do corpus

Com os dados da taxonomia e as relações entre entidades anotadas na base de dados relacional exportamos um documento no formato OWL (*Ontology Web Language*) (Chaves, 2009, p. 62) para processar no Protege (Musen, 2015), software para o desenvolvimento de ontologias e bases de conhecimento. O objetivo principal desta implementação foi revisar as relações de pertença entre as entidades e comprovar que o conjunto funciona como um todo coerente. Com esta finalidade, declararam-se as relações *é_Parte_de* e a inversa *Contém*. Para definir as classes, usou-se a taxonomia de tipos descrita no capítulo 8. As principais dificuldades no instanciamento das entidades geográficas mencionadas do corpus consistiram em:

• Definição de um holónimo para as entidades mais abrangentes, aquelas sem um holónimo anotado como entidade geográfica mencionada no corpus. Assim, *África*, *Ásia* e *Europa* são entidades mencionadas no corpus, adscritas ao tipo geográfico *continente*, mas não têm um objeto para satisfazerem a relação *é_Parte_de*, necessária para prover uma definição no modelo semântico proposto. Criou-se assim um objeto, *Terra*, ao que se ligam os tipos mais abrangentes. Por inferência a partir da propriedade transitiva da relação *é_Parte_de* e a sua inversa, todas as entidades mencionadas são parte da *Terra*.

• Homologação de objetos holónimos para permitir uma representação espacial mais equilibrada das entidades. Um problema surge para as entidades dispersas em pontos afastados do planeta, para as quais não há um holónimo mencionado. É o caso das três entidades mencionadas pertencentes a América *(Brasil*, *noua Espanha* e *Panamá)*. Ao não aparecer o continente como sendo mencionado no corpus, teriam de ficar como parte da *Terra* diretamente, portanto, ao mesmo nível que os continentes. Outro tanto sucede, num nível de granularidade inferior, com as entidades da zona da Insulíndia e Ásia Central: por não terem um holónimo citado, ficariam todas agrupadas ao mesmo nível. Introduziram-se, portanto, os objetos *América*, *Ásia Ocidental*, *Ásia Oriental*, *Índico Ocidental*, *Índico Oriental*, *Indochina* e *Insulíndia* como holónimos não mencionados no corpus, mas convenientes para a taxonomia, porquanto os cohipónimos ficam melhor agrupados numa relação de proximidade.

A ontologia permite a obtenção de novas relações por inferência, de utilidade na análise crítica e para a observação dos dados como um conjunto ordenado de relações. Não obstante, para os objetivos deste trabalho, teve como objetivo principal corrigir e sistematizar a relação de holonímia para a obtenção de um índice do total das entidades geográficas mencionadas (§9.4).

9.3 Captura das relações no corpus

Na ontologia consideramos duas relações, a hiponímia (membro de uma classe) e a meronímia (parte de um todo), que nos permitem ligar o problema do instanciamento ao das relações semânticas e, dum modo mais amplo, à similitude semântica entre termos, suscetível de uma aproximação automática. Machado e Lima (2015) revêem o estado da arte e apresentam uma solução baseada em regras a partir de padrões sintáticos para o português. Outra solução comum é a

aplicação de vetores para medir as coocorrências (Manning & Schütze, 1999; Dinu, Thater, & Laue, 2012; Clark, 2015; Jurafsky & Martin, 2015). A assunção básica é recolhida pelo modelo distribucional: termos relacionados partilham contextos relacionados (Mitchell & Lapata, 2010; Baroni, Bernardi & Zamparelli, 2014). Se comparamos como dois termos mudam relativamente ao contexto, obtemos o seu grau de correlação. Mintz, Bills, Snow e Jurafsky (2009) extraem relações previamente resolvidas numa base de dados para criarem e analisarem corpora em que valores estatísticos são utilizados como atributos preditivos e assim instanciarem entidades (com exemplos das geográficas) num modelo de aprendizado de máquina. Hu e Janowicz (2016) experimentam com relações previamente resolvidas numa ontologia com variáveis de tipo geográfico como alternativa ao PLN. O estado da arte para modelos baseados em vetores é revisado por Gamallo (2016), obtendo os melhores resultados pelo filtrado dos termos, ordenados em matrizes de coocorrências, com uma medida de similitude (o cosseno) para computar a relação.

Nesta secção desenvolvemos as entidades mencionadas relativamente a dois termos do domínio geográfico para examinarmos como condicionam semanticamente uma entidade mencionada. Usamos as coocorrências entre entidades mencionadas e tipos geográficos da taxonomia como dados num modelo de aprendizado de máquina (§9.3.2.1).

9.3.1 Captura de relações com traços semânticos alvo

Segundo o nosso modelo, uma entidade mencionada vem definida por um atributo geográfico e uma relação de pertença a outra entidade, consequentemente, a primeira relação que definimos é a da entidade mencionada com um tipo geográfico. Por outro lado, um termo que entra em relação com a entidade e coocorre com ela, segundo o modelo distribucional, também é parte do seu significado. Continuando com esta argumentação, podemos usar os tipos geográficos (que coocorrem com as entidades, como vimos nos testes de extração de termos em §8.4) como traços semânticos. No exemplo a seguir usamos os termos *ilha* e *cidade* como traços semânticos para avaliar a similitude semântica de uma lista dada de entidades mencionadas, selecionadas com atenção aos seguintes critérios:

• Serem instância das subclasses *cidade* e *ilha* dentro da ontologia.

• A adscrição à subclasse é feita pela descrição do corpus e não unicamente por conhecimento prévio.

• Dentro da subclasse, terem as frequências mais altas na lista de frequências do corpus.

• Serem as menos ambíguas como representantes do tipo geográfico dentro das entidades com frequências mais altas. Assim, para *cidade* escolhemos as três entidades mencionadas com maior frequência e atributo secundário *metrópole*, porquanto supõe uma maior proximidade ao protótipo. Para o caso das ilhas, as três entidades mencionadas com maior frequência cujo único atributo é *ilha*.

Deste modo obtemos três entidades geográficas para cada tipo geográfico, aquelas que menos

ambiguamente representam o protótipo segundo a sua descrição no corpus e, dentro destas, as três com mais ocorrências no corpus (tabela 9.1).

EM	Freq.	Tipo geográfico	Referência atual
Çamatra	(14)	Ilha	Sumatra, Indonesia (AS).
Iaoa	(26)	Ilha	Java, Indonesia (AS).
Martauão	(35)	Cidade (metrópole)	Martaban, Myanmar [Burma] (AS)
Odiaa	(19)	Cidade (metrópole)	Phra Nakhon Si Ayutthaya, Thailand (AS)
Pequim	(47)	Cidade (metrópole)	Beijing, China (AS).
Tanixumaa	(18)	Ilha	Tanega Shima, Japan (AS)

Tabela 9.1: Entidades mencionadas selecionadas como protótipos de ilha e cidade segundo a sua frequência absoluta e tipo geográfico adscrito no índice.

A tabela 9.2 abaixo mostra o número de vezes que cada entidade mencionada coocorre na mesma oração com a expressão do traço semântico associado a *ilha* e *cidade*. Para a recuperação de concordâncias consideramos apenas as variantes com forma de topónimo.

	Çamatra	Iaoa	Martauão	Odiaa	Pequim	Tanixumaa
+ CIDADE	1	7	18	17	38	5
+ ILHA	13	11	1	1	1	11

Tabela 9.2: Valores de coocorrência no corpus com uma expressão dos traços semânticos CIDADE e ILHA para as 6 entidades selecionadas como protótipos dos tipos geográficos *cidade* e *ilha*

Usamos a noção de traço semântico já que um traço pode ser ativado por mais de uma expressão (por exemplo, para +ILHA podemos considerar expressões como *ilhéu*, *ilhota* ou *arquipélago* como ativadoras do traço). Neste exemplo, simplificado, cada traço é ativado por apenas uma expressão.

Por outro lado, as coocorrências também podem ser representadas num sistema de coordenadas (fig. 9.1) a partir de dois traços: um, a abscissa (ILHA), o outro, a ordenada (CIDADE). Deste modo cada entidade mencionada é definida como um vetor, as suas coordenadas são os componentes da posição que podemos representar em forma geométrica com um valor de magnitude e direção.

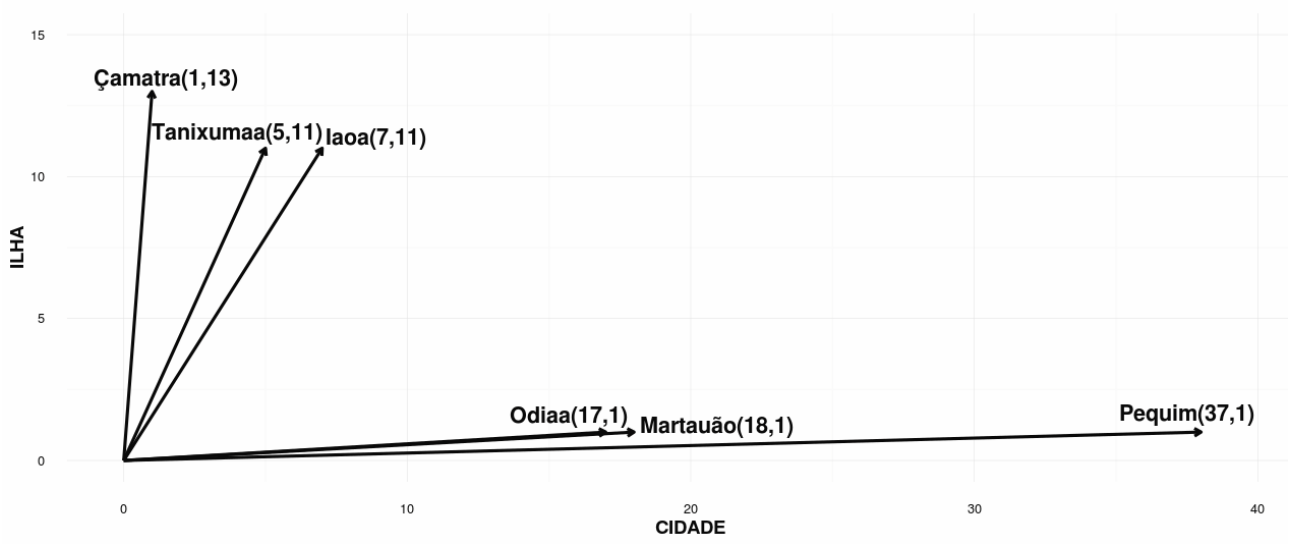


Figura 9.1: Representação das coocorrências num diagrama cartesiano.

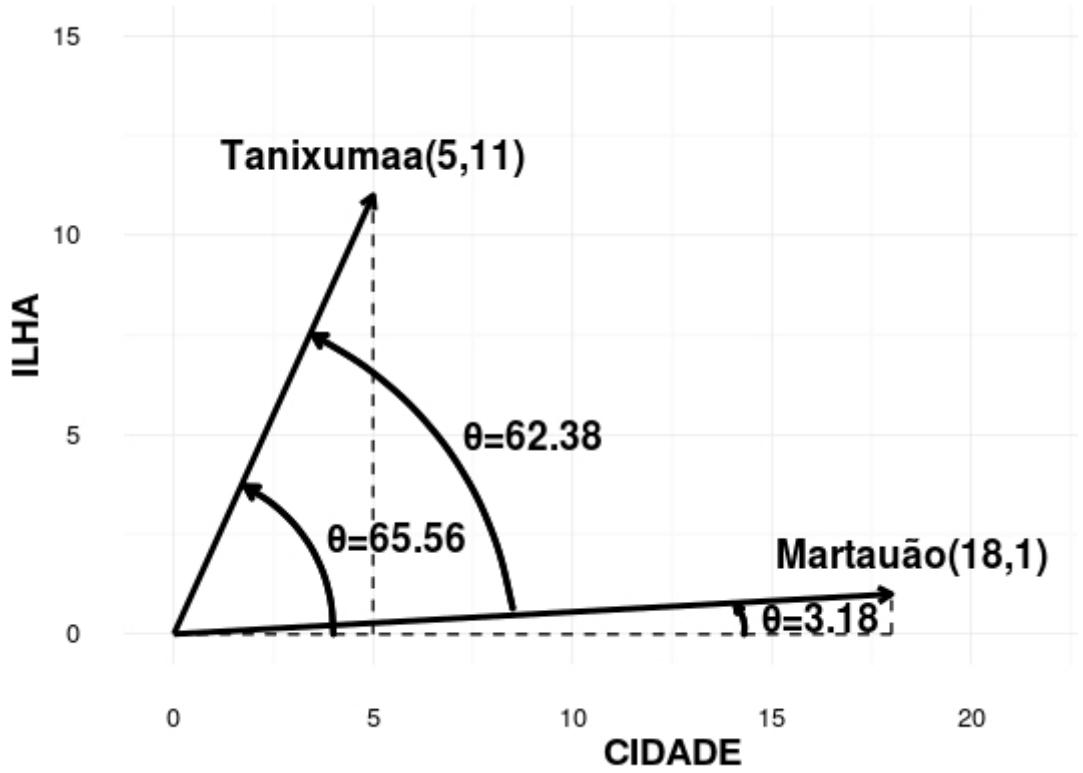


Figura 9.2: Resolução geométrica dos traços CIDADE e ILHA nas entidades mencionadas *Ainão* e *Cantão*.

Considerando o eixo *x* positivo das coordenadas (em direção contrária às agulhas do relógio) obtemos o ângulo da direção do vetor a partir dos componentes da posição. A figura 9.2 acima seleciona as expressões *Tanixumaa* e *Martauão* para ilustrar o procedimento.

Recolhendo os dados da figura 9.2 obtemos os ângulos:

$$\theta \text{ (Tanixumaa)} = \arctan (|11 / 5|) = 65.56°$$

$$\theta \text{ (Martauão)} = \arctan (|1 / 18|) = 3.18°$$

Consequentemente, o valor do ângulo definido pelos traços CIDADE E ILHA para as entidades geográficas mencionadas *Tanixumaa* e *Martauão* é:

$$|\theta \text{ (Tanixumaa)} - \theta \text{ (Martauão)}| = |65.56° - 3.18°| = 62.38°$$

Independentemente da sua magnitude, a direção dos vetores fica entre os 0° e 90°. O cosseno do ângulo fica deste modo como sendo medida da similitude semântica entre as entidades mencionadas relativamente aos tipos geográficos. O apêndice VI mostra os resultados obtidos para todas as entidades mencionadas consideradas. Trazemos como exemplo *Tanixumaa*:

$$\cos(|\theta \text{ (Tanixumaa)} - \theta \text{ (Çamatra)}|) = \cos (20.05°) = 0.94$$

$$\cos(|\theta \text{ (Tanixumaa)} - \theta \text{ (Iaoa)}|) = \cos(0.03°) = 0.99$$

$$\cos(|\theta \text{ (Tanixumaa)} - \theta \text{ (Martauão)}|) = \cos(62.38°) = 0.46$$

$$\cos(|\theta \text{ (Tanixumaa)} - \theta \text{ (Odiaa)}|) = \cos(62.19°) = 0.47$$

$$\cos(|\theta \text{ (Tanixumaa)} - \theta \text{ (Pequim)}|) = \cos(64.05°) = 0.44$$

Os valores mais altos correspondem com a maior similitude (1 o valor máximo), que achamos com *Çamatra* e *Iaoa*, ambas as duas classificadas como ilhas. Quanto mais baixo for o valor do cosseno, menor similitude (caso das cidades *Martauão*, *Odiaa* e *Pequim*).

9.3.2 Captura de relações no conjunto do corpus

Uma das limitações da medida de similitude apresentada na secção anterior é a necessidade de contarmos com um corpus suficientemente representativo da entidade para podermos determinar a variação no seu contexto. Quanto maior for a frequência dos termos, melhor esperaremos capturar a proporção de coocorrências representativas de uma relação, em oposição a aquelas outras simples coincidência de expressões numa mesma oração. Nesse sentido, o corpus objeto de estudo neste trabalho aparece limitado para resolver as entidades com menores frequências. Necessitamos que os termos tenham uma frequência estatisticamente significativa para estabelecermos comparações. Partindo desta limitação inicial, examinamos os resultados obtidos da consideração do conjunto do corpus como um espaço multidimensional, em que cada termo ou entidade é representado por um vetor, como recolhido na tabela 9.2, mas desta vez para avaliar a ocorrência com o conjunto de termos (geográficos ou doutro tipo) e entidades do corpus.

9.3.2.1 Aplicação de um modelo de aprendizado de máquina

O aprendizado de máquina permite obter uma boa aproximação a um resultado a partir de volumes importantes de dados (Alpaydin, 2014). Definido um problema, usando uma parte dos dados como experiência, treinamos um sistema para criar um modelo que, confrontado com o mesmo tipo de

variáveis, tenha capacidade de predizer uma resposta (Mitchel, 1997). No caso que nos ocupa, o problema a resolver é a classificação das entidades mencionadas usando como dados as suas coocorrências no corpus. Continuamos o problema classificatório em que definimos umas entidades como protótipos representativos da classe como exemplo exploratório. Para a aprendizagem usamos os dados do índice (estudo crítico) em que as relações foram resolvidas numa combinatória de conhecimento prévio e revisão manual. A análise que pretendemos é determinar se é possível, para um corpus como o que nos ocupa, treinar um sistema capaz de resolver, ainda que seja de modo aproximado mas suficientemente significativo, uma relação semântica para uma entidade mencionada. O apêndice VII mostra os resultados obtidos da aplicação de um modelo classificatório Random Forest no pacote Caret (Kuhn, 2008; 2016) de R (R Core Team, 2016). Como atributos escolhidos, em vez de apenas dois tipos (fig. 9.2) consideramos agora todos aqueles que contribuam para as relações de meronímia (entidades geográficas) e hiponímia (tipos geográficos). Descrevemos os resultados a seguir.

9.3.2.2 Exemplo de classificação supervisada para as instâncias de classe

A nível exploratório consideramos, em primeiro lugar, a classificação das entidades geográficas relativamente às classes no topo da hierarquia da ontologia (Apêndice VII, Ronda 1), a continuação realizamos testes de classificação binária de uma classe relativamente ao resto (Apêndice VII, Ronda 2 e Ronda 3). Em todos os testes obtivemos resultados significativos, os melhores índices para as classes com entidades de maior frequência absoluta, a exceção a classe *Construções* cujas instâncias têm frequências mais baixas e maior índice de subclasses na ontologia (a sua classificação teve de facto maior intervenção a respeito dos atributos do corpus, isto é, maior variação entrópica na redução dos tipos, vid. tabela 8.1).

Realizamos também uma bateria de testes para comparar o efeito da classificação das entidades em relação aos tipos *ilha* e *cidade* considerando uma matriz de ocorrências ao modo da tabela (9.3), mas, desta vez, com todas as entidades mencionadas e termos geográficos extraídos do corpus. No modelo de aprendizado de máquina adicionamos como preditores três variáveis da análise do corpus: a frequência absoluta e o seu tipo gramatical (topónimo ou gentílico).

A figura 9.3 mostra os resultados obtidos do treino de modelos preditivos a partir da base de dados de análise do corpus e o cômputo de matrizes de coocorrências de termos e entidades. Os dados são repartidos em 70% para aprendizagem e 30% restante para testar o desempenho do modelo inferido. Observamos como em todas as medidas os desempenhos médios ficam por cima de 70%. A classe *ilha* apresenta um caso de valor atípico, no entanto, salientamos que se produz na abrangência (na precisão, para o mesmo atributo, o melhor resultado é de 100%). O atributo *cidade* (com maiores frequências no corpus), obtém resultados médios aproximados a 90%, o pior nunca por debaixo de 70%.

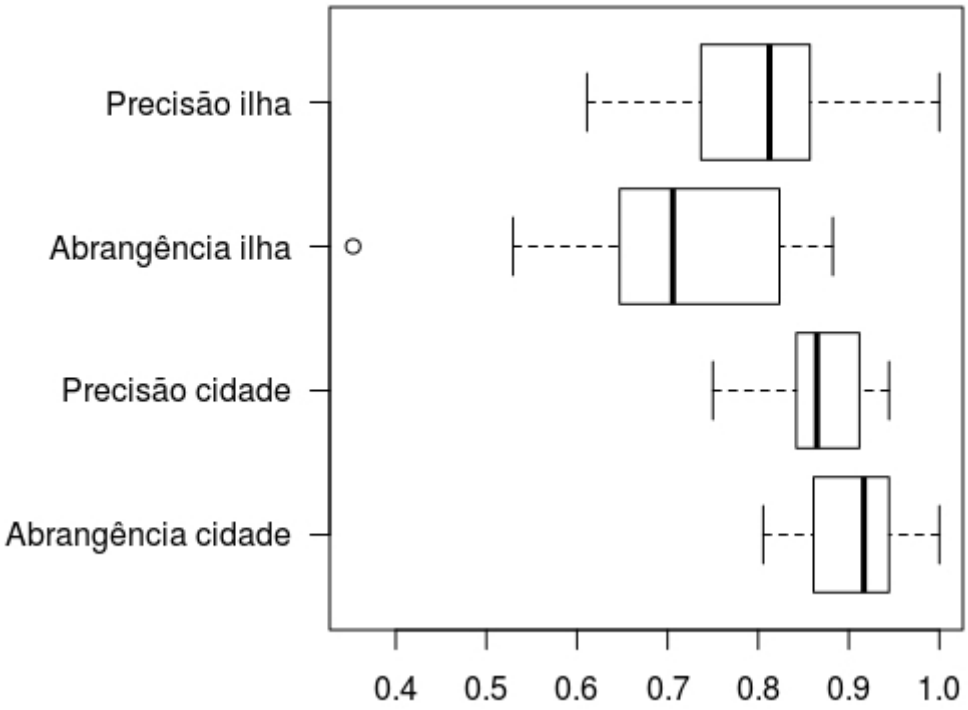


Figura 9.3: Resultados da classificação de entidades geográficas com atributos *cidade* e *ilha* obtidos do treino (25 testes) sobre a base de dados das entidades e da análise de coocorrências do conjunto de termos do corpus.

A análise de erros do teste com melhores resultados (Apêndice VII, Ronda 4) mostra casos que poderiam ser facilmente resolvidos pela incorporação ao modelo de atributos preditivos morfológicos (formas *Pulo* ou *ilha* dentro do topónimo).

9.3.2.3 Exemplo de classificação para a relação *é_Parte_de*

A nível exploratório consideramos as associações de cada entidade geográfica mencionada a partir da análise das suas coocorrências.

Pequim: China (17), Tartaria (10), Nanquim (3), Batampina (3), Calaminhan (2), Cantão (2), Cauchenchina (2), Lançame (2), Nixiamcoo (2), Pegù (2), Quansy (2), Siammon (2), Sornau (2), Amadabad (1), Anapleu (1), Auaa (1), Bagou (1), Banda (1), Bengala (1), Bisnagà (1), Cambaya (1), Capimper (1), Cayro (1), Chaleu (1), Comhay (1), Constantinopla (1), Demaa (1), Egypto (1), Europa (1), Famstir (1), Gouro (1), Guijampee (1), Guimpel (1), Hifaticau (1), Hiquegens (1), Huzamguee (1), Iaoa (1), Iapaõ (1), Lautimey (1), Mecuy (1), Miacoo (1), Nacau (1), Narsinga (1), Odiaa (1), Pacão (1), Palemxitau (1), Paris (1), Passaruão (1), Persia (1), Pocasser (1), Pommitay (1), Pongor (1), Quaygatrum (1), Roma (1), Sansy (1), Seuilha (1), Sileyjacau (1), Taurys (1), Timplão (1), Tinlau (1), Tuymicão (1), Veneza (1), Xinamguibaleu (1).

Tabela 9.3: Entidades mencionadas associadas a *Pequim* pela sua coocorrência no corpus.

A tabela 9.3 mostra as entidades associadas a *Pequim*. *China* aparece em primeira posição com o valor mais alto. A associação a partir das coocorrências intui-se, portanto, como solução para a resolução da relação *é_Parte_de*. No entanto, o corpus apresenta várias dificuldades. A primeira, a limitação nas ocorrências: as entidades com menores frequências ficam limitadas no número de associações e na sua relevância (o corpus não é suficientemente representativo para a captura de relações). Mais outra dificuldade é o facto de a ontologia apresentar uma taxonomia mista que incorpora termos e entidades geográficas não presentes no corpus (Ex. Á*sia Oriental*, *Índico Ocidental*, *Índico Oriental*, *Indochina* e *Insulíndia*) para cobrir ocos nos holónimos da ontologia (§9.2.5). Mesmo se o corpus for suficientemente representativo a nível de frequências, a solução da relação *é_Parte_de* requererá, nestes casos, algum tipo de inferência a partir da ontologia, ou um maior processamento dos dados para reconhecer classes cuja expressão não aparece mencionada no corpus.

Como exemplo da resolução automática usando unicamente critérios estatísticos, consideramos as duas entidades com maior número de instâncias na relação de meronímia (*é_Parte_de*). No apêndice VII registamos os resultados de 25 segmentações aleatórias sobre as entidades geográficas mencionadas resolvidas na base de dados relacional como pertencentes à *China* e à *India*. A diferença da relação de hiperonímia em que considerávamos os tipos geográficos como preditores, usamos agora apenas as entidades geográficas (reduzimos a dimensão da matriz para o treino do modelo). A figura 9.4 mostra os resultados do teste (Apêndice VII, Ronda 5).

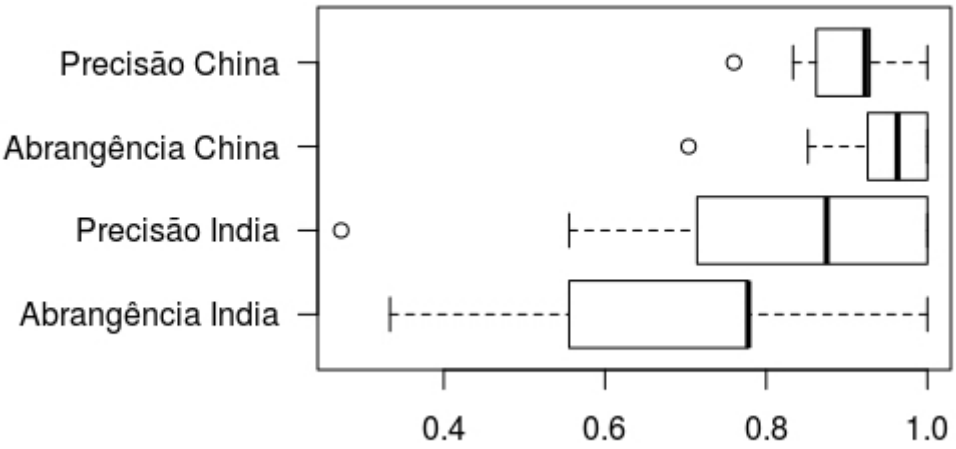


Figura 9.4: Resultados da classificação de entidades geográficas *é_Parte_de* para *China* e *India* obtidos do treino (25 testes) sobre a base de dados das entidades e a análise de coocorrências de entidades geográficas mencionadas no corpus (667 preditores).

Para um problema de classificação binário, sobre as entidades com mais instâncias dentro da relação de meronímia, obtemos resultados médios altos e sempre significativos, com uma média de precisão de 91% para a entidade geográfica com mais instâncias (*China*). Não obstante, na classificação da *India*, ainda tendo resultados significativos, há uma maior oscilação, com um caso

de um teste que mostra um resultado negativo. Neste sentido, há que considerar a menor relevância estatística da amostra (há apenas 7 entidades a classificar) que justifica uma variação alta nos resultados, no entanto a diferença real no desempenho seja mínima. Muito mais relevante é o facto de acharmos a entidade mencionada *Pequim* classificada erradamente (Apêndice VII, Ronda 5, Teste 23). O exemplo serve para ilustrar a importância dos preditores e o seu peso.

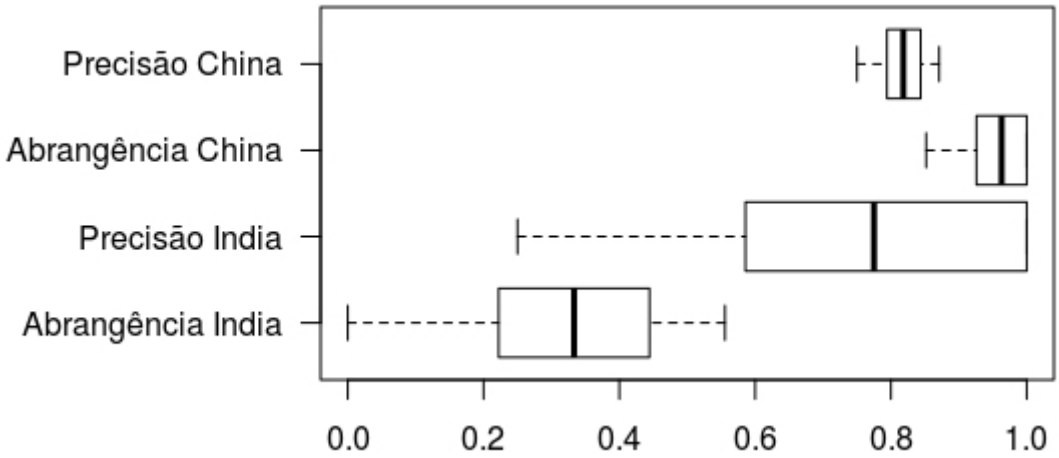


Figura 9.5: Resultados da classificação de entidades geográficas *é_Parte_de* para *China* e *India* obtidos do treino (25 testes) sobre a base de dados das entidades e a análise de coocorrências do conjunto de termos do corpus (4 preditores).

Uma nova ronda de testes com os mesmos parâmetros (Apêndice VII, Ronda 6) e apenas dois preditores como entidades mencionadas (*India* e *China*) classifica *Pequim* corretamente. Na tabela de dados pormenorizados de cada teste (Apêndice VII, Ronda 6) observamos que diminuem os casos de máxima precisão (100%), no entanto, a figura 9.5 mostra uns resultados similares e mesmo melhora na presença de valores atípicos a respeito do teste anterior em que usamos todas as entidades como preditores (fig. 9.4). Os resultados apontam para uma maior estabilidade, com menor variação, conservando um desempenho médio similar. A importância e qualidade aparecem neste caso como mais importantes que o número de variáveis usadas como preditores.

9.3.3 Considerações sobre a captura automática de relações

Os requerimentos de volume de dados requeridos pelos modelos explorados limita a sua aplicação no corpus. Usamo-los de modo exploratório, para testarmos soluções atuais na captura de relações semânticas. Confirmarmos os seus limites, no entanto, também a sua aplicabilidade quando os dados forem suficientemente representativos. Nos testes iniciais consideramos todas as classes para finalmente elaborarmos um teste classificatório de tipo binário. Para as instâncias da relação de classe (atributo geográfico) realizamos um teste sobre os atributos *cidade* e *ilha*. Mesmo para um corpus reduzido, como é o caso de estudo, os resultados obtidos num trabalho deste tipo têm desempenhos médios por cima de 70% e, no caso do atributo com maiores frequências, à volta de 90%. Na relação *é_Parte_de*, observamos uma limitação ainda maior pelo número de coocorrências no corpus, mesmo escolhendo os dois casos com mais instâncias (*China* e *India*), encontramos

resultados que podem ser explicados pela menor significatividade estatística da amostra (*India*), não obstante, para o caso com maior frequência no corpus (*China*) obtemos resultados por cima de 80%. Finalmente, a observação de erros na classificação, também colhendo como exemplo um caso com frequência alta no corpus (*Pequim*), mostrou que, para um mesmo problema, a redução dos preditores, deixando apenas os mais relevantes, manteve uns resultados médios similares e corrigiu a distorção classificatória num caso prototípico. O alto índice de representatividade dos exemplos escolhidos (maiores frequências) confirma que a aplicação de técnicas estatísticas contribui para solucionar as relações precisas para o nosso modelo de georreferenciação quando o volume de dados (frequência das ocorrências das entidades) é suficiente. Os melhores resultados dos exemplos (China) têm ocorrências do lexema > 300 (fig. 7.19). Porém, os testes aplicados não resolvem o conjunto das entidades do corpus, apenas mostram que, com um número de ocorrências suficientes, é possível automatizar e resolver, com resultados muito satisfatórios, trabalhos de instanciamento de entidades geográficas mencionadas. Contribui, nesse sentido, para alargar os potenciais usos do modelo concetual para além do uso restrito do corpus do caso prático.

9.4 Elaboração de um índice de entidades geográficas mencionadas

Uma vez definida uma tipologia e sistematizada uma ontologia que relacione as entidades, uma aplicação prática é elaborarmos um índice para ordenar os dados mais representativos obtidos do trabalho com as entidades geográficas do corpus. Para a *Peregrinação 1614* temos um texto sem normalizar, com expressões sem correspondentes conhecidos ou facilmente reconhecíveis na atualidade, às que lhe há que adicionar o facto de considerarmos os gentílicos como variantes com o mesmo valor que os nomes próprios (os gentílicos sim têm entrada nos dicionários, como adjetivos ou nomes comuns no português). Temos assim variações gráficas, morfológicas ou mesmo expressões em distintas línguas para uma mesma entidade geográfica. Surge, portanto, o problema de selecionar a forma padrão que represente o conjunto de expressões com um mesmo objeto geográfico como referente.

A solução que propomos é escolher como forma representativa a variante com maior frequência no corpus. Deste modo aplicamos um critério objetivo para trabalhar com todas as expressões e agrupá-las, facilitando o estudo das variantes correspondentes a uma mesma entidade geográfica. No entanto, a ordenação dos elementos da lista pode estar condicionada por critérios de presentação (elaboração de índices) que condicionam uma escolha independentemente do valor da frequência. Por exemplo, para as entidades geográficas conhecidas, a convenção é usar o topónimo (isto é, o nome próprio é preferido ainda quando o gentílico tenha uma frequência maior). Outro caso de dificuldade acontece quando duas variantes têm a mesma frequência absoluta no corpus, consequentemente, é preciso adicionar um critério diferenciador para a escolha da forma padrão.

Como caso prático, elaboramos um índice com todos os lexemas identificados do corpus a que adicionamos as ocorrências por capítulos, a relação *é_Parte_de* e mais outras relações com valor espacial não consideradas nesta versão da ontologia, mas anotadas no estudo crítico e sistematizadas de um modo formal na base de dados relacional como sendo pontos de referência

complementares do holónimo que possam contribuir para uma solução da georreferência exata: direção, distância percorrida (em tempo e unidades de comprimento), pertença e proximidade.

9.4.1 Elaboração das listas

Em §6.2.1.1 definimos:

W= {expressões das entidades geográficas mencionadas} = {w_1, w_2, …, w_n} onde *w* é um nome de lugar que usamos para nos referir a uma entidade geográfica mencionada contida no corpus e W abrange o conjunto de todas as expressões com que operamos, isto é, todos os elementos anotados como entidades geográficas no corpus.

Em §6.2.1.2:

G = {referentes para as entidades geográficas mencionadas de W}

é uma lista que contém a descrição dos referentes das entidades geográficas mencionadas.

Com os elementos definidos em §6.2.1 podemos operar formalmente para criarmos a lista de expressões representativas.

9.4.1.1 Lista de expressões representativas

Em §6.3.2.3 introduzimos a noção de *expressão representativa* como aquela que agrupa todas as expressões de um mesmo referente.

Seja P ={lista de expressões representativas} = {p_1, p_2, …, p_n} em que *p* é a expressão que representa todas as expressões (§6.2.1.1) com um mesmo referente (§6.2.1.2).

Definamos a relação:

Tem_expressão_representativa(w,p) = "*w* é uma expressão representada por *p*" (9.1)

Para criarmos um representante $p \in$ P é necessária como mínimo uma expressão $w \in$ W associada a um referente $g \in$ G:

$\forall\, w \in$ W ($\exists\, g \in$ G, Tem_georreferente(w,g)) $\rightarrow \exists\, p \in$ P, Tem_expressão_representativa(w,p) (9.2)

Isto é, toda entidade geográfica mencionada referenciada tem uma expressão representativa.

9.4.1.2 Definição de uma lista de expressões com mesmo representante

Denominemos de W' ao subconjunto de expressões $\subset$ W delimitado por um referente *g*, sendo as expressões obtidas da função definida em (6.2) que devolve uma expressão dado um referente:

$$W' = \{\text{Topónimo}(g) \mid g \in G\} \quad (9.3)$$

W' é agora a lista que contém todas as expressões agrupadas baixo uma mesma expressão representativa.

9.4.1.3 Definição formal de expressão representativa

Chegamos assim a uma definição formal de expressão representativa como o elemento $p \in$ P, P ={lista de representantes} que representa o subconjunto W' $\subset$ W definido em (9.3).

A expressão representativa vem finalmente dada pela função:

$$\text{Expressão\_representativa}(\arg\max_i \text{freq}(w_i)) = \text{p} \qquad (9.4)$$

que outorga o valor da expressão *w* a *p* ao selecionar o elemento $w_i \in$ W' com a frequência absoluta mais alta no corpus segundo as listas obtidas por uma das fórmulas para as frequências de variantes em §4.4. A fig 9.6 mostra um exemplo da relação entre a expressão representativa e as expressões do corpus que representa.

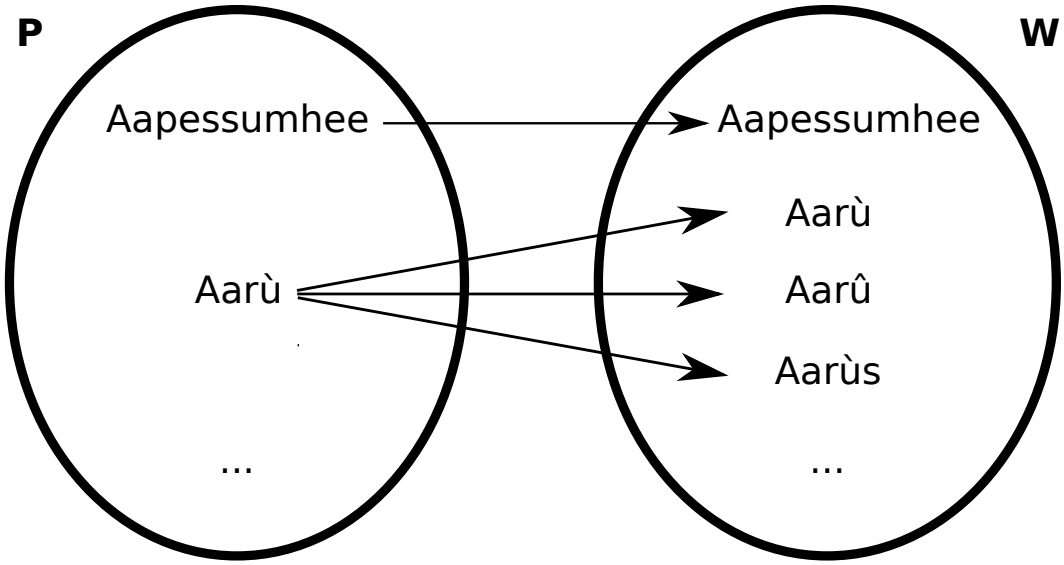


Figura 9.6: Diagrama com os dois primeiros representantes da lista P e as suas correspondentes expressões na lista que contém todas as expressões, W.

Exemplo 1 Expressão representativa de um lexema sem variantes

Aapessumhee é uma expressão de entidade geográfica anotada no corpus da *Peregrinação 1614* com uma única ocorrência, sem nenhuma outra variante relacionada em todo o corpus.

> "& chegando a hum lugar que se dizia **Aapessumhee**, quatro legoas do rio de Puneticão ..." (*PR*, 32).

Neste caso na lista W temos a expressão w_1= "Aapessumhee".

A expressão tem um referente explícito no corpus:

g= "*a quatro légoas do rio de Puneticão"*,

Para *w* = "Aapessumhee" e *g* = "*a quatro légoas do rio de Puneticão"* cumpre-se a relação (6.1):

Tem_georreferente(w,g) = Tem_georreferente(Aapessumhee,“*a quatro légoas do rio de Puneticão*”) = *Aapessumhe* tem o georreferente *a quatro légoas do rio de Puneticão*

Dado que só há uma expressão com este referente, o subconjunto W' ⊂ W, W' = {*Aapessumhe* | *g* = “*a quatro légoas do rio de Puneticão*”} (9.3) tem um único elemento. Fica assim:

Expressão_representativa(arg max$_i$ freq(w_1)) = w_1 = “Aapessumhe”.

Na lista de lexemas esta é a primeira entrada, portanto resolvemos p_1= “Aapessumhe”.

Exemplo 2 Expressão representativa de um lexema com variantes

Aarù, *Aarû*, *Aarùs* são expressões que referem a mesma área geográfica, situada como um ponto com coordenadas na georreferenciação por conhecimento prévio em 4º 10' N 98º 08' E”. Isto é, para g = “4º 10' N 98º 08' E” obtemos (9.3):

W' = {*Aarù*, *Aarû*, *Aarùs* | *g* = “4º 10' N 98º 08' E”}

Dizemos que as expressões *Aarù*, *Aarû* e *Aarùs*, todas elas contidas em W, correspondem-se com um mesmo nome *p* que as representa em P. Isto é, *Aarù*, *Aarû* e *Aarùs* são expressões com um mesmo objeto como referente.

Para obtermos a expressão representativa usamos a função (9.4) que seleciona a expressão com a frequência mais alta. As tabelas de frequências foram obtidas em §4.4.

Variante	Frequência absoluta
Aarù	21
Aarû	10
Aarùs	2

Tabela 9.4: Frequências absolutas das expressões *Aarù*, *Aarû* e *Aarùs*

De onde se tira, para W' = {*Aarù*, *Aarû*, *Aarùs*} segundo (9.4):

Expressão_representativa(arg max$_i$ freq(w_i))

= Expressão_representativa(w_1) = *Aarù*

A expressão selecionada é, portanto, *Aarù*.

9.4.2 Implementação da lista de expressões representativas para a elaboração de um índice de entidades geográficas

Uma aplicação prática da ordenação das entidades geográficas mencionadas e os seus referentes em listas é a elaboração de um índice. Nesta secção descrevemos o processo de ordenação das variantes para mostrar um índice completo de todas as entidades mencionadas no corpus, recuperando ademais informação adicional sobre a sua ocorrência e georreferência.

9.4.2.1 Descrição das regras de ordenação

Para a extração da expressão representativa a partir da lista ordenada de expressões e referentes aplicamos uma série de critérios restritivos:

> Seleciona como candidatos para a expressão representativa apenas aquelas variantes que sejam topónimos.
>
> > De entre todos os topónimos, seleciona aquele que tenha a frequência mais alta:
> >
> > > Se mais de um topónimo tem a mesma frequência máxima, aplica um critério de ordem alfabética.
> >
> > Se só houver gentílicos para esta entidade geográfica, seleciona o gentílico com a frequências mais alta:
> >
> > > Se mais de um gentílico tem a mesma frequência máxima, aplica um critério de ordem alfabética.

9.4.2.2 Configuração dos dados do índice

Um índice associa uma entrada a um conteúdo. Para o caso prático do corpus escolhemos como dados de saída os de um índice de obra convencional, que aponta para o capítulo em que a entidade geográfica é citada (sendo assim válido para qualquer edição). Como dados complementares anexamos:

Variantes. Em itálico, todas as variantes representadas pela forma padrão. Em caso de haver apenas uma expressão, é a mesma que a forma padrão.

Capítulo. Depois de cada variante, o capítulo ou capítulos em que aparece.

Ocorrência no capítulo. Número de vezes que a variante ocorre no capítulo. Entre parêntese depois de cada capítulo.

Tipo geográfico. A categoria principal da entidade na forma recolhida no estudo crítico do corpus (cap. 7).

Tipos geográficos alternativos. Outras categorias registadas no corpus para esta entidade geográfica usadas no estudo crítico como georreferências relativas, entre parêntese depois do tipo principal.

Relação Parte_de. A classe a que a entidade pertence pela relação *é_Parte_de*.

Relacionada com. Outras relações georreferenciadoras da entidade. Tem uma convenção própria:

> **# #** expressa a medida usada para assinalar a distância ao ponto de referência. Se não expressa distância, implica uma situação de proximidade a respeito do ponto de referência (isto é, a georreferenciação faz-se dentro de um polígono):
>
> **@ @** expressa o ponto de referência sobre o qual se aplica a medida que o precede.

A georreferência relativa expressa relações não consideradas no conceito central desta tese, mas que servem também para criar uma referência, isto é, cumprem a relação:

Tem_georreferente(w,g)

Assim para:

w=Tanauquir

g= “#40 léguas#@Mutipinão”

temos a relação:

Tem_georreferente(Tanauquir, “#40 léguas#@Mutipinão”)

= Tanauquir tem o georreferente “a 40 léguas de Mutipinão”.

Nome internacional contemporâneo. A forma contemporânea segundo o identificativo de GeoNames, se a entidade geográfica for georreferenciada por coordenadas.

Entidade administrativa de primeira ordem atual. A entidade administrativa de referência conforme a uma forma internacional, quando a entidade tiver um nome contemporâneo identificado.

Continente. Entre parêntese e maiúsculas, abreviatura do continente da entidade administrativa de primeira ordem.

Coordenadas. Latitude e longitude.

Como resultado obtemos um índice das entidades geográficas mencionadas (Apêndice VIII) que agrupa as variantes sob uma mesma forma, o topónimo com a forma mais comum no corpus. Se não houver topónimo, o gentílico.

9.5 Conclusão

Ordenamos as relações entre as entidades geográficas através de uma ontologia que caracterizamos como uma taxonomia com capacidade de inferência. Introduzimos exemplos de classes, subclasses e relações das entidades geográficas mencionadas do corpus para mostrarmos o modo em que ordenamos as entidades.

Limitamos a captura das relações para a ontologia a duas relações semânticas, as usadas no modelo de conceito (cap. 6), não obstante, no processo do estudo crítico (cap. 7) anotamos mais relações de tipo espacial que podem ser aproveitadas num desenvolvimento futuro da georreferência relativa. Continuando com o método de aproveitamento do corpus para a pesquisa do desempenho de ferramentas e métodos de PLN, exploramos a captura automática das relações semânticas que estruturam inicialmente a ontologia. Usamos os casos de entidades com maior frequência para uma análise exploratória de tipo preditivo em trabalhos de classificação. Os resultados obtidos foram altamente significativos mostrando o potencial do modelo concetual proposto para trabalho futuro com corpora de maior volume.

Uma vez resolvidas as relações numa ontologia, usamos os resultados para ordenar os dados da base relacional, de onde extraímos os termos da definição, mais dados complementares de corpus e

espaciais. Surge o problema de escolher qual é a variante para a entrada num índice que ordene as entidades do mesmo modo que um vocabulário, atlas ou dicionário, com a dificuldade de, no nosso caso prático, trabalharmos com topónimos e gentílicos cuja forma padrão é desconhecida. Usamos como critério inicial de ordenação a frequência absoluta no corpus, neutral relativamente às propriedades morfológicas das variantes. Não obstante, critérios de apresentação podem requerer ordenações alternativas. Os próprios condicionantes de saída alteram a ordenação inicial a partir de regras de seleção.

Resolvida a expressão representativa que serve de entrada, enriquecermos o índice com dados elaborados no processo de análise crítica da base documental (cap. 7). Como resultado obtemos um glossário com a numeração dos capítulos originais, georreferências para todas as entidades segundo o modelo de conceito (cap. 6) <tipo geográfico, parte_de>, e relações anotadas no estudo crítico tais como áreas e distâncias a outras entidades indexadas. No caso das entidades georreferenciadas por conhecimento prévio (cap. 7) oferecemos também um nome contemporâneo, uma entidade administrativa de referência, o código do continente e as coordenadas geográficas consideradas o seu centroide.

O índice obtido é apenas um dos possíveis, as listas conforme definidas podem ser reordenadas e ajustadas segundo novas regras.

9.6 Sumário de objetivos

Objetivos desta secção foram:

- Concluir a definição da entidade geográfica mencionada com a consideração da relação *é_Parte_de*.
- Integrar a taxonomia e as relações usadas para a definição do conceito numa ontologia.
- Aplicar os resultados da análise de frequências do corpus e da georreferência no estudo crítico para explorar a extração automática das relações usadas na definição das entidades geográficas mencionadas.
- Extrair um índice de entidades geográficas mencionadas da base de dados relacional.

Capítulo 10

Síntese de resultados, principais contributos e trabalho futuro

Examinamos os principais resultados deste trabalho, os materiais elaborados, as limitações e possíveis aplicações, e linhas de trabalho consideradas prioritárias para o desenvolvimento do modelo de georreferenciação proposto.

10.1 Síntese de resultados

Começamos este trabalho definindo o problema da identificação e referenciação das entidades geográficas mencionadas e centramos a proposta no desenvolvimento de um modelo para georreferenciar aqueles casos não solucionáveis pela simples recuperação de coordenadas. Na hipótese central, um mesmo modelo concetual integra as entidades cuja georrefência é conhecida com aquelas outras desconhecidas, contribuindo deste modo para o georreferenciamento relativo das segundas. Definimos duas grandes áreas na resolução do problema: primeiro, a identificação das expressões das entidades num texto. Segundo, uma vez anotada a entidade como geográfica, o georreferenciamento propriamente, cujo fim é apontar para um objeto geográfico no mundo real.

Situamos o trabalho sobre as entidades geográficas mencionadas dentro de um âmbito interdisciplinar. A hipótese secundária da tese, metodológica, prioriza as técnicas de PLN como proposta para a consecução dos objetivos práticos. Enquadramos como geográficos os resultados finais do índice de entidades geográficas mencionadas (útil de geografia histórica) e a base de dados SIG das entidades de conhecimento prévio. Durante todo o desenvolvimento da tese, foram usadas ferramentas geográficas, especialmente para a geovisualização e, em menor medida, para a análise e representação cartográfica. Apareceram como problemas metodológicos aspectos estudados no âmbito NERC e disciplinas mais computacionais, a sua solução foi atendida como procedimento para avançar num fim (identificação de entidades para a anotação do corpus). Na aplicação do caso prático, o estudo das georreferências foi em primeiro lugar abordado manualmente com aparato crítico procedente das disciplinas da geografia histórica, a história e a filologia.

Elaboramos um corpus padrão dourado para o estudo das entidades geográficas mencionadas. A partir de uma transcrição digital, defeituosa, da primeira edição da *Peregrinação*, realizamos sucessivas melhorias sobre o texto comparando-o com edições impressas e o original em versão fac-similar. Paralelamente, levantamos manualmente um índice de entidades geográficas mencionadas que nos serviram para anotar de modo semiautomático o corpus. O resultado foi um corpus anotado

que fomos melhorando (correção de erros de anotação, ampliação do marcado) e mesmo expandimos com a criação de um corpus paralelo, em que os capítulos com unidade temática da Tartária foram alinhados relativamente à primeira edição em inglês. O trabalho com o corpus seguiu uma pauta cíclica, de questionamento e processamento, para responder a novos objetivos segundo avançávamos nos labores de georreferenciação. A partir das análises exploratórias iniciais para introduzir a lei de Zipf e observar como a frequência se mostrava relevante na distribuição geográfica dos capítulos, sucessivos *scripts* foram criando subcorpora e selecionando apenas aqueles aspectos da anotação requeridos nos testes. Deste modo, na pesquisa sobre trabalhos NERC, para além do corpus paralelo, aproveitamos a anotação para distinguir gentílicos de topónimos. Mais adiante, a anotação serviu para criar um subcorpus das orações com valor geoespacial como material experimental para os testes de recuperação de termos geográficos. Na extração de relações trabalhamos com o conjunto do corpus e aproveitamos as anotações como variável de predição. Durante a fase de estudo crítico e geovisualização, o corpus permitiu recuperar as concordâncias das entidades, agilizando a consulta do texto e permitindo pesquisas seletivas para a recuperação das descrições dos referentes.

O corpus foi, assim, o principal material de apoio deste trabalho, o padrão com que se conferiram os resultados obtidos da aplicação das técnicas de PLN na identificação de entidades. Definidos os processos para a anotação, ensaiamos três métodos de automatização. Em primeiro lugar, mostramos as dificuldades surgidas mesmo no melhor dos cenários, quando operamos como uma lista *ad hoc* com todas as entidades mencionadas. Posteriormente empregamos uma ferramenta de anotação automática para, configurada com a lista *ad hoc*, compararmos resultados relativamente ao corpus. Tínhamos um duplo objetivo: primeiro, metodológico, apresentarmos as métricas usadas para a avaliação de resultados no resto da tese e, segundo, material, melhorarmos o corpus mediante a deteção de erros. A avaliação da divergência entre os resultados da ferramenta e o corpus permitiu melhorar a anotação, detetando novas expressões de entidades geográficas mencionadas pela inspeção de apenas 5% das respostas. Finalmente, usamos um corpus paralelo elaborado a partir do alinhamento dos capítulos com unidade temática no espaço geográfico da Tartária nas primeiras edições da *Peregrinação* em português e inglês para considerarmos a automatização do processo completo de anotação. Usamos dois tipos de soluções, um modelo estatístico e outro de regras, em forma de três ferramentas de livre disposição (uma estatística, duas de regras), com que avaliamos resultados em função dos objetivos específicos do nosso corpus, diferentes em parte de aqueles para que foram concebidos os úteis: anotar tanto gentílicos como topónimos e usar uma variante de língua distinta do padrão contemporâneo. Ainda com estes condicionantes, os melhores resultados conseguiram ultrapassar a barreira de 60% na medida-F, com o melhor desempenho próximo a 70%. Como conclusão apresentamos um esquema de procedimento que combina processos automáticos e revisão manual para reduzir o tempo e melhorar a qualidade da anotação de textos não normalizados.

Introduzimos um modelo concetual com que geoferrenciamos todas as entidades mencionadas no

corpus. A partir de uma aproximação semântica referencial, consideramos as ligações entre a expressão, o conceito e o referente, para distinguirmos dois tipos de georreferenciamento. No primeiro, ostensivo, a entidade geográfica mencionada é apontada diretamente por meio de umas coordenadas. No segundo, mais elaborado (intensivo), o referente é denotado através do conceito. Aproveitamos uma noção da semântica cognitiva, em que o conceito se estrutura a partir de regras e atributos, para elaborarmos um esquema simples, com apenas dois componentes, em que o referente é denotado pela adscrição de um tipo geográfico e uma relação espacial com outra entidade. Uma particularidade do nosso modelo, a diferença dos cognitivos, é mantermos o referente como ente físico, objeto geográfico. O resto da tese desenvolveu estas duas possibilidades de georreferenciação.

No modo de georreferenciamento ostensivo encontramos um paralelismo entre as entidades referenciadas por coordenadas e aquelas outras que deixamos para a denotação por definição. As primeiras são entidades que conhecemos previamente, reconhecíveis pela aplicação de instrumentos geográficos (SIG, atlas, glossários e estudos específicos). As segundas, entidades não referenciadas por coordenadas, são desconhecidas na documentação ou apresentam alguma dúvida que nos impede de dar a sua localização como segura. Esta distinção tem implicações no procedimento da georreferenciação. As entidades de conhecimento prévio são as primeiras em ser georreferenciadas. O seu georreferenciamento no caso prático da *Peregrinação* começou pela elaboração de uma base documental em que recolhemos quanto trabalho achamos disponível para contextualizar e oferecer referentes das entidades mencionadas. Elaboramos um estudo crítico e visualizamos e anotamos os objetos geográficos em aplicações SIG. Adotamos uma medida conservadora para a avaliação crítica da georreferência. Apenas aquelas entidades mencionadas para as quais não achamos nenhum tipo de contradição têm a máxima probabilidade, P(georreferência)=1, na atribuição do referente e são finalmente incluídas numa lista que chamamos de conhecimento prévio. A comparação da classificação por tipo de conhecimento (prévio ou descrito) com os dados de frequência no corpus mostra como as entidades com maior frequência são também as mais conhecidas, no entanto, aquelas cuja georreferenciação tem de vir dada pela descrição ocupam uma escala 10^{-1} menor numa distribuição de Zipf. A frequência aparece como um elemento determinante para o georreferenciamento da entidade.

Na georreferenciação por descrição, atendendo ao esquema do conceito proposto, o primeiro elemento a solucionar foi a atribuição de um tipo geográfico e a ordenação dos tipos numa taxonomia que classificasse as entidades numa ontologia. Consideramos duas soluções, a aplicação de uma taxonomia prévia, como fizemos no caso de referenciação das entidades de conhecimento prévio (em que um referente tem um tipo geográfico assimilável ao objeto na atualidade), e a criação de um vocabulário a partir dos termos presentes no corpus (procedimento aplicado na descrição das entidades no estudo crítico). A primeira apresenta o problema da adequação dos termos ao corpus, isto é, a terminologia da taxonomia externa pode classificar corretamente a entidade, mas, se não houver um termo equivalente no corpus, dificultará a recuperação de

concordâncias e entidades relacionadas. A elaboração de uma taxonomia *ad hoc* tem a vantagem de descrever mais de perto as entidades e oferecer um vocabulário procedente do próprio corpus, no entanto, deixa vazios na classificação quando a entidade não tem o tipo declarado explicitamente, e resulta difícil de organizar nos níveis superiores da hierarquia (os tipos mais abstratos). A solução adotada foi uma taxonomia híbrida, em que os tipos são extraídos do corpus e integrados no esquema classificatório já usado para as entidades de conhecimento prévio. Outro problema, o da densidade nas classes, foi resolvido pelo cálculo da entropia e o agrupamento dos tipos próximos. Como resultado final obtivemos uma taxonomia com que classificamos as entidades mencionadas do corpus (procedimento manual). Elaboramos assim uma lista de entidades e tipos geográficos com que testarmos procedimentos de extração de terminologia de modo automático. O primeiro teste aplicado teve uma base mais teórica e mostrou como o efeito da frequência na recuperação de candidatos (a ser incluídos como termos da taxonomia) obriga a recuperar mais termos dos necessários pelo carácter exponencial da distribuição do vocabulário (primeira lei de Zipf). Considerada esta limitação, procuramos métricas e métodos de filtrado para limitar o número de candidatos. Com carácter prévio, preparamos um subcorpus geoespacial, formado por orações expressão de proposições com valor geográfico. Sobre este subcorpus realizamos trabalhos de processamento de linguagem natural: substituição da entidade mencionada por uma expressão genérica, anotação da categoria gramatical dos tokens examinados, subsegmentação em cláusulas e frases e limitação do número de tokens a processar (janelas com a entidade geográfica mencionada como centro). Os distintos resultados foram classificados como modos do corpus. Sobre cada modo aplicamos variantes de filtros e métricas para a recuperação dos termos do domínio. Conseguimos os melhores resultados com o maior nível de PLN. Tentamos ainda melhorá-los aplicando um útil mais elaborado, uma base de conhecimento lexical difusa, CLIP2.1 (Gonçalo Oliveira & Gomes, 2014; 2016), de recente elaboração e sem aplicações similares por nós conhecidas, sobre a qual realizamos mais uma bateria de testes até conseguirmos uma configuração com que atingimos uma medida-F por cima de 80% e resultados de precisão de 100% sem diminuirmos consideravelmente a abrangência relativamente aos métodos de filtrado e métricas dos testes anteriores. Como conclusão, consideramos a aplicação de este tipo de base lexicais um recurso que incrementa notavelmente o desempenho nos trabalhos de extração de termos e situamos aqui um dos principais contributos metodológicos desta tese.

Uma vez conseguida uma taxonomia para o caso da *Peregrinação* e classificadas as entidades geográficas mencionadas no seu tipo, ficou apenas por desenvolver a relação espacial para obter a definição da entidade que, no caso das entidades não conhecidas previamente, representa o seu georreferenciamento relativo. Como parte do estudo crítico, as entidades foram anotadas para várias relações, as mais importantes: a distância (em medidas de longitude e tempo) e proximidade a outra entidade. No modelo concetual aplicado nesta tese apenas usamos a relação *é_Parte_de*, que se corresponde com a meronímia em termos semânticos. Por outro lado, a taxonomia dos tipos geográficos tem o seu paralelo na hiponímia (membro da classe) e hiperonímia (a classe). Estas relações semânticas organizam as entidades e os tipos numa ontologia, um modo de ordenar todos

os elementos do modelo (entidades, tipos e relações). Para avaliarmos as possibilidades da captura automática no corpus, consideramos a hipótese do modelo distribucional em que o significado de um termo vem dado pelo seu contexto. Os estudos revistos para a contextualização do modelo vetorial, aplicado neste tipo de problemas, usam preferencialmente grandes volumes de dados com frequências altas para os termos objeto de análise. Considerada esta limitação do nosso corpus, começamos por ilustrar o método com um teste de tipo binário: selecionadas as entidades mencionadas com maior frequência para os tipos *ilha* e *cidade*, formamos para cada uma o vetor das suas coocorrências e observamos a sua distribuição num diagrama cartesiano. Posteriormente usamos esta representação para demonstrar o funcionamento da medida de proximidade do cosseno na classificação das entidades dentro de um dos dois tipos. Exemplificado o método, procedemos a classificar o conjunto das entidades em função de todos os tipos no topo da taxonomia. Por meio de aprendizado de máquina, treinamos um modelo baseado numa matriz composta pelos vetores de coocorrência de entidades geográficas e tipos para avaliar os seus resultados sobre 20% da lista da taxonomia que deixamos fora do treino. Para aproveitarmos melhor o corpus, realizamos o mesmo teste alterando de modo aleatório a segmentação da lista em treino e avaliação. Os resultados foram significativos para as classes no topo dos tipos administrativos, terrestres e hidrológicos, as mais relevantes em termos de frequência no corpus. Atendendo aos resultados, realizamos outro teste exploratório para avaliar a relação de hiponímia nos tipos *ilha* e *cidade* (os dois com frequência alta), mas desta volta considerando o conjunto de entidades do corpus. Com o mesmo critério e objetivos, consideramos a relação *é_Parte_de*. Em ambos os casos, os resultados obtidos foram altamente significativos, com níveis de precisão média para os tipos e entidades mais comuns por volta de 90%. Ainda sendo testes exploratórios, ilustrativos de um método, aplicados apenas a tipos selecionados por terem as frequências mais altas, contribuem para confirmar a aproximação metodológica da tese, baseada na pesquisa de variáveis quantitativas para o modelado do corpus e, em última instância, das georreferências.

Para o objetivo deste trabalho, a ontologia permitiu-nos revisar as entidades geográficas mencionadas e as suas relações e assim criarmos um índice das entidades mencionadas no corpus. A elaboração do índice requereu a escolha da expressão que representasse todas as variantes associadas, tal e como fazemos na elaboração de um dicionário ou índice geográfico num atlas. Aos dados do modelo concetual adicionamos outros de frequências e do estudo crítico da base de dados relacional. Obtivemos assim um índice esquemático das entidades geográficas mencionadas com os dados de ocorrência (por capítulos para permitir a sua recuperação em qualquer edição) e da georreferência, assim como outras relações espaciais que denotam o objeto geográfico e não foram consideradas no esquema concetual proposto neste trabalho.

10.2 Principais contributos e aproveitamento dos materiais

No desenvolvimento da tese produzimos materiais que podem ser reutilizados em trabalhos futuros.

Corpus anotado de entidades geográficas mencionadas da *Peregrinação*

A utilidade dos corpora para labores de PLN constitui um dos argumentos centrais da tese. O resultado final do trabalho com a *Peregrinação* permite disponibilizar um corpus de tamanho médio, com as entidades mencionadas anotadas e classificadas em topónimos e gentílicos. No decurso da tese, desenvolvemos também ferramentas de trabalho num ambiente web, com painel de consultas para recuperar as concordâncias de cada entidade.

Corpus paralelo da Tartária

Os corpora paralelos têm utilidade no treinamento de ferramentas de tradução automática. No caso do corpus da Tartária, pelo carácter de texto histórico, tem uma finalidade mais contrastiva, orientada para a análise crítica das edições. O corpus tem saída num ambiente web, em que as orações são alinhadas para serem comparadas. A sua relevância como texto histórico para uma área geográfica pouco estudada e com escassez de fontes escritas (área linguística do mongol), o feito de recolher expressões e mesmo orações que representam a língua original e permitem o contraste com a língua de produção do autor (português), mais a tradução para uma língua relativamente próxima (inglês), fazem do corpus um recurso particularmente relevante para o estudo da história da Ásia e da receção da *Peregrinação* (como são interpretadas grafias e termos nas distintas línguas).

Índice de entidades geográficas mencionadas

O mais completo índice de entidades geográficas da *Peregrinação* de que tenhamos conhecimento é o publicado organizado por Biederman (2010) na obra coletiva editada por Alves (2010), em que cada entidade aparece representada pelo nome padrão contemporâneo, com uma referência a um tipo e uma localização numa área (semelhante à definição usada neste texto). O índice desta tese amplia o trabalho anterior:

1) Organiza todas as entidades, incluindo topónimos e gentílicos, a partir da expressão do corpus. Todas as variantes, mesmo que gralhas evidentes, são consideradas.

2) Referencia as ocorrências por capítulos, de modo que sejam recuperáveis em qualquer edição da obra (não só a primeira edição).

3) Inclui o nome padrão contemporâneo e as coordenadas geográficas para as entidades sobre as quais há consenso no seu georreferenciamento.

4) Inclui um tipo geográfico e uma entidade geográfica superior (holónimo) para cada entidade mencionada.

5) Adiciona relações de distância e proximidade para a georreferência relativa nos casos em que esta apareça expressada de modo quantitativo ou numa relação de proximidade.

Base de dados SIG de entidades de conhecimento prévio

Fruto da classificação das entidades geográficas em conhecimento prévio e descrito, e da restrição da primeira categoria apenas a aquelas entidades sobre as que achamos consenso na atribuição de coordenadas, obtemos um índice de entidades georreferenciadas que pode ser aproveitado para a elaboração de um mapa vetorial num formato padrão SIG. Deste modo pode ser combinado com

outras bases de dados e visualizado sobre novas capas.

10.3 Publicações e outras atividades divulgativas relacionadas

Durante o processo de investigação redigiram-se artigos, disponibilizaram-se materiais on-line e realizaram-se atividades divulgativas sobre os métodos e resultados obtidos. Comentamos aquelas que explicam ou usam diretamente materiais contidos neste trabalho.

Revistas

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

Descreve os principais resultados deste trabalho. Contém excertos das secções §1.4, §2.3 e §10.2.

• Canosa Rodrigues, A. X. (2015). Estudo, selecção e classificação de entidades geográficas para um mapa global da "Peregrinação". *Boletim da Academia Galega da Língua Portuguesa*, *8.* (Em processo de publicação).

Descreve a representação cartográfica e bibliografia usada no estudo crítico das entidades geográficas mencionadas no corpus. Contém excertos da secção §7.2.

• Canosa Rodrigues, A. X. (2013). Notas biográficas e estudo das referências documentais de Fernão Mendes Pinto. *Veredas*, *20*, 9-34. Disponível em http://ojs.lusitanistasail.org/index.php/Veredas/article/view/2

Analisa a receção histórica como repertório geográfico do corpus estudado neste trabalho.

Plataformas geográficas

• Canosa Rodrigues, A. X. (2016). *Asia in the 16th century*. Arquivos SIG. Disponível em http://worldmap.harvard.edu/maps/8478

Arquivo vetorial em formato *shapefile* das entidades geográficas da Ásia referenciadas por conhecimento prévio para uso com ferramentas SIG. Os mapas das figuras 7.17 e 7.21 foram produzidos ao processar o arquivo com o SIG QGIS.

• Canosa Rodrigues, A. X. (2011). *Mapa com localidades da Peregrinação. (Versão 1.0) / The travels of Mendes Pinto. Map of mentioned places. (Version 1.0)*. Documento KML. Disponível em http://goo.gl/iqc3P

Representação cartográfica e estudo crítico em português e inglês usando as referências bibliográficas da base documental citadas na secção §7.2.1. As figuras 7.7, 7.12, 7.13 e 7.14 foram produzidas nesta plataforma. As imagens satélite do capítulo 7 são resultado de visualizar o documento KML no software GoogleEarth.

O estudo crítico em português foi também publicado em formato e-livro:
Canosa Rodrigues, A. X. (2013). *As viagens de Mendes Pinto: guia para um mapa*. E-book: Yr haul

ar y ffenigl.

Conferências

• Text alignment for a parallel corpus. A case study and some applications. *MIU Colloquium Series. Spring 2015*. Ulaanbaatar: MIU.

Conferência para descrever o processo de elaboração do corpus paralelo descrito em §5.6.4.

• The geography of early transoceanic navigations. A positivist approach to place names in The Travels of Mendes Pinto. *MIU Colloquium Series. Fall 2014*. Ulaanbaatar: MIU. Presentação disponível em:

https://www.academia.edu/29936526/The_geography_of_early_transoceanic_navigations._A_positivist_approach_to_place_names_in_The_Travels_of_Mendes_Pinto_Slides_from_a_presentation_at_MIU_Colloquim_Series

Conferência para contextualizar e descrever o procedimento analítico aplicado no estudo crítico das georreferências descrito no capítulo 7 deste trabalho.

• Word frequency and sentence length for a basic model of sentence complexity. *MIU Colloquium Series. Spring 2014*. Ulaanbaatar: MIU.

Conferência sobre as bases frequentísticas da língua e análise de corpus descrita em http://www.canosarodriguez.net/pwyll/. As ferramentas e métodos comentados foram aplicados na preparação e anotação do corpus descrita no capítulo 4 deste trabalho.

10.4 Trabalho futuro

Na consideração das entidades de conhecimento prévio deixamos fora qualquer entidade que tenha apresentado discrepâncias na base documental. No entanto, a combinatória da análise crítica com a descrição do corpus permite resolver, em termos de coordenadas e com um alto nível de probabilidade, um bom número de entidades descritas apenas como relativas. Por outro lado, no estudo crítico avaliamos as entidades segundo uma escala de probabilidade de possível a muito provável, critério não aplicado neste trabalho e que pode ser desenvolvido para a ampliação da base de dados do SIG, usando a escala como indicativo da certeza na atribuição das coordenadas.

Dentro da linha de automatização do processo de georreferenciação, a captura de relações é um aspecto que ficou ilustrado, mas menos desenvolvido. Trabalho futuro consiste no avanço na automatização da captura de relações para o conjunto das entidades, procurando soluções para os casos de frequências baixas, avaliando as possibilidades de melhoria pelo incremento de PLN (maior relevância dos preditores) e técnicas de corpus para o aumento relevância estatística das ocorrências.

O modelo concetual integra relações e entidades de maneira que facilita a sua integração numa ontologia. Ainda que foi apontado, e resulta factível com a ontologia elaborada, ficou por

desenvolver a inferência automática da georreferência a partir da relação de meronímia. Deste modo desenvolvemos a definição de modo automático, substituindo a entidade mais próxima na relação por aquela, primeira das imediatamente superiores, cuja georreferência seja conhecida por conhecimento prévio. Asseguramos assim que a definição venha sempre dada relativamente a umas coordenadas geográficas.

10.5 Conclusões sobre os contributos da tese

Sobre a hipótese principal

Consideramos um modelo de georreferenciação que opera quer com entidades geográficas conhecidas, quer com aquelas que, nos métodos mais convencionais, seriam deixadas como não georreferenciáveis. A partir da definição intensiva da entidade (face ao carácter extensivo da definição por coordenadas, aproveitadas para servirem de ponto de referência) georreferenciamos de modo relativo 100% das entidades extraídas no índice.

Exploramos também a operatividade do modelo concetual para trabalhar com métodos de mineração aplicáveis a grandes volumes de dados. Mesmo que a proposta avaliada neste trabalho contempla apenas duas relações, ela supõe um primeiro passo para a resolução do referente por uma via alternativa à ostensão, baseada na inferência a partir de processos dedutivos apriorísticos combinados com dados induzidos pela análise de corpus, um avanço na direção da Inteligência Artificial.

A caracterização epistemológica das entidades geográficas mencionadas contribui também a organizar e visualizar os resultados. Quando as entidades cujas coordenadas conhecemos previamente foram processadas num SIG, da representação cartográfica emerge uma distribuição espacial característica que debuxa os perfis costeiros da Ásia e parte da Insulíndia. A ordenação das entidades numa ontologia produz um índice ordenado, em que as entidades se relacionam coerentemente segundo as relações de hiponíma e meronímia propostas.

Sobre a hipótese secundária

Avaliamos métodos e sistemas de georreferenciação para um caso pouco estudado, o de um texto histórico, escrito numa variedade distinta ao padrão contemporâneo. Conseguimos resultados que, como mínimo, se aproximam e em ocasiões mesmo superam aqueles obtidos para textos atuais. Nos trabalhos de identificação e georreferenciação, o auxílio do PLN na operação com o corpus facilitou a análise e extração de dados que produziram como resultado um índice de entidades.

Demonstramos que a aplicação de variáveis quantitativas produz resultados significativos, no entanto, o critério de seleção das variáveis aparece como relevante e a elaboração do corpus por métodos de PLN (em que a variável principal é mais comummente qualitativa) proporcionou os maiores incrementos nos níveis de desempenho. Contributo secundário desta tese é melhorar na compreensão dos métodos quantitativos aplicados em trabalhos de análise de corpus e PLN.

Sobre os resultados materiais

A elaboração do corpus e material de trabalho base desta tese requereu um intenso estudo e anotação manual que decorreu durante mais de média década. Os produtos finais podem ser usados como padrões para o teste de métodos que simplifiquem estes trabalhos e avançar assim na sua automatização.

Usamos como caso prático um texto de grande relevo histórico e geográfico. Mediante a elaboração dos corpora, bases de dados e índice, pensamos ter contribuído para a sua compreensão. Disponibilizamos materiais de apoio para futuras pesquisas num âmbito global.

EXPERIMENTOS

Apêndice I

Anotação automática de um corpus reduzido (Tartária 1653)

Corpus

Texto em inglês (Tartária 1653) de um corpus paralelo da *Peregrinação* de Fernão Mendes Pinto.

Unidades processadas	Tokens	Formas únicas
Lexicais	18975	2729
Georreferências anotadas	270	147
Topónimos anotados	203	121
Gentílicos anotados	67	26

Tabela I.1: Características do corpus anotado *Tartaria (1653)*

Configuração dos sistemas NERC

Para cada sistema usamos as configurações predefinidas com a instalação. No caso do Stanford NER, três classificadores instalados nas duas modalidades oferecidas: sensível a maiúsculas e sem considerar maiúsculas e minúsculas, em total 6 configurações usando as opções predefinidas pelo sistema. Para Linguakit e o Edinburgh Geoparser, 3 configurações para cada um, 6 em total para os sistemas de regras: com a lista predefinida pelo sistema, com as listas alargadas com todas as variantes da *Peregrinação* (1614) e mais uma outra de variantes sem os gentílicos.

Combinatória da configuração do corpus com os sistemas

O corpus tem duas modalidades: anotação só de topónimos e uma outra de topónimos e gentílicos.

Os sistemas têm 12 configurações segundo amostrado na figura I.1.

Deste modo obtemos 24 avaliações ao comparar os resultados das 12 configurações com as duas modalidades do padrão.

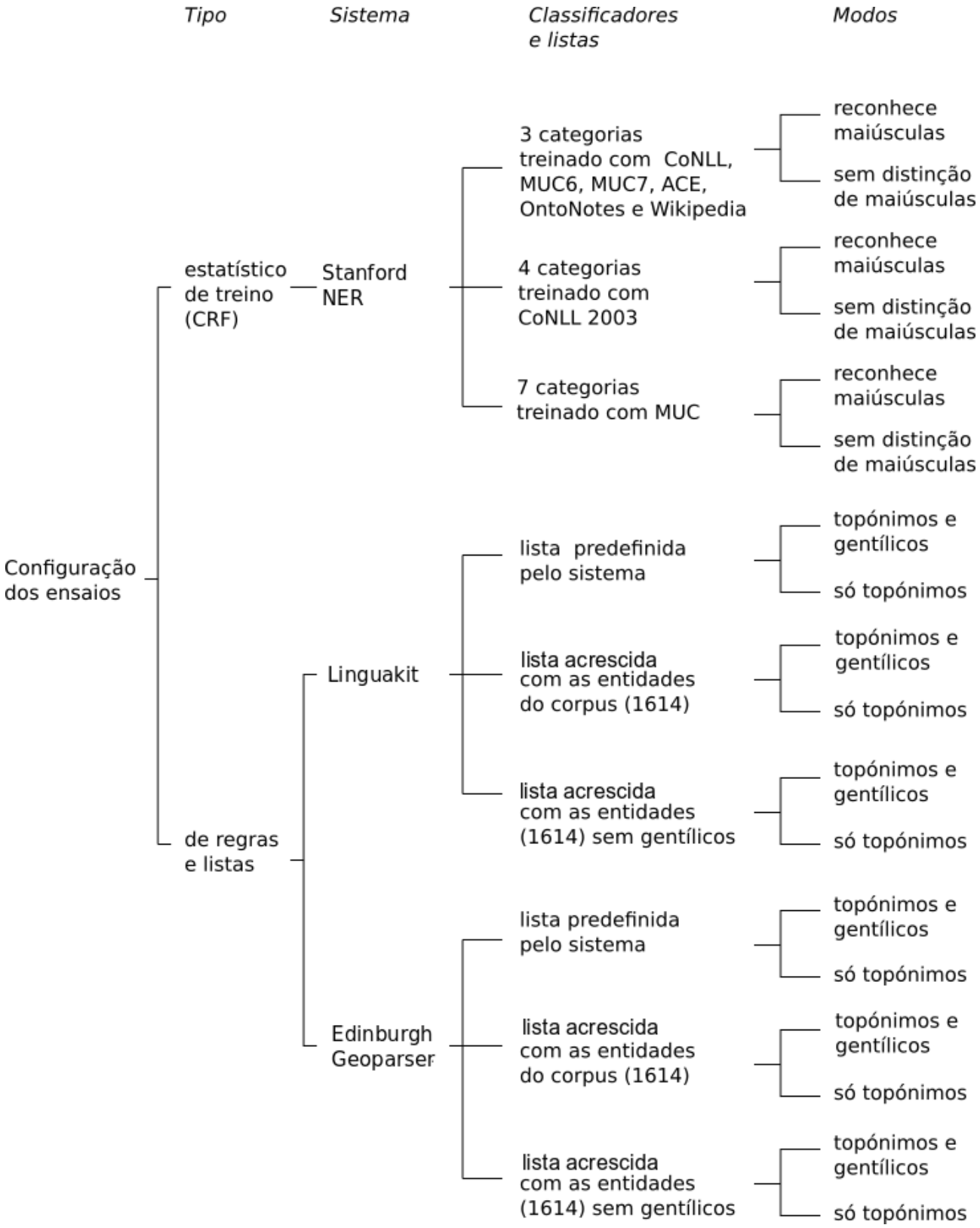


Figura I.1: Diagrama com o código dos ensaios para a anotação de um corpus com sistemas NERC

Resultados dos ensaios NERC

Sistema	Ocor_NERC	unicas_NERC	pos_plenos	pos_parciais	positivos	fal_pos	fal_negativ	Corpus	Tipo_lista	anot_padrão	Precisão	Abrangência	Medida-F
Stanford_3M	166	111	126	7	133	33	137	G	3class	270	80,12	49,26	61,01
Stanford_3M	166	111	120	7	127	39	76	T	3class	203	76,51	62,56	68,84
Stanford_4M	247	140	105	10	115	132	155	G	CoNLL	270	46,56	42,59	44,49
Stanford_4M	247	140	88	10	98	149	105	T	CoNLL	203	39,68	48,28	43,56
Stanford_7M	119	78	74	6	80	39	190	G	MUC	270	67,23	29,63	41,13
Stanford_7M	119	78	69	6	75	44	128	T	MUC	203	63,03	36,95	46,59
Stanford_3N	583	322	185	4	189	394	80	G	3class_nocase	270	32,42	70	44,32
Stanford_3N	583	322	149	3	152	431	51	T	3class_nocase	203	26,07	74,88	36,68
Stanford_4N	467	258	135	21	156	311	113	G	CoNLL_nocase	270	33,4	57,78	42,33
Stanford_4N	467	258	117	19	136	331	67	T	CoNLL_nocase	203	29,12	67	40,6
Stanford_7N	539	299	204	2	206	333	64	G	MUC_nocase	270	38,22	76,3	50,93
Stanford_7N	539	299	161	2	163	376	40	T	MUC_nocase	203	30,24	80,3	43,93
Linguakit_D	249	102	41	22	63	186	207	G	D	270	25,3	23,33	24,28
Linguakit_D	249	102	30	21	51	198	152	T	D	203	20,48	25,12	22,56
Linguakit_P	149	96	73	22	95	54	175	G	P	270	63,76	35,19	45,35
Linguakit_P	149	96	59	21	80	69	123	T	P	203	53,69	39,41	45,45
Linguakit_TP	140	88	65	22	87	53	183	G	TP	270	62,14	32,22	42,44
CitiusTools_T	140	88	59	21	80	60	123	T	TP	203	57,14	39,41	46,65
Geoparser_D	85	54	37	17	54	31	216	G	D	270	65,53	20	30,42
Geoparser_D	85	54	29	17	46	39	157	T	D	203	54,12	22,66	31,94
Geoparser_P	171	115	120	17	137	35	133	G	D+P	270	80,12	50,74	62,13
Geoparser_P	171	115	100	17	117	54	86	T	D+P	203	68,42	57,64	62,57
Geoparser_TF	153	100	103	17	120	33	150	G	D+TP	270	78,43	44,44	56,73
Geoparser_TF	153	100	95	17	112	41	91	T	D+TP	203	73,2	55,17	62,92

Tabela I.2: Resultados dos ensaios para a anotação de um corpus com sistemas NERC.

Exemplo de resultados

Listas usadas num ensaio NERC: Stanford_3M

Ocorrências das entidades mencionadas anotadas no padrão (270)

Tartar, Quincay, Nixiamcoo, Portugals, Tartary, Pequin, Luançama, Famstir, Mecuy, Batampina, China, Nanquin, Tartar, Malincataran, Quinçay, Laura, Carncha, Tartaria, Petilau Nameioo, Tartars, Pequin, Chineses, Nixianicoo, Luançama, Chinese, Quinçay, Tartar, Chineses, Tartar, Chineses, Tartar, Tartar, Ainan, Tartars, Tartars, Tartar, Nixiamcoo, Lancama, Pequin, Tartars, Pequin, Chineses, Tartars, Portugals, Japan, Chineses, Chineses, Mogores, Tartars, Chineses, Tartars, Portugals, Nixiamcoo, Tartary, Pequin, Pequin, Lantimay, Quinçay, Nixiamcoo, Pommitay, Pequin, Palamxitan, Tartar, Lautir, Tartar, Mixiancoo, Lechuna, China, Chineses, Turkish, Nixiamcoo, Mogores, Persians, Bordies, Calaminhams, Bramaas, Tartars, Tartary, Angicamoy, Xipaton, China, China, Europe, Angicamoy, Tartars, Tartaria, Pafua, Mecuy, Capinper, China, Portugal, Pequin, Tartaria, Pequin, Tartaria, Tartaria, Pequim, Chineses, Tartaria, Pequin, Quaytragun, Guiiampea, Liampeu, Guauxitim, Caixiloo, Mogors, Cauchins, Champaas, Chineses, Singrachirau, China, Tartaria, Panquinor, Psipator, Lançame, Lançame, Lançame, Tuymicoa, Persia, Gueos, Bramaa, Tanguu, Calaminham, Odiaa, Siam, Tanauserin, Champaa, Malayos, Berdios, Patanes, Passioloqua, Capioper, Chiammay, Lauhos, Gueos, Mogor, Corazones, Persiu, Dely, Chitor, Caran, Goncalidau, Moscovites, Flemings, Europe, Caran, Caran, Tartar, China, Tartaria, Tartaria, Cauchinchina, Caran, Tartar, Pequin, Pontiveu, Cauchenchina, Pequin, Pequim, Uzamguee, Cauchenchina, Cauchenchina,

Tuymican, Guatipanior, Almains, Muscovy, Gaytor, Denmark, Tartaria, Enxcau, Singuafatur, Tartar, Fanius, Quanginau, Echuna, China, Cauchinchina, Portugals, Portugal, Portugal, Quanginau, Lechuna, Rome, Tartaria, Pafua, Rendacalem, Tartaria, Xinaleygrau, Voulem, Catencur, Singapamor, Cunebetea, Singapamor, Ventrau, Sornau, Siam, Chiamtabuu, Jangumaa, Chiammay, Laos, Gueos, Danbambur, Martabano, Pegu, Pamphileu, Capimper, Sacotay, Monginoco, Meleytay, Sovady, Cosmim, Arracan, Ganges, Bengala, Caleyputa, Tarem, Cauchin, Xolor, China, Cauchin, Xolor, Ventinau, Manaquileu, Chomay, China, Cauchenchina, Quinancaxi, Tinocouhos, Fanaugrem, Usamguee, Cauchim, Usamguee, Fanaugrem, Taraudachit, Lindau Panoo, Cauchenchina, Fanaugrem, Latiparau, Agimpur, Fanaugrem, Tartarian, Cauchenchina, Uzemguee, Tartaria, Agimpur, Tartaria, Tartar, Tartaria, Tartaria, Uzamguee, Tartar, Cauchin, Uzamguee, Uzamguee, Tartar, Tartaria, Fanaugrem, Benau, Pamgatur, Lingator, Baguetor, Natibasoy, Uzamguee, Tinocouhos, Cauchin, Tartar, China, Portugal, Malaca, Indiaes.

Tipos únicos das entidades mencionadas únicas anotadas no padrão (147)

Tartar, Quincay, Nixiamcoo, Portugals, Tartary, Pequin, Luançama, Famstir, Mecuy, Batampina, China, Nanquin, Malincataran, Quinçay, Laura, Carncha, Tartaria, Petilau Nameioo, Tartars, Chineses, Nixianicoo, Chinese, Ainan, Lancama, Japan, Mogores, Lantimay, Pommitay, Palamxitan, Lautir, Mixiancoo, Lechuna, Turkish, Persians, Bordies, Calaminhams, Bramaas, Angicamoy, Xipaton, Europe, Pafua, Capinper, Portugal, Pequim, Quaytragun, Guiiampea, Liampeu, Guauxitim, Caixiloo, Mogors, Cauchins, Champaas, Singrachirau, Panquinor, Psipator, Lançame, Tuymicoa, Persia, Gueos, Bramaa, Tanguu, Calaminham, Odiaa, Siam, Tanauserin, Champaa, Malayos, Berdios, Patanes, Passioloqua, Capioper, Chiammay, Lauhos, Mogor, Corazones, Persiu, Dely, Chitor, Caran, Goncalidau, Moscovites, Flemings, Cauchinchina, Pontiveu, Cauchenchina, Uzamguee, Tuymican, Guatipanior, Almains, Muscovy, Gaytor, Denmark, Enxcau, Singuafatur, Fanius, Quanginau, Echuna, Rome, Rendacalem, Xinaleygrau, Voulem, Catencur, Singapamor, Cunebetea, Ventrau, Sornau, Chiamtabuu, Jangumaa, Laos, Danbambur, Martabano, Pegu, Pamphileu, Capimper, Sacotay, Monginoco, Meleytay, Sovady, Cosmim, Arracan, Ganges, Bengala, Caleyputa, Tarem, Cauchin, Xolor, Ventinau, Manaquileu, Chomay, Quinancaxi, Tinocouhos, Fanaugrem, Usamguee, Cauchim, Taraudachit, Lindau Panoo, Latiparau, Agimpur, Tartarian, Uzemguee, Benau, Pamgatur, Lingator, Baguetor, Natibasoy, Malaca, Indiaes.

Anotações processadas com o NERC (166)

Quincay, Nixiamcoo, Famstir, Batampina, China, Nanquin, Quinçay, Tartaria, Pagode, Petilau Nameioo, Luançama, Ainan, Tileymay, Councel, Lancama, Spades, Vantguard, Bavins, Spades, Japan, Nixiamcoo, Quinçay, Palamxitan, Lautir, Mitaquer, Mixiancoo, Kingdom, Lechuna, China, Nixiamcoo, Bordies, Bramaas, Xipaton, China, China, Europe, Champhire, Angicamoy, Tartaria, Pafua, Mecuy, Capinper, Anchesacotay, China, Portugal, Pucau, Tartaria, Tartaria, Pequim, Tartaria, Liampeu, Caixiloo, Champaas, Singrachirau, China, Tartaria, Panquinor, Lançame, Tuymicoa,

Persia, Gueos, Bramaa, Calaminham, Earth, Odiaa, Siam, Tanauserin, Corazones, Goncalidau, Europe, Arras, Christendom, Sea of China, Tartaria, Tartaria, Kingdom, Cauchinchina, Pontiveu, Mitaquer, Pequim, Uzamguee, Cauchenchina, Pagode, Falconets, Almains, Denmark, Tartaria, Singuafatur, Migama, Fanius, Chappels, Echuna, China, Cauchinchina, Pagode, Pontimaqueu, Vanguenarau, Portugal, Pagod, Portugal, Quanginau, Lechuna, Rome, Tartaria, Chappels, Rendacalem, Kingdom of Tartaria, Xinaleygrau, Pirot, East, East, Lake of Singapamor, Sornau, Siam, Chiamtabuu, Kingdom of Chiammay, Laos, Danbambur, Martabano, Kingdom of Pegu, Capimper, Meleytay, Cosmim, Arracan, Ganges, Kingdom of Bengala, Lake, Xolor, China, Cauchin, Xolor, Ventinau, Manaquileu, Chomay, China, Cauchenchina, Fanaugrem, Usamguee, Cauchim, Usamguee, Fanaugrem, Cauchenchina, Fanaugrem, Latiparau, Fanaugrem, Cauchenchina, Uzemguee, Tartaria, Tartaria, Broquem, Xinarau of Tartaria, Tartaria, Uzamguee, Earth, Uzamguee, Uzamguee, Tartaria, Fanaugrem, Lingator, Baguetor, Laulees, Uzamguee, Cauchin, China, Portugal, Malaca.

Lista completa de positivos reconhecidos (133)

Quincay, Nixiamcoo, Famstir, Batampina, China, Nanquin, Quinçay, Tartaria, Petilau Nameioo, Luançama, Ainan, Lancama, Japan, Nixiamcoo, Quinçay, Palamxitan, Lautir, Mixiancoo, Lechuna, China, Nixiamcoo, Bordies, Bramaas, Xipaton, China, China, Europe, Angicamoy, Tartaria, Pafua, Mecuy, Capinper, China, Portugal, Tartaria, Tartaria, Pequim, Tartaria, Liampeu, Caixiloo, Champaas, Singrachirau, China, Tartaria, Panquinor, Lançame, Tuymicoa, Persia, Gueos, Bramaa, Calaminham, Odiaa, Siam, Tanauserin, Corazones, Goncalidau, Europe, Sea of China, Tartaria, Tartaria, Cauchinchina, Pontiveu, Pequim, Uzamguee, Cauchenchina, Almains, Denmark, Tartaria, Singuafatur, Fanius, Echuna, China, Cauchinchina, Portugal, Portugal, Quanginau, Lechuna, Rome, Tartaria, Rendacalem, Kingdom of Tartaria, Xinaleygrau, Lake of Singapamor, Sornau, Siam, Chiamtabuu, Kingdom of Chiammay, Laos, Danbambur, Martabano, Kingdom of Pegu, Capimper, Meleytay, Cosmim, Arracan, Ganges, Kingdom of Bengala, Xolor, China, Cauchin, Xolor, Ventinau, Manaquileu, Chomay, China, Cauchenchina, Fanaugrem, Usamguee, Cauchim, Usamguee, Fanaugrem, Cauchenchina, Fanaugrem, Latiparau, Fanaugrem, Cauchenchina, Uzemguee, Tartaria, Tartaria, Xinarau of Tartaria, Tartaria, Uzamguee, Uzamguee, Uzamguee, Tartaria, Fanaugrem, Lingator, Baguetor, Uzamguee, Cauchin, China, Portugal, Malaca.

Positivos plenos reconhecidos (126)

Quincay, Nixiamcoo, Famstir, Batampina, China, Nanquin, Quinçay, Tartaria, Petilau Nameioo, Luançama, Ainan, Lancama, Japan, Nixiamcoo, Quinçay, Palamxitan, Lautir, Mixiancoo, Lechuna, China, Nixiamcoo, Bordies, Bramaas, Xipaton, China, China, Europe, Angicamoy, Tartaria, Pafua, Mecuy, Capinper, China, Portugal, Tartaria, Tartaria, Pequim, Tartaria, Liampeu, Caixiloo, Champaas, Singrachirau, China, Tartaria, Panquinor, Lançame, Tuymicoa, Persia, Gueos, Bramaa, Calaminham, Odiaa, Siam, Tanauserin, Corazones, Goncalidau, Europe, Tartaria, Tartaria,

Cauchinchina, Pontiveu, Pequim, Uzamguee, Cauchenchina, Almains, Denmark, Tartaria, Singuafatur, Fanius, Echuna, China, Cauchinchina, Portugal, Portugal, Quanginau, Lechuna, Rome, Tartaria, Rendacalem, Xinaleygrau, Sornau, Siam, Chiamtabuu, Laos, Danbambur, Martabano, Capimper, Meleytay, Cosmim, Arracan, Ganges, Xolor, China, Cauchin, Xolor, Ventinau, Manaquileu, Chomay, China, Cauchenchina, Fanaugrem, Usamguee, Cauchim, Usamguee, Fanaugrem, Cauchenchina, Fanaugrem, Latiparau, Fanaugrem, Cauchenchina, Uzemguee, Tartaria, Tartaria, Tartaria, Uzamguee, Uzamguee, Uzamguee, Tartaria, Fanaugrem, Lingator, Baguetor, Uzamguee, Cauchin, China, Portugal, Malaca.

Positivos parciais reconhecidos (7)

Sea of China, Kingdom of Tartaria, Lake of Singapamor, Kingdom of Chiammay, Kingdom of Pegu, Kingdom of Bengala, Xinarau of Tartaria.

Falsos positivos (33)

Pagode, Tileymay, Councel, Spades, Vantguard, Bavins, Spades, Mitaquer, Kingdom, Champhire, Anchesacotay, Pucau, Earth, Arras, Christendom, Kingdom, Mitaquer, Pagode, Falconets, Migama, Chappels, Pagode, Pontimaqueu, Vanguenarau, Pagod, Chappels, Pirot, East, East, Lake, Broquem, Earth, Laulees.

Tipos com ao menos uma ocorrência falso negativo (81)

Tartar, Nixiamcoo, Portugals, Tartary, Pequin, Luançama, Mecuy, China, Malincataran, Quinçay, Laura, Carncha, Tartaria, Tartars, Chineses, Nixianicoo, Chinese, Mogores, Lantimay, Pommitay, Turkish, Persians, Calaminhams, Angicamoy, Pafua, Quaytragun, Guiiampea, Guauxitim, Mogors, Cauchins, Psipator, Lançame, Gueos, Tanguu, Champaa, Malayos, Berdios, Patanes, Passioloqua, Capioper, Chiammay, Lauhos, Mogor, Persiu, Dely, Chitor, Caran, Moscovites, Flemings, Cauchenchina, Tuymican, Guatipanior, Muscovy, Gaytor, Enxcau, Quanginau, Voulem, Catencur, Singapamor, Cunebetea, Ventrau, Jangumaa, Pegu, Pamphileu, Sacotay, Monginoco, Sovady, Bengala, Caleyputa, Tarem, Cauchin, Quinancaxi, Tinocouhos, Taraudachit, Lindau Panoo, Agimpur, Tartarian, Benau, Pamgatur, Natibasoy, Indiaes.

Tipos com todas as ocorrências falso negativo (68)

Tartar, Portugals, Tartary, Pequin, Malincataran, Laura, Carncha, Tartars, Chineses, Nixianicoo, Chinese, Mogores, Lantimay, Pommitay, Turkish, Persians, Calaminhams, Quaytragun, Guiiampea, Guauxitim, Mogors, Cauchins, Psipator, Tanguu, Champaa, Malayos, Berdios, Patanes, Passioloqua, Capioper, Chiammay, Lauhos, Mogor, Persiu, Dely, Chitor, Caran, Moscovites,

Flemings, Tuymican, Guatipanior, Muscovy, Gaytor, Enxcau, Voulem, Catencur, Singapamor, Cunebetea, Ventrau, Jangumaa, Pegu, Pamphileu, Sacotay, Monginoco, Sovady, Bengala, Caleyputa, Tarem, Quinancaxi, Tinocouhos, Taraudachit, Lindau Panoo, Agimpur, Tartarian, Benau, Pamgatur, Natibasoy, Indiaes.

Apêndice II

Elaboração de uma taxonomia para a classificação das entidades geográficas mencionadas do corpus

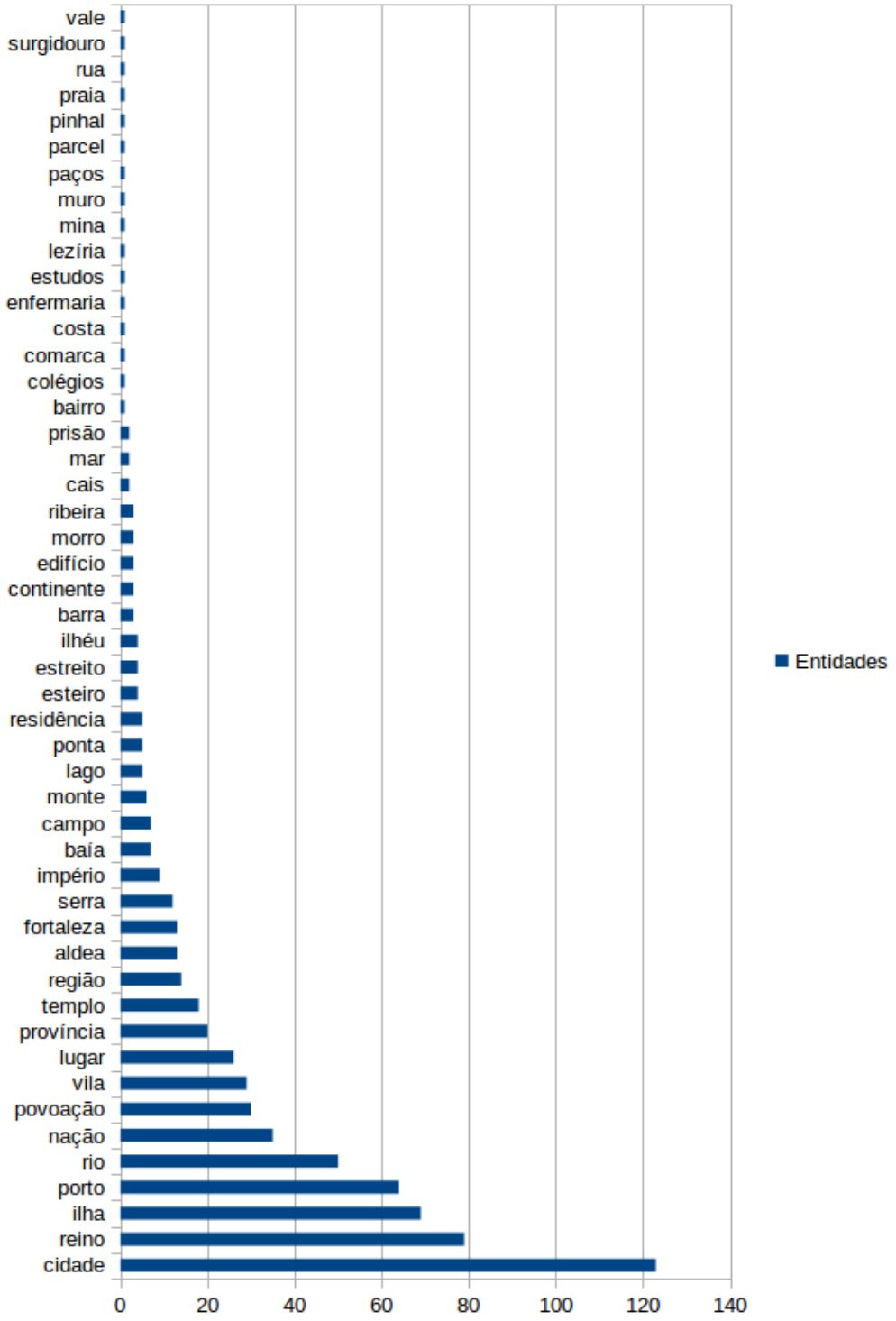


Figura II.1: Lista final após a aplicação de regras e redução de tipos por similitude morfológica e semântica

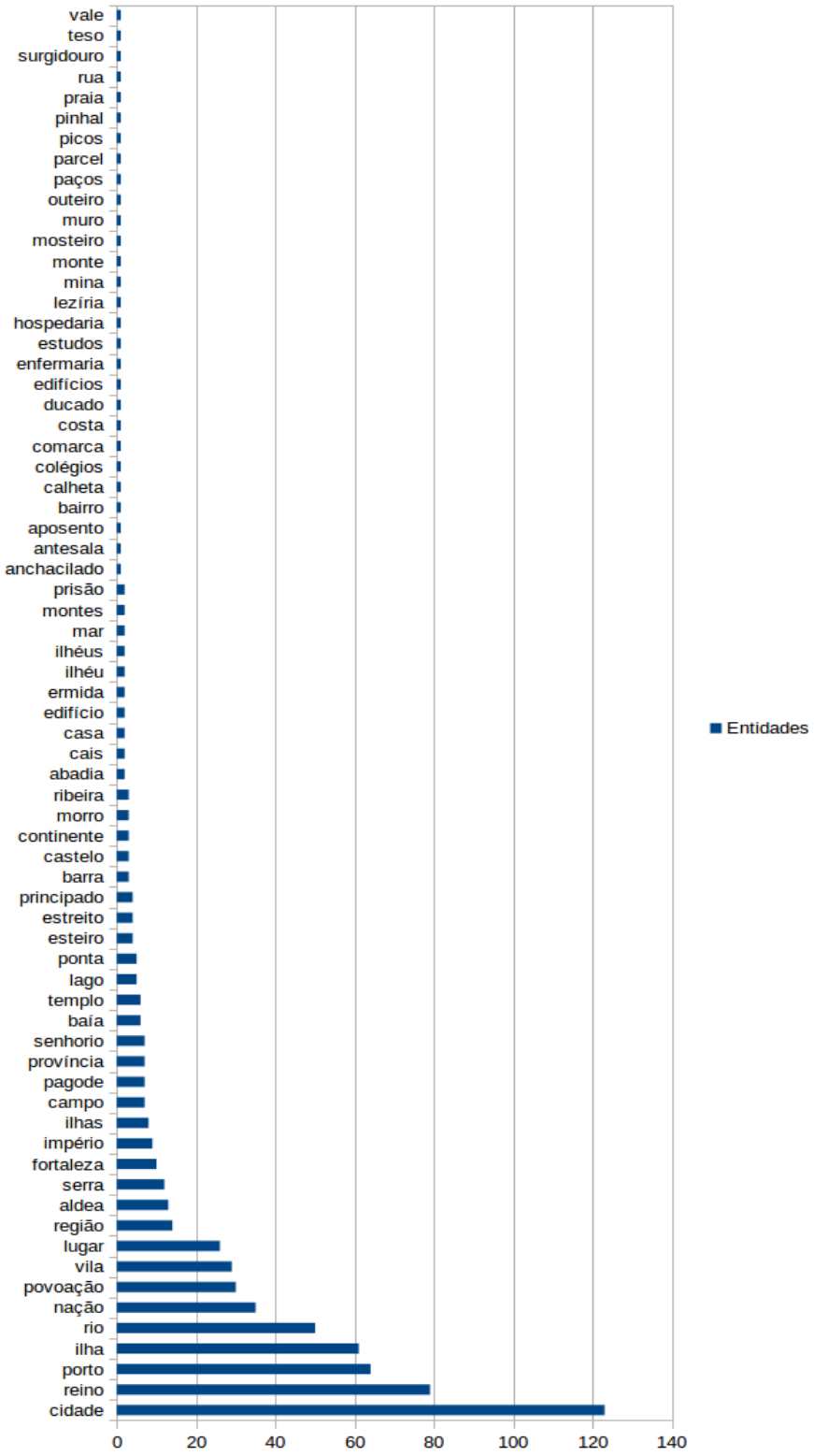


Figura II.2: Lista anterior à redução de tipos

Apêndice III

Testes de extração de termos de domínio geográfico por elaboração do corpus, métricas e filtros

Testes de avaliação dos efeitos do modo do corpus e métodos de filtrado de candidatos para a extração de termos do domínio geográfico.

Esquemas TF-IDF

A formulação da medida TF-IDF contempla variantes na consideração das ocorrências do termo, dos documentos e normalização (Mannig & Schütze, 1999). Para os testes aplicamos variantes que consideram a variação na normalização e base logarítmica. Usamos inicialmente uma formulação que permite variações segundo o sistema SMART (Singhal, Salton & Buckley, 1996):

$$tf-idf(i,j)=\begin{cases}(1+\log_{10}(tf_{i,j}))\dfrac{N}{df_i} & \text{quando } tf_{i,j}\geq 1\\ 0 & \text{quando } tf_{i,j}=0\end{cases} \quad (1)$$

em que $tf_{i,j}$ é a frequência do termo t_i no documento d_j, N o número de segmentos do corpus e df_i o número de segmentos do corpus em que t_i ocorre. Sobre este esquema, aplicamos em dois testes a normalização pelo cosseno (Singhal, Salton & Buckley, 1996):

$$\sqrt{(1+\log(tf_1))^2+(1+\log(tf_2))^2+...+(1+\log(tf_t))^2}$$

em que t é o número de termos no documento.

Usamos mais uma variante que contempla a normalização da frequência pelo uso da frequência relativa:

$$tf-idf(i,j)=\begin{cases}(fr_{i,j})\log_2\dfrac{N}{df_i} & \text{quando } tf_{i,j}\geq 1\\ 0 & \text{quando } tf_{i,j}=0\end{cases} \quad (2)$$

em que $fr_{i,j}$ é a frequência relativa do termo t_i no documento d_j, N o número de segmentos do corpus e df_i o número de segmentos do corpus em que t_i ocorre.

Implementação

Aplicamos os esquemas TF-IDF usando a base do software estatístico R (R Core Team, 2016). Contrastamos os resultados com os oferecidos pelo pacote TM (Feinerer, Hornik, & Meyer, 2008); as configurações T1TF_0, T10RF_LOG2IDF_0 e T14_1LOG2TF_LOG2IDF_COS são obtidas diretamente pela aplicação deste pacote; as restantes, com *scripts* de elaboração própria com base R.

Modos do corpus

DATA2. Orações normalizadas com, no mínimo, uma entidade geográfica mencionada com categoria de topónimo representada por um tipo único.

DATA3. Orações normalizadas com, no mínimo, uma entidade geográfica mencionada que contém unicamente os nomes comuns.

DATA4. Unidade menor à oração (cláusula ou frase) com os nomes comuns que coocorrem com um topónimo.

DATA5. Vetor com a lista dos nomes comuns mais próximos à entidade geográfica mencionada.

Listas filtro

1. Não se aplica nenhuma lista filtro sobre os dados. Não se consideram os termos com menos de 3 caracteres, isto é, longitude(termo)>2.

2. Palavras comuns e correcção do processamento de normalização para o padrão contemporâneo e anotação morfossintática:

agora, ahi, ainda, algum, alguma, algumas, alguns, ali, ambos, antes, aos, aquela, aqui, assaz, assim, até, atrás, auendo, avia, bem, boa, bom, cada, cem, cem, cento, cento, chamava, chegamos, chegou, cinco, cinco, cinquenta, cinquenta, com, como, Como, comsigo, cos, daí, dali, daqui, dar, das, del, dela, dele, deles, dentro, deram, deram, despois, desta, deste, deu, dez, dez, dezassete, dezassete, dezia, dezoito, dezoito , diante, disse, dizer, dois, donde, dos, dous, dous, doze, doze, duas, duas, duzentos, duzentos, ela, ele, eles, então, entre, era, eram, esta, estava, estavam, este, estes, faria, faria, fazer, fez, fizera, foi, foram, forão, havia, huns, isso, isto, lhe, lhes, logo, maior, mais, mas, meio, menos, meu, mil, mil, mim, minha, muita, muitas, muito, muitos, não, nas, nela, nele, nem, nenhum, nenhuma, nesta, neste, nos, nòs, nossa, nosso, noue, oitenta, oitenta, oito, oito, onde, outra, outras, outras, outro, outros, para, partio, pela, pelo, pero, por, porem, porque, pouco, primeiro, primeiro, quais, qual, quando, quanto, quarenta, quarenta, quasi, quatorze, quatorze, quatro, quatro, quatrocentos, quatrocentos, que, quem, quinze, quinze, saira, são, segundo, segundo, seis, seis, sem, sempre, sendo, ser, sessenta, sessenta, sete, sete, setenta, setenta, setima, setima, seu, seus, sobre, sua, suas, tambem, também, tanto, tão, tem, ter, tinha, tinham, tinhão, toda, todas, todo, todos, tres,

tres, treze, treze, trezentas, trezentas, trinta, trinta, tudo, uma, vendo, ver, vieram, vinha, vinte, vinte, vos.

Configuração dos esquemas

Nome do esquema	**TF-IDF**	**Modos do corpus**	**Filtros**
T1TF_0	TF	DATA2, DATA3, DATA4, DATA5	1. Longitude (termo)>2 2. Lista de palavras comuns e correcção do processamento, longitude (termo)>2
T10RF_LOG2IDF_0	(2)	DATA2, DATA3, DATA4	1. Longitude (termo)>2 2. lista de palavras comuns e correcção do processamento, longitude (termo)>2
T11RF_LOGIDF_1	(2) ($\log_e$)	DATA2, DATA3, DATA4	1. Longitude (termo)>2 2. lista de palavras comuns e correcção do processamento, longitude (termo)>2
T13_1LOG10TF_LOG10IDF	(1)	DATA2, DATA3, DATA4	1. Longitude (termo)>2 2. lista de palavras comuns e correcção do processamento, longitude (termos)>2
T13_1LOG10TF_LOG10IDF_COS	(1) normalizada pelo cosseno	DATA2, DATA3, DATA4	1. Longitude (termo)>2 2. lista de palavras comuns e correcção do processamento, longitude (termo)>2
T14_1LOG2TF_LOG2IDF_COS	(1) $\log_2$ e normalizada pelo cosseno	DATA2, DATA3, DATA4	1. Longitude (termo)>2 2. lista de palavras comuns e correcção do processamento, longitude (termo)>2

Tabela III.1: Esquemas das configurações para o ensaio de recuperação de termos de domínio.

Rondas

R1 Todos os esquemas sobre DATA2 sem lista filtro nenhuma.

R2 Todos os esquemas sobre DATA2 com lista filtro.

R3 Todos os esquemas sobre DATA3 sem lista filtro nenhuma.

R4 Todos os esquemas sobre DATA3 com lista filtro.

R5 Todos os esquemas sobre DATA4 sem lista filtro nenhuma.

R6 Todos os esquemas sobre DATA4 com lista filtro.

R7 T1TF_0 sobre DATA5 sem lista filtro nenhuma.

R8 T1TF_0 sobre DATA5 com lista filtro.

Listas de candidatos recuperados por rondas

R1

antonio, assim, até, capitão, cidade, com, como, das, despois, desta, deste, dia, dias, dos, ele, então, era, esta, estava, este, faria, fez, foi, fora, gente, grande, havia, ilha, lhe, logo, mais, mil, muito, não, nos, onde, padre, para, passei, passou, pelo, por, porque, quais, qual, que, rei, reino, rio, sem, ser, seu, seus, sua, tão, tempo, terra, tinha, toda, todo, todos, tres, uma.

R2

aconteceu, ano, anos, antonio, armada, caminho, capitão, casa, causa, chamava, chegamos, chegar, chegarmos, chegou, cidade, costa, cousa, cousas, dar, deu, deus, dezia, dia, dias, disse, embaixador, faria, fazenda, fomos, fora, fortaleza, gente, grande, grandes, homem, homens, ilha, legoas, lugar, mandou, maneira, mar, menham, morte, nao, noite, nome, padre, parte, partimos, partiu, passamos, passei, passou, perdemos, porto, rainha, rei, reino, rio, senhor, socedeu, tempo, terra, verdade, vezes, vimos, vinha.

R3

ano, armada, barra, caminho, capitão, carta, casa, caso, causa, cidade, conta, costa, cousa, despois, dia, dois, dom, duas, embaixador, fazenda, filho, fortaleza, gente, homem, hora, ilha, inimigo, junco, legoas, lugar, maneira, mão, mar, mercador, modo, morte, nao, noite, nome, padre, parte, partes, pessoa, porto, razão, reino, rio, santo, senhor, successo, tempo, terra, trabalho, tres, vela, verdade, vez, viagem, vida, vinte.

R4

ano, armada, barra, caminho, campo, capitão, carta, casa, caso, causa, cidade, conta, costa, cousa, dia, dom, embaixador, embarcação, fazenda, filho, fortaleza, gente, homem, hora, ilha, inimigo, irmão, junco, legoas, lugar, maneira, mão, mar, menham, mercador, modo, morte, nao, noite, nome, padre, parte, partes, pessoa, porto, razão, reino, rio, santo, senhor, successo, tempo, terra, trabalho, vela, verdade, vez, viagem, vida.

R5

ano, armada, barra, caminho, campo, capitão, carta, casa, caso, cidade, costa, cousa, despois, dia, dois, dom, duas, embaixador, enseada, entrada, estreito, fazenda, filho, fim, fortaleza, gente, homem, ilha, imperio, junco, legoas, longo, lugar, maneira, mar, menham, mercador, morte, nao, naos, noite, nome, padre, pagode, parte, partes, partido, ponta, porta, porto, reino, rio, senhor, serra, tempo, terra, tres, vez, viagem, vila, vinte.

R6

ano, armada, barra, caminho, campo, capitão, carta, casa, caso, cidade, companhia, costa, cousa, dia, dom, embaixador, enseada, entrada, estreito, fazenda, filho, fim, fortaleza, gente, homem, ilha, imperio, junco, legoas, longo, lugar, maneira, mar, menham, mercador, mês, morte, nao, naos, noite, nome, padre, pagode, parte, partes, partido, ponta, porta, porto, povoação, reino, rio, senhor, serra, tempo, terra, vez, viagem, vila.

R7

cidade, reino, ilha, porto, rio, nome, lugar, fortaleza, terra, costa, dia, barra, enseada, ano, capitão, tempo, vila, filho, gente, casa, estreito, imperio, mar, partes, dom, estado, morte, naos, parte, povoação, armada, mercador, nao, del, fazenda, padre, senhor, serra, caminho, campo, embaixador, junco, lago, legoas, viagem, cousa, despois, guerra, monte, partido.

R8

cidade, reino, ilha, porto, rio, nome, lugar, fortaleza, terra, costa, dia, barra, enseada, ano, capitão, tempo, vila, filho, gente, casa, estreito, imperio, mar, partes, dom, estado, morte, naos, parte, povoação, armada, mercador, nao, fazenda, padre, senhor, serra, caminho, campo, embaixador, junco, lago, legoas, viagem, cousa, guerra, monte, partido, ponta, castelo.

Verdadeiros positivos totais

Lista de termos avaliados como verdadeiros positivos (28):

barra, caminho, campo, casa, castelo, cidade, costa, enseada, estado, estreito, fazenda, fortaleza, ilha, imperio, lago, lugar, mar, monte, pagode, ponta, porta, porto, povoação, reino, rio, serra, terra, vila.

Resultados verdadeiros positivos por teste

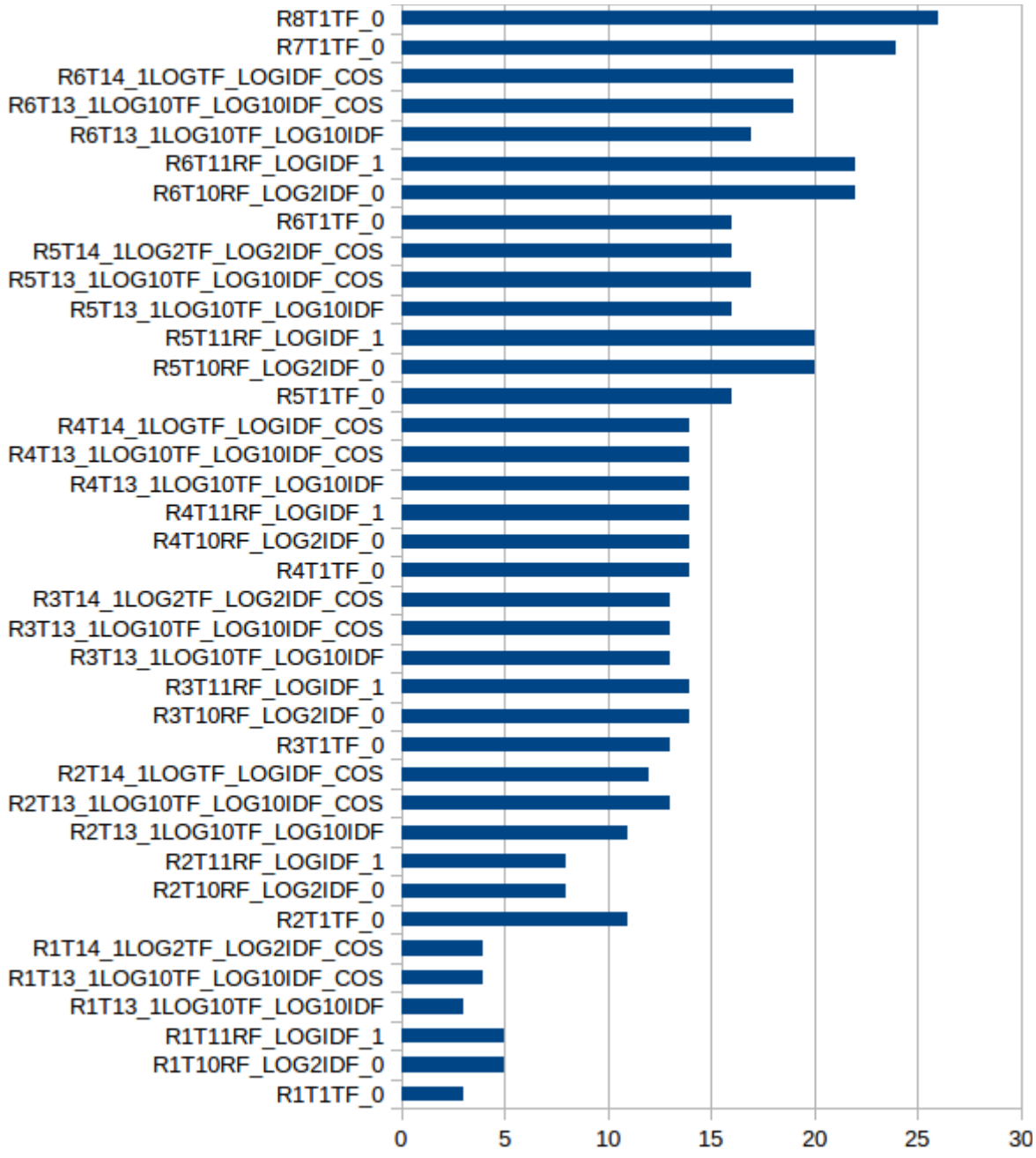


Figura III.1: Verdadeiros positivos recuperados nos testes de extração de termos de domínio geográfico

Apêndice IV

Validação de listas de termos a partir de glossários geográficos

Descrição geral do teste

Desempenho dos glossários geográficos na validação de termos do domínio no corpus.

Dados do corpus

Lista a validar

TR8 (melhor resultado no labor de extração de termos do domínio):

cidade, reino, ilha, porto, rio, nome, lugar, fortaleza, terra, costa, dia, barra, enseada, ano, capitão, tempo, vila, filho, gente, casa, estreito, imperio, mar, partes, dom, estado, morte, naos, parte, povoação, armada, mercador, nao, fazenda, padre, senhor, serra, caminho, campo, embaixador, junco, lago, legoas, viagem, cousa, guerra, monte, partido, ponta, castelo.

Verdadeiros positivos do corpus

Avaliação manual de todos os termos extraídos do conjunto de testes no trabalho de extração de termos de domínio. A abrangência sobre o corpus mede-se sobre esta lista:

cidade, ilha, reino, rio, terra, caminho, casa, costa, fazenda, fortaleza, lugar, mar, porto, barra, campo, enseada, estreito, imperio, pagode, ponta, porta, serra, vila, povoação, estado, lago, monte, castelo.

Características da lista a validar

Avaliação de TR8 a respeito da lista de verdadeiros positivos do corpus:

VERDADEIROS POSITIVOS (26): *cidade, ilha, reino, rio, terra, caminho, casa, costa, fazenda, fortaleza, lugar, mar, porto, barra, campo, enseada, estreito, imperio, ponta, serra, vila, povoação, estado, lago, monte, castelo.*

FALSOS POSITIVOS (24): *nome, dia, ano, capitão, tempo, filho, gente, partes, dom, morte, naos, parte, armada, mercador, nao, padre, senhor, embaixador, junco, legoas, viagem, cousa, guerra, partido.*

FALSOS NEGATIVOS(2): *pagode, porta.*

Desempenho sobre a lista de verdadeiros positivos do corpus

PRECISÃO : 0.52 ABRANGÊNCIA : 0.93 medida-F : 0.67

Glossários de termos geográficos

1. (IBGE, 2015)

Obtido do processamento do pdf do *Glossário dos Termos Genéricos dos Nomes Geográficos Utilizados no Mapeamento Sistemático do Brasil. Vol. 2.* (IBGE, 2015).

127 termos:

aeroclube, aeroporto, aglomerado rural, aglomerado rural de extensão urbana, aglomerado rural isolado, agrovila, aguinha, aldeia, área militar, arroio, atol, baixa, baixo, banhado, barra, barragem, boca, brejo, cabeceira, cabo, cachoeira, cachoeirinha, canal, canaleta, capela, capital, capital federal, castanhal, chapada, cidade, colocação, comunidade, corguinho, corixinho, corixo, corredeira, corregozinho, destacamento, enseada, escola, esgotinho, esgoto, estância, estirão, estrada, estreito, farol, fazenda, ferrovia, furo, garimpo, gleba, granja, grota, igarapezinho, igreja, ilha, ilhas, ilhota, ilhote, ipixuna, ipueira, lagamar, lago, lagoa, lagoinha, laguinho, lajeadinho, lajeado, linha, lugarejo, maloca, mar, marimbu, monte, montes, morraria, morro, morros, nome local, núcleo, oceano, oleoduto, parada, pedra, pico, poliduto, ponta, pontal, ponte, porto, posto indígena, povoado, praia, projeto, ramal, rancho, recife, represa, ressaca, restinga, retiro, riachinho, riacho, rio, riozinho, rocha, rodovia, saco, salto, sanga, sangrador, sangradouro, seringal, serra, serrania, serraria, sítio, terra indígena, terrestres, travessão, usina, vazante, vazantinha, vereda, vila, volta.

2. Lista de atributos geográficos da ontologia de GeoNames

667 termos obtidos da aplicação da página <http://www.GeoNames.org/export/codes.html> sobre o sistema de tradução automática de Google <https://translate.google.com/?hl=pt-PT>.

3. IBGE união GeoNames

725 termos, isto é, os tipos únicos da união de IBGE e GeoNames.

Desempenho dos glossários

1.1 IBGE. Quantos termos são validados como geográficos em TR8?

VERDADEIROS POSITIVOS (14): *barra, cidade, enseada, estreito, fazenda, ilha, lago, mar, monte, ponta, porto, rio, serra, vila.*

FALSOS POSITIVOS (0):

FALSOS NEGATIVOS (12): *caminho, campo, casa, castelo, costa, estado, fortaleza, imperio, lugar, povoação, reino, terra.*

PRECISÃO : 1 ABRANGÊNCIA : 0.54 Medida-F : 0.7

1.2 Avaliação da validação de TR8 pelo glossário IBGE a respeito a lista de verdadeiros positivos do corpus

VERDADEIROS POSITIVOS (14): *barra, cidade, enseada, estreito, fazenda, ilha, lago, mar, monte, ponta, porto, rio, serra, vila.*

FALSOS POSITIVOS (0):

FALSOS NEGATIVOS (14): *caminho, campo, casa, castelo, costa, estado, fortaleza, imperio, lugar, pagode, porta, povoação, reino, terra.*

PRECISÃO : 1 ABRANGÊNCIA : 0.5 Medida-F : 0.67

2.1 GeoNames. Quantos termos são validados como geográficos em TR8?

VERDADEIROS POSITIVOS (11): *barra, casa, castelo, costa, estreito, fazenda, ilha, lago, mar, monte, porto.*

FALSOS POSITIVOS (0):

FALSOS NEGATIVOS(15): *caminho, campo, cidade, enseada, estado, fortaleza, imperio, lugar, ponta, povoação, reino, rio, serra, terra, vila.*

PRECISÃO : 1 ABRANGÊNCIA : 0.42 Medida-F : 0.59

2.2 Avaliação da validação de TR8 pelo glossário GeoNames relativamente à lista de corretos positivos do corpus

VERDADEIROS POSITIVOS (11): *barra, casa, castelo, costa, estreito, fazenda, ilha, lago, mar, monte, porto.*

FALSOS POSITIVOS (0):

FALSOS NEGATIVOS (17): *caminho, campo, cidade, enseada, estado, fortaleza, imperio, lugar, pagode, ponta, porta, povoação, reino, rio, serra, terra, vila.*

PRECISÃO : 1 ABRANGÊNCIA : 0.39 Medida-F : 0.56

3.1 IBGE união GeoNames. Quantos termos são validados como geográficos em TR8?

VERDADEIROS POSITIVOS (17): barra, casa, castelo, cidade, costa, enseada, estreito, fazenda, ilha, lago, mar, monte, ponta, porto, rio, serra, vila

FALSOS POSITIVOS (0):

FALSOS NEGATIVOS (9): *caminho, campo, estado, fortaleza, imperio, lugar, povoação, reino,*

terra.

PRECISÃO : 1 ABRANGÊNCIA : 0.65 Medida-F : 0.79

3.2 Avaliação da validação de TR8 pelo glossário BGE união GeoNames relativamente à lista de corretos positivos do corpus

VERDADEIROS POSITIVOS (17): *barra, casa, castelo, cidade, costa, enseada, estreito, fazenda, ilha, lago, mar, monte, ponta, porto, rio, serra, vila.*

FALSOS POSITIVOS (0):

FALSOS NEGATIVOS (11): *caminho, campo, estado, fortaleza, imperio, lugar, pagode, porta, povoação, reino, terra.*

PRECISÃO : 1 ABRANGÊNCIA : 0.61 Medida-F : 0.76

Apêndice V

Validação de listas de termos a partir de uma base lexical difusa

Descrição geral dos testes

Desempenho da base lexical difusa CLIP2.1 (Gonçalo Oliveira & Gomes, 2014; 2016) na validação de termos do domínio no corpus.

1. Selecionamos uma lista de termos do domínio geográfico como semente.

2. Pesquisamos os synsets do CLIP2.1 que contêm como mínimo um termo da lista semente segundo um valor mínimo de corte.

3. Cada synset selecionado contém todos os termos associados que o compõem. Os termos recuperados são filtrados aplicando uma nova medida de corte. O resultado final é uma lista de termos associados ao domínio dos termos inicializados pela lista semente.

4. A lista obtida é aplicada como medida validadora repetindo o procedimento usado no APÊNDICE IV.

5. O desempenho da lista é avaliado segundo a lista de verdadeiros positivos totais obtida nos testes do APÊNDICE III.

Códigos das listas do domínio geográfico usadas como semente

IBGE: 127 termos (vid. Apêndice IV). Lista de máxima qualidade na pertença dos termos ao domínio, com abrangência média nos tipos.

GeoNames_trad: 667 (vid. Apêndice IV). Lista de grande abrangência nos tipos (ontologia de GeoNames), qualidade limitada pelo efeito da tradução automática.

IBGEGeoNames: 725 termos, IBGE U GeoNames_trad. (vid. Apêndice IV)

TR8IBGEGeoNames: 17 termos resultado do teste TR8 (vid. Apêndice III) validados por IBGEGeoNames (vid. Apêndice IV): *barra, casa, castelo, cidade, costa, enseada, estreito, fazenda, ilha, lago, mar, monte, ponta, porto, rio, serra, vila.*

Vetor de valores de corte sobre as medidas de associação do CLIP2.1

(0,0.01,0.05,0.1,0.15,0.2,0.25,0.5,1,1.5)

Configuração dos testes

Cada teste recebe um nome segundo o esquema:

Nome da lista a validar _

Nome da lista de termos do domínio geográfico _

Valor de corte para a recuperação da lista no CLIP2.1 _

Valor de corte para a seleção de termos das listas recuperadas do CLIP2.1

Ex. TR8_IBGEGeoNames_1_1.5

Valida a lista de candidatos TR8 a partir do glossário de termos geográficos IBGEGeoNames estabelecendo uma medida de associação mínima por cima do 1 para a recuperação de listas no CLIP2.1 e um valor 1.5 para a seleção dos termos associados.

Esquemas

Sobre cada lista de domínio geográfico (4 glossários) aplicamos o vetor de valores de corte (10 elementos) para a seleção de synsets.

Sobre cada resultado de seleção de synsets aplicamos um novo vetor de corte (10 elementos) para a seleção de termos.

Obtemos 4 glossários x 10 rondas de recuperação de synsets x 10 rondas de seleção de termos = 400 esquemas = 400 testes.

Exemplo de validação e avaliação de resultados

Nome do teste: TR8_TR8IBGEGeonames_0.15_0.25

Esquema: TR8IBGEGeoNames_0.15_0.25

Lista de termos a validar: TR8

cidade, reino, ilha, porto, rio, nome, lugar, fortaleza, terra, costa, dia, barra, enseada, ano, capitão, tempo, vila, filho, gente, casa, estreito, imperio, mar, partes, dom, estado, morte, naos, parte, povoação, armada, mercador, nao, fazenda, padre, senhor, serra, caminho, campo, embaixador, junco, lago, legoas, viagem, cousa, guerra, monte, partido, ponta, castelo.

Desempenho de TR8 avaliado com a lista de verdadeiros positivos do corpus

VERDADEIROS POSITIVOS (26): *barra, caminho, campo, casa, castelo, cidade, costa, enseada, estado, estreito, fazenda, fortaleza, ilha, imperio, lago, lugar, mar, monte, ponta, porto, povoação, reino, rio, serra, terra, vila.*

FALSOS POSITIVOS (24): *nome, dia, ano, capitão, tempo, filho, gente, partes, dom, morte, naos, parte, armada, mercador, nao, padre, senhor, embaixador, junco, legoas, viagem, cousa, guerra, partido*.

FALSOS NEGATIVOS(2): *pagode, porta.*

PRECISÃO : 0.52 ABRANGÊNCIA : 0.93 medida-F : 0.67

Lista de termos geográficos de referência para a consulta do CLIP2.1: TR8IBGEGeoNames

cidade, ilha, porto, rio, costa, barra, enseada, vila, casa, estreito, mar, fazenda, serra, lago, monte, ponta, castelo.

Desempenho de TR8IBGEGeoNames sobre TR8:

PRECISÃO : 1 ABRANGÊNCIA : 0.65 medida-F : 0.79

Lista de termos avaliados como positivos do domínio ao aplicar os synsets selecionados em CLIP2.1

Valor da associação para a recuperação de termos do CLIP2.1: > 0.15 .

Valor da associação para a validação da lista de candidatos: > 0.25

barra, campo, casa, castelo, cidade, costa, embaixador, enseada, estreito, fazenda, fortaleza, ilha, lago, lugar, mar, monte, ponta, porto, povoação, rio, serra, terra, vila.

Avaliação da validação

VERDADEIROS POSITIVOS (21): *barra, campo, casa, castelo, cidade, costa, enseada, estreito, fazenda, fortaleza, ilha, lago, lugar, mar, monte, ponta, porto, povoação, rio, serra, terra.*

FALSOS POSITIVOS (0):

FALSOS NEGATIVOS (5): *caminho, estado, imperio, reino, vila.*

PRECISÃO : 1 ABRANGÊNCIA : 0.81 medida-F : 0.9

Comparativa de resultados segundo lista semente: Precisão

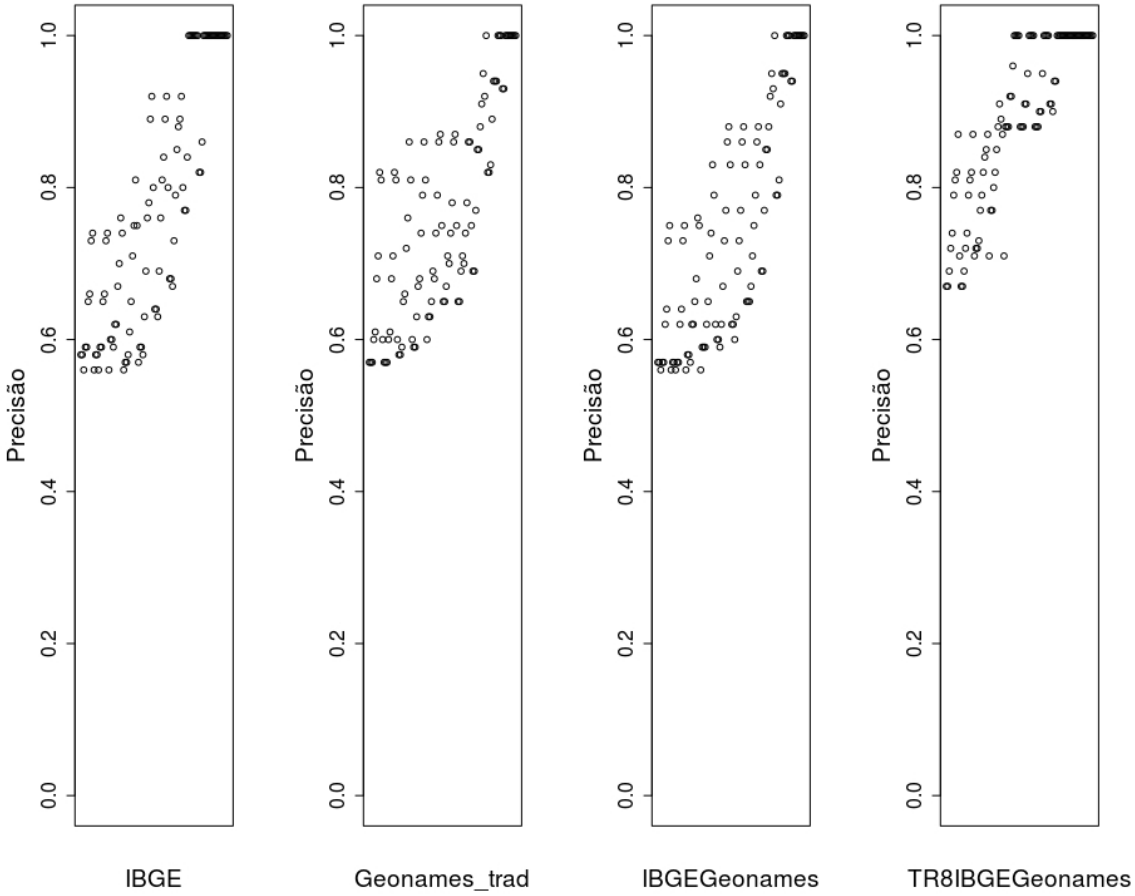


Figura IV.1: Comparativa da Precisão para os 400 testes agrupados pela lista inicial do teste

Melhores resultados Precisão

	Num	Teste	CP	FP	FN	Precisão	Abrangência	Medida_F	VCluster	Vtermos	Glossário
1	47	TR8_TR8IBGEGeonames_0.15_0.25	21	0	5	1	0.81	0.90	0.15	0.25	TR8IBGEGeonames
2	57	TR8_TR8IBGEGeonames_0.2_0.25	20	0	6	1	0.77	0.87	0.20	0.25	TR8IBGEGeonames
3	74	TR8_IBGE_0.5_0.1	18	0	8	1	0.69	0.82	0.50	0.10	IBGE
4	67	TR8_TR8IBGEGeonames_0.25_0.25	18	0	8	1	0.69	0.82	0.25	0.25	TR8IBGEGeonames
5	81	TR8_TR8IBGEGeonames_1_0	18	0	8	1	0.69	0.82	1.00	0.00	TR8IBGEGeonames
6	82	TR8_TR8IBGEGeonames_1_0.01	18	0	8	1	0.69	0.82	1.00	0.01	TR8IBGEGeonames
7	75	TR8_IBGE_0.5_0.15	17	0	9	1	0.65	0.79	0.50	0.15	IBGE
8	76	TR8_IBGE_0.5_0.2	17	0	9	1	0.65	0.79	0.50	0.20	IBGE
9	88	TR8_IBGEGeonames_1_0.5	17	0	9	1	0.65	0.79	1.00	0.50	IBGEGeonames
10	83	TR8_TR8IBGEGeonames_1_0.05	17	0	9	1	0.65	0.79	1.00	0.05	TR8IBGEGeonames

Num: número de teste. **Teste**: código do teste. **CP**: Corretos (verdadeiros) Positivos. **FP**: Falsos Positivos. **Vcluster**: valor da associação na seleção do synset. **Vtermos**:Valor da associação no filtro final dos termos.

Comparativa de resultados segundo lista semente: Abrangência

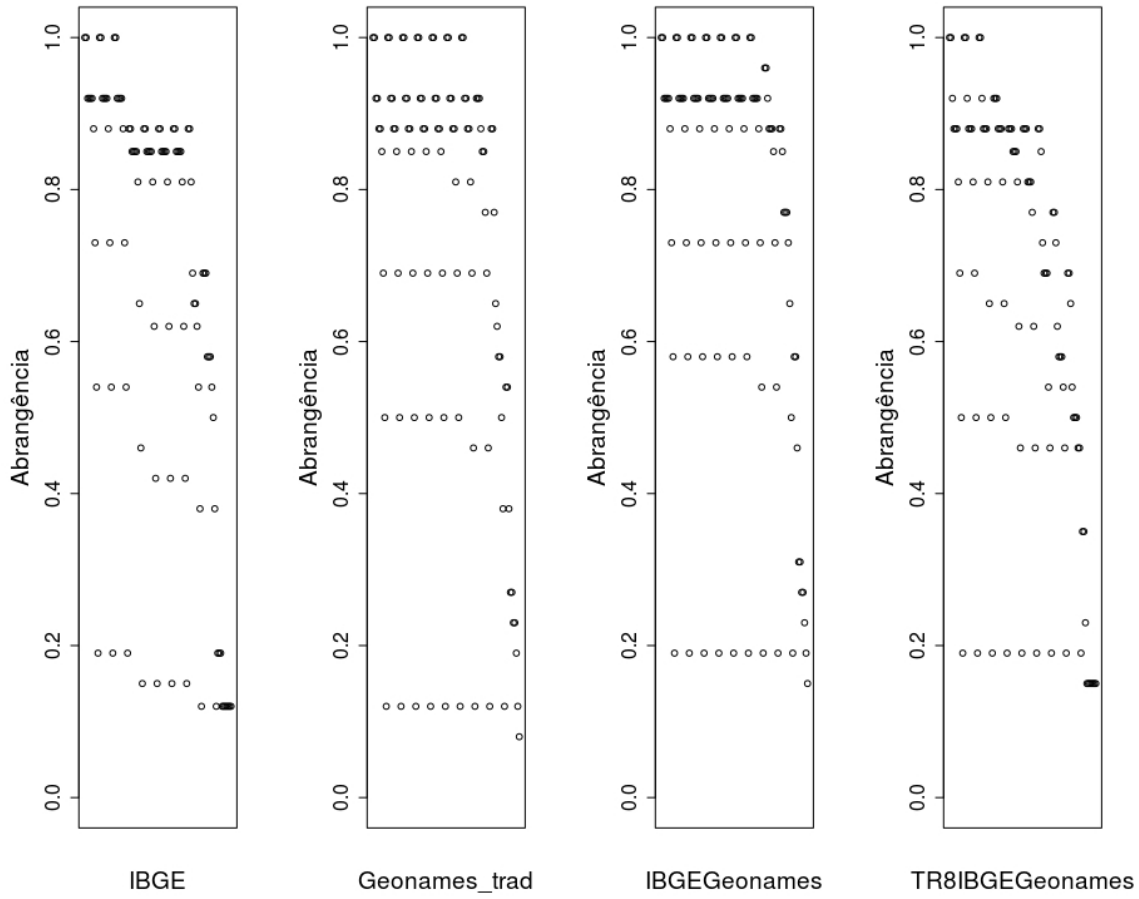


Figura IV.2: Comparativa da abrangência para os 400 testes agrupados pela lista inicial do teste

Melhores resultados Abrangência

	Num	Teste	CP	FP	FN	Precisão	Abrangência	Medida_F	VCluster	Vtermos	Glossário
1	21	TR8_TR8IBGEGeonames_0.05_0	26	10	0	0.72	1	0.84	0.05	0.00	TR8IBGEGeonames
2	22	TR8_TR8IBGEGeonames_0.05_0.01	26	10	0	0.72	1	0.84	0.05	0.01	TR8IBGEGeonames
3	1	TR8_TR8IBGEGeonames_0_0	26	13	0	0.67	1	0.80	0.00	0.00	TR8IBGEGeonames
4	2	TR8_TR8IBGEGeonames_0_0.01	26	13	0	0.67	1	0.80	0.00	0.01	TR8IBGEGeonames
5	11	TR8_TR8IBGEGeonames_0.01_0	26	13	0	0.67	1	0.80	0.01	0.00	TR8IBGEGeonames
6	12	TR8_TR8IBGEGeonames_0.01_0.01	26	13	0	0.67	1	0.80	0.01	0.01	TR8IBGEGeonames
7	51	TR8_Geonames_trad_0.2_0	26	14	0	0.65	1	0.79	0.20	0.00	Geonames_trad
8	52	TR8_Geonames_trad_0.2_0.01	26	14	0	0.65	1	0.79	0.20	0.01	Geonames_trad
9	61	TR8_Geonames_trad_0.25_0	26	14	0	0.65	1	0.79	0.25	0.00	Geonames_trad
10	62	TR8_Geonames_trad_0.25_0.01	26	14	0	0.65	1	0.79	0.25	0.01	Geonames_trad

Num: número de teste. **Teste**: código do teste. **CP**: Corretos (verdadeiros) Positivos. **FP**: Falsos Positivos. **Vcluster**: valor da associação na seleção do synset. **Vtermos**:Valor da associação no filtro final dos termos.

Comparativa de resultados segundo lista semente: Medida-F

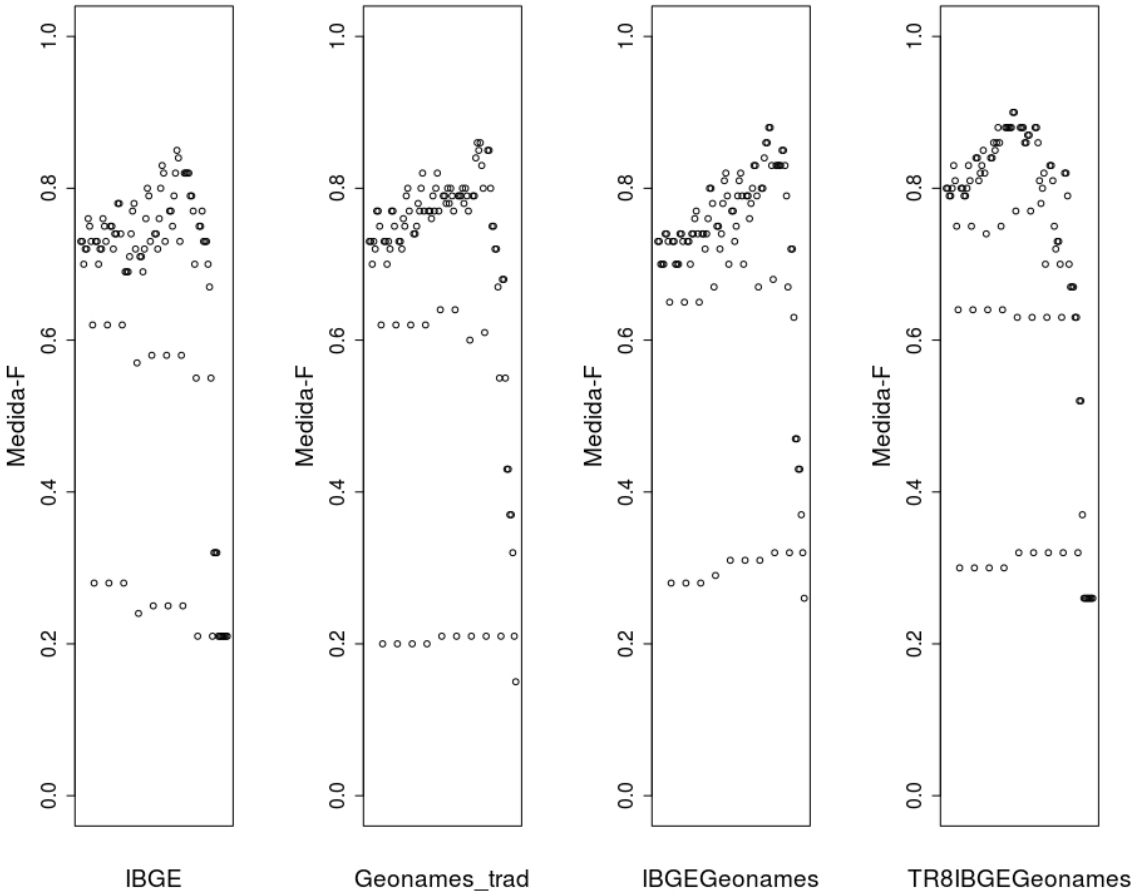


Figura IV.3: Comparativa da medida-F para os 400 testes agrupados pela lista inicial do teste

Melhores resultados Medida-F

	Num	Teste	CP	FP	FN	Precisão	Abrangência	Medida_F	VCluster	Vtermos	Glossário
1	47	TR8_TR8IBGEGeonames_0.15_0.25	21	0	5	1.00	0.81	0.90	0.15	0.25	TR8IBGEGeonames
2	46	TR8_TR8IBGEGeonames_0.15_0.2	22	1	4	0.96	0.85	0.90	0.15	0.20	TR8IBGEGeonames
3	77	TR8_IBGEGeonames_0.5_0.25	22	2	4	0.92	0.85	0.88	0.50	0.25	IBGEGeonames
4	44	TR8_TR8IBGEGeonames_0.15_0.1	22	2	4	0.92	0.85	0.88	0.15	0.10	TR8IBGEGeonames
5	45	TR8_TR8IBGEGeonames_0.15_0.15	22	2	4	0.92	0.85	0.88	0.15	0.15	TR8IBGEGeonames
6	76	TR8_IBGEGeonames_0.5_0.2	23	3	3	0.88	0.88	0.88	0.50	0.20	IBGEGeonames
7	36	TR8_TR8IBGEGeonames_0.1_0.2	23	3	3	0.88	0.88	0.88	0.10	0.20	TR8IBGEGeonames
8	41	TR8_TR8IBGEGeonames_0.15_0	23	3	3	0.88	0.88	0.88	0.15	0.00	TR8IBGEGeonames
9	42	TR8_TR8IBGEGeonames_0.15_0.01	23	3	3	0.88	0.88	0.88	0.15	0.01	TR8IBGEGeonames
10	43	TR8_TR8IBGEGeonames_0.15_0.05	23	3	3	0.88	0.88	0.88	0.15	0.05	TR8IBGEGeonames

Num: número de teste. **Teste**: código do teste. **CP**: Corretos (verdadeiros) Positivos. **FP**: Falsos Positivos. **Vcluster**: valor da associação na seleção do synset. **Vtermos**:Valor da associação no filtro final dos termos.

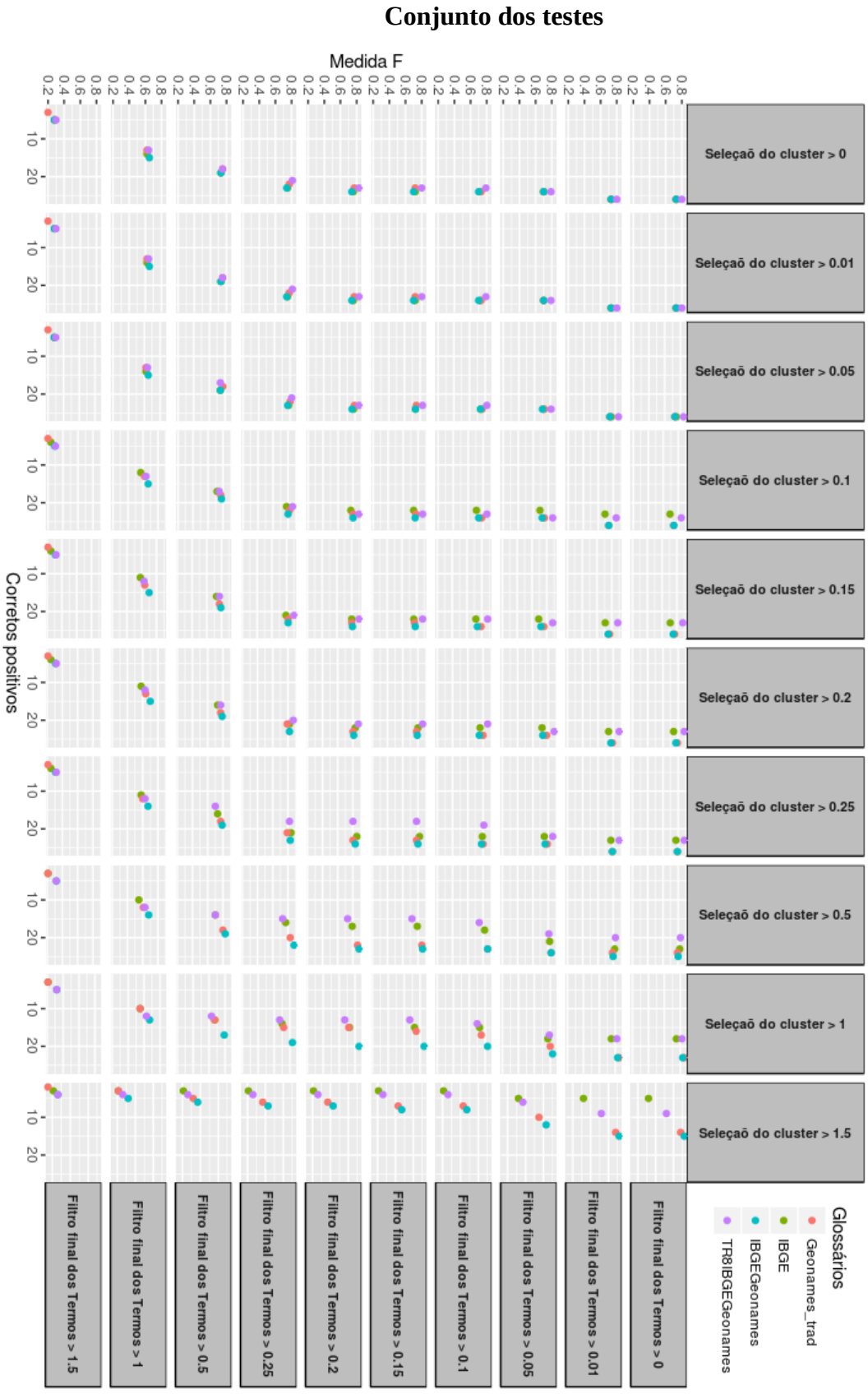


Figura IV.4: Resultados na medida-F e verdadeiros positivos para o conjunto dos testes

Apêndice VI

Exemplo de classificação de entidades geográficas em relação a dois tipos geográficos

Dadas as entidades mencionadas selecionadas pela sua maior frequência no corpus relativamente aos atributos geográficos **cidade** (*Pequim*, *Martauão*, *Odiaa*) e **ilha** (*Iaoa*, *Tanixumaa*, *Çamatra*) obtemos as coordenadas das suas concordâncias com as expressões *cidade* e *ilha* de onde calculamos a direção pela forma geométrica dos seus vetores:

θ (Çamatra) = arctan (|13 / 1|) = 85.60°

θ (Iaoa) = arctan (|11 / 7|) = 57.53°

θ (Martauão) = arctan (|1 / 18|) = 3.18°

θ (Odiaa) = arctan (|1 / 17|) = 3.37°

θ (Pequim) = arctan (|1 / 38|) = 1.51°

θ (Tanixumaa) = arctan (|11 / 5|) = 65.56°

Cada vetor forma aliás um ângulo a respeito de outra entidade geográfica mencionada, o valor pode ser calculado pela diferença entre os ângulos de direção:

	Çamatra	Iaoa	Martauão	Odiaa	Pequim	Tanixumaa
Çamatra	0.00000	28.072487	82.4214645	82.2348340	84.093859	20.045249
Iaoa	28.07249	0.000000	54.3489776	54.1623470	56.021372	8.027238
Martauão	82.42146	54.348978	0.0000000	0.1866305	1.672394	62.376215
Odiaa	82.23483	54.162347	0.1866305	0.0000000	1.859025	62.189585
Pequim	84.09386	56.021372	1.6723944	1.8590249	0.000000	64.048609
Tanixumaa	20.04525	8.027238	62.3762151	62.1895846	64.048609	0.000000

Tabela VI.1. Distâncias entre os ângulos dos vetores de coocorrências

Os ângulos por sua vez podem ser expressados pelo seu cosseno, o valor final usado para medir a diferença a respeito dos traços semânticos CIDADE e ILHA entre as entidades mencionadas:

	Çamatra	Iaoa	Martauão	Odiaa	Pequim	Tanixumaa
Çamatra	1.0000000	0.8823529	0.1318850	0.1351132	0.1028992	0.9394222
Iaoa	0.8823529	1.0000000	0.5828468	0.5854906	0.5588836	0.9902018
Martauão	0.1318850	0.5828468	1.0000000	0.9999947	0.9995740	0.4636639
Odiaa	0.1351132	0.5854906	0.9999947	1.0000000	0.9994737	0.4665474
Pequim	0.1028992	0.5588836	0.9995740	0.9994737	1.0000000	0.4376085
Tanixumaa	0.9394222	0.9902018	0.4636639	0.4665474	0.4376085	1.0000000

Tabela VI.2. Distância semântica entre entidades mencionadas para os traços +CIDADE e +ILHA

Apêndice VII

Aprendizado de máquina para a classificação de entidades geográficas mencionadas

Ronda 1

Modo do Corpus

Selecionam-se apenas as orações com topónimos. Lematizamos as entidades geográficas mencionadas (topónimos e gentílicos) de modo que todas as variantes fiquem representadas por um mesmo lexema. O texto é semi-normalizado nas grafias, formas comuns e morfologia verbal mais regular.

Matriz de coocorrências

Criamos uma matriz de coocorrências termo a termo (10043 x 10043), finalmente reduzida a entidades geográficas mencionadas e tipos geográficos reconhecidos como atributos no índice (719x719).

Seleção de observações e atributos de predição

Convertida a matriz num *data frame*, as linhas ficam reduzidas às entidades geográficas mencionadas (os atributos geográficos ficam apenas como variáveis nas colunas) para lhe adicionar os dados da base de dados do corpus, na tabela de índice de lexemas:

Frequência absoluta: da entidade no corpus.

Classe: a que pertence a entidade geográfica.

Tipo: Topónimo ou Gentílico.

Labor a resolver

Classificação das entidades geográficas numa classe dos tipos geográficos: A, H, L, N, P, S e T.

Segmentação dos dados e treino

Usamos um modelo Random Forest no pacote Caret de R. Aproveitamos dois subconjuntos do data frame, segmentados de modo aleatório, 80% para treino e 20% para teste (validação).

Distribuição dos tipos no data frame

```
A   H    L   N   P   S   T
```

210

```
109 153  52  26 196  50 102
```

a repartir em dois grupos, treino e teste:

Treino

```
A   H   L   N   P   S   T
88 123  42  21 157  40  82
```

Teste

```
A  H  L  N  P  S  T
21 30 10  5 39 10 20
```

Resultados do modelo obtido do treino sobre os dados reservados para teste

		Aguardados						
Predição		A	H	L	N	**P**	S	T
	A	**9**	2	1	0	1	0	1
	H	1	**18**	0	3	4	2	2
	L	2	0	**4**	0	1	0	1
	N	0	0	0	**0**	0	0	0
	P	6	8	1	2	**28**	7	5
	S	0	0	0	0	2	**1**	0
	T	2	1	0	0	2	0	**10**

	A	H	L	N	P	S	T
Abrangência	0.45000	0.6207	0.66667	0.00000	0.7368	0.100000	0.52632
Precisão	**0.64286**	0.6000	0.50000	NaN	0.4912	0.333333	**0.66667**

Ronda 2

Labor a desenvolver

Classificação das entidades geográficas pertencentes à categoria Povoação.

O mesmo modo de corpus e matriz de coocorrências que na Ronda 1.

Seleção de observações e atributos de predição

Entidades geográficas mencionadas pertencentes à classe de Povoação. Reclassificamos todas as demais numa única classe (Outras).

```
 O   P
469 195
```

O: Outro tipo que povoação

P: Povoação

Segmentação dos dados e treino

Usamos um modelo Random Forest no pacote Caret de R. Aproveitamos dois subconjuntos do data frame, segmentados de modo aleatório, 80% para treino e 20% para teste (validação).

Treino

O	P
282	117

Teste

O	P
187	78

Resultados do modelo obtido do treino sobre os dados reservados para teste

		Aguardados	
Predição			
		O	P
	O	**76**	13
	P	17	**26**

Ronda 3

Labor a desenvolver

Classificação das entidades geográficas pertencentes à categoria *Povoação* ou *Terrestre*.

Modo do Corpus

Selecionam-se apenas as orações com topónimos. Lematizamos as entidades geográficas mencionadas (topónimos e gentílicos) de modo que todas as variantes fiquem representadas por um mesmo lexema. O texto é semi-normalizado nas grafias e morfologia verbal mais regular.

Matriz de coocorrências

Criamos uma matriz de coocorrências termo a termo (10043 x 10043).

Seleção de observações e atributos de predição

347 entidades geográficas mencionadas pertencentes à classe de acidentes geográficos terrestres e povoacionais como observações.

Lista de 721 entidades geográficas mencionadas (todas as classes) e tipos geográficos (de todas as classes) como preditores.

Treino

Usamos um modelo Random Forest no pacote Caret de R. Aproveitamos dois subconjuntos do data frame, segmentados de modo aleatório, 80% para treino e 20% para teste (validação).

Resultados do modelo obtido do treino sobre os dados reservados para teste

	Aguardados	
Predição		
	P	T
P	37	8
T	2	12

Lista de entidades classificadas usadas no teste (sem destacar entidades corretamente classificadas, erros em negrito):

[1]	Abedalcuria	**Ainão**	Alcouchete	Anchepisaõ	Angenia	Apefingau
[7]	**Bandou**	Bitonto	Caleypute	Canguexumaa	Fucanxi	Fucheo
[13]	Fumbana	Fumbau	Goa	**Guampalaor**	Guanxiroo	Hiquegens
[19]	Iacur	**Ilher**	Iunquinilau	Lantor	Laura	Lechune
[25]	Liampoo	**Machão**	Madura	Manaquileu	Medina	Meleitor
[31]	**Miaygimaa**	Mindoo	Minhacutem	Moncalor	Muria	Nazarè
[37]	Nicubar	Odiaa	Pequim	**Pilaunera**	platarias	Poutel
[43]	**PulloCondor**	PulloQuirim	Quanginau	Seuilha	**Sinay**	Sorocataõ
[49]	**Sundaypatir**	Taurys	Taypor	Taysiraõ	Timplão	Tinlau
[55]	Xaraa	Xianguulee	Xinligau	Xipatom	Xipatom	

Ronda 4

Labor a desenvolver

Classificação das entidades geográficas pertencentes aos atributos *ilha* e *cidade*.

Modo do Corpus

Selecionam-se apenas as orações com topónimos. Lematizamos as entidades geográficas mencionadas (topónimos e gentílicos) de modo que todas as variantes fiquem representadas por um mesmo lexema. O texto é semi-normalizado nas grafias e morfologia verbal mais regular.

Matriz de coocorrências

Criamos uma matriz de coocorrências termo a termo (10043 x 10043).

Seleção de observações e atributos de predição

182 x 721

182 entidades geográficas mencionadas com atributos cidade e ilha como observações.

721 atributos para a predição. Lista de entidades geográficas mencionadas (todas as classes) e tipos geográficos (atributos) mais os dados da base de dados do corpus, na tabela de índice de lexemas:

Frequência absoluta: da entidade no corpus

Tipo: Topónimo ou Gentílico.

Treino

Usamos um modelo Random Forest no pacote Caret de R. Aproveitamos dois subconjuntos do data frame, segmentados de modo aleatório, 70% para treino e 30% para teste (validação). Repetimos o teste em 25 ocasiões, modificando de cada vez a partição dos dados de treino e de teste, sempre com a mesma proporção 70% treino e 30% teste.

Resultados dos modelos obtido do treino sobre os dados reservados para teste

	cidade_A	cidade_P	ilha_A	ilha_P
1	0.8611111	0.8157895	0.5882353	0.6666667
2	0.9166667	0.9428571	**0.8823529**	0.8333333
3	0.8888889	0.8205128	0.5882353	0.7142857
4	0.8611111	0.9117647	0.8235294	0.7368421
5	0.8611111	0.9117647	0.8235294	0.7368421
6	0.9444444	0.8292683	0.5882353	0.8333333
7	0.8888889	0.8421053	0.6470588	0.7333333
8	0.9444444	**0.9444444**	**0.8823529**	0.8823529
9	0.9166667	0.8918919	0.7647059	0.8125000
10	0.8611111	0.9117647	0.8235294	0.7368421
11	0.9722222	0.8974359	0.7647059	0.9285714
12	0.9166667	0.9428571	**0.8823529**	0.8333333
13	0.9722222	0.8536585	0.6470588	0.9166667
14	0.8888889	0.8648649	0.7058824	0.7500000
15	0.8611111	0.8611111	0.7058824	0.7058824
16	**1.0000000**	0.8571429	0.6470588	**1.0000000**
17	0.9722222	0.9210526	0.8235294	0.9333333
18	0.9444444	0.8500000	0.6470588	0.8461538
19	0.9444444	0.8095238	0.5294118	0.8181818
20	0.9444444	0.8717949	0.7058824	0.8571429
21	0.8611111	0.9117647	0.8235294	0.7368421
22	0.8055556	0.8285714	0.6470588	0.6111111
23	0.9166667	0.7500000	0.3529412	0.6666667
24	0.8888889	0.8648649	0.7058824	0.7500000
25	0.9444444	0.9189189	0.8235294	0.8750000

Cidade_A: Abrangência na classificação de entidades com o atributo *cidade*.

Cidade_P: Precisão na classificação de entidades com o atributo *cidade*.

Ilha_A: Abrangência na classificação de entidades com o atributo *ilha*.

Ilha_P: Precisão na classificação de entidades com o atributo *ilha*.

Exemplo de resultados. Teste 16.

	Aguardados	
Predição		
	Cidade	Ilha
Cidade	36	6
Ilha	0	11

Lista de entidades usadas no teste (em negrito classificações erradas):

[1]	Abedalcuria	Alexandrino	Amadabad	Anapleu	Bale
[6]	**Banchaa**	**Borneo**	Buda	Caixiloo	Calempluy
[11]	Canguexumaa	**Ceilão**	Cerdenha	Constantinopla	Cordoua
[16]	Corpilem	Cosmim	Digum	Fucheo	Geilolo
[21]	Guijampee	Gumbim	Guntaleu	**Iaoa**	Iapara
[26]	Ierusalem	**ilha do fogo**	ilha dos cocos	Iunquinilau	Iuropisaõ
[31]	Lisboa	Madeyra	Martauão	Medina	Meleitay
[36]	Minapau	Muria	Nanquim	Pamquinor	Pequim
[41]	Ponquilor	**Pullo Catão**	Pullo Hinhor	Pullo Tiquòs	Quoansy
[46]	Roma	Sicay	Surião	Timplão	Toro
[51]	Vpe	Xipatom	Xipator		

Ronda 5

Labor a desenvolver

Classificação das entidades geográficas parte da *China* e da *India*.

Modo do Corpus

Selecionam-se apenas as orações com topónimos. Lematizamos as entidades geográficas mencionadas (topónimos e gentílicos), todas as variantes ficam representadas por um mesmo lexema. O texto é normalizado (terminações e grafias).

Matriz de coocorrências

Criamos uma matriz de coocorrências termo a termo (10043 x 10043).

Seleção de observações e atributos de predição

121 x 667

121 entidades geográficas mencionadas classificadas como *é_Parte_de Índia* ou *é_Parte_de China* no índice como observações.

667 atributos para a predição. Lista de entidades geográficas mencionadas mais os dados da base de dados do corpus, na tabela de índice de lexemas:

Frequência absoluta: da entidade no corpus.

Tipo: Topónimo ou Gentílico.

Treino

Usamos um modelo Random Forest no pacote Caret de R. Aproveitamos dois subconjuntos do data frame, segmentados de modo aleatório, 70% para treino e 30% para teste (validação).

Repetimos em 25 ocasiões, modificando de cada vez a partição dos dados de treino e de teste, sempre com a mesma proporção 70% treino e 30% teste.

Resultados dos modelos obtido do treino sobre os dados reservados para teste

	India_A	India_P	China_A	China_P
1	0.4444444	**1.0000000**	**1.0000000**	0.8437500
2	0.4444444	**1.0000000**	**1.0000000**	0.8437500
3	0.6666667	0.8571429	0.9629630	0.8965517
4	0.6666667	0.7500000	0.9259259	0.8928571
5	0.5555556	**1.0000000**	**1.0000000**	0.8709677
6	0.7777778	0.8750000	0.9629630	0.9285714
7	0.6666667	**1.0000000**	**1.0000000**	0.9000000
8	0.6666667	0.8571429	0.9629630	0.8965517
9	0.6666667	**1.0000000**	**1.0000000**	0.9000000
10	**1.0000000**	**1.0000000**	**1.0000000**	**1.0000000**
11	0.7777778	**1.0000000**	**1.0000000**	0.9310345
12	0.5555556	**1.0000000**	**1.0000000**	0.8709677
13	0.5555556	**1.0000000**	**1.0000000**	0.8709677
14	0.6666667	0.7500000	0.9259259	0.8928571
15	0.5555556	**1.0000000**	**1.0000000**	0.8709677
16	0.4444444	**1.0000000**	**1.0000000**	0.8437500
17	0.5555556	0.8333333	0.9629630	0.8666667
18	0.6666667	0.8571429	0.9629630	0.8965517
19	0.7777778	**1.0000000**	**1.0000000**	0.9310345
20	0.5555556	**1.0000000**	**1.0000000**	0.8709677
21	0.6666667	0.8571429	0.9629630	0.8965517
22	0.5555556	0.7142857	0.9259259	0.8620690
23	0.3333333	0.7500000	0.9629630	0.8125000
24	0.6666667	0.8571429	0.9629630	0.8965517
25	0.5555556	**1.0000000**	**1.0000000**	0.8709677

India_A: Abrangência na classificação de entidades *é_Parte_de Índia*

India_P: Precisão na classificação de entidades *é_Parte_de Índia*

China_A: Abrangência na classificação de entidades *é_Parte_de China*

China_P: Precisão na classificação de entidades *é_Parte_de China*

Exemplo de resultados. Teste 10.

	Aguardados	
Predição		
	China	Índia
China	27	0
Índia	0	9

Lista de entidades usadas no teste (em negrito classificações erradas se as houver) :

```
[1]  Anay        Angicamoy  Cantão  Carapatão Chaul   Chitor     Comhay
[8]  Conxinacau  Dabul      Diu     Fucheo    Goa     Guinapalir  Guinaytaraõ
[15] Guintoo    Lautimey   Lautir  Malauares Micuy   Mogores     Narsinga
[22] Ochileuday Paatebenam Pacão   Pocasser  PulloQuirim Quansy  Quaytragum
[29] Quoansy    Sampitay   Sansy   Sileupamor  Sumbor     Suzoanganee Tuxenguim
[36] Xinligau
```

Exemplo de resultados. Teste 23.

	Aguardados	
Predição		
	China	Índia
China	26	6
Índia	1	3

Lista de entidades usadas no teste (em negrinha classificações erradas se as houver)

```
 [1] Anay      Ancolaa      Angicamoy    Angitur      Buncalou     Cambaya
 [7] Cananor   Carapatão    Comorim      Corpilem     Damaõ        Dely
[13] Fanjus    Finginilau   Fiunganorsee Gotom        Guijampee    Lamau
[19] Liampoo   Manicataraõ  Masulepatão  Nacataas     Nangafau     Nixiamcoo
[25] Ochileuday Orixaa      Pequim       Ponquilor    PulloQuirim  Quaytragum
[31] Sansy     Sumbor       Susoquerim   Tautaa       Xingrau      Xinligau
```

Ronda 6

Labor a desenvolver

Classificação das entidades geográficas parte da *China* e da *India*.

Modo do Corpus

Selecionam-se apenas as orações com topónimos. Lematizamos as entidades geográficas mencionadas (topónimos e gentílicos) de modo que todas as variantes fiquem representadas por um mesmo lexema. O texto é semi-normalizado nas grafias e morfologia verbal mais regular.

Matriz de coocorrências

Criamos uma matriz de coocorrências termo a termo (10043 x 10043).

Seleção de observações e atributos de predição

121 x 4

121 entidades geográficas mencionadas classificadas como *é_Parte_de Índia* ou *é_Parte_de China* no índice como observações.

4 atributos para a predição. Coocorrências com duas entidades geográficas mencionadas (China e India) mais os dados da base de dados do corpus, na tabela de índice de lexemas:

Frequência absoluta: da entidade no corpus

Tipo: Topónimo ou Gentílico

Treino

Usamos um modelo Random Forest no pacote Caret de R. Aproveitamos dois subconjuntos do data frame, segmentados de modo aleatório, um 70% para treino e 30% para teste (validação).

Repetimos o teste em 25 ocasiões, modificando de cada vez a partição dos dados de treino e de teste, sempre com a mesma proporção 70% treino e 30% teste.

Resultados dos modelos obtido do treino sobre os dados reservados para teste

	India_A	**India_P**	**China_A**	**China_P**
1	0.3333333	0.6000000	0.9259259	0.8064516
2	0.2222222	0.5000000	0.9259259	0.7812500
3	0.3333333	0.5000000	0.8888889	0.8000000
4	0.3333333	**1.0000000**	**1.0000000**	0.8181818
5	0.4444444	0.8000000	0.9629630	0.8387097
6	0.2222222	0.6666667	0.9629630	0.7878788
7	0.4444444	**1.0000000**	**1.0000000**	0.8437500
8	0.2222222	**1.0000000**	**1.0000000**	0.7941176
9	0.3333333	**1.0000000**	**1.0000000**	0.8181818
10	**0.5555556**	**1.0000000**	**1.0000000**	**0.8709677**
11	0.3333333	0.7500000	0.9629630	0.8125000
12	0.3333333	0.6000000	0.9259259	0.8064516
13	0.1111111	0.5000000	0.9629630	0.7647059

14	0.2222222	**1.0000000**	**1.0000000**	0.7941176
15	0.3333333	0.7500000	0.9629630	0.8125000
16	**0.5555556**	0.5555556	0.8518519	0.8518519
17	0.5555556	0.8333333	0.9629630	0.8666667
18	0.4444444	**1.0000000**	**1.0000000**	0.8437500
19	0.0000000	NaN	**1.0000000**	0.7500000
20	**0.5555556**	0.7142857	0.9259259	0.8620690
21	**0.5555556**	**1.0000000**	**1.0000000**	**0.8709677**
22	0.3333333	**1.0000000**	**1.0000000**	0.8181818
23	0.1111111	0.2500000	0.8888889	0.7500000
24	0.4444444	0.5714286	0.8888889	0.8275862
25	0.4444444	**1.0000000**	**1.0000000**	0.8437500

India_A: Abrangência na classificação de entidades *é_Parte_de Índia*

India_P: Precisão na classificação de entidades *é_Parte_de Índia*

China_A: Abrangência na classificação de entidades *é_Parte_de China*

China_P: Precisão na classificação de entidades *é_Parte_de China*

Exemplo de resultados. Teste 21.

	Aguardados	
Predição		
	China	Índia
China	27	4
Índia	0	5

Lista de entidades usadas no teste (em negrito classificações erradas se as houver):

[1]	Anay	Angitur	Baçaim	**Batecalaa**	Bigaypotim	Cambaya
[7]	**Cananor**	Chabaquee	Chaul	Cohilouza	Conxinacau	**Dely**
[13]	Diu	Goa	Guauxitim	Guintoo	Lantau	Liampoo
[19]	Ochileuday	Paatebenam	Pocasser	Ponquilor	Potimleu	Quaytragum
[25]	Quoamão	Quoansy	**Sategão**	Sileupamor	Sumbor	Tautaa
[31]	Tinlau	Xianguulee	Xilendau	Xingrau	Xinxipou	Xipatom

Exemplo de resultados. Teste 22.

	Aguardados	
Predição		
	China	Índia
China	27	6
Índia	0	3

Lista de entidades usadas no teste (em negrito classificações erradas se as houver):

[1]	**Ancolaa**	Angitur	Bigaypotim	Buncalou	**Chitor**	Choromandel
[7]	**Comorim**	Conxinacau	Coretumbagâ	Corpilem	Çurrate	**Dabul**
[13]	**Dely**	Fumbana	Iunquileu	Lamau	Lautimey	Lautir
[19]	Lequimpau	Liampeu	Malauares	Pequim	Pocasser	Ponquilor
[25]	PulloQuirim	Quoansy	**Raizbutos**	riodosal	Sileyjacau	Sumbor
[31]	Susoquerim	Tanquilem	Taypor	Tinlau	Xingrau	Xinxipou

RESULTADOS

Apêndice VIII

Índice de entidades geográficas mencionadas na *Peregrinação*

Aapessumhee. *Aapessumhee* 32(1). Lugar (lugar). Parte de: Aarù. Outras relações: #4 léguas#@Puneticão@.

Aarù. *Aarù* 14(1), 18(1), 21(3), 22(1), 23(1), 24(1), 26(3), 27(1), 28(2), 31(4), 32(3). *Aarû* 21(1), 22(1), 23(1), 26(2), 27(1), 28(1), 29(1), 30(1), 31(1). *Aarùs* 26(1), 27(1). *Aarûs* 22(1). Reino (cidade, costa). Parte de: Çamatra.
Teluk Aru, Indonesia (AS). Lat. 4, long. 98.2102.

Abedalcuria. *Abedalcuria* 3(1). Ilha. Parte de: Arabia felix. Outras relações: #area#@Curia@; #area#@Muria@; #area#@Çacotorà@.

Achem. *Aachem* 27(1), 148(1), 151(1). *Aachẽs* 147(1). *Achẽ* 15(1), 16(1), 17(1), 26(1), 203(1). *Achem* 13(8), 14(2), 15(4), 16(5), 17(3), 18(2), 21(3), 22(1), 26(3), 27(1), 28(1), 30(1), 31(7), 32(8), 144(2), 203(2), 205(2), 206(1), 207(1). *Achẽs* 13(1), 17(2), 18(1), 22(3), 26(2), 28(1), 32(2), 33(1), 40(2), 144(1), 145(2), 146(2), 148(1), 149(1), 173(1), 175(1), 178(1), 185(1), 203(2), 204(1), 205(2), 206(2), 207(2). Reino (cidade). Parte de: Çamatra. Outras relações: #18 léguas#@Turbaõ@.
Nanggroe Aceh Darussalam Province, Indonesia (AS). Lat. 4, long. 97.

Adẽ. *Adẽ* 20(1). *Adem* 3(1). Porto (lugar). Parte de: Arabia Felix.
Muḫāfaẕat ‘Adan, Yemen (AS). Lat. 12.83333, long. 44.91667.

Africa. *Africa* 68(1), 118(1). Continente. Parte de: Terra.
Africa, (AF). Lat. 7.1881, long. 21.09375.

Agimpur. *Agimpur* 129(1), 130(1). Vila. Parte de: Cauchenchina. Outras relações: #3 léguas#@Latiparau@; #1 léguas#@Fanaugrem@.

Ainão. *Ainão* 39(1), 40(1), 42(2), 44(2), 45(5), 48(3), 49(1), 50(3), 52(1), 53(1), 55(1), 56(1), 95(1), 112(1), 118(1), 189(1). *Ainaõ* 37(1), 46(1). Ilha (enseada, costa). Parte de: China.
Hainan Sheng, China (AS). Lat. 19.25, long. 109.75.

Alcocer. *Alcocer* 158(1). *Alcosser* 20(1). Porto. Parte de: Arabia Felix.

Alcouchete. *Alcouchete* 91(1), 116(1). Povoação (lugar). Parte de: Portugal.
Alcochete, Portugal (EU). Lat. 38.73827, long. -8.97936.

Alemanha. *Alemanha* 92(1). *Alemão* 85(1). *Alimania* 72(1). *Alimanis* 126(1). Província (nação). Parte de: Moscouia.

Alexandrino. *Alexandrino* 225(1). Cidade. Parte de: Egypto.
Alexandria, Egypt (AF). Lat. 31.21564, long. 29.95527.

Alfama. *Alfama* 1(1). Bairro. Parte de: Lisboa.
Alfama, Portugal (EU). Lat. 38.7, long. -9.11667.

Algarauio. *Algarauio* 173(1). Província. Parte de: Portugal.

Amadabad. *Amadabad* 107(1). Cidade. Parte de: Cambaya.
Ahmedabad, India (AS). Lat. 23.02579, long. 72.58727.

Amboyno. *Amboyno* 207(2). Ilha. Parte de: Insulíndia.
Pulau Ambon, Indonesia (AS). Lat. -3.65947, long. 128.16976.

Anapleu. *Anapleu* 185(1). Cidade. Parte de: Bramá.

Anay. *Anay* 56(1), 57(1). Rio. Parte de: China. Outras relações: #area#@Lamau, costa@; #area#@Chincheo@.
Anhai, China (AS). Lat. 24.71958, long. 118.47079.

Anchepisaõ. *Anchepisaõ* 23(1). Ilhéus (ilhéus). Parte de: Çamatra.

Ancolaa. *Ancolaa* 217(1). Barra. Parte de: India.

Andamoens. *Andamoens* 149(1). Nação. Parte de: Ásia.
Andaman Islands, India (AS). Lat. 12.50029, long. 92.75004.

Andraguiree. *Andraguire* 16(1). *Andraguirè* 19(1). *Andraguiree* 22(1), 31(1), 207(2). *Andraguirees* 149(1). *Andraguires* 16(1). Reino (porto). Parte de: Çamatra.
Batang Indragiri, Indonesia (AS). Lat. -0.3622, long. 103.4397.

Angegumaa. *Angegumaa* 158(1), 159(1). *Iangumaa* 128(1). Rio (reino). Parte de: Índico Oriental. Outras relações: #area#@Auaa@; #area#@Gumbim@; #area#@Catammaas@; #area#@Singilapau@; #area#@Timplão@.

Angenia. *Angenia* 172(1). *Angenio* 189(1). Ilha. Parte de: Insulíndia. Outras relações: #area#@Iaoa@; #area#@Bale@; #area#@Madura@.

Angicamoy. *Angicamoy* 122(2). Templo. Parte de: China.

Angitur. *Angitur* 71(1). Ilhas (ilhas). Parte de: China. Outras relações: #5 dias nauticos#@Nanquim@.

Angunee. *Angunee* 202(1). Porto. Parte de: Iapaõ.
Akune, Japan (AS). Lat. 32.01667, long. 130.2.

Ansedaa. *Ansedaa* 150(2), 155(1), 174(1), 178(4), 188(1), 190(2), 191(1). *Ansesedaa* 153(1). Reino (rio). Parte de: Pegù. Outras relações: #area#@Danapluu@; #area#@Digum@; #area#@Meydoo@.
Henzada District, Myanmar [Burma] (AS). Lat. 18, long. 95.

Apefingau. *Apefingau* 18(1). *Fingau* 19(1). Ilhéu. Parte de: Çamatra. Outras relações: #2 horas náuticas#@Panaajù@; #27 léguas#@Minhagaruu@; #area#@Oceano@.

Arabia Felix. *Arabia felix* 1(1). *Arabia Felix* 20(1), 149(1). Região. Parte de: Índico Ocidental.
Arabia, (AS). Lat. 25, long. 45.

Arimaa. *Arimaa* 200(3). Reino. Parte de: Iapaõ.
Harima, Japan (AS). Lat. 35, long. 134.66667.

Arissumhee. *Arissumhee* 24(1). Rio. Parte de: Çamatra. Outras relações: #1 dia náutico#@Siaca@.

Armenia. *Armenia* 221(2). *Armenio* 43(3), 221(3). *Armenios* 149(2). Reino. Parte de: Ásia.
Republic of Armenia, Armenia (AS). Lat. 40.25, long. 45.

Arquico. *Arquico* 4(1), 5(2). *Arquîco* 4(1). Porto. Parte de: Etiopia.
Hirgīgo, Eritrea (AF). Lat. 15.53774, long. 39.45294.

Arracão. *Arracão* 128(1), 197(1). *Arracoẽs* 149(1). *Racão* 146(1), 147(1). Reino. Parte de: Índico Oriental.
Rakhine State, Myanmar [Burma] (AS). Lat. 19, long. 94.25.

Ásia. *Ásia* 1(1), 99(1), 221(1). *Oriente* 113(1). Continente. Parte de: Terra.
Ásia, (AS). Lat. 29.84064, long. 89.29688.

Auaa. *Auaa* 107(1), 153(2), 155(1), 156(4), 157(9), 158(1), 164(1), 165(1), 170(2), 197(1), 198(1), 199(1). *Auaas* 156(1), 194(1). Reino (cidade). Parte de: Índico Oriental.
Ava Nandaw, Myanmar [Burma] (AS). Lat. 21.85479, long. 95.97635.

a varella. *a varella* 41(1). *rio da varella* 42(1). *Varella* 41(1). Rio. Parte de: Indochina. Outras relações: #area#@Champaa@; #area#@Cunebetee@; #area#@Taiquilleu@; #area#@Pilaucacem@.

Azebibe. *Azebibe* 4(1). Lugar. Parte de: Índico Ocidental. Outras relações: #area#@Etiopia@; #area#@Cayro@; #area#@Iudeu@.

Baarròs. *Baarrós* 26(1). *Baarròs* 28(2), 31(1). Reino. Parte de: Çamatra.
Barus, Indonesia (AS). Lat. 2.0106, long. 98.3982.

Babylonia. *Babylonia* 6(1), 148(1). Cidade. Parte de: Egypto.
Babylon, Iraq (AS). Lat. 32.54083, long. 44.42417.

Baçaim. *Baçaim* 43(1). Porto. Parte de: India.

Bagou. *Bagou* 107(1), 153(1). Cidade (capital). Parte de: Pegù.
Bago, Myanmar [Burma] (AS). Lat. 17.33521, long. 96.48135.

Baguetor. *Baguetor* 131(1). Rio. Parte de: Tartaria. Outras relações: #area#@Famstir@; #area#@Natibasoy@; #14 dias fluviais#@Huzamguee@; #area#@Lingator@.

Balambuaõ. *Balambuaõ* 178(1). Principado (principado). Parte de: Panaruca.

Bale. *Baile* 189(1). *Bale* 172(1), 177(1), 179(1). Ilha. Parte de: Insulíndia. Outras relações: #area#@Iaoa@; #area#@Madura@; #area#@Timor@.
Bali, Indonesia (AS). Lat. -8.33333, long. 115.

Banchâ. *Banchâ* 38(1). *Banchà* 33(1). Comarca. Parte de: Sião. Outras relações: #area#@Lequios@.

Banchaa. *Banchaa* 146(2), 183(4). *ilha de Banchaa* 141(1). Ilha. Parte de: Ásia Oriental. Outras relações: #area#@Lequios@.

Banda. *Bãda* 26(1). *Banda* 20(1), 21(1), 26(1), 203(1). Ilha. Parte de: Insulíndia. Outras relações: #area#@Maluco@; #area#@Çunda@.
Pulau-Pulau Banda, Indonesia (AS). Lat. -4.4551, long. 129.8502.

Bandou. *Bãdou* 210(1). *Bandou* 112(1), 135(1). Cidade. Parte de: Iapaõ.
Kantō-chihō, Japan (AS). Lat. 36.25, long. 139.5.

Banta. *Banta* 172(2), 173(1), 179(2). Cidade (porto). Parte de: Çunda. Outras relações: #17 dias náuticos#@Malaca@; #15 dias náuticos#@Iapara@.
Banten, Indonesia (AS). Lat. -6.5, long. 106.25.

Bardees. *Bardees* 8(1). Barra. Parte de: Goa.
Bardez, India (AS). Lat. 15.64441, long. 73.83308.

Bargança. *Bargança* 20(1), 195(1). *Bragãça* 35(1), 167(1). *Bragança* 206(1). Cidade. Parte de: Portugal.
Distrito de Bragança, Portugal (EU). Lat. 41.83615, long. -6.76346.

Barruhaas. *Barruhaas* 144(1). Rio. Parte de: Malayo. Outras relações: #area#@Salangor@; #area#@Panaagim@; #area#@Quedaa@.
Kuala Beruas, Malaysia (AS). Lat. 4.45, long. 100.61667.

Basoy. *Basoy* 150(1). Nação. Parte de: Índico Oriental. Outras relações: #area#@Bramá@.

Bata. *Bata* 13(10), 14(1), 15(1), 16(6), 17(6), 18(1). *Batas* 13(6), 14(2), 15(2), 16(1), 17(5), 20(3), 21(1), 31(1). Reino. Parte de: Çamatra.
Raneue Alue Batak, Indonesia (AS). Lat. 4.6748, long. 95.5421.

Batampina. *Batampina* 85(2), 88(3), 92(1), 97(1), 100(1), 111(1), 117(1). *Bitampina* 98(1). *Batãpina* 103(1). Rio. Parte de: China. Outras relações: #area#@Pequim@; #area#@Nanquim@.

Batecalaa. *Batecalaa* 217(1). Porto. Parte de: India.
Bhatkal, India (AS). Lat. 13.98534, long. 74.55531.

Batelor. *Batelor* 188(1). Fortaleza. Parte de: Pegù. Outras relações: #6 horas,(max)#@Pegù@; #area#@Ansedaa@.

Batobasoy. *Batobasoy* 88(1). Rio. Parte de: Índico Oriental. Outras relações: #area#@Sansy@;

#area#@Cosmim@.

Batoquirim. *Batoquirim* 32(1). Ponta. Parte de: Çamatra. Outras relações: #area#@Puneticão@.

Batorrendão. *Batorrendão* 15(1). Povoação. Parte de: Bata. Outras relações: #0.25 léguas#@Panaajù@.

Beidao. *Beidao* 151(1). Outeiro (outeiro). Parte de: Martauão. Outras relações: #2 tiros de falção#@Martauão@.

Benão. *Benão* 48(1), 122(2), 128(1). Reino. Parte de: Ásia Oriental. Outras relações: #area#@Pafua@; #area#@Mecuy@; ; #area#@Capimper@.

Benau. *Benau* 131(1). Vila. Parte de: Cauchenchina. Outras relações: #6 léguas#@ Fanaugrem@; #area#@Mecuy@.

Bengala. *Bẽgala* 17(1), 107(1), 147(1), 167(1). *Bengala* 20(1), 43(1), 128(1), 144(1), 146(1), 147(1), 148(1), 171(2), 205(2). *Bengalas* 149(1). Reino. Parte de: India.
West Bengal, India (AS). Lat. 24, long. 88.

Berdio. *Berdio* 19(1), 150(1), 189(1). *Berdios* 122(1), 124(1). Reino. Parte de: Indochina. Outras relações: #area#@Sião@; area#@Mõpolocota@; #area#@Cuy@; #area#@Lugor@; #area#@Chintabu@.

Betenigus. *Betenigus* 4(1). Paços (casas). Parte de: Etiopia. Outras relações: #5 léguas#@Satilgão@.

Bidor. *Bidor* 165(1), 166(1). Vila. Parte de: Calaminhan. Outras relações: #area#@Timplão@; #area#@Pituy@; #8 dias fluviais#@Pauel@.

Bigay potim. *Bigay potim* 96(1). Pagode (pagode). Parte de: China. Outras relações: #area#@Batampina@; #3 léguas#@Mindoo@; #7 léguas#@Fiunganorsee@.

Binagorem. *Binagorem* 166(1). Lugar (terra). Parte de: Ásia Oriental. Outras relações: #area#@Pituy@; #200 léguas#@Pauel@.

Bintão. *Bintão* 22(1), 30(2), 31(1), 32(2), 59(1), 176(1), 207(2). *Bintaõ* 51(1). Reino (porto). Parte de: Insulíndia. Outras relações: #area#@Malaca@; #area#@Iantana@.
Pulau Bintan, Indonesia (AS). Lat. 1.10708, long. 104.4728.

Bintor. *Bintor* 142(2). Povoação (lugar). Parte de: Lequios. Outras relações: #6 léguas#@Pongor@.

Bisnagà. *Bisnagà* 107(1). Cidade (capital). Parte de: Narsinga.
Vijayanagara, India (AS). Lat. 15.325, long. 76.465.

Bitonto. *Bitonto* 4(1). Povoação (lugar). Parte de: Etiopia. Outras relações: #5 léguas#@Satilgão@.
Debre Bīzen Gedam, Eritrea (AF). Lat. 15.3307, long. 39.08397.

Borneo. *Borneo* 21(1), 26(2), 35(2), 36(1), 149(1), 173(1), 174(1), 189(1). *Borneos* 16(1), 17(1), 57(1), 59(1), 68(1), 186(1). Ilha (mar). Parte de: Insulíndia.

Borneo, Indonesia (AS). Lat. 1, long. 114.

Botinafau. *Botinafau* 73(1). Serra. Parte de: Ásia Oriental. Outras relações: #area#@Paatebenam@; #area#@Calindão@; #area#@Gangitanou@; #area#@Gigauhos@.

Bralapisaõ. *Bralapisaõ* 39(1). Surgidouro. Parte de: Pullo Condor.

Bramá. *Bramâ* 151(1), 153(1), 188(1), 194(1). *Bramaa* 124(1), 146(1), 147(1), 148(5), 149(5), 150(5), 151(2), 152(2), 153(9), 154(5), 155(5), 156(7), 157(6), 158(1), 162(1), 163(2), 164(1), 165(3), 167(5), 168(2), 170(2), 185(7), 186(8), 187(4), 188(6), 189(1), 190(12), 191(1), 192(2), 194(6), 195(7), 196(1), 197(7), 198(5), 199(2). *Bramaas* 114(1), 122(1), 149(5), 150(3), 151(4), 152(1), 155(1), 156(1), 163(2), 164(1), 165(1), 166(1), 167(1), 188(3), 190(11), 194(2), 195(2), 196(1), 197(1), 198(1). *Bramas* 153(1), 170(1), 190(1), 195(2). *Bramâs* 149(1), 152(1). *Bramás* 150(1), 159(1). *Bramàs* 165(1). *Bramau* 198(1). *Bramá* 200(2). *Bramà* 200(1). Reino. Parte de: Índico Oriental.
Union of Burma, Myanmar [Burma] (AS). Lat. 21, long. 96.

Brasil. *Brasil* 37(1), 215(1). Região. Parte de: America.
Federative Republic of Brazil, Brazil (AM). Lat. -10, long. -55.

Broteo. *Broteo* 20(1). Rio. Parte de: Campar. Outras relações: #area#@Iambè@.

Buaquirim. *Buaquirim* 39(1). Cidade (povoação). Parte de: Quitiruão. Outras relações: #area#@Pinator@.

Buatẽdoo. *Buatẽdoo* 81(1). Casa (casa). Parte de: Catihorau.

Buda. *Buda* 96(1). Cidade (lugar). Parte de: Vngaro.
Budapest, Hungary (EU). Lat. 47.49801, long. 19.03991.

Bugẽ. *Bugẽ* 166(1). Nação (povo). Parte de: Ásia Oriental. Outras relações: #area#@Calouhos@; #area#@Timpates@.

Buhaquirim. *Buhaquirim* 44(1). Povoação. Parte de: Ainão. Outras relações: #7 léguas#@Guamboy@.

Bumioens. *Bumioens* 166(1). Nação. Parte de: Ásia Oriental. Outras relações: #area#@Pauel@.

Bumxay. *Bumxay* 120(1). Campo. Parte de: China. Outras relações: #area#@Lautimey@; #area#@Pequim@.

Buncalou. *Buncalou* 66(1). Ilha. Parte de: China. Outras relações: #area#@Comolem@; #6 dias náuticos#@Liampoo@.

Bungo. *Bũgo* 135(3), 200(2), 208(1), 211(1). *Bungo* 134(1), 135(4), 208(1), 209(1), 210(1), 215(1), 218(2), 223(3), 224(1), 225(2). Reino. Parte de: Iapaõ. Outras relações: #100 léguas#@Tanixumaa@; #60 léguas#@Omanguche@.
Oita Prefecture, Japan (AS). Lat. 33.19899, long. 131.43353.

Buxipalem. *Buxipalem* 72(2). Baía. Parte de: Ásia Oriental. Outras relações: #13 dias náuticos#@Fanjus@; #area#@Calindão@.

Çacotorà. *Çacotorà* 3(1). Ilha. Parte de: Arabia Felix. Outras relações: #area#@Abedalcuria@; #area#@Curia@; #area#@Muria@.
Socotra, Yemen (AS). Lat. 12.5, long. 54.

Cagerrendaõ. *Cagerrendaõ* 13(1). Serra. Parte de: Çamatra. Outras relações: #area#@Bata@; #area#@Achem@.

Caixiloo. *Caixiloo* 123(1). Cidade. Parte de: China. Outras relações: #160 léguas#@Pequim@; #area#@Guauxitim@; #area#@Singrachirau@.

Calaminhan. *Calaminhan* 41(1), 107(1), 124(1), 153(1), 157(2), 158(4), 162(6), 163(6), 164(6), 165(8), 167(2), 170(1), 185(1). *Calaminhans* 163(1). *Calaminhás* 194(1). *Calaminhãs* 114(1), 122(1), 149(1). Império. Parte de: Ásia. Outras relações: #area#@Sião@.

Calandor. *Calandor* 20(1). Rio. Parte de: Çamatra.

Calantão. *Calantão* 35(2). Rio. Parte de: Malayo. Outras relações: #18 léguas#@Patane@.
Sungai Kelantan, Malaysia (AS). Lat. 6.18837, long. 102.2354.

Calapa. *Calapa* 180(1). Reino. Parte de: Iaoa.
Jakarta, Indonesia (AS). Lat. -6.21462, long. 106.84513.

Calecù. *Calecù* 7(1), 12(1). *Calecut* 176(1). Reino. Parte de: India.
Kozhikode, India (AS). Lat. 11.24802, long. 75.7804.

Calempluy. *Calempluy* 71(2), 72(1), 74(1), 75(1), 111(1). *Calẽpluy* 70(1), 74(2), 100(1). Ilha. Parte de: China. Outras relações: #10 léguas#@Tanquilem@; #area#@Guinaytaraõ@; #area#@Corpilem@; #area#@Fumbana@; #area#@Paatebenam@.

Caleypute. *Caleypute* 128(1). Povoação (lugar). Parte de: Ásia Oriental. Outras relações: #7 dias fluviais#@Singapamor@; #9 dias fluviais#@Tarem@; #area#@Cauchenchina@.

Calindão. *Calindão* 72(1). Baía. Parte de: Ásia Oriental. Outras relações: #area#@Paatebenam@; #15 léguas#@Buxipalem@; #260 léguas#@Nanquim,enseada@.

Calogẽs. *Calogẽs* 166(1). Nação (povo). Parte de: Ásia Oriental. Outras relações: #area#@Fungaos@; #area#@Friucaranjaa@; #area#@Pauel@.

Calouhos. *Calouhos* 166(1). Nação (povo). Parte de: Ásia Oriental. Outras relações: #area#@Timpates@; #area#@Bugẽ@; #area#@Oqueus@; #area#@Magores@; #area#@Friucaranjaa@; #area#@Pauel@.

Çamatra. *Çamatra* 13(1), 14(1), 20(2), 21(1), 23(2), 39(1), 48(1), 143(1), 144(1), 166(1), 207(1). *Samatra* 1(1). Ilha. Parte de: Insulíndia. Outras relações: #area#@mediterraneo@; #area#@Oceano@.

Sumatra, Indonesia (AS). Lat. 0, long. 102.

Cambaya. *Cambaya* 2(2), 7(1), 20(2), 43(1), 107(1), 151(1). Reino. Parte de: India. Outras relações: #area#@Amadabad@.
Khambhāt, India (AS). Lat. 22.31744, long. 72.61916.

Çambilão. *Çambilão* 146(1). Ilha. Parte de: Sião. Outras relações: #area#@Taubasoy@; #area#@Pullo Hinhor@; #area#@Tanauçarim@; #area#@Pisanduree@.

Camboja. *Cãboja* 112(1). *Camboja* 39(2), 92(1), 179(1), 184(1), 185(1), 189(1). Reino (senhorio, barra). Parte de: Indochina.
Kingdom of Cambodia, Cambodia (AS). Lat. 13, long. 105.

Camoy. *Camoy* 44(1). Baía. Parte de: Ainão. Outras relações: #7 léguas#@Buhaquirim@; #area#@Guamboy@; ; #2 dias nauticos#@Tanauquir@.

Campalagor. *Campalagor* 158(1). Fortaleza. Parte de: Calaminhan. Outras relações: #1 dia fluvial#@Catammaas@; #area#@Tinlau@; #area#@Angegumaa@; #13 dias rio acima#@Chipanocão@.

Campalagrau. *Campalagrau* 162(1). Castelo (castelo). Parte de: Calaminhan. Outras relações: #1 légua#@Timplão@; #12 horas fluvial#@Singilapau@.

Campalarau. *Campalarau* 34(1). Povoação. Parte de: Pão.

Campalarraja. *Campalarraja* 163(1). Cais. Parte de: Timplão.

Campar. *Campar* 31(1), 32(1). *Cãpar* 20(1), 32(1). Reino (ilha, porto). Parte de: Çamatra. Outras relações: #area#@Iambee@; #area#@Broteo@.
Sungai Kampar, Indonesia (AS). Lat. 0.469, long. 103.1446.

Canafama. *Canafama* 209(1), 211(1), 223(1). *Canafamaa* 200(1). Teso (teso). Parte de: Iapaõ. Outras relações: #area#@Fucheo@; #area#@Osquy@; #area#@Fucheo@.

Cananor. *Cananor* 153(1). Porto. Parte de: India.

Canguexumaa. *Canguexumâ* 203(1). *Canguexumaa* 202(2), 208(1). *Quãguixumaa* 135(1). Cidade (enseada, baía, porto). Parte de: Iapaõ. Outras relações: #area#@Hiamangoo@; #area#@Tanixumaa@.

Cantão. *Cantão* 44(1), 67(1), 81(1), 82(1), 84(1), 89(1), 91(1), 116(1), 212(1), 220(1), 222(5). *Cantaõ* 82(1), 215(1), 221(1). *Cãtão* 221(1). Cidade (porto, ilhas). Parte de: China.
Guangzhou, China (AS). Lat. 23.11667, long. 113.25.

Cãpalator. *Cãpalator* 15(1). Cais. Parte de: Panaajù.

Capimper. *Capimper* 122(1), 124(1), 128(1), 181(2), 182(1), 185(1). Reino. Parte de: Ásia. Outras relações: #area#@Sião@; #area#@Passiloco@; #area#@Chiammay@; #area#@Sacotay@; #area#@Pumfileu@.

Kamphaeng Phet, Thailand (AS). Lat. 16.48344, long. 99.52153.

Carapatão. *Carapatão* 8(1). Rio. Parte de: India.
Vāghotan River, India (AS). Lat. 16.55704, long. 73.34146.

Castella. *Castelhano* 208(2). *Castelhanos* 68(1), 143(1), 176(1). *Castella* 68(1). Nação. Parte de: Espanha.
Castilla-La Mancha, Spain (EU). Lat. 39.5, long. -3.

Catammaas. *Catammaas* 158(1). Cidade. Parte de: Tinlau. Outras relações: #area#@Angegumaa@; #7 dias fluviais#@Gumbim@; #area#@Campalagor@; #area#@Meidur@.

Çatão. *Çatão* 150(1), 190(5), 191(2), 193(2), 194(2), 198(1). *Çataõ* 193(1), 194(1), 204(1). *çatão* 190(1). Cidade. Parte de: Pegù. Outras relações: #area#@Pegù, cidade@; #5 léguas#@Moucham@.
Sittang, Myanmar [Burma] (AS). Lat. 17.45474, long. 96.88209.

Catebasoy. *Catebasoy* 93(1). Lugar (mítico). Parte de: Ásia Oriental. Outras relações: #47 dias fluviais#@Pequim@; #area#@Pilaunera@; #area#@Guantipocau@.

Catebenão. *Catebenão* 88(1). Senhorio (senhorio). Parte de: Indochina. Outras relações: #area#@Chiammay@; #area#@Cauchenchina@; #area#@Lechune, rio@.

Catihorau. *Catihorau* 81(1). Aldea (aldea). Parte de: China. Outras relações: #area#@Sileyjacau@; #area#@Nanquim, enseada@; #area#@Conxinacau@.

Catimparù. *Catimparù* 39(1). Povoação. Parte de: Indochina. Outras relações: #area#@Pullo Cambim@; #area#@Champaa@; #area#@Camboja@.

Cauchenchina. *Cauchenchina* 42(1), 43(1), 44(1), 46(1), 50(1), 55(1), 56(1), 92(1), 95(1), 112(1), 125(2), 127(1), 129(2), 130(1), 220(2). *Cauchim* 45(1), 68(1), 88(1), 107(1), 125(1), 128(2), 129(2), 224(1). *Cauchins* 45(2), 48(1), 67(1), 113(1), 114(1), 123(1), 159(1), 184(1). Reino (enseada, costa, mar). Parte de: Indochina.
Tonkin, Vietnam (AS). Lat. 22, long. 105.

Cayro. *Cairo* 107(1). *Cayro* 3(1), 4(3), 7(1), 17(1), 20(1), 21(1), 26(2), 32(1), 43(1), 146(1). Cidade (metrópole). Parte de: Egypto.
Cairo, Egypt (AF). Lat. 30.06263, long. 31.24967.

Cayxem. *Cayxem* 6(1). Porto. Parte de: Índico Ocidental.
Qeshm, Iran (AS). Lat. 26.78333, long. 55.86667.

Ceilão. *Ceilão* 20(1), 146(1), 151(1), 213(1). Ilha. Parte de: Índico Ocidental.
Sri Lanka, Sri Lanka (AS). Lat. 7.5, long. 80.5.

Cerdenha. *Cerdenha* 3(1). Ilha. Parte de: Europa.
Sardinia, Italy (EU). Lat. 40, long. 9.

Cezimbra. *Cezimbra* 1(1). Vila. Parte de: Portugal.

Sesimbra, Portugal (EU). Lat. 38.44451, long. -9.10149.

Chabaquee. *Chabaqué* 115(1). *Chabaquee* 132(1), 143(1), 179(1). *Choaboquec* 51(1). Porto (rio, montes). Parte de: China. Outras relações: #area#@Lamau,costa@; #5 léguas#@rio do sal@.
Chaozhou, China (AS). Lat. 23.65396, long. 116.62262.

Chacomaas. *Chacomaas* 149(1). Nação. Parte de: Ásia.

Chaleu. *Chalẽs* 153(1). *Chaleu* 107(1), 114(1), 150(1), 157(1), 165(1). *Chaleus* 149(1), 153(1), 186(1), 194(1). *Chaloẽs* 149(1), 195(1). Reino (cidade). Parte de: Índico Oriental. Outras relações: #area#@Bramá@.
Salin, Myanmar [Burma] (AS). Lat. 20.57923, long. 94.65834.

Champaa. *Champaa* 39(2), 40(1), 41(1), 70(1), 112(1), 124(1), 174(1), 220(1). *Champaas* 57(1), 68(1), 123(1), 143(1), 186(1). *Champana* 51(1). *Champas* 95(1). *Chãpá* 189(1). *Chãpaa* 88(1). Reino (costa). Parte de: Indochina.
Cochin China, Vietnam (AS). Lat. 11, long. 107.

Champeiloo. *Champeiloo* 221(2). *Champeyloo* 220(1). *Pullo Champeiloo* 42(1). *Pullo Chãpeiloo* 220(1). Ilha. Parte de: Cauchenchina. Outras relações: #area#@Cauchenchina, enseada@; #coordenadas#@14.33N@; #5 dias náuticos#@Sanchaõ@.
Cù Lao Chàm, Vietnam (AS). Lat. 15.95242, long. 108.52037.

Chatigaõ. *Chatigaõ* 171(1). *Ghatigaõ* 147(1). Porto. Parte de: Bengala.
Chittagong, Bangladesh (AS). Lat. 22.3384, long. 91.83168.

Chatir. *Chatir* 183(1). Porto. Parte de: Malayo. Outras relações: #5 léguas, abaixo#@Lugor@.

Chaul. *Chaul* 2(1), 7(1), 8(2), 12(1), 20(1). Porto (barra). Parte de: India.
Chaul, India (AS). Lat. 18.56723, long. 72.93177.

Chauseroo. *Chauseroo* 188(1). Reino. Parte de: Índico Oriental. Outras relações: #area#@Peguu@; #area#@Bramá@.

Chautir. *Chautir* 84(1). Povoação (lugar). Parte de: China. Outras relações: #area#@Sileyjacau@.

Chem ahicogim. *Chenchico* 211(1). *Chenchicogim* 134(1), 218(1), 225(1). *Chenchicogins* 135(1). *Chenchicogis* 133(1). *Chem ahicogim* 209(1). Reino. Parte de: Portugal.
Portuguese Republic, Portugal (EU). Lat. 39.6945, long. -8.13057.

Cherbom. *Cherbom* 178(1), 180(1). Aldea (povoação). Parte de: Iaoa.
Cirebon, Indonesia (AS). Lat. -6.7063, long. 108.557.

Chiammay. *Chiãmay* 181(1), 182(1), 185(1). *Chiammay* 41(1), 70(1), 112(1), 124(1), 128(1), 166(1), 181(1), 182(3), 184(1). Reino (lago). Parte de: Indochina. Outras relações: #area#@Iangomaa@; #area#@Sião@; #area#@Quitiruão@.
Chiang Mai, Thailand (AS). Lat. 18.79038, long. 98.98468.

Chiãtabuu. *Chiãtabuu* 128(1). *Chintabu* 189(1). Costa (barra). Parte de: Sião.
Changwat Chanthaburi, Thailand (AS). Lat. 12.85798, long. 102.15434.

China. *Chim* 39(2), 45(2), 46(1), 50(1), 56(2), 58(2), 63(1), 65(1), 70(2), 71(1), 76(1), 83(1), 85(1), 89(1), 90(1), 91(1), 92(2), 94(2), 95(6), 96(1), 105(1), 113(1), 116(3), 119(1), 122(1), 133(2), 134(1), 137(1), 140(1), 180(1), 208(1), 215(1). *China* 1(1), 21(1), 22(1), 25(1), 26(2), 35(1), 36(1), 41(1), 43(1), 44(3), 45(2), 46(2), 47(1), 51(1), 52(1), 55(1), 57(2), 58(1), 62(1), 63(2), 64(2), 65(1), 66(1), 67(3), 70(2), 73(2), 77(1), 85(1), 88(3), 89(2), 90(1), 91(2), 92(2), 94(3), 95(4), 96(1), 97(1), 98(1), 99(6), 101(1), 103(1), 105(5), 107(1), 108(2), 110(1), 111(5), 112(2), 114(2), 116(1), 117(1), 121(1), 122(2), 123(2), 124(1), 127(1), 128(1), 129(1), 131(1), 132(2), 133(3), 134(1), 135(1), 136(1), 137(3), 140(3), 143(4), 151(1), 153(1), 158(1), 165(1), 166(1), 171(1), 172(1), 179(4), 180(1), 181(2), 183(1), 185(1), 189(3), 202(4), 203(2), 208(4), 210(1), 211(1), 213(1), 214(3), 215(16), 216(2), 217(1), 220(3), 221(4), 222(2). *Chins* 1(1), 40(1), 42(1), 44(1), 45(1), 46(1), 47(1), 48(1), 50(2), 52(1), 54(1), 55(3), 57(1), 58(2), 63(4), 64(4), 65(2), 66(2), 68(4), 70(2), 71(1), 72(1), 74(4), 75(1), 76(1), 78(2), 79(2), 81(1), 83(1), 88(4), 89(7), 90(2), 91(2), 92(2), 93(1), 94(2), 95(2), 96(5), 97(1), 103(2), 105(4), 107(2), 108(2), 109(5), 111(2), 117(2), 118(2), 119(3), 121(1), 123(1), 133(3), 134(2), 137(2), 140(2), 142(3), 143(2), 159(1), 166(2), 178(1), 179(4), 180(1), 202(1), 215(1), 220(1), 221(4). Império. Parte de: Ásia Oriental.
People’s Republic of China, China (AS). Lat. 35, long. 105.

Chincheo. *Chincheo* 44(1), 46(2), 55(1), 57(3), 66(1), 132(2), 179(1), 181(1), 203(1), 221(2). Porto. Parte de: China.
Zhangzhou, China (AS). Lat. 24.51333, long. 117.65556.

Chintaleuhos. *Chintaleuhos* 41(1). Reino. Parte de: Ásia. Outras relações: #area#@Tinacoreu@; #area#@Chiammay@; #area#@Moncalor@.

Chipanocão. *Chipanocão* 159(1). Enfermaria (enfermaria). Parte de: Calaminhan. Outras relações: #12 léguas rio acima#@Meidur@; #area#@Angegumaa@; #13 dias rio acima#@Singilapau@; #13 dias rio acima a partir de#@Catammaas@.

Chitor. *Chitor* 124(1). Reino. Parte de: India. Outras relações: #area#@Mogores@.
Chittaurgarh, India (AS). Lat. 24.88963, long. 74.62403.

Choromandel. *Choromandel* 3(1), 149(1), 162(1). Região. Parte de: India.
Coromandel Coast, India (AS). Lat. 13.39938, long. 80.34736.

Cincaapura. *Cincaapura* 43(2). *Cincapura* 26(1). *Sincaapura* 220(1). Estreito. Parte de: Insulíndia. Outras relações: #area#@Malaca@; #area#@Sabaom@.
Republic of Singapore, Singapore (AS). Lat. 1.36667, long. 103.8.

Cochim. *Cochim* 2(1), 12(1), 217(1), 223(1), 225(1). *Coochim* 218(1). Porto. Parte de: India.
Cochin, India (AS). Lat. 9.93988, long. 76.26022.

Çofalla. *Çofalla* 185(1). Fortaleza. Parte de: Moçambique.

Cofilem guaxy. *Cofilem guaxy* 100(1). Casa (casa de misericórdia). Parte de: Pequim.

Cohilouza. *Cohilouza* 96(2). *Cohilouzaa* 96(1). Cidade (ruínas). Parte de: China. Outras relações: #area#@Batampina@; #10 léguas#@Mindoo@; #1 légua#@Bigay potim@.

Colem. *Colem* 47(1). Povoação. Parte de: Cauchenchina. Outras relações: #area#@Panduree@; #area#@Mutipinão@; #area#@Tilaumera@.

Comhay. *Comhay* 44(2), 55(1), 81(1), 82(1), 129(1), 189(1). Porto (montes). Parte de: China. Outras relações: #area#@Manaquileu@; #area#@China@; #area#@Cauchenchina@; #area#@Cauchenchina, enseada@; #area#@ilha dos ladroẽs@; #area#@Quoamão@.
Guanghai, China (AS). Lat. 21.96164, long. 112.79307.

Comolem. *Comolem* 66(1). Ilhas (ilhas). Parte de: China. Outras relações: #5 dias náuticos#@Nouday@; #6 dias náuticos#@Liampoo@.

Comorim. *Comorim* 213(1). Porto. Parte de: India.
Kanniyākumāri, India (AS). Lat. 8.09008, long. 77.53841.

Congrau. *Congrau* 41(1). Lugar. Parte de: Champaa. Outras relações: #area#@Tinacoreu@; #area#@Taiquilleu@.

Constantinopla. *Constantinopla* 3(1), 107(1). Cidade. Parte de: Turcos.
İstanbul, Turkey (AS). Lat. 41.01384, long. 28.94966.

Conxinacau. *Conxinacau* 79(1). Mina. Parte de: China. Outras relações: #coordenadas#@41.66@; #area#@Nanquim, enseada@.

Coraçone. *Coraçone* 124(1). *Coraçones* 95(1), 114(1), 149(1). Nação. Parte de: Ásia. Outras relações: #area#@Persia@; #area#@Mogores@.
Ostān-e Khorāsān-e Raẕavī, Iran (AS). Lat. 35.25, long. 59.

Cordoua. *Cordoua* 208(1). Cidade. Parte de: Castella.
Córdoba, Spain (EU). Lat. 37.89155, long. -4.77275.

Coretumbagâ. *Coretumbagâ* 96(1). Serra. Parte de: China. Outras relações: #area#@Batampina@; #2 léguas#@Mindoo@; #1 légua#@Bigay potim@.

Corpilem. *Corpilem* 78(1). Cidade. Parte de: China. Outras relações: #area#@Calempluy@; #area#@Sileupamor@; #area#@Tanquilem@; #area#@Fumbana@.

Cosmim. *Cosmim* 88(1), 128(1), 147(1), 150(1), 168(1), 171(2), 190(1), 200(1). Cidade (rio, barra, porto). Parte de: Pegù. Outras relações: #area#@Batobasoy@.
Pathein, Myanmar [Burma] (AS). Lat. 16.77919, long. 94.73212.

Couilham. *Couilham* 215(1). Vila. Parte de: Portugal.
Covilhã, Portugal (EU). Lat. 40.28601, long. -7.50396.

Coulaam. *Coulaam* 150(1). *Coulão* 193(1). Cidade. Parte de: Pegù.

Coutasarem. *Coutasarem* 190(1). Campo. Parte de: Pegù. Outras relações: #1 légua#@Pegù, cidade@.

Çunda. *Çũda* 26(1), 178(1), 215(1). *Çunda* 26(1), 46(1), 55(1), 57(2), 66(1), 172(5), 173(3), 174(1), 175(1), 176(3), 179(3), 180(1), 181(3), 220(1). *Sunda* 21(1). *çunda* 20(1). Reino (porto). Parte de: Iaoa.
Banten, Indonesia (AS). Lat. -6.5, long. 106.25.

Cunebetee. *Cunebetee* 41(1), 128(1). Lago. Parte de: Ásia. Outras relações: #area#@Tinacoreu@; #area#@Chiammay@.

Curia. *Curia* 3(1). Ilha. Parte de: Arabia Felix. Outras relações: #area#@Abedalcuria; #area#@Muria@; #area#@Çacotorà@.

Çurrate. *Çurrate* 43(1). Porto. Parte de: India.
Sūrat, India (AS). Lat. 21.19594, long. 72.83023.

Curuche. *Curuche* 117(1). Vila. Parte de: Portugal.

Cutamuilau. *Cutamuilau* 163(1). Antesala (antesala). Parte de: Timplão.

Cuy. *Cuy* 88(1), 95(1), 189(1), 220(1), 222(1). Barra. Parte de: Sião. Outras relações: #area#@Tauquiday@; #130 léguas#@Patane@.

Daanuu. *Daanuu* 12(1). Picos. Parte de: India. Outras relações: #area#@Chaul@; #area#@Diu@.

Dãbambuu. *Dãbambuu* 128(1). *Dambambuu* 150(1). Região. Parte de: Índico Oriental. Outras relações: #area#@Merguim@; #area#@Angegumaa@; #area#@Lauhos@; #area#@Gueos@.

Dabul. *Dabul* 8(2), 12(1), 37(1). Porto (barra). Parte de: India.
Dābhol, India (AS). Lat. 17.58971, long. 73.18001.

Dalaa. *Dalaa* 188(1), 190(2), 193(1). Cidade. Parte de: Pegù.
Dala, Myanmar [Burma] (AS). Lat. 16.75829, long. 96.16088.

Damaõ. *Damaõ* 2(1). Cidade. Parte de: India.
Daman, India (AS). Lat. 20.41432, long. 72.83236.

Danapluu. *Danapluu* 153(1), 188(1), 190(2). Cidade (comarca). Parte de: Pegù. Outras relações: #area#@Ansedaa@.
Danubyu, Myanmar [Burma] (AS). Lat. 17.25684, long. 95.59234.

Dayaa. *Dayaa* 31(1). Senhorio (senhorio). Parte de: Achem. Outras relações: #area#@Achem@.
Lampu Daya, Indonesia (AS). Lat. 5.5397, long. 95.3014.

Dely. *Dely* 124(1). Reino. Parte de: India.
Delhi, India (AS). Lat. 28.65195, long. 77.2315.

Demaa. *Demâ* 178(1). *Demaa* 36(1), 107(1), 172(1), 173(2), 174(1), 175(1), 177(4), 178(1), 179(1). Reino (cidade, porto). Parte de: Iaoa.

Demak, Indonesia (AS). Lat. -6.8909, long. 110.6396.

Digum. *Degum* 148(1), 168(1). *Digum* 164(1), 188(1), 190(2), 193(1). Cidade (rio). Parte de: Pegù. Outras relações: #area#@Meidoo@; #area#@Dalaa@.
Yangon, Myanmar [Burma] (AS). Lat. 16.80528, long. 96.15611.

Dinamarca. *Dinamarca* 126(1). Reino. Parte de: Europa.
Kingdom of Denmark, Denmark (EU). Lat. 56, long. 10.

Diu. *Diu* 2(5), 3(3), 4(1), 7(1), 8(1), 12(4), 20(2), 43(1). Fortaleza (fortaleza). Parte de: India.
Diu, India (AS). Lat. 20.71405, long. 70.98224.

Dumclee. *Dumclee* 164(1). Lugar. Parte de: Ásia. Outras relações: #area#@Calaminhan@.

Egypto. *Egypto* 107(1). Império. Parte de: Africa.
Arab Republic of Egypt, Egypt (AF). Lat. 27, long. 30.

Espanha. *Espanha* 68(1). Reino. Parte de: Europa.
Kingdom of Spain, Spain (EU). Lat. 40, long. -4.

Etiopia. *Abexim* 5(2), 20(1), 26(3), 27(1). *Abexins* 4(2), 16(1), 26(1), 149(1), 161(1), 178(1), 185(1), 186(1), 225(1). *Ethiopia* 20(1), 89(1). *Ethyopia* 4(1). *Etiopia* 1(1), 161(1), 225(1). Império. Parte de: Africa.
Federal Democratic Republic of Ethiopia, Ethiopia (AF). Lat. 9, long. 39.5.

Euora. *Euora* 185(1), 216(1). Vila. Parte de: Portugal.
Distrito de Évora, Portugal (EU). Lat. 38.58333, long. -7.83333.

Europa. *Europa* 99(1), 107(2), 122(1), 159(1), 224(1). Continente. Parte de: Terra.
Western Europe, (EU). Lat. 50.21909, long. 7.42676.

Facataa. *Facataa* 135(1), 136(2), 202(1), 218(1). Cidade (porto). Parte de: Iapaõ. Outras relações: #area#@Bungo@.
Hakata Ku, Japan (AS). Lat. 33.57989, long. 130.44351.

Faleu. *Faleu* 164(1). Serra. Parte de: Auaa. Outras relações: #area#@Iatir@; #area#@Pontau@.

Famstir. *Famstir* 117(1), 131(1). *Fãostir* 88(1). Cidade (lago). Parte de: Tartaria. Outras relações: #9 léguas#@Lançame@.

Fanaugrem. *Fanaugrẽ* 129(1), 131(1). *Fanaugrem* 129(3), 130(1), 131(2). Vila. Parte de: Cauchenchina. Outras relações: #106 léguas#@Tinamquaxy@; #6 léguas#@Benau@; #1 légua#@Agimpur@#15 léguas#@Lindau panoo@.

Fancleu. *Fancleu* 197(1). Povoação. Parte de: Pegù. Outras relações: #1 légua#@Potem@; #area#@Arracão@.

Fanjus. *Fanjus* 71(1), 87(1). *Fanjùs* 85(1), 126(1). Reino (enseada, morro). Parte de: China. Outras relações: #metropoli#@Nanquim@.

Fiancima. *Fiancima* 135(1), 210(1), 211(1). Colégios. Parte de: Iapaõ. Outras relações: #area#@Bungo@.

Fingau. *Fingau* 223(1). Povoação (lugar). Parte de: Iapaõ. Outras relações: #0.25 léguas#@Osquy@.

Finge. *Finge* 209(3). Porto (rio). Parte de: Bungo. Outras relações: #area#@Fucheo@.
Hiji, Japan (AS). Lat. 33.37081, long. 131.53025.

Finginilau. *Finginilau* 84(1). Lugar. Parte de: China. Outras relações: #4 léguas#@Xianguulee@; #area#@Nanquim@; #area#@Taypor@; #area#@Chautir@; #area#@Guinapalir@.

Firando. *Firãdo* 225(1). *Firando* 66(1), 208(4). Reino (cidade). Parte de: Iapaõ. Outras relações: #100 léguas#@Canguexumaa@.
Hirado, Japan (AS). Lat. 33.36853, long. 129.55247.

Fiungaa. *Fiũgaa* 223(1). *Fiungaa* 135(1), 218(1). *Fiunguaa* 202(1). Porto. Parte de: Iapaõ. Outras relações: #1dia nautico#@Minato@; #area#@Osquy@; #area#@Tanoraa@.

Fiunganorsee. *Fiunganorsee* 96(1), 97(1). Cidade (ruínas). Parte de: China. Outras relações: #area#@Batampina@; #10 léguas#@Mindoo@; #1 légua#@Bigay potim@.

Florença. *Florença* 164(1). *Florentino* 164(1). Cidade. Parte de: Europa.
Florence, Italy (EU). Lat. 43.77925, long. 11.24626.

Framengos. *Framengos* 124(1). Nação. Parte de: Europa.
Flanders, Belgium (EU). Lat. 51, long. 4.5.

França. *Frãcesa* 20(1). *França* 1(1), 16(1), 158(1). *Frances* 1(1). *Franceses* 16(1), 20(1). Reino. Parte de: Europa.
Republic of France, France (EU). Lat. 46, long. 2.

Freixo de espada cinta. *Freixo de espada cinta* 200(1). Vila. Parte de: Portugal.
Freixo de Espada À Cinta, Portugal (EU). Lat. 41.08398, long. -6.82992.

Frenojama. *Frenojama* 211(3). Lugar. Parte de: Iapaõ.

Friucaranjaa. *Friucaranjaa* 166(1). Província. Parte de: Ásia. Outras relações: #area#@Calaminhan@; #area#@Pauel@; #area#@Calogẽs@#; #area#@Fungaos@#.

Fucanxi. *Fucanxi* 143(1). Ilha. Parte de: Iapaõ. Outras relações: #area#@Sesirau@; #area#@Goto@; #area#@Pollem@.

Fucheo. *Fucheo* 134(1), 135(1), 136(1), 137(1), 200(1), 201(1), 202(1), 208(1), 209(1), 211(1), 213(1), 214(1), 223(1), 224(1), 225(2). Porto (costa). Parte de: Iapaõ. Outras relações: #area#@Sumbor@.

Fucheo. *Fucheo* 60(1). Cidade (metrópole). Parte de: China. Outras relações: #capital#@Bungo@; #7 léguas#@Osquy@; #70 léguas#@Facataa@; #7 dias náuticos#@Tanixumaa@.

Fumbacor. *Fumbacor* 182(1). Lugar. Parte de: Guibem.

Fumbana. *Fumbana* 78(1). Cidade. Parte de: China. Outras relações: #area#@Calempluy@; #area#@Sileupamor@; #area#@Tanquilem@; #area#@Corpilem@.

Fumbau. *Fumbau* 4(1). Povoação. Parte de: Etiopia. Outras relações: #2 léguas#@Gileytor@; #2.5 dias#@Vangaleu@.

Fungaos. *Fungaos* 166(1). Nação. Parte de: Ásia. Outras relações: #area#@Calogẽs@; #area#@Friucaranjaa@; #area#@Pauel@.

Gaborem. *Gaborem* 221(1). Lugar. Parte de: Armenia.

Gãges. *Gãges* 128(1). Rio. Parte de: Bengala.
Ganges River, Bangladesh (AS). Lat. 23.36667, long. 90.53333.

Gàle. *Gàle* 20(1). Porto. Parte de: Ceilão.
Galle, Sri Lanka (AS). Lat. 6.0367, long. 80.217.

Galego. *Galega* 2(1). *Galego* 105(1), 185(1), 191(1), 204(2), 217(1). Reino. Parte de: Europa.
Galicia, Spain (EU). Lat. 42.75508, long. -7.86621.

Gangitanou. *Gangitanou* 73(2). Serra. Parte de: Ásia Oriental. Outras relações: #area#@Gigauhos@; #area#@Botinafau@; #area#@Paatebenam@; #area#@Nanquim, enseada@.

Geilolo. *Geilolo* 20(1). Ilha. Parte de: Insulíndia. Outras relações: #area#@Ternate@; #area#@Banda@.
Pulau Halmahera, Indonesia (AS). Lat. 0.73558, long. 127.9515.

Gigauhos. *Gigaos* 126(1). *Gigauhos* 73(3). Reino. Parte de: Ásia Oriental. Outras relações: #area#@Gangitanou@; #area#@Fanjus@; #area#@Paatebenam@; #area#@Nanquim, enseada@.

Gileytor. *Gileytor* 4(3). Fortaleza. Parte de: Etiopia. Outras relações: #2 léguas#@Fumbau@; #area#@Vangaleu@; #area#@Arquico@.

Ginafógaos. *Ginafógaos* 166(1). Nação. Parte de: Ásia. Outras relações: #area#@Surobasoy@.

Ginocoginana. *Ginocoginana* 96(1). Pagode (pagode). Parte de: Pegù.

Gizares. *Gizares* 95(1), 149(1). Nação (barra). Parte de: Persia.
Shatt al Arab, Iran (AS). Lat. 29.94036, long. 48.59596.

Gizom. *Gizom* 92(1). Templo (mítico). Parte de: Guantipocau.

Goa. *Goa* 2(3), 4(1), 7(2), 8(5), 10(1), 11(1), 12(4), 20(1), 26(1), 147(1), 171(1), 172(1), 176(1), 203(1), 208(3), 215(1), 216(1), 217(7), 218(3), 219(1), 223(1), 225(1), 226(1). Cidade (barra). Parte de: India.
Goa, India (AS). Lat. 15.33333, long. 74.08333.

Gocão. *Gocão* 5(1). Ponta. Parte de: Índico Ocidental. Outras relações: #area#@Arquico@;

#area#@Mocaa@.

Gofanjauserca. *Gofanjauserca* 100(1). Prisão. Parte de: Pequim.

Goncalidau. *Goncalidau* 124(1). Montes. Parte de: Ásia. Outras relações: #area#@Mogores@; #area#@Moscouia@.

Goto. *Goto* 57(1), 135(1), 143(1). Ilhas (ilhas). Parte de: Iapaõ. Outras relações: #area#@Sesirau@; #area#@Fucanxi@; #area#@Pollem@.
Gotō Rettō, Japan (AS). Lat. 32.83333, long. 129.

Gotom. *Gotom* 137(1). Parcel. Parte de: China. Outras relações: #area#@Liampoo@; #area#@Lequios@.

Gotor. *Gotor* 4(1). Porto. Parte de: Etiopia. Outras relações: #1 légua#@Massuaa@.

Gouro. *Gouro* 107(1). Cidade. Parte de: Bengala. Outras relações: #capital#@Bengala@.
Gaur, India (AS). Lat. 24.86878, long. 88.12639.

Gregos. *Grega* 3(1). *Grego* 6(1), 187(1). *Gregos* 146(1), 149(1), 206(1). Nação. Parte de: Europa.
Hellenic Republic, Greece (EU). Lat. 39, long. 22.

Guamboy. *Guamboy* 44(1). Porto. Parte de: Ainão. Outras relações: #area#@Buhaquirim@; #area#@Camoy@.

Guampalaor. *Guampalaor* 170(1). Campo. Parte de: Sauady.

Guampanoo. *Guampanoo* 158(2). Esteiro. Parte de: Índico Oriental. Outras relações: #area#@Angegumaa@; #area#@Queitor@; #area#@Auaa@; #area#@Guatelday@.

Guantipocau. *Guantipocau* 92(1). Reino (mítico). Parte de: Ásia.

Guanxiroo. *Guanxiroo* 200(1). Cidade. Parte de: Tanixumaa.

Guateamgim. *Guateamgim* 14(1). Rio. Parte de: Çamatra. Outras relações: #area#@Panaajù@; #area#@ Batorrendão@.

Guatelday. *Guatelday* 158(1). Povoação. Parte de: Índico Oriental. Outras relações: #area#@Guampanoo@; #area#@Angegumaa@; #area#@Auaa@; ; #area#@Queitor@.

Guatipamor. *Guatipamor* 126(1). Estudos (estudos). Parte de: Tartaria. Outras relações: #1 dia#@Tuymicão@; #1 dia fluvial#@Puxanguim@.

Guauxitim. *Guauxitim* 123(1). Cidade. Parte de: China. Outras relações: #160 léguas#@Pequim@; #12 horas#@ Caixiloo@; #area#@Singrachirau@.

Guaytor. *Guaytor* 126(1). Lugar (mítico). Parte de: Moscouia. Outras relações: #area#@Tartaria@.

Gueos. *Gucos* 112(1). *Gueos* 1(1), 41(1), 47(1), 124(2), 128(1), 181(1). Reino. Parte de: Indochina. Outras relações: #area#@Tanguu@; #area#@Sião@.

Guibem. *Guibem* 182(3). Reino. Parte de: Índico Oriental. Outras relações: #15 léguas#@Quitiruão@; #area#@Sião@; #area#@Capimper@; #area#@Chiammay@.

Guijampee. *Guijampee* 123(1). Cidade. Parte de: China. Outras relações: #2 dias#@Pequim@; #6 horas#@Liampeu@; #area#@Quaytragum@.

Guimaraẽs. *Guimaraẽs* 2(1). Cidade. Parte de: Portugal.
Guimarães, Portugal (EU). Lat. 41.44444, long. -8.29619.

Guimpel. *Guimpel* 107(1). Cidade. Parte de: Siammon.

Guinacoutel. *Guinacoutel* 190(1). Povoação (lugar). Parte de: Pegù. Outras relações: #40 léguas#@Pegù, cidade@.

Guinapalir. *Guinapalir* 84(1). Lugar. Parte de: China. Outras relações: #area#@Nanquim@; #area#@Taypor@; #area#@Xianguulee@; #area#@Chautir@.

Guinaytaraõ. *Guinaytaraõ* 74(1), 75(1). Ponta. Parte de: China. Outras relações: #area#@Calempluy@; #area#@Tanquilem@.

Guintoo. *Guintoo* 55(1). Ilhéu. Parte de: China. Outras relações: #12 horas#@ilha dos ladroẽs@; #18 léguas#@Xinguau@; #18 léguas#@Xamoy@.

Guitor. *Guitor* 182(1). Cidade (metrópole). Parte de: Guibem. Outras relações: #capital#@Guibem@.

Gumbim. *Gumbim* 158(1). Cidade. Parte de: Iangomaa. Outras relações: #area#@Queitor@; #7 dias fluviais acima#@Angegumaa@; #7 dias fluviais acima#@Catammaas@.

Gundexilau. *Gundexilau* 139(1). Vila. Parte de: Lequios. Outras relações: #6 horas#@Pongor@.

Guntaleu. *Guntaleu* 183(1). Cidade. Parte de: Sião.

Guzarates. *Guzarate* 37(1), 57(1). *Guzarates* 16(1), 26(1), 27(1), 146(2), 149(1). Nação. Parte de: India.
State of Gujarāt, India (AS). Lat. 23, long. 71.75.

Hiamangoo. *Hiamãgoo* 202(1). *Hiamangoo* 135(1), 202(2), 203(1). Porto (rio, angra). Parte de: Iapaõ. Outras relações: #area#@Canguexumaa@.
Yamagawa Kō, Japan (AS). Lat. 31.20832, long. 130.63245.

Hicanduré. *Hicanduré* 14(1). Rio. Parte de: Çamatra. Outras relações: #area#@Mediterraneo@; #area#@Peedir@.

Hicanduree. *Hicanduree* 173(1). Rio. Parte de: Iaoa. Outras relações: #area#@Passaruão@.

Hifaticau. *Hifaticau* 222(1). Templo. Parte de: Cantão.

Hiquegens. *Hiquegens* 85(1). Cidade. Parte de: Moscouia. Outras relações: #area#@Alemanha@.

Huzamguee. *Huzamguee* 129(2), 131(1), 132(2). *Huzanguee* 130(1), 131(2), 132(1). *Huzanquee*

130(1). *Vzanguee* 107(1), 112(1), 125(1). Cidade (império, metrópole). Parte de: Cauchenchina. Outras relações: #capital#@Cauchenchina@.
Hanoi, Vietnam (AS). Lat. 21.0245, long. 105.84117.

Iacuçalão. *Iacuçalão* 157(1). *Iaquesaloẽs* 149(1). Nação. Parte de: Índico Oriental. Outras relações: #area#@Chaleu@; #area#@Queitor@.

Iacur. *Iacur* 13(2), 14(1). Povoação (lugar). Parte de: Çamatra. Outras relações: #area#@Bata@; #area#@Achem@.

Iambee. *Iambè* 19(1). *Iambee* 20(1), 24(1). *Iambes* 16(1). Reino (rio). Parte de: Çamatra. Outras relações: #area#@Campar@.
Provinsi Jambi, Indonesia (AS). Lat. -1.5, long. 103.

Iangomaa. *Iangomaa* 150(1), 158(2). *Iangumaa* 167(1). Reino. Parte de: Índico Oriental. Outras relações: #area#@Angegumaa@.

Iantana. *Iantana* 30(2), 31(2), 32(5), 51(1), 220(1). *Iãtana* 32(1), 207(1). Reino. Parte de: Insulíndia. Outras relações: #area#@Malaca@; #area#@Aarù@; #area#@Achem@.

Iaoa. *Iaoa* 20(1), 26(1), 33(1), 36(1), 38(1), 39(1), 107(1), 144(1), 172(1), 173(2), 176(1), 177(1), 179(2), 180(1), 189(1). *Iaos* 48(1), 57(1), 59(1), 145(1), 149(1), 177(1), 180(1), 185(1), 186(1). Ilha. Parte de: Insulíndia.
Java, Indonesia (AS). Lat. -7.49167, long. 110.00444.

Iapaõ. *Iapão* 40(1), 92(1), 133(1), 134(2), 200(2), 201(1), 208(1), 211(1), 215(3), 219(2), 220(1). *Iapaõ* 21(1), 26(1), 60(1), 66(1), 107(1), 112(1), 132(1), 136(1), 137(1), 143(3), 161(1), 171(1), 181(1), 201(1), 202(5), 203(5), 208(3), 210(1), 211(1), 212(1), 213(1), 214(1), 218(2), 220(2), 221(2), 222(1), 223(2), 225(1), 226(2). *Iapoa* 208(1), 212(1), 213(1). *Iapoẽs* 119(1), 134(2), 135(3), 159(1), 166(1), 178(1), 179(1), 200(1), 203(1), 208(1), 209(1), 211(1), 223(1). *Japão* 132(1). *Japaõ* 200(1). Reino (terra, ilha). Parte de: Ásia Oriental.
Japan, Japan (AS). Lat. 35.68536, long. 139.75309.

Iapara. *Iapara* 36(1), 172(1), 173(2), 179(1). Cidade (porto). Parte de: Iaoa. Outras relações: #5 léguas#@Demaa@.
Jepara, Indonesia (AS). Lat. -6.5924, long. 110.671.

Iatir. *Iatir* 164(1). Lugar (serra). Parte de: Auaa. Outras relações: #area#@Faleu@; #area#@Pontau@.

Ierusalem. *Ierusalem* 4(1), 5(1), 20(1). Cidade. Parte de: Ásia.
Jerusalem, Israel (AS). Lat. 31.75, long. 35.

ilha das naos. *ilha das naos* 203(1). Ilha. Parte de: Malaca.
Pulau Jawa, Malaysia (AS). Lat. 2.1797, long. 102.2503.

ilha do fogo. *ilha do fogo* 138(1), 140(1). Ilha. Parte de: Lequios.

ilha do ouro. *ilha do ouro* 13(1), 14(1), 20(1). Ilha. Parte de: Oceano.

ilha dos cocos. *ilha dos cocos* 50(1). Ilha. Parte de: Ásia Oriental. Outras relações: #12 dias náuticos#@Mutipinão@; #area#@Ainão, enseada@; #area#@Cauchenchina, enseada@; #area#@Madel@.

ilha dos ladroẽs. *ilha dos ladroẽs* 53(1), 55(1). *ilha que se dezia dos ladroẽs* 53(1). Ilha. Parte de: Ásia. Outras relações: #area#@Cauchenchina, enseada@; #area#@Quoamão@; #area#@Comhay@; #260 léguas#@Liampoo@; #12 horas#@Guintoo@.

Ilher. *Ilher* 29(1). Campo. Parte de: Malaca.
Kampung Hilir, Malaysia (AS). Lat. 2.1871, long. 102.2614.

India. *India* 1(4), 2(4), 3(4), 4(5), 6(1), 7(3), 12(1), 14(1), 15(2), 16(2), 18(2), 20(2), 21(1), 22(1), 26(2), 30(1), 32(1), 36(1), 41(1), 43(1), 52(2), 53(1), 57(1), 65(1), 75(1), 96(2), 131(1), 137(1), 143(2), 146(1), 147(3), 148(2), 151(1), 158(1), 162(1), 172(1), 176(1), 183(1), 185(1), 189(2), 196(1), 200(2), 203(4), 208(4), 215(5), 216(3), 218(2), 219(7), 221(2), 223(1), 224(1), 225(1), 226(5). Região. Parte de: Ásia.
Republic of India, India (AS). Lat. 22, long. 79.

Iudaa. *Iudà* 26(1). *Iudaa* 3(1), 20(2), 37(1), 43(1), 158(1), 206(1). Porto. Parte de: Arabia Felix.
Jeddah, Saudi Arabia (AS). Lat. 21.54238, long. 39.19797.

Iudeu. *Iudeu* 4(1), 6(2), 203(1). *Iudeus* 4(1), 149(1). Nação. Parte de: Ásia. Outras relações: #area#@Vangaleu@; #area#@Toro@; #area#@Cayro@.
State of Israel, Israel (AS). Lat. 31.5, long. 34.75.

Iunçalão. *Iũçalão* 146(1). *Iuncalão* 185(1). *Iunçalão* 19(1), 153(1), 189(1), 205(1). *Iunçalaõ* 144(1). Porto (costa). Parte de: Sião. Outras relações: #area#@Tanauçarim@.
Phuket, Thailand (AS). Lat. 7.89059, long. 98.3981.

Iuncay. *Iuncay* 147(1). Porto. Parte de: Sião. Outras relações: #area#@Merguim@; #area#@Tanauçarim@; #area#@Vagaruu@; #area#@Pullo Camude@.

Iunquileu. *Iunquileu* 90(2). Vila. Parte de: China. Outras relações: #area#@Batampina@; #1 dia#@Xinligau@; #area#@Sampitay@.

Iunquinilau. *Iunquinilau* 97(2). Cidade. Parte de: China. Outras relações: #area#@Batampina@; #area#@Pequim@.

Iuropisaõ. *Iuropisaõ* 185(1). Cidade. Parte de: Sião. Outras relações: #area#@Iunçalão@.

Laa. *Laa* 150(1). Senhorio (senhorio). Parte de: Índico Oriental. Outras relações: #area#@Bramá@; #area#@Sauady@; #area#@Iangomaa@; #area#@Çatão@; #area#@Martauão@.

Lailoo. *Lailoo* 55(1), 57(1), 58(3), 132(1). Porto. Parte de: China. Outras relações: #5 dias náuticos, - 8 léguas#@Chincheo@; #3dias náuticos#@Lailoo@.
Xiamen Gang, China (AS). Lat. 24.41861, long. 118.07694.

Lamau. *Lamau* 44(1), 51(1), 52(1), 56(2), 62(1), 86(1), 115(1), 132(1), 140(1), 143(1), 203(1). Porto (costa, ilhéus). Parte de: China. Outras relações: #area#@Sanchão@; #area#@Lailoo@.
Nan'ao Dao, China (AS). Lat. 23.44043, long. 117.08464.

Lampacau. *Lampacau* 132(1), 133(1), 221(3), 226(2). *Lãpacau* 223(1). Porto (ilha). Parte de: China. Outras relações: #7 léguas#@Sanchão@; #40 dias#@Goa@.

Lampom. *Lampom* 20(1). Rio. Parte de: Çamatra. Outras relações: #area#@Campar@; #area#@Menancabo@.

Lançame. *Lãçame* 119(1). *Lançame* 88(1), 107(1), 117(2), 118(3), 123(2), 124(2). Cidade. Parte de: Tartaria. Outras relações: #capital#@Tartaria@; #9 léguas#@Famstir@; #6 dias fluviais#@Xipator@.

Lantau. *Lantau* 73(1). Cidade (alfandega). Parte de: China.

Lantor. *Lantor* 183(2). Cidade (fortaleza). Parte de: Sião. Outras relações: #area#@Passiloco@.

Larache. *Larache* 1(1). Porto. Parte de: Africa.
Larache, Morocco (AF). Lat. 35.19321, long. -6.15572.

Lasaparà. *Lasaparà* 38(1). Lugar. Parte de: Iaoa. Outras relações: #area#@Quaijuão@.

Latinas. *Latinas* 7(2). *latinas* 204(1). *Latinos* 124(1). Região. Parte de: Europa.
Lazio, Italy (EU). Lat. 42.07762, long. 12.77878.

Latiparau. *Latiparau* 129(1). Abadia (abadia). Parte de: Cauchenchina. Outras relações: #12 horas#@Lindau panoo@; #3 léguas#@Agimpur@; #4 léguas#@Fanaugrem@.

Laue. *Laue* 31(1). Rio. Parte de: Aarù.

Laue. *Laue* 35(1), 36(1), 39(1). Porto. Parte de: Iaoa. Outras relações: #area#@Tajampura@.

Lauhos. *Laos* 112(1), 128(1), 181(1). *Lauhos* 124(1), 146(2), 166(1), 167(1). *Lauhós* 95(1). *Lauhòs* 41(1), 47(1). Reino. Parte de: Indochina. Outras relações: #area#@Angegumaa@; #area#@Chiammay@; #area#@Surobasoy@; #area#@Gueos@; #area#@Sião@; #area#@Xenxinapau@.
Lao People's Democratic Republic, Laos (AS). Lat. 18, long. 105.

Laura. *Laura* 117(1). Vila. Parte de: Portugal.
Lavra, Portugal (EU). Lat. 41.25936, long. -8.71849.

Lautimey. *Lautimey* 120(1). Povoação. Parte de: China. Outras relações: #area#@Pequim@; #area#@Bumxay@; #area#@Pommitay@; #area#@Quansy@.

Lautir. *Lautir* 120(1). Castelo (castelo). Parte de: China. Outras relações: #area#@Pequim@; #area#@Bumxay@; #area#@Pommitay@; #area#@Palemxitau@.

Lechune. *Lechune* 88(1), 121(1), 127(1), 128(1). Cidade (rio). Parte de: Tartaria. Outras relações: #4 dias fluviais#@Quanginau@; #5 dias fluviais#@Rendacalem@; #area#@Famstir@.

Leibrau. *Leibrau* 185(1). Rio. Parte de: Indochina. Outras relações: #area#@Chiammay@; #area#@Sacotay@.

Lequimpau. *Lequimpau* 92(1). Povoação (lugar). Parte de: China. Outras relações: #area#@Batampina@; #area#@Sampitay@; #1 dia fluvial#@Pacão@; #1 dia fluvial#@Nacau@; #5 léguas#@Tuxenguim@.

Lequios. *Elequios* 1(1). *Lequia* 133(1), 140(1), 143(6). *Lequias* 143(1). *Lequîo* 107(1). *Lequio* 143(1). *Lequios* 26(1), 39(1), 41(1), 56(1), 68(1), 112(1), 132(1), 133(1), 134(1), 137(1), 138(1), 141(1), 143(2), 151(1), 159(1), 178(1), 184(1), 189(1), 224(1), 225(1). *Lequio grande* 138(1). Reino (ilha). Parte de: Ásia Oriental. Outras relações: #area#@China@; #area#@Banchaa@.
Taiwan, Taiwan (AS). Lat. 26.5, long. 128.

Leysacotay. *Leysacotay* 88(1). Rio. Parte de: Ásia Oriental. Outras relações: #area#@Xinxipou@; #area#@Moscouia@.

Liampeu. *Liampeu* 123(1). Serra. Parte de: China. Outras relações: #2 dias#@Pequim@; #6 horas#@Guijampee@; #area#@Quaytragum@.

Liampoo. *Lãpoo* 45(1). *Liampoo* 40(1), 44(1), 51(1), 52(1), 55(4), 56(2), 57(3), 60(4), 61(2), 66(5), 67(2), 68(2), 70(2), 71(3), 76(1), 81(1), 85(1), 86(1), 91(1), 100(1), 137(1), 143(2), 144(1), 208(1), 221(2). *Liãpoo* 40(1), 46(1), 57(1), 60(1), 64(1), 67(1), 68(1), 71(1), 74(1), 137(1), 140(1), 144(1), 221(1). Cidade (porto, mar). Parte de: China.
Ningbo, China (AS). Lat. 29.87819, long. 121.54945.

Lindau. *Lindau* 39(1). Lugar. Parte de: Tosa.

Lindau panoo. *Lindau panoo* 129(1). Vila. Parte de: Cauchenchina. Outras relações: #12 horas#@Taraudachit@; #15 léguas#@Fanaugrem@; #12 horas#@Latiparau@.

Lingator. *Lingator* 131(1). Cidade. Parte de: Cauchenchina. Outras relações: #54 léguas#@Mecuy@; #area#@Baguetor@.

Lingau. *Lingau* 13(2), 14(1). Povoação (lugar). Parte de: Çamatra. Outras relações: #area#@Bata@; #area#@Achem@; #area#@Iacur@.

Lingua. *Lingaa* 31(1). *Lingua* 176(1), 179(1). Ilha. Parte de: Insulíndia. Outras relações: #area#@Bintão@; #area#@Iantana@; #area#@Andraguiree@.
Pulau Lingga, Indonesia (AS). Lat. -0.17493, long. 104.69526.

Linxau. *Linxau* 126(1). Cidade. Parte de: Tartaria. Outras relações: #1 dia fluvial#@Puxanquim@; #5 dias fluviais#@Singuafatur@.

Lisboa. *Lisboa* 1(3), 2(1), 66(1), 67(1), 89(1), 107(1), 108(1), 226(2). Cidade (metrópole). Parte de: Portugal.
Lisbon, Portugal (EU). Lat. 38.71667, long. -9.13333.

Londres. *Londres* 107(1). Cidade (metrópole). Parte de: Europa.

London, United Kingdom (EU). Lat. 51.50853, long. -0.12574.

Lugor. *Lugor* 36(4), 38(1), 50(1), 132(1), 183(1), 189(1). *Lúgor* 220(1). Cidade (barra). Parte de: Sião.
Nakhon Si Thammarat, Thailand (AS). Lat. 8.43333, long. 99.96667.

Lunçor. *Lunçor* 167(1). Aldea. Parte de: Calaminhan. Outras relações: #1 dia fluvial#@Pauel@; #9 dias fluviais#@Ventinau@.

Lusoẽs. *Lusaõ* 28(1). *Lusoẽs* 16(1), 17(1), 26(1), 57(1), 59(1), 145(1), 149(1), 173(1), 186(1). Nação. Parte de: Borneo.
Luzon, Philippines (AS). Lat. 16, long. 121.

Luxitay. *Luxitay* 55(1). Ilha. Parte de: China. Outras relações: #3 dias#@Pullo Quirim@; #area#@Xamoy@; #area#@Luxitay@; #area#@Lamau@.

Macao. *Macao* 221(1). Porto. Parte de: China.
Macau, Macao (AS). Lat. 22.20056, long. 113.54611.

Macassar. *Macassar* 1(1). Reino. Parte de: Insulíndia.
Makassar, Indonesia (AS). Lat. -5.14861, long. 119.43194.

Machão. *Machão* 188(1). Campo. Parte de: Pegù. Outras relações: #2 léguas#@Pegù, cidade@.

Madel. *Madel* 50(3), 52(2). Porto (rio). Parte de: Ainão.

Madeyra. *Madeyra* 2(1), 20(1), 35(1). Ilha. Parte de: Portugal.
Ilha da Madeira, Portugal (EU). Lat. 32.73333, long. -17.

Madur. *Madur* 167(1). Esteiro. Parte de: Ventinau. Outras relações: #area#@Magadaleu@; #5 dias fluviais#@Mouchel@; #5 dias fluviais#@Pegù@; #8 dias fluviais#@Martauão@.

Madura. *Madura* 172(1), 177(1), 179(1), 189(1). Ilha. Parte de: Insulíndia. Outras relações: #area#@Iaoa@; #area#@Angenia@; #area#@Bale@.
Madura, Indonesia (AS). Lat. -7, long. 113.33333.

Magadaleu. *Magadaleu* 167(1). Cidade. Parte de: Índico Oriental. Outras relações: #5 dias fluviais#@Penauchim@; #area#@Ventinau@; #area#@Iangomaa@; #13 dias fluviais#@Martauão@.

Magores. *Magores* 166(1). Nação. Parte de: Ásia Oriental. Outras relações: #area#@Oqueus@; #area#@Calouhos@; #area#@Timpates@; #area#@Bugẽ@; #area#@Friucaranjaa@; #area#@Pauel@.

Malaca. *Malaca* 12(3), 13(2), 14(5), 15(5), 17(2), 18(4), 19(7), 20(2), 21(4), 22(3), 23(1), 24(8), 25(1), 26(3), 27(1), 28(3), 29(1), 30(3), 31(2), 32(1), 33(4), 34(5), 35(5), 36(3), 37(1), 38(3), 43(1), 46(1), 48(1), 51(2), 55(2), 56(1), 57(3), 60(1), 64(1), 86(1), 90(3), 131(1), 132(2), 133(1), 140(2), 143(1), 144(6), 145(2), 146(1), 147(1), 148(3), 153(8), 171(1), 172(1), 176(2), 179(1), 180(1),

183(5), 185(1), 200(2), 203(4), 204(2), 205(4), 206(1), 207(6), 208(7), 214(1), 215(8), 216(7), 217(3), 218(3), 219(8), 220(7), 221(4), 223(1), 225(1). *Malata* 12(1). Porto. Parte de: Insulíndia.
Malacca, Malaysia (AS). Lat. 2.196, long. 102.2405.

Malacou. *Malacou* 150(1), 155(1), 190(1). Senhorio (senhorio). Parte de: Pegù. Outras relações: #area#@Ansedaa@.

Malauares. *Malauar* 27(1). *Malauares* 8(1), 16(2), 22(1), 26(2), 27(1), 146(1), 149(1), 175(1), 185(1). Região (costa). Parte de: India.
Malabār Coast, India (AS). Lat. 11.24306, long. 75.88257.

Malayo. *Malaya* 13(1), 220(1). *Malayo* 20(1), 36(1), 41(1), 46(1), 57(1), 132(1), 144(1). *Malayos* 21(1), 48(1), 68(1), 124(1), 153(1), 185(1). Região (costa). Parte de: Ásia. Outras relações: #area#@Pullo Çambilão@; #area#@Barruhaas@; #area#@Salangor@; #area#@Panaagim@; #area#@Quedaa@; #area#@Parlés@; #area#@Pendão@; #area#@Sambilão Sião@.
Malay Peninsula, (AS). Lat. 4.03962, long. 102.12891.

Malhorquy. *Malhorquy* 3(1), 175(1). Ilha. Parte de: Europa.
Mallorca, Spain (EU). Lat. 39.6078, long. 3.01197.

Maluco. *Maluco* 17(1), 20(1), 21(1), 26(2), 33(1), 43(1), 143(1), 176(1), 203(2), 207(1), 208(1). Ilhas (ilhas, mar). Parte de: Insulíndia.
Provinsi Maluku Utara, Indonesia (AS). Lat. -0.25, long. 127.5.

Manamotapa. *Manamotapa* 215(1). Região (terra). Parte de: Africa.
Republic of Zimbabwe, Zimbabwe (AF). Lat. -19, long. 29.75.

Manaquileu. *Manaquileu* 129(1). Cidade. Parte de: Cauchenchina. Outras relações: #area#@Comhay, montes@; #area#@China@; #12 horas fluviais abaixo#@Tinamquaxy@; #5 dias fluviais#@Xolor@.

Manauedee. *Manauedee* 162(1). Cidade. Parte de: Calaminhan. Outras relações: #area#@Angegumaa@; #13 dias rio acima a partir de#@Chipanocão@; #1 légua#@Timplão@.

Mandouim. *Mandouim* 26(1). Rio (alfândega). Parte de: Goa.

Manicafaraõ. *Manicafaraõ* 162(1). Hospedaria (hospedaria). Parte de: Calaminhan. Outras relações: #area#@Angegumaa@; #1 légua#@Campalagrau@; #1 légua#@Timplão@; #area#@Vrpanesendoo@.

Manicataraõ. *Manicataraõ* 117(1). Pinhal. Parte de: China. Outras relações: #1.5 léguas#@Quansy@.

Martauão. *Martauão* 107(1), 128(1), 144(3), 146(1), 147(3), 148(2), 151(2), 152(1), 153(5), 154(1), 158(2), 167(3), 168(1), 169(2), 171(1), 185(2), 188(2), 190(2). Cidade (metrópole, barra). Parte de: Pegù. Outras relações: #area#@Angegumaa@.
Martaban, Myanmar [Burma] (AS). Lat. 16.52834, long. 97.6157.

Massuaa. *Massuaa* 3(1), 4(2). Porto. Parte de: Etiopia.
Massawa, Eritrea (AF). Lat. 15.60811, long. 39.47455.

Masulepatão. *Masulepatão* 158(1). Porto. Parte de: India.
Machilīpatnam, India (AS). Lat. 16.18747, long. 81.13888.

Meca. *Meca* 3(2), 6(2), 13(1), 16(2), 18(1), 20(2), 21(1), 27(1), 31(1), 37(1), 59(1), 151(1), 158(1), 203(1), 225(1). Cidade (estreito, casa). Parte de: Arabia Felix.
Mecca, Saudi Arabia (AS). Lat. 21.42664, long. 39.82563.

Mecuy. *Mecuy* 117(1), 122(1), 131(1). Reino (cidade). Parte de: Cauchenchina. Outras relações: #area#@Benau@; #54 léguas#@Lingator@.

Medina. *Medina* 6(1). Cidade. Parte de: Arabia Felix.
Medina, Saudi Arabia (AS). Lat. 24.46861, long. 39.61417.

mediterraneo. *Mediterraneo* 20(1), 31(1). *mediterraneo* 14(1), 18(1), 19(1), 179(1). Mar. Parte de: Insulíndia.

Meidoo. *Meidoo* 191(1), 193(1), 196(1). *Meydoo* 190(1). Cidade. Parte de: Pegù. Outras relações: #area#@Ansedaa@; #area#@Digum@.

Meidur. *Meidur* 159(1). Cidade. Parte de: Calaminhan. Outras relações: #area#@Angegumaa@; #12 léguas#@Chipanocão@; #area#@Catammaas@; #area#@Singilapau@.

Meigauotau. *Meigauotau* 153(1). Campo. Parte de: Prom. Outras relações: #2 léguas#@Prom, cidade@.
Shwedaung, Myanmar [Burma] (AS). Lat. 18.70138, long. 95.20748.

Meleitay. *Meleitais* 195(1). *Meleitay* 150(1), 155(1), 156(5), 157(1), 194(1). *Meleytay* 128(1), 156(1), 157(2). Cidade (fortaleza, rio). Parte de: Chaleu. Outras relações: #18 léguas#@Prom@; #area#@Pumfileu@.

Meleitor. *Meleitor* 214(1). Ilha. Parte de: Minâcoo.

Melides. *Melides* 1(1). Praia. Parte de: Portugal.
Praia de Melides, Portugal (EU). Lat. 38.12928, long. -8.79407.

Melinde. *Melinde* 151(1). Porto. Parte de: Africa.
Malindi, Kenya (AF). Lat. -3.21799, long. 40.11692.

Menancabo. *Menãcabos* 16(1). *Menamcabo* 31(1). *Menancabo* 20(1), 35(1), 39(1), 166(1). *Menancabos* 149(1), 153(1), 186(1). Região (minas). Parte de: Çamatra. Outras relações: #area#@Iambee@; #area#@Broteo@; #area#@Campar@.
Provinsi Sumatera Barat, Indonesia (AS). Lat. -1, long. 100.5.

Merguim. *Merguim* 146(1), 147(1), 150(1). Porto. Parte de: Sião. Outras relações: #area#@Martauão@; #area#@Tanauçarim@; #area#@Touay@; #area#@Iuncay@; #area#@Pullo

Camude@; #area#@Vagaruu@.
Myeik, Myanmar [Burma] (AS). Lat. 12.43954, long. 98.60028.

Miacoo. *Miacoo* 112(1), 143(1), 210(1), 223(2). *Miocoo* 107(1), 208(3), 224(1). Cidade (metrópole, porto). Parte de: Iapaõ. Outras relações: #18 léguas#@Sicay@.
Kyoto Prefecture, Japan (AS). Lat. 35.25, long. 135.43333.

Miaygimaa. *Miaygimaa* 132(1), 133(1), 218(1). Cidade (povoação). Parte de: Tanixumaa.

Miaygimaa. *Miaygimaa* 211(1), 212(1), 213(1). Templo. Parte de: Bungo. Outras relações: #12 léguas#@Fucheo@.

Micuy. *Micuy* 61(1). Ponta. Parte de: China. Outras relações: #1.5 léguas#@Nouday@; #1.5 léguas#@Xilendau@; #1.5 léguas#@Nipafau@; #area#@Tinlau@; #5 dias#@Comolem@.

Minacáleu. *Minacáleu* 17(1). Morro. Parte de: Achem.

Minâcoo. *Minâcoo* 214(1). Reino. Parte de: Iapaõ. Outras relações: #area#@Fucheo@.

Minapau. *Minapau* 114(1). Cidade. Parte de: Pequim.

Minatoo. *Minato* 135(1). *Minatoo* 202(1), 223(1). Porto (lugar). Parte de: Iapaõ. Outras relações: #1dia nautico#@Tanoraa@; #2 dias náuticos#@Canguexumaa@; #area#@Bungo@.

Mindanaos. *Mindanaos* 143(1), 149(1). *Mindanous* 214(1). Ilha. Parte de: Insulíndia. Outras relações: #area#@Papuaas@; #area#@Selebres@.
Mindanao, Philippines (AS). Lat. 8, long. 125.

Mindoo. *Mindoo* 96(1). Cidade. Parte de: China. Outras relações: #area#@Batampina@; #area#@Nacau@; #area#@Pacão@; #3 léguas#@Bigay potim@; #10 léguas#@Fiunganorsee@.

Minhaçumbaa. *Minhaçumbaa* 28(1). Rio. Parte de: Aarù. Outras relações: #5 léguas#@Aarù@#5 léguas#@Paneticão@.

Minhacutem. *Minhacutem* 88(1). Aldea. Parte de: Nanquim. Outras relações: #area#@Batampina@.

Minhagaruu. *Minhagaruu* 19(1). Estreito. Parte de: Çamatra. Outras relações: #area#@Oceano@; #area#@Mediterraneo@; #area#@#area#@Apefingau@; #area#@Pullo Bugay@.

Minhatoley. *Minhatoley* 14(1). Baía. Parte de: Peedir. Outras relações: #5 dias náuticos#@Hicanduré@; #area#@Mediterraneo@; #23 léguas#@Oceano@.

Mocaa. *Mocaa* 5(1). Cidade. Parte de: Arabia Felix.
Al Mukhā’, Yemen (AS). Lat. 13.31563, long. 43.26044.

Moçambique. *Moçambique* 2(2). Porto. Parte de: Africa.
Republic of Mozambique, Mozambique (AF). Lat. -18.25, long. 35.

Moẽs. *Moẽs* 114(1), 149(1), 153(1), 156(6), 157(2), 162(1), 186(1), 194(1), 195(1). Nação. Parte

de: Índico Oriental. Outras relações: #area#@Pondaleu@; #area#@Siammon@; #area#@Auaa@.
Bago, Myanmar [Burma] (AS). Lat. 17.33521, long. 96.48135.

Mogores. *Mogor* 43(1), 185(1). *Mogores* 20(1), 45(1), 95(1), 103(1), 106(1), 114(1), 119(1), 122(1), 123(1), 124(1), 146(1), 149(1). Império. Parte de: India. Outras relações: #area#@Dely@; #area#@Coraçone@; #area#@Persia@.
Āgra, India (AS). Lat. 27.18333, long. 78.01667.

Mompollacota. *Mompollacota* 46(1). *Mõpolocota* 189(1). Lugar. Parte de: Sião. Outras relações: #area#@Sião@; #area#@Cuy@; #area#@Lugor@; #area#@Chintabu@.

Moncalor. *Moncalor* 41(1). Serra. Parte de: Indochina. Outras relações: #80 léguas#@Taiquilleu@; #area#@Tinacoreu@; #area#@Champaa@.

Monginoco. *Monginoco* 128(1). Império. Parte de: Índico Oriental. Outras relações: #area#@Pumsileu@; #area#@Capimper@; #area#@Sacotay@; #area#@Meleitay@; #area#@Sauady@.

Montemór o velho. *Montemór o velho* 1(1). Vila. Parte de: Portugal.
Montemor-o-Velho, Portugal (EU). Lat. 40.17287, long. -8.68616.

Moscouia. *Moscoby* 124(1). *Moscouia* 85(1). *Moscouitas* 88(1). *Muscoo* 126(1). Região (terra). Parte de: Europa. Outras relações: #area#@Xinxipou@; #area#@Leysacotay@.
Moscow, Russia (EU). Lat. 55.75222, long. 37.61556.

Moscumbià. *Moscumbià* 72(1). Lago. Parte de: Ásia. Outras relações: #area#@Alemanha@; #area#@Paatebenam@.

Moucham. *Moucham* 190(1). *Mouchão* 188(1). Vila. Parte de: Pegù. Outras relações: #5 léguas#@Çatão@.

Mouchel. *Mouchel* 167(1). Aldea. Parte de: Pegù. Outras relações: #3 dias fluviais#@Martauão@; #5 dias fluviais#@Madur@.

Mounay. *Mounay* 148(1), 150(2), 152(1), 167(3), 168(2), 169(3). Ilha (ponta). Parte de: Martauão. Outras relações: #1.5 léguas#@Martauão@; #Pegù#@20 léguas@; #5 léguas#@Tagalaa@.

Muchiparom. *Muxiparaõ* 109(1). *Muchiparom* 109(1). Edifício. Parte de: Pequim.

Muhar. *Muhar* 207(1). Rio. Parte de: Insulíndia. Outras relações: #6 léguas#@Malaca@.
Kuala Muar, Malaysia (AS). Lat. 2.05139, long. 102.53389.

Muria. *Muria* 3(1). Ilha. Parte de: Arabia Felix. Outras relações: #area#@Abedalcuria; #area#@Curia@; #area#@Çacotorà@.

Mutipinão. *Mutepinão* 50(1). *Mutipinão* 47(2). Porto (rio). Parte de: Cauchenchina. Outras relações: #40 léguas#@Tanauquir@; #3 dias náuticos#@Tilaumera@; #area#@Colem@; #area#@Panduree@; #12 dias#@Quangepaarù@.

Nacapirau. *Nacapirau* 110(2), 222(1). Edifício (templo). Parte de: Pequim.

Nacataas. *Nacataas* 88(1). Reino. Parte de: China. Outras relações: #area#@Tauquiday@.

Nacau. *Nacau* 92(1), 94(2), 96(1). Cidade. Parte de: China. Outras relações: #area#@Batampina@; #area#@Pacão@; #1 dia fluvial#@Lequimpau@; #area#@Mindoo@.

Nangafau. *Nangafau* 71(1). Serra. Parte de: China. Outras relações: #40 léguas#@Nanquim, enseada@.

Nanquim. *Nanquim* 71(4), 73(1), 74(2), 79(3), 81(2), 82(2), 84(3), 85(3), 87(1), 88(3), 90(1), 91(1), 94(2), 97(1), 100(3), 103(1), 111(1), 117(1), 143(1), 222(1). *Nãquim* 71(1), 72(1), 88(1). Cidade (enseada). Parte de: China. Outras relações: #metropoli#@Sumbor@; #metropoli#@Fanjus@; #metropoli#@Liampo@; #170 léguas#@Pequim@.
Nanjing, China (AS). Lat. 32.06167, long. 118.77778.

Narsinga. *Narsinga* 9(1), 107(1), 149(1), 158(1). Reino. Parte de: India.
Vijayanagara, India (AS). Lat. 15.325, long. 76.465.

Natibasoy. *Natibasoy* 131(1). Povoação (lugar). Parte de: Cauchenchina. Outras relações: #area#@Baguetor@; #area#@Lingator@; #98 léguas#@Huzamguee@.

Naypatim. *Naypatim* 126(1). Pagode (pagode). Parte de: Guatipamor.

Nazarè. *Nazarè* 63(1). Cidade. Parte de: Ásia.
Nazareth, Israel (AS). Lat. 32.69925, long. 35.30483.

Negrais. *Negrais* 150(1). Senhorio (senhorio). Parte de: Pegù.
Mawdin, Myanmar [Burma] (AS). Lat. 15.95527, long. 94.24502.

Neytor. *Neytor* 45(1). Aldea. Parte de: Tanauquir.

Nicubar. *Nicubar* 20(1). Ilhas (ilhas). Parte de: India.
Nicobar Islands, India (AS). Lat. 8, long. 93.5.

Nipafau. *Nipafau* 63(1). Calheta (calheta). Parte de: Nouday.

Nixiamcoo. *Nixiamcoo* 117(1), 118(1), 120(1). *Nixiancoo* 120(1), 121(2). Castelo (castelo). Parte de: China. Outras relações: #2 dias#@Quansy@.

Nixihumflaõ. *Nixihumflaõ* 71(1). Lugar. Parte de: Ásia Oriental. Outras relações: #area#@Tartaria@; #area#@Nanquim@.

Nobins. *Nobins* 149(1). Nação. Parte de: Ásia. Outras relações: #area#@Bramá@.

nossa Senhora do outeyro. *nossa Senhora do outeyro* 48(1), 203(1), 204(1), 215(1), 216(1). Ermida (ermida). Parte de: Malaca.

noua Espanha. *noua Espanha* 208(1). Província. Parte de: America.
Mexico City, Mexico (AM). Lat. 19.42847, long. -99.12766.

Nouday. *Nouday* 62(1), 63(2), 64(1), 65(1), 66(3), 67(3). Cidade (porto, angra, lugar). Parte de: China. Outras relações: #area#@Xilendau@; #area#@Nipafau@; #area#@Mecuy@; #5 dias#@Comolem@; #area#@Tinlau@; #area#@Liampoo@.

Obidos. *Obidos* 4(1), 148(1). Vila. Parte de: Portugal.
Óbidos, Portugal (EU). Lat. 39.36055, long. -9.1567.

Oceano. *Oceano* 13(1), 14(1), 19(1), 20(2), 31(1). Mar. Parte de: Insulíndia.

Ochileuday. *Ochileuday* 90(1). Rio (mitico). Parte de: China. Outras relações: #area#@Batampina@; #area#@Iunquileu@.

Ocumchaleu. *Ocumchaleu* 170(1). Cidade (mítica). Parte de: Oregantor. Outras relações: #area#@Sauady@.

Odiaa. *Odiâ* 181(1). *Odiaa* 36(1), 107(1), 124(1), 146(2), 181(1), 182(1), 183(3), 185(4), 186(1), 189(2), 190(1). Cidade (metrópole). Parte de: Sião. Outras relações: #capital#@Sião@; #capital#@Sornau@.
Phra Nakhon Si Ayutthaya, Thailand (AS). Lat. 14.35167, long. 100.57739.

Oeyras. *Oeyras* 67(1). Vila. Parte de: Portugal.
Oeiras, Portugal (EU). Lat. 38.69105, long. -9.31085.

Omanguche. *Omãguche* 210(1). *Omanguche* 208(2), 209(2), 210(1), 215(1), 218(2). *Omanguchè* 208(1). Reino (cidade). Parte de: Iapaõ.
Yamaguchi, Japan (AS). Lat. 34.18333, long. 131.46667.

Onor. *Onor* 8(2), 9(1), 12(1). Reino (porto). Parte de: India.
Honāvar, India (AS). Lat. 14.28088, long. 74.44497.

Oqueus. *Oqueus* 166(1). Nação. Parte de: Ásia Oriental. Outras relações: #area#@Magores@; #area#@Calouhos@; #area#@Timpates@; #area#@Bugẽ@; #area#@Friucaranjaa@; #area#@Pauel@.

Oregantor. *Oregantor* 170(1). Lago. Parte de: Sauady. Outras relações: #area#@Auaa@.

Orixaa. *Orixaa* 158(1). Reino. Parte de: India.
State of Odisha, India (AS). Lat. 20.5, long. 84.41667.

Orliens. *Orliens* 158(1). Cidade. Parte de: França.
Orléans, France (EU). Lat. 47.90289, long. 1.90389.

Ormuz. *Ormuz* 2(1), 6(2), 7(2), 176(1). Porto (fortaleza). Parte de: Persia.
Hormoz, Iran (AS). Lat. 27.095, long. 56.4529.

Osquy. *Osquy* 135(1), 201(2), 223(3), 224(1), 225(3). Fortaleza. Parte de: Bungo. Outras relações: #7 léguas#@Fucheo@.
Usuki, Japan (AS). Lat. 33.12342, long. 131.80401.

Paacem. *Paacem* 17(2), 18(1), 22(1), 26(1), 31(1), 206(1). *Pacem* 13(1), 20(1). Porto (fortaleza). Parte de: Çamatra. Outras relações: #area#@Bata@; #area#@Achem@.
Kota Lhokseumawe, Indonesia (AS). Lat. 5.13333, long. 97.06667.

Paatebenam. *Paatebenam* 72(1). *Patebenão* 72(1). Rio. Parte de: China. Outras relações: #area#@Nanquim@; #area#@Calindão@.

Pacão. *Pacão* 92(1), 94(2), 96(1). Cidade. Parte de: China. Outras relações: #area#@Batampina@; #area#@Nacau@; #1 dia fluvial#@Lequimpau@; #area#@Mindoo@.

Pacarou. *Pacarou* 188(1). Ribeira. Parte de: Odiaa.

Pachissarù. *Pachissarù* 17(1). Povoação (lugar). Parte de: Bata. Outras relações: #area#@Guateamgim@; #area#@Panaajù@.

Pafuaas. *Pafua* 122(1). *Pafuâ* 128(1). *Pafuaas* 41(1), 47(1), 70(1). *Pafuas* 112(1). *Pafuâs* 159(1). Principado (principado). Parte de: Indochina. Outras relações: #area#@Benão@; #area#@Gueos@; #area#@Lauhos@; #area#@Mutipinão@.

Palemxitau. *Palemxitau* 120(1). Ribeira. Parte de: Pequim. Outras relações: #area#@Pommitay@.

Palimbão. *Palimbão* 24(1). Ilha. Parte de: Çamatra. Outras relações: #area#@Iambee@; #area#@Sorobaya@; #area#@Malaca@.
Palembang, Indonesia (AS). Lat. -2.91673, long. 104.7458.

Pamquinor. *Pamquinor* 123(1). Cidade. Parte de: Tartaria. Outras relações: #3 léguas#@Singrachirau@; #1 dia#@Xipator@.

Panaagim. *Panaagim* 144(1). Rio. Parte de: Malayo. Outras relações: #area#@Barruhaas@; #area#@Salangor@; #area#@Quedaa@; #area#@Parlés@; #area#@Pendão@; #area#@Sambilão Sião@.
Sungai Linggi, Malaysia (AS). Lat. 2.3911, long. 101.9732.

Panaajù. *Panaajù* 15(2), 17(2), 18(1), 20(1). *Panaajû* 15(1), 18(1). *Panajù* 13(1). Cidade (porto). Parte de: Bata. Outras relações: #capital#@Bata@; #7 léguas#@Oceano@; #area#@Guateamgim@; #1 légua#@Batorrendão@.

Panamá. *Panamá* 208(1). Porto. Parte de: America.
Panamá, Panama (AM). Lat. 8.9936, long. -79.51973.

Panaruca. *Panaruca* 36(1), 173(2), 178(6), 179(2). Reino (porto). Parte de: Iaoa.
Panarukan, Indonesia (AS). Lat. -7.70181, long. 113.91844.

Pancanor. *Pancanor* 167(1). Principado (principado). Parte de: Iangomaa. Outras relações: #area#@Ventinau@; #area#@Penauchim@.

Pancruum. *Pamcrùs* 194(1). *Pancrùs* 95(1). *Pancruum* 88(1). Nação (montes). Parte de: Indochina. Outras relações: #area#@Cauchenchina@; #area#@Catebenão@.

Panduree. *Panduree* 47(2). Aldea. Parte de: Cauchenchina. Outras relações: #9 léguas#@Tilaumera@; #area#@Colem@.

Paneticão. *Paneticão* 26(2), 31(1). Rio. Parte de: Aarù.

Panguassirau. *Panguassirau* 165(1). Região (terra). Parte de: Índico Oriental. Outras relações: #area#@Bramá@.

Pão. *Pão* 33(3), 34(2), 35(3). *Paõ* 35(3), 51(1). Reino. Parte de: Insulíndia. Outras relações: #11 léguas#@Pullo timão@; #100 léguas#@Patane@.
Pahang, Malaysia (AS). Lat. 3.5, long. 102.75.

Papuaas. *Papuaas* 143(1), 149(1), 214(1). *Papuas* 180(1). Ilhas (ilhas). Parte de: Insulíndia. Outras relações: #area#@Selebres@; #area#@Mindanaos@.
Provinsi Papua Barat, Indonesia (AS). Lat. -0.86531, long. 134.06118.

Paris. *Paris* 107(1). Cidade (metrópole). Parte de: França.
Paris, France (EU). Lat. 48.85341, long. 2.3488.

Parlés. *Parlés* 144(1), 153(1), 206(2). *Parlês* 20(1), 205(1). *Parlès* 19(1), 20(1), 205(1). Reino (rio). Parte de: Quedaa. Outras relações: #2.5 dias náuticos#@Pullo Çambilão@; #2.5 dias náuticos#@Iunçalão@; #area#@Barruhaas@; #area#@Salangor@; #area#@Panaagim@; #area#@Pendão@; #area#@Sambilão Sião@.
Perlis, Malaysia (AS). Lat. 6.5, long. 100.25.

Passaruão. *Passaruão* 107(1), 173(3), 175(2), 179(1). *Passaruaõ* 36(1). *Passaruoẽs* 174(2), 175(2), 177(1), 178(1). *Passerução* 172(1). Reino (cidade, porto). Parte de: Iaoa. Outras relações: #2 léguas#@Hicanduree@.
Pasuruan, Indonesia (AS). Lat. -7.6453, long. 112.9075.

Passiloco. *Passiloco* 41(1), 124(1), 181(1), 182(2), 183(1), 184(1), 185(2), 186(2). Reino. Parte de: Sião. Outras relações: #area#@Capimper@; #area#@Sião@; #area#@Chiammay@; #area#@Lauhos@; #area#@Gueos@; #area#@Timocouhós@.

Patane. *Patane* 19(2), 32(1), 33(1), 34(2), 35(3), 36(1), 38(4), 39(2), 42(1), 46(1), 47(1), 50(1), 51(1), 55(1), 56(2), 57(2), 66(1), 71(1), 88(1), 132(1), 181(1), 183(1), 204(1), 205(1), 207(1), 220(2). *Patanes* 124(1). Porto. Parte de: Malayo. Outras relações: #130 léguas#@Cuy@; #100 léguas#@Pão@; #85 léguas#@Lugor@.
Pattani, Thailand (AS). Lat. 6.86814, long. 101.25009.

Pauel. *Pauel* 166(2), 167(1). Cidade. Parte de: Calaminhan. Outras relações: #8 dias fluviais#@Timplão@; #1 dia fluvial#@Lunçor@; #area#@Pituy@.

Pauileus. *Pauileus* 166(1). Nação. Parte de: Binagorem.

Peedir. *Peedir* 14(1), 17(1), 20(1), 31(1), 148(1), 206(1). Reino. Parte de: Çamatra. Outras relações: #9 léguas#@Minhatoley@; #area#@Achem@; #area#@mediterraneo@.

Kabupaten Pidie, Indonesia (AS). Lat. 5.08, long. 96.11.

Pegù. *Pegù* 20(1), 88(1), 96(1), 107(1), 112(1), 146(1), 148(1), 150(2), 153(3), 155(2), 162(1), 164(1), 165(3), 167(4), 169(1), 170(1), 171(1), 185(2), 188(1), 189(1), 190(8), 191(1). *Pègù* 193(1), 194(1), 200(1). *Pégù* 192(1), 195(1), 199(1), 200(2), 204(1), 205(1). *Pegu* 168(1). *Pegú* 167(1). *Pegû* 151(1), 153(1), 167(1), 168(1), 170(1), 190(1), 198(1). *Pegua* 188(1), 190(1). *Pégua* 194(1), 198(1). *Pegùs* 149(1), 152(1), 155(1), 156(1), 188(1), 189(1), 190(2), 194(1). *Pégûs* 194(1). *Pêguu* 188(1), 194(1), 197(1). *Pèguu* 196(1). *Péguu* 193(3), 194(3), 196(1), 197(1). *Peguu* 128(1), 147(1), 148(1), 152(1), 153(3), 158(1), 163(2), 165(1), 188(4), 191(2). *Pèguus* 194(1), 195(1), 197(1).
Reino. Parte de: Índico Oriental.
Bago, Myanmar [Burma] (AS). Lat. 17.33521, long. 96.48135.

Penacão. *Penacão* 16(1), 17(1). Rio. Parte de: Achem. Outras relações: #area#@Minacáleu@.

Penamacor. *Penamacor* 176(1). Vila. Parte de: Portugal.
Penamacor, Portugal (EU). Lat. 40.17112, long. -7.12574.

Penauchim. *Penauchim* 167(1). Povoação (lugar). Parte de: Iangomaa. Outras relações: #12 horas fluviais#@Rauditẽs@; #area#@Pancanor@; #area#@Ventinau@.

Pendão. *Pendão* 144(1). Rio. Parte de: Malayo. Outras relações: #area#@Barruhaas@; #area#@Salangor@; #area#@Panaagim@; #area#@Quedaa@; #area#@Parlés@; #area#@Sambilão Sião@.
Sungai Pendang, Malaysia (AS). Lat. 6.03757, long. 100.47677.

Pequim. *Paquim* 52(1), 81(1), 105(1). *Pequim* 45(1), 85(1), 86(1), 87(1), 88(2), 93(1), 94(5), 95(2), 97(1), 100(2), 105(3), 107(3), 108(1), 112(1), 113(1), 114(1), 115(1), 117(2), 119(2), 120(3), 122(1), 123(3), 125(3), 185(1), 222(1). Cidade (metrópole). Parte de: China. Outras relações: #170 léguas#@Nanquim@; #area#@Batampina@.
Beijing, China (AS). Lat. 39.9075, long. 116.39723.

Pera. *Pera* 32(1), 207(1). Porto (barra). Parte de: Çamatra. Outras relações: #area#@Aarù@; #area#@Bintão@; #area#@Achem@.
Perak, Malaysia (AS). Lat. 5, long. 101.

Persia. *Parsio* 43(1). *Persas* 88(1), 124(1), 149(1). *Persia* 20(1), 95(1), 107(1), 120(1), 124(1), 165(1). *Persios* 114(1), 122(1). Reino. Parte de: Índico Ocidental.
Islamic Republic of Iran, Iran (AS). Lat. 32, long. 53.

Petilau Namejoo. *Petilau Namejoo* 117(1). Pagode (pagode). Parte de: Quansy.

Pichau malacou. *Pichau malacou* 153(1). Rio. Parte de: Prom.

Pilaucacem. *Pilaucacem* 41(1). Cidade (metropoli). Parte de: Champaa. Outras relações: #area#@Tinacoreu@; #area#@Congrau@; #area#@Taiquilleu@.

Pilaunera. *Pilaunera* 92(1). Lezíria (mítico). Parte de: Ásia. Outras relações: #46 dias

fluviais#@Pequim@; #area#@Catebasoy@; #area#@Guantipocau@.

Pimlaxau. *Pimlaxau* 208(1). Lugar. Parte de: Bungo. Outras relações: #2 léguas#@Fucheo@.

Pinator. *Pinator* 39(1). Lago. Parte de: Quitiruão. Outras relações: #260 léguas#@Pullo Cambim@.

Pisammanes. *Pisammanes* 179(1). Lugar. Parte de: Iaoa. Outras relações: #12 léguas#@Passaruão@.

Pisanduree. *Pisanduree* 144(1). Ilha. Parte de: Malayo.

Pituy. *Pituy* 165(1), 166(1). Rio. Parte de: Calaminhan. Outras relações: #area#@Timplão@; #area#@Pauel@.

platarias. *platarias* 143(1). Ilhas (ilhas). Parte de: Ásia Oriental. Outras relações: #area,NNE#@Lequios@.

Pocasser. *Pocasser* 73(1), 89(2), 90(1), 199(1), 222(1). Cidade. Parte de: China. Outras relações: #area#@Batampina@; #12 horas#@Xinligau@; #4 dias fluviais#@Minhacutem@; #area#@Nanquim@.

Pocausilim. *Pocausilim* 23(1). Aldea (lugar). Parte de: Aarù. Outras relações: #area#@Puneticão@.

Pollem. *Pollem* 143(1). Ilha. Parte de: Iapaõ. Outras relações: #area#@Sesirau@; #area#@Goto@; #area#@Fucanxi@.

Pomgatur. *Pomgatur* 131(1). Abadia (abadia). Parte de: Benau.

Pommiseray. *Pommiseray* 171(1). Aldea. Parte de: Sauady. Outras relações: #1.5 dias#@Quiay Vogarem@; #1.5 dias#@Oregantor@.

Pommitay. *Pommitay* 120(1). Serra. Parte de: China. Outras relações: #7 léguas#@Pequim@; #area#@Palemxitau@; #area#@Lautimey@; #area#@Bumxay@.

Pondaleu. *Pondaleu* 157(1). Montes. Parte de: Índico Oriental. Outras relações: #area#@Siammon@; #area#@Auaa@; #area#@Moẽs@.

Pongor. *Põgor* 140(1). *Pongor* 139(2), 142(2), 143(1). *Pangor* 107(1). Cidade (metropoli). Parte de: Lequios. Outras relações: #capital#@Lequios@.

Ponquilor. *Ponquilor* 95(1). Cidade. Parte de: China. Outras relações: #area#@Tartaria@.

Pontareu. *Pontareu* 194(1). Rio. Parte de: Pegù. Outras relações: #2 léguas#@Pegú@.

Pontau. *Pontau* 164(1). Lugar. Parte de: Auaa. Outras relações: #area#@Iatir@; #area#@Faleu@.

Pontir. *Pontir* 57(1). Reino. Parte de: Goto.

Ponxileytay. *Ponxileytay* 101(1). Região (lugares). Parte de: China.

Portugal. *Portugal* 4(1), 8(1), 9(2), 11(1), 13(2), 14(1), 21(3), 22(1), 29(2), 64(2), 67(1), 68(2),

Pullo Cambim. *Pullo Cambim* 39(1), 40(1), 220(1). Rio. Parte de: Indochina. Outras relações: #coordenadas#@9@; #260 léguas#@Pinator@; #area#@Camboja@; #area#@Champaa@.

Pullo Camude. *Pullo Camude* 147(1), 153(1). Ilha. Parte de: Sião. Outras relações: #area#@Martauão@; #area#@Tanauçarim@; #area#@Touay@; #area#@Iuncay@; #area#@Merguim@; #area#@Vagaruu@.

Pullo Capás. *Pullo Capás* 42(1). Morro. Parte de: Ainão.

Pullo Catão. *Pullo Catão* 46(1). *Pullo catão* 183(1). Ilha. Parte de: Sião. Outras relações: #area#@Sião@; #area#@China@.

Pullo Condor. *Pullo Condor* 39(1), 179(1). Ilha. Parte de: Indochina. Outras relações: #coordenadas#@8.33@.
Côn Son, Vietnam (AS). Lat. 8.70833, long. 106.60833.

Pullo Hinhor. *Pullo Hinhor* 66(1), 67(1). Ilha. Parte de: Liampoo. Outras relações: #15 léguas#@Liampoo@.

Pullo Hinhor. *Pullo Hinhor* 145(2), 146(2), 147(1). Ilha. Parte de: Sião. Outras relações: #7 léguas#@Taubasoy@; #area#@Tanauçarim@; #area#@Pisanduree@; #area#@Çambilão@.

Pullo Hinhor. *Pullo Hinhor* 50(1). Morro. Parte de: ilha dos cocos.

Pullo pisaõ. *Pullo pisaõ* 220(1). Ilha. Parte de: Insulíndia. Outras relações: #area#@Cincaapura@; #area#@Malaca@.
Pulau Pisang, Malaysia (AS). Lat. 1.4675, long. 103.2546.

Pullo Quenim. *Pullo Quenim* 20(1). Ilha. Parte de: Çamatra. Outras relações: #area#@Bata@; #area#@Pullo Tiquòs@.

Pullo Quirim. *Pullo Quirim* 55(1). Ilha. Parte de: China. Outras relações: #9 léguas#@Xamoy@; #3 dias#@Luxitay@.

Pullo timão. *Pullo Timão* 33(1). *Pullo timão* 220(2). Ilha. Parte de: Malayo. Outras relações: #11 léguas#@Pão@; #90 léguas#@Malacca@.
Pulau Tioman, Malaysia (AS). Lat. 2.7972, long. 104.166.

Pullo Tiquòs. *Pullo Tiquòs* 20(1). Ilha. Parte de: Çamatra. Outras relações: #area#@Bata@; #area#@Pullo Quenim@.
Tiku, Indonesia (AS). Lat. -0.3958, long. 99.92307.

Pullopracelar. *Pullopracelar* 144(1). Lugar. Parte de: Insulíndia. Outras relações: #area#@Malaca@; #area#@Çamatra@.
Bukit Jugra, Malaysia (AS). Lat. 2.83333, long. 101.41667.

Pumfileu. *Pumfileu* 128(1). Rio. Parte de: Índico Oriental. Outras relações: #area#@Singapamor@; #area#@Capimper@; #area#@Sacotay@; #area#@Monginoco@; #area#@Meleitay@;

#area#@Sauady@.

Puneticão. *Puneticão* 21(1), 22(1), 26(1), 32(3). Rio (fortaleza). Parte de: Çamatra. Outras relações: #area#@Aarù@.

Puxanguim. *Puxanguim* 126(1). Cidade. Parte de: Tartaria. Outras relações: #1 dia fluvial#@Guatipamor@; #1 dia fluvial#@Linxau@; #area#@Tuymicão@.

Quaijuão. *Quaijuão* 38(1). Reino. Parte de: Iaoa. Outras relações: #area#@Lasaparà@.

Quangepaarù. *Quangepaarù* 48(1). *Quangeparuu* 132(1). *Quãogeparu* 61(1). *Quoangeparù* 70(1). *Quoanjaparù* 56(1). *Quangiparù* 52(1). Cidade. Parte de: Cauchenchina.
Quảng Yên, Vietnam (AS). Lat. 20.9421, long. 106.8027.

Quanginau. *Quanginau* 127(2), 128(2). Cidade. Parte de: Tartaria. Outras relações: #1 dia fluvial#@Singuafatur@; #4 dias fluviais#@Lechune@.

Quansy. *Quansy* 103(1), 104(2), 115(2), 117(2), 118(1), 120(2). *Quãsy* 114(1). Cidade. Parte de: China.

Quatanqur. *Quatanqur* 128(1). Esteiro. Parte de: Ásia Oriental. Outras relações: #7 dias fluviais a partir de#@Voulem@; #area#@Sigapamor@; #area#@Xinaleygrau@.

Quaytragum. *Quaytragum* 123(1). Ribeira. Parte de: China. Outras relações: #12 horas#@Pequim@; #12 horas#@Guijampee@.

Quedaa. *Quedà* 17(1), 19(1). *Quedâ* 19(1). *Quedaa* 144(1), 153(1), 185(1), 205(1), 207(1). Reino (rio, costa). Parte de: Malayo. Outras relações: #area#@Parlés@; #area#@Iunçalão@; #area#@Malaca@.
Kedah, Malaysia (AS). Lat. 6, long. 100.66667.

Queitor. *Queitor* 156(1), 157(1), 158(1). Rio. Parte de: Índico Oriental. Outras relações: #area#@Auaa@; #area#@Prom@; #area#@Guampanoo@.
Irrawaddy River, Myanmar [Burma] (AS). Lat. 15.83333, long. 95.1.

Quiay Hifarom. *Quiay Hifarom* 195(1). Pagode (pagode). Parte de: Pegù.

Quiay Hinarel. *Quiay Hinarel* 171(1). Pagode (pagode). Parte de: Sauady. Outras relações: #8 dias fluviais#@Quiay Hinarel@.

Quiay Vogarem. *Quiay Vogarem* 170(1). Ermida (ermida). Parte de: Oregantor.

Quilem. *Quilem* 16(1). Rio. Parte de: Achem.

Quitiruão. *Quitiruão* 39(1), 181(4), 182(1). Reino. Parte de: Índico Oriental. Outras relações: #260 léguas#@Pullo Cambim@; #area#@Sião@; #15 léguas#@Guibem@; #area#@Chiammay@; #12 léguas#@Suropisem@; #area#@Capimper@; #12 léguas#@Siputay@.
Kamphaeng Phet, Thailand (AS). Lat. 16.48344, long. 99.52153.

Quoamão. *Quoamão* 55(1). Povoação (lugar). Parte de: China. Outras relações:

#area#@Cauchenchina, enseada@; #area#@ilha dos ladroẽs@; #area#@Comhay@.

Quoansy. *Quoansy* 67(1). Cidade. Parte de: China. Outras relações: #area#@Cauchenchina@.

Raizbutos. *Raizbutos* 149(1). Região. Parte de: India.

Rates. *Rates* 12(2). Vila. Parte de: Portugal.
Rates, Portugal (EU). Lat. 41.42834, long. -8.67925.

Rauditẽs. *Rauditẽs* 167(1). Fortaleza (fortalezas). Parte de: Pancanor. Outras relações: #12 horas fluviais#@Penauchim@; #area#@Ventinau@; #5 dias fluviais#@Magadaleu@.

Rebandar. *Rebandar* 217(1). Povoação. Parte de: Goa. Outras relações: #0.5 léguas#@Goa@.

Rendacalem. *Rendacalem* 128(1). Cidade. Parte de: Tartaria. Outras relações: #5 dias fluviais#@Lechune@; #area#@Xinaleygrau@.

rio das serpes. *rio das serpes* 72(1). Baía. Parte de: Ásia Oriental. Outras relações: #13 dias náuticos#@Fanjus@; #area#@Calindão@.

rio do sal. *rio do sal* 132(1). Rio. Parte de: China. Outras relações: #5 léguas#@Chabaquee@; #area#@Lamau@.
Hanjiang Xixi, China (AS). Lat. 23.36144, long. 116.66125.

Roma. *Roma* 5(1), 107(1), 116(1), 128(1), 169(1). *Romanos* 92(1), 113(1). Cidade. Parte de: Europa. Outras relações: #area#@Latinos@.
Rome, Italy (EU). Lat. 41.89193, long. 12.51133.

Roparoẽs. *Roparoẽs* 163(1). Nação. Parte de: Calaminhan. Outras relações: #area#@Calaminhan@.

S. Tome. *S. Tome* 1(1). Ilha. Parte de: Africa.

Sabà. *Sabà* 20(1). *Sabaa* 4(1). Reino. Parte de: Arabia Felix. Outras relações: #area#@Etiopia@.
Sheba, Yemen (AS). Lat. 15.37954, long. 44.70886.

Sabambainhaa. *Sabambainhaa* 196(1), 198(2). Rua (porta). Parte de: Pegù. Outras relações: #area#@Pegù, cidade@.

Sabaom. *Sabaom* 26(1). Estreito. Parte de: Insulíndia. Outras relações: #area#@Cincaapura@.

Sacotay. *Çacotais* 159(1). *Sacotay* 128(1), 185(1). Cidade (cidade). Parte de: Indochina. Outras relações: #9 léguas#@Tapurau@; #area#@Pumfileu@; #area#@Lebrau@; #area#@Sião@.
Sukhothai, Thailand (AS). Lat. 17.00778, long. 99.823.

Salangor. *Salangor* 17(1), 144(1), 207(1). Porto (rio). Parte de: Malayo. Outras relações: #area#@Barruhaas@; #area#@Panaagim@; #area#@Quedaa@; #area#@Parlés@; #area#@Pendão@; #area#@Sambilão Sião@.
Selangor, Malaysia (AS). Lat. 3.16667, long. 101.5.

Saleyjacau. *Saleyjacau* 40(1). Baía. Parte de: Champaa. Outras relações: #17 léguas#@Pullo Cambim@; #12 horas#@Toobasoy@.

Sambilão Sião. *Sambilão Sião* 144(1). Rio. Parte de: Malayo. Outras relações: #area#@Barruhaas@; #area#@Panaagim@; #area#@Quedaa@; #area#@Parlés@; #area#@Pendão@; #area#@Salangor@.

Sampitay. *Sampitay* 91(2), 92(1). Cidade. Parte de: China. Outras relações: #area#@Batampina@; #area#@Lequimpau@; #1 dia fluvial#@Iunquileu@.

Sanchaõ. *Sanchão* 132(1), 215(1). *Sanchaõ* 215(1), 216(1), 221(3). Porto. Parte de: China. Outras relações: #7 léguas#@Lampacau@; #26 léguas#@Cantão@.
Shangchuan Dao, China (AS). Lat. 21.70964, long. 112.78544.

Sansy. *Sansy* 222(3). *Sansim* 88(1). Província. Parte de: China. Outras relações: #area#@Batobasoy @; #area#@Cantão@.
Shaanxi, China (AS). Lat. 36, long. 109.

Santarem. *Santarem* 66(1). Vila. Parte de: Portugal.
Santarém, Portugal (EU). Lat. 39.23333, long. -8.68333.

Santiago de Cacem. *Santiago de Cacem* 1(1). Vila. Parte de: Portugal.
Santiago do Cacém, Portugal (EU). Lat. 38.01693, long. -8.69475.

Santiago de Galiza. *Santiago de Galiza* 5(1). Cidade. Parte de: Galego.
Santiago de Compostela, Spain (EU). Lat. 42.88052, long. -8.54569.

Sao Miguel. *Sao Miguel* 4(1). *São Miguel* 4(1). Povoação (lugar). Parte de: Bitonto.

São Tomè. *São Tomè* 147(1). Porto. Parte de: India.
Mylapore, India (AS). Lat. 13.02917, long. 80.27083.

Sategão. *Sategão* 128(1). Cidade. Parte de: India.
Chittagong, Bangladesh (AS). Lat. 22.3384, long. 91.83168.

Satilgão. *Satilgão* 4(1). Mosteiro. Parte de: Etiopia. Outras relações: #1 dia#@Massuaa@.
Saati, Eritrea (AF). Lat. 15.5793, long. 39.25598.

Sauady. *Sauadijs* 153(1). *Sauadis* 149(1), 170(2), 194(1), 195(1). *Sauady* 41(1), 150(2), 164(1), 165(1), 168(1), 170(4), 199(1). *Souady* 128(1). Reino (cidade). Parte de: Índico Oriental. Outras relações: #130 léguas#@Pegù@; #9 dias fluviais acima a partir de#@Pegù@; ##area#@Calaminhan@; #area#@ Sião@; #area#@Bramá@; #area#@Auaa@; #area#@Tanguu@; #area#@Chaleu@.

Selebres. *Selebres* 143(1), 149(1), 180(1), 214(1). Ilha. Parte de: Insulíndia. Outras relações: #area#@Papuas@; #area#@Mindanaos@.
Sulawesi, Indonesia (AS). Lat. -2.13154, long. 120.28888.

Sesirau. *Sesirau* 143(1). Ilha. Parte de: Iapaõ. Outras relações: #area#@Sesirau@; #area#@Goto@; #area#@Pollem@.

Setuuel. *Setuual* 1(1). *Setuuel* 1(1), 193(1). Porto (vila). Parte de: Portugal.
Setúbal, Portugal (EU). Lat. 38.5244, long. -8.8882.

Seuilha. *Seuilha* 107(1). Cidade. Parte de: Espanha.
Sevilla, Spain (EU). Lat. 37.38283, long. -5.97317.

Siaca. *Siaca* 18(1), 19(1), 24(2), 32(1), 33(1). Cidade (rio). Parte de: Iambee. Outras relações: #1 dia náutico#@Arissumhee@; #5 léguas#@Sorobaya@.
Siak Sri Indrapura, Indonesia (AS). Lat. 0.79317, long. 102.0511.

Siammon. *Siammon* 107(1), 153(1), 157(4), 158(1), 165(1), 185(1). *Siammõ* 156(1), 170(1). *Siãmon* 162(1). *Siammom* 124(1), 151(1). Império. Parte de: Indochina. Outras relações: #area#@Gueos@; #area#@Tanguu@; #area#@Calaminhan@; #area#@Auaa@; #area#@Queitor@.

Sião. *Siame* 165(1), 183(1), 184(1), 185(1). *Siames* 1(1), 48(1), 57(1), 68(1), 149(1), 151(1), 159(1), 185(1), 186(1). *Sião* 17(1), 33(1), 36(5), 37(1), 39(1), 41(3), 44(1), 53(1), 55(2), 57(1), 58(1), 66(1), 76(1), 81(1), 82(1), 84(1), 86(1), 88(1), 91(1), 92(1), 95(1), 96(1), 100(1), 112(1), 122(1), 124(1), 128(1), 143(1), 144(1), 146(3), 148(1), 167(1), 181(8), 182(5), 183(5), 184(1), 185(2), 186(2), 189(3), 190(1), 200(4), 215(1), 220(1). *Siaõ* 46(2), 48(1), 181(1), 183(1). *Sioẽs* 150(1), 162(1). Reino. Parte de: Sornau. Outras relações: #area#@Tauquiday@.
Kingdom of Thailand, Thailand (AS). Lat. 15.5, long. 101.

Sicay. *Sicay* 208(1). Cidade. Parte de: Iapaõ. Outras relações: #18 léguas#@Miocoo@.
Sakai, Japan (AS). Lat. 34.58333, long. 135.46667.

Sidayo. *Sidayo* 36(1), 179(1). Porto. Parte de: Iaoa.

Sileupamor. *Sileupamor* 74(1). Cidade. Parte de: China. Outras relações: #area#@Nanquim, enseada@; #7 dias nauticos#@Tanquilem@.

Sileupaquim. *Sileupaquim* 71(1). Estreito. Parte de: Ásia Oriental. Outras relações: #5 dias#@Fanjus@.

Sileyjacau. *Sileyjacau* 81(1), 82(1). Vila. Parte de: China. Outras relações: #1 dia caminho#@Nanquim, enseada@; #147 léguas#@Nanquim@; #5 léguas#@Suzoanganee@; #7 léguas#@Xiangulee@; #area#@Taypor@; #area#@Conxinacau@.

Sinay. *Sinay* 6(1), 43(1), 96(1). Monte. Parte de: Egypto.
Jabal Mūsá, Egypt (AF). Lat. 28.53914, long. 33.97496.

Singapamor. *Singapamor* 128(2). *Singuapamor* 182(1). Lago. Parte de: Ásia Oriental. Outras relações: #area#@Xinaleygrau@; #area#@Ventinau@; #7 dias fluviais#@Caleypute@; #6 dias#@Taysirão@; #area#@Quitiruão@.

Singilapau. *Singilapau* 162(2). Cidade. Parte de: Calaminhan. Outras relações: #area#@Angegumaa@; #13 dias rio acima a partir de#@Chipanocão@; #1 légua#@Timplão@.

Singrachirau. *Singrachirau* 123(2). Muro. Parte de: China. Outras relações: #area#Tartaria@; #area#@Caixiloo@; #area#@Xipator@.
Great Wall, China (AS). Lat. 39.96675, long. 119.79533.

Singuafatur. *Singuafatur* 126(1). Templo (templo). Parte de: Tartaria. Outras relações: #5 dias fluviais#@Linxau@; #1 dia fluvial#@Quanginau@.

Sipautor. *Sipautor* 138(1). Povoação (lugar). Parte de: Lequios. Outras relações: #1 dia#@Gundexilau@.

Siputay. *Siputay* 181(1). Vale. Parte de: Sião. Outras relações: #1.5 léguas#@Quitiruão@; #area#@Suropisem@.

Solor. *Solor* 35(1), 36(1), 189(1). Ilha (porto). Parte de: Insulíndia. Outras relações: #area#@Borneo@.

Sornau. *Sornau* 17(1), 36(2), 41(1), 88(1), 95(1), 107(1), 122(1), 124(1), 128(1), 146(4), 148(1), 181(1), 182(1), 184(2), 185(2), 189(1). Império (reino). Parte de: Indochina. Outras relações: #area#@Tauquiday@.
Kingdom of Thailand, Thailand (AS). Lat. 15.5, long. 101.

Sorobaya. *Sorobaya* 25(1). *Surobaya* 33(1). *Surobayaa* 177(1). *Surubayaa* 179(1). Principado (ponta, povoação, principado). Parte de: Çamatra. Outras relações: #5 léguas#@Sorobaya@#area#@Iambee@.

Sorocataõ. *Sorocataõ* 162(1). Cidade (mítica). Parte de: Calaminhan.

Suez. *Suez* 2(1), 3(2), 146(1). Porto. Parte de: Egypto.
Suez, Egypt (AF). Lat. 29.97371, long. 32.52627.

Sumbor. *Sumbor* 44(1), 55(1), 60(1), 62(1), 85(1), 87(1), 115(1), 143(1). *Çumbor* 57(1). Reino (porto). Parte de: China. Outras relações: #metropoli#@Nanquim@.
Songmenzhen, China (AS). Lat. 28.34417, long. 121.60361.

Sumheehitão. *Sumheehitão* 38(1). Rio. Parte de: Malayo. Outras relações: #area#@Lugor@; #area#@Patane@.

Sumhepadaõ. *Sumhepadaõ* 71(1). Rio. Parte de: Ásia Oriental. Outras relações: #170 léguas#@Sileupaquim@.

Sunday patir. *Sunday patir* 195(1). Campo. Parte de: Pegù. Outras relações: #1 légua#@Pegù@.

Surião. *Surião* 188(1). Cidade. Parte de: Pegù. Outras relações: #area#@Digum@; #area#@Dalaa@.
Syriam, Myanmar [Burma] (AS). Lat. 16.76887, long. 96.24503.

Surobasoy. *Surobasoy* 166(1). Província. Parte de: Ásia. Outras relações: #area#@Lauhos@; #area#@Chiammay@; #area#@Pauel@.

Suropisem. *Suropisem* 181(2). Vila (lugar). Parte de: Sião. Outras relações: @#12 léguas#@Quitiruão@; #9 dia fluviais acima#@Odiaa@#16 dias caminho#@Odiaa@; #area#@Siputay@; .

Surotilau. *Surotilau* 14(1). Porto. Parte de: Aarù.

Susoquerim. *Susoquerim* 79(1). Aldea. Parte de: China. Outras relações: @Nanquim, enseada@; #area#@Xalingau@.

Suzoanganee. *Suzoanganee* 82(2). Povoação (lugar). Parte de: China. Outras relações: #5 léguas#@Sileyjacau@; #2 léguas#@Xiangulee@.

Tagalaa. *Tagalaa* 148(1), 168(1), 190(1). Fortaleza (lugar). Parte de: Pegù. Outras relações: #6 léguas#@Martauão@; #5 léguas#@Mounay@.

Taiquilleu. *Taiquilleu* 41(1). Vila. Parte de: Champaa. Outras relações: #area#@Tinacoreu@; #area#@Congrau@.

Talãgame. *Talãgame* 33(1). Porto. Parte de: Ternate.
Sango, Indonesia (AS). Lat. 0.81667, long. 127.38333.

Tanamadel. *Tanamadel* 74(1). Edifícios (edifícios). Parte de: Sileupamor.

Tanauçarim. *Tanaçarim* 96(1). *Tanauçarim* 17(1), 19(1), 20(1), 41(1), 124(1), 144(3), 145(1), 146(7), 147(3), 148(1), 189(1), 205(1). Porto (costa). Parte de: Sião.
Tanintharyi, Myanmar [Burma] (AS). Lat. 12.09032, long. 99.01165.

Tanauquir. *Tanauquir* 45(2), 46(2), 47(1), 49(2). Rio (cidade). Parte de: Cauchenchina. Outras relações: #40 léguas#@Mutipinão@; #2 dias nauticos#@Camoy@.

Tanguu. *Tãgù* 151(1). *Tãgus* 149(1). *Tãguu* 198(1). *Tangù* 41(1). *Tanguu* 114(1), 124(1), 152(1), 153(1), 155(1), 162(1), 163(1), 190(1), 194(1), 195(3), 196(1), 199(1). Reino. Parte de: Índico Oriental. Outras relações: #area#@Bramá@; #area#@Auaa@; #area#@Sauady@.
Taunggyi, Myanmar [Burma] (AS). Lat. 20.78919, long. 97.03776.

Tanixumaa. *Tanixumâ* 137(1). *Tanixumaa* 132(2), 133(2), 134(1), 135(3), 136(2), 137(2), 140(1), 143(1), 200(1), 223(2). Ilha. Parte de: Iapaõ.
Tanega Shima, Japan (AS). Lat. 30.66667, long. 131.

Tanjampura. *Tanjampura* 35(1), 36(1), 39(1). Porto. Parte de: Insulíndia. Outras relações: #area#@Laue@; #area#@Iaoa@.

Tanocos. *Tanocos* 149(1). Nação. Parte de: Arabia Felix.

Tanoraa. *Tanorâ* 135(1). *Tanoraa* 202(1), 223(1). Porto. Parte de: Iapaõ. Outras relações: #1dia náutico#@Tanoraa@; #1 dia náutico#@Canguexumaa@; #area#@Fiungaa@.

Tanquilem. *Tanquilem* 74(1). Lugar (terra). Parte de: China. Outras relações: #10 léguas#@Calempluy@; #7 dias náuticos#@Sileupamor@; #area#@Guinaytaraõ@.

Tapurau. *Tapurau* 185(1). Fortaleza. Parte de: Sião. Outras relações: #9 léguas#@Sacotay@.

Taraudachit. *Taraudachit* 129(1). Aposento (aposento). Parte de: Cauchenchina. Outras relações: #86 léguas#@Tinamquaxy@; #12 horas#@Lindau panoo@; #area#@Fanaugrem@.

Taraulachim. *Taraulachim* 41(1). Rio. Parte de: Indochina. Outras relações: #area#@Champaa@; #area#@Cunebetee@; #area#@Taiquilleu@; #area#@Pilaucacem@.

Tarees. *Tarees* 153(1). Nação. Parte de: Índico Oriental. Outras relações: #area#@Moẽs@; #area#@Chaleu@.

Tarem. *Tarem* 128(1). Povoação. Parte de: Cauchenchina. Outras relações: #9 dias fluviais#@Caleypute@; #7 dias#@Xolor@.

Tartaria. *Tartaria* 1(1), 67(1), 71(1), 88(1), 89(1), 95(2), 107(1), 108(2), 117(2), 123(1), 124(1), 128(1), 130(5), 131(2). *Tartaro* 45(1), 68(1), 95(3), 117(2), 118(2), 120(1), 122(2), 123(3), 124(2), 125(2), 126(2), 130(2), 131(1), 185(1). *Tartaros* 88(1), 104(1), 110(1), 113(1), 114(2), 118(3), 119(4), 120(1), 122(2), 131(1), 149(1), 178(1). Império. Parte de: Ásia Oriental.

Tauay. *Tauay* 146(1). *Touay* 147(1). Porto. Parte de: Sião. Outras relações: #area#@Martauão@; #area#@Tanauçarim@; #area#@Vagaruu@; #area#@Iuncay@; #area#@Pullo Camude@; #area#@Merguim@.
Dawei, Myanmar [Burma] (AS). Lat. 14.0823, long. 98.19151.

Taubasoy. *Taubasoy* 146(1). *Tobasoy* 146(1). Ilha. Parte de: Sião. Outras relações: #7 léguas#@Pullo Hinhor@; #area#@Çambilão@.

Tauquiday. *Tauquiday* 88(1). Rio. Parte de: Sião. Outras relações: #area#@Sião@; #area#@Cuy@.
Tauquiday, (). Lat. 0, long. 0.

Taurys. *Taurys* 107(1). Cidade. Parte de: Persia.
Tabriz, Iran (AS). Lat. 38.08, long. 46.2919.

Tautaa. *Tautaa* 87(1). Ilhéus (ilhéus). Parte de: China. Outras relações: #area#@Nanquim@; #area#@Sumbor@; #area#@Fanjus@.

Taydacão. *Taidacão* 140(1). *Taydacão* 138(1), 140(1). Serra. Parte de: Lequios.

Taypor. *Taipor* 86(1), 100(1), 103(1). *Taypor* 84(2), 85(2), 87(1). Vila (lugar). Parte de: China. Outras relações: #area#@Nanquim@; #area#@Xianguulee@; #area#@Chautir@; #area#@Guinapalir@.

Taysiraõ. *Taysiraõ* 182(1). Cidade. Parte de: Chiammay. Outras relações: #6 dias#@Singapamor@; #area#@Guibem@; #area#@Sião@.

Ternate. *Ternate* 20(1), 33(1), 143(1). Ilha (fortaleza). Parte de: Insulíndia. Outras relações:

#area#@Maluco@.
Ternate Island, Indonesia (AS). Lat. 0.80562, long. 127.34044.

Tigremahom. *Tigremahom* 4(1), 225(1). Reino. Parte de: Etiopia.
Tigray Region, Ethiopia (AF). Lat. 14.16667, long. 38.83333.

Tilau. *Tilau* 185(1). Lugar. Parte de: Iunçalão. Outras relações: #140 léguas#@Malaca@; #area#@Quedaa@.

Tilaumera. *Tilaumera* 47(3). Ponta (morro). Parte de: Cauchenchina. Outras relações: #9 léguas#@Panduree@; #3 dias náuticos#@Mutipinão@; #area#@Colem@.

Timocouhós. *Timocouhós* 130(1), 181(1). *Timocouhos* 131(1). *Tinocouhós* 129(1). Nação. Parte de: Ásia. Outras relações: #area#@Chiammay@; #area#@Lauhos@; #area#@Gueos@; #area#@Sião@; #area#@Calaminhan@.

Timor. *Timor* 21(1), 26(1), 57(1), 177(1). Ilha. Parte de: Insulíndia.
Timor Sea, (AS). Lat. -10.5, long. 126.

Timpates. *Timpates* 166(1). Nação. Parte de: Ásia Oriental. Outras relações: #area#@Calouhos@; #area#@Bugẽ@; #area#@Fungaos@; #area#@Calogẽs@; #area#@Pauel@.

Timplão. *Timplão* 158(2), 162(2), 165(2). *Timplaõ* 107(1). Cidade (metrópole). Parte de: Calaminhan. Outras relações: #area#@Pituy@; #area#@Singilapau@; #area#@Bidor@.

Tinacoreu. *Tinacoreu* 41(3). *Tinaçoreu* 42(1). Rio. Parte de: Indochina. Outras relações: #area#@Champaa@; #area#@Cunebetee@; #area#@Taiquilleu@; #area#@Pilaucacem@.

Tinagoogoo. *Tinagoogoo* 158(2), 159(3), 161(2), 162(1). Pagode (edifício, pagode). Parte de: Chipanocão.

Tinamquaxy. *Tinamquaxy* 129(1). Cidade. Parte de: Cauchenchina. Outras relações: #12 horas#@Manaquileu@; #86 léguas#@Taraudachit@; #107 léguas#@Fanaugrem@.

Tinlau. *Tinlau* 58(2), 61(2). Rio. Parte de: China. Outras relações: #3 dias náuticos#@Lailoo@.

Tinlau. *Tinlau* 107(1), 158(1). Cidade (senhorio). Parte de: Siammon. Outras relações: #area#@Angegumaa@; #area#@Iangomaa@; #area#@Calaminhan@; #area#@Siammon@; #7 dias rio acima a partir de#@Gumbim@.

Tõdacur. *Tõdacur* 16(1). Lugar. Parte de: Achem. Outras relações: #2 léguas#@Achem, cidade@; #area#@Penacão@; #area#@Quilem@.

Toobasoy. *Toobasoy* 40(1), 41(1). Rio. Parte de: Champaa. Outras relações: #12 horas#@Saleyjacau@; #2 dias#@Tinacoreu@.

Toro. *Toro* 6(1). Cidade. Parte de: Arabia felix.
El-Tor, Egypt (AF). Lat. 28.24168, long. 33.6222.

Tosa. *Tosa* 39(1), 135(1), 143(1), 218(1), 224(1). Ilha. Parte de: Iapaõ. Outras relações:

#coordenadas#@36@.
Shikoku Chihō, Japan (AS). Lat. 33.75, long. 133.5.

Toscana. *Toscana* 164(1). Ducado (ducado). Parte de: Europa.
Toscana, Italy (EU). Lat. 43.41667, long. 11.

Tudescos. *Tudescos* 124(1). Nação. Parte de: Europa.
Federal Republic of Germany, Germany (EU). Lat. 51.5, long. 10.5.

Tuparaas. *Tuparaas* 149(1). Nação. Parte de: Ásia.

Tuparahos. *Tuparahos* 183(1). Reino. Parte de: Ásia. Outras relações: #area#@Passiloco@; #area#@Sião@.

Tuparoens. *Tuparoens* 166(1). Nação. Parte de: Ásia Oriental. Outras relações: #area#@Pauel@; #area#@Pituy@.

Turbão. *Turbão* 15(1), 16(1). *Turbaõ* 16(1). Lugar. Parte de: Bata. Outras relações: #18 léguas#@Achem@; #5 léguas#@Panaajù@.

Turcos. *Turco* 2(2), 3(3), 4(1), 5(1), 8(2), 17(1), 18(1), 21(1), 26(2), 27(2), 32(1), 40(2), 43(2), 146(7), 186(1), 191(2), 225(2). *Turcos* 3(1), 4(1), 5(6), 6(1), 7(4), 8(2), 9(5), 10(3), 11(2), 12(5), 13(1), 16(3), 20(1), 21(1), 22(1), 26(2), 32(1), 36(1), 43(1), 145(1), 146(4), 149(2), 155(1), 173(1), 175(1), 178(1), 181(1), 185(1), 186(4), 206(4). *Turquesca* 122(1). *Turquesco* 188(1). *turquescos* 206(1). Nação. Parte de: Índico Ocidental.
Republic of Turkey, Turkey (AS). Lat. 39, long. 35.

Tuxenguim. *Tuxenguim* 92(1). Serra. Parte de: China. Outras relações: #5 léguas#@Lequimpau@; #area#@Batampina@; #1 dia fluvial#@Pacão@; #1 dia fluvial#@Nacau@; #area#@Sampitay@.

Tuymicão. *Taymicão* 125(1). *Tuymicão* 45(1), 124(2), 126(2). Cidade (metrópole). Parte de: Tartaria. Outras relações: #capital#@Tartaria@.

Vagaruu. *Vagarù* 150(1). *Vagaruu* 146(1), 147(1). Porto. Parte de: Sião. Outras relações: #area#@Martauão@; #area#@Tanauçarim@; #area#@Touay@; #area#@Iuncay@; #area#@Pullo Camude@; #area#@Merguim@.

Valeutay. *Valeutay* 170(1). Lugar. Parte de: Sauady. Outras relações: #9 dias#@Oregantor@.

Vangaleu. *Vangaleu* 4(1). Serra. Parte de: Etiopia. Outras relações: #area#@Betenigus@; #15 léguas#@Gileytor@; #12.5 léguas#@Fumbau@.

Vbra. *Vbra* 202(1). Porto. Parte de: Iapaõ.
Aburatsu, Japan (AS). Lat. 31.58006, long. 131.40908.

Veneza. *Veneza* 5(1), 107(1). *Venezeanos* 149(1). Cidade. Parte de: Europa.
Venice, Italy (EU). Lat. 45.43713, long. 12.33265.

Ventinau. *Ventinau* 129(1). *Ventrau* 128(1). *Vẽtrau* 167(1). Rio. Parte de: Indochina.

villa de Conde. *villa de Conde* 16(1). *Villa de Conde* 1(1). Vila. Parte de: Portugal. Outras relações: #area#@Singapamor@; #area#@Sião@; #area#@Chiãtabuu@; #area#@Timplão@; #area#@Pituy@.
Vila do Conde, Portugal (EU). Lat. 41.60407, long. -7.53272.

Villanoua. *Villanoua* 1(1). Vila. Parte de: Portugal.
Portimão, Portugal (EU). Lat. 37.13856, long. -8.53775.

Vngaro. *Vngaro* 96(1). Nação. Parte de: Europa.
Hungary, Hungary (EU). Lat. 47, long. 20.

Voulem. *Voulem* 128(1). Povoação. Parte de: Xinaleygrau.

Vpe. *Vpe* 14(1), 203(1). Ilha. Parte de: Malaca. Outras relações: #0.33 léguas#@Malaca, fortaleza@.
Pulau Upeh, Malaysia (AS). Lat. 2.1926, long. 102.2031.

Vrpanesendoo. *Vrpanesendoo* 162(1). Templo. Parte de: Calaminhan. Outras relações: #area#@Angegumaa@; #1 légua#@Campalagrau@; #1 légua#@Timplão@; #area#@Manicafaraõ@.

Xael. *Xael* 4(2). Porto. Parte de: Arabia felix.
Ash Shiḩr, Yemen (AS). Lat. 14.76026, long. 49.60537.

Xalingau. *Xalingau* 79(1). Esteiro. Parte de: China. Outras relações: #area#@Nanquim, enseada@.

Xamoy. *Xamoy* 55(3). Porto (aldea, povoação). Parte de: Xingrau. Outras relações: #18 léguas#@Guintoo@; #9 léguas#@Pullo Quirim@; #area#@ilha dos ladroẽs@; #area#@Luxitay@.

Xaraa. *Xaraa* 190(1), 193(1). Cidade. Parte de: Pegù.

Xemenaxeque. *Xemenaxeque* 218(2). *Xamanaxeque* 135(1). Ilha. Parte de: Iapaõ. Outras relações: #area#@Goto@; #area#@Bungo@.

Xenxinapau. *Xenxinapau* 95(1). Província. Parte de: Ásia Oriental. Outras relações: #area#@Lauhos@; #area#@Tartaria@; #area#@China@.

Xeque. *Xeque* 223(2), 225(1), 226(2). Ilha (porto). Parte de: Bungo. Outras relações: #2 léguas#@Osquy@.
Saganoseki, Japan (AS). Lat. 33.2478, long. 131.87083.

Xianguulee. *Xianguulee* 82(1), 83(1). Aldea (lugar). Parte de: China. Outras relações: #140 léguas#@Nanquim@; #2 léguas#@Suzoanganee@; #7 léguas#@Sileyjacau@; #area#@Taypor@; #area#@Chautir@; #area#@Guinapalir@.

Xifangau. *Xifangau* 96(1). Aldea. Parte de: Batampina. Outras relações: #area#@Cohilouza@; #10 léguas#@Mindoo@.

Xilendau. *Xilendau* 61(1). Rio. Parte de: China. Outras relações: #1.5 léguas#@Mecuy@;

#area#@Xilendau@; #area#@Nipafau@; #area#@Tinlau@; #area#@Liampoo@.

Xinaleygrau. *Xinaleygrau* 128(1). Senhorio (senhorio). Parte de: Ásia Oriental. Outras relações: #4 dias fluviais#@Rendacalem@; #7 dias fluviais#@Quatanqur@; #area#@Singapamor@.

Xinamguibaleu. *Xinamguibaleu* 108(3), 109(1). *Xinanguibaleu* 108(2). Prisão. Parte de: Pequim.

Xincaleu. *Xincaleu* 39(1). Cidade. Parte de: Quitiruão. Outras relações: #area#@Pinator@.

Xingrau. *Xingrau* 55(1). *Xinguau* 55(1). Rio. Parte de: China. Outras relações: #18 léguas#@Guintoo@; #9 léguas#@Pullo Quirim@; #area#@ilha dos ladroẽs@; #area#@Luxitay@.

Xinligau. *Xinligau* 90(1). Cidade. Parte de: China. Outras relações: #area#@Batampina@; #1 dia#@Iunquileu@; #12 horas#@Pocasser@.

Xinxipou. *Xinxipou* 88(1). Anchacilado (anchacilado). Parte de: China. Outras relações: #area#@Leysacotay@; #area#@Moscouia@.

Xipatom. *Xipatom* 221(1). Aldea. Parte de: Liampoo. Outras relações: #2 léguas#@Liampoo@.

Xipatom. *Xipatom* 122(1). Cidade. Parte de: China. Outras relações: #area#@Tartaria@.

Xipator. *Xipator* 123(1). Cidade. Parte de: Tartaria. Outras relações: #area#@Singrachirau@; #6 dias fluviais#@Lançame@; #area#@Caixiloo@.

Xiuau. *Xiuau* 183(1). Fortaleza. Parte de: Sião. Outras relações: #area#@Lantor@; #area#@Passiloco@.

Xolor. *Xolor* 128(2), 129(2). Cidade. Parte de: Cauchenchina. Outras relações: #7 dias#@Tarem@; #5 dias#@Manaquileu@.

Zeila. *Zeila* 225(1). Reino. Parte de: Africa. Outras relações: #area#@Etiopia@.

Bibliografia

Geral

Agnew, J. (2011). Space and place. In A. Agnew & D. Livingstone (Eds.), *The SAGE handbook of geographical knowledge* (pp. 16-330). London: SAGE. Disponível em: http://www.sscnet.ucla.edu/geog/downloads/856/416.pdf

Alex, B., Byrne, K., Grover, C., & Tobin, R. (2015). Adapting the Edinburgh Geoparser for Historical Georeferencing. *International Journal of Humanities and Arts Computing*, *9*(1), 15-35. Disponível em: http://www.euppublishing.com/doi/full/10.3366/ijhac.2015.0136

Almeida, G. M. B., Aluísio, S. M., & Oliveira, L. H. M. (2001). O método em terminologia: revendo alguns procedimentos. In A. N. Isquerdo & I. M. Alves (Orgs.), *Ciências Do Léxico: lexicologia, lexicografia, terminologia* (Vol. 3, pp. 409-420). Campo Grande/São Paulo: Editora da UFMS/Humanitas. Disponível em: http://www.geterm.ufscar.br/textospublicados/o_metodo_em_terminologia_%20revendo_alguns_procedimentos.pdf

Alpaydin, E. (2014). *Introduction to Machine Learning* (3rd ed.). Massachusetts: MIT.

Alves, D., & Queiroz, A. I. (2013). Studying urban space and literary representations using GIS: Lisbon, Portugal, 1852–2009. *Social Science History*, *37*(04), 457-481. Disponível em: https://www.researchgate.net/profile/Ana_Queiroz6/publication/293796558_Studying_Urban_Space_and_Literary_Representations_Using_GIS/links/56c1ab8b08aee5caccf8415c.pdf

Alves, D., & Queiroz, A. I. (2015). Exploring literary landscapes: From texts to spatiotemporal analysis through collaborative work and GIS. *International Journal of Humanities and Arts Computing*, *9*(1), 57-73. Disponível em: http://www.euppublishing.com/doi/full/10.3366/ijhac.2015.0138

Amaral, D. O. (2013). *O reconhecimento de entidades nomeadas por meio de conditional random fields para a língua portuguesa* (Tese de mestrado, Universidade Católica do Rio Grande do Sul, Brasil). Disponível em: http://tede.pucrs.br/tde_arquivos/4/TDE-2014-04-24T051906Z-4975/Publico/457280.pdf

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), 2014, Reykjavik* (pp. 2554-2558). European Language Resources Association (ELRA). Disponível em: http://www.lrec-conf.org/proceedings/lrec2014/pdf/513_Paper.pdf

Anastácio, I., Martins, B., & Calado, P. (2009). A comparison of different approaches for assigning geographic scopes to documents. *Proceedings of the 1st INForum-Simpósio de Informática* (pp. 285-296). Disponível em: http://xldb.lasige.di.fc.ul.pt/xldb/publications/Anastacio.etal:AComparisonOf:2009_document.pdf

Anastácio, I., Martins, B., & Calado, P. (2011). Supervised learning for linking named entities to knowledge base entries. *Proceedings of the Text Analysis Conference*. Disponível em: http://web.ist.utl.pt/bruno.g.martins/papers/paper.pdf

Ballatore, A. (2016). Prolegomena for an Ontology of Place. In H. Onsrud & W. Kuhn (Eds.), *Advancing Geographic Information Science: The Past and Next Twenty Years* (pp. 91-104). Needham, MA : GSDI.

Biderman, M. T. C. (1998). A face quantitativa da linguagem: um dicionário de freqüências do português. *ALFA: Revista de Linguística*, *42*(1). Disponível em: http://seer.fclar.unesp.br/alfa/article/download/4049/3713

Baroni, M., Bernardi, R., & Zamparelli, R. (2014). Frege in space: A program of compositional distributional semantics. *Linguistic Issues in Language Technology*, *9*(6), 5-110. Disponível em: http://csli-lilt.stanford.edu/ojs/index.php/LiLT/article/download/6/5

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University.

Barros, João de. (1540). *Gramática da língua portuguesa.* Olyssipone : apud Lodouicum Rotorigiu[m], Typographum. Ed. Digital Biblioteca Nacional de Portugal. Disponível em: http://purl.pt/12148

Benguigui, L., & Blumenfeld-Lieberthal, E. (2011). The end of a paradigm: is Zipf's law universal?. *Journal of Geographical Systems*, 13(1), 87-100.

Bian, C., Lin, R., Zhang, X., Ma, Q. D., & Ivanov, P. C. (2016). Scaling laws and model of words organization in spoken and written language. *Europhysics Letters*, 113(1), 18002. Disponível em: https://www.researchgate.net/profile/Plamen_Ivanov/publication/292206119_Scaling_laws_and_model_of_words_organization_in_spoken_and_written_language/links/56aff74a08ae8e37214d146d.pdf

Bod, R. (2003). Introduction to elementary probability theory and formal stochastic language theory. In R. Bod, J. Hay & S. Jannedy (Eds.), *Probabilistic Linguistics* (pp. 11-38). Massachusetts: MIT.

Bol, P., & Ge, J.(2005). China Historical GIS. *Historical Geography*, *33*,150-2.

Bol, P. K. (2007). Creating a GIS for the History of China. In A. K. Knowles (Ed.), *Placing History: How Maps, Spatial Data, and GIS Are Changing Historical Scholarship* (pp. 27-58). New York: ESRI.

Bol, P. K. (2013). On the Cyberinfrastructure for GIS-Enabled Historiography: Space–Time Integration in Geography and GIScience. *Annals of the Association of American Geographers*, *103*(5), 1087-1092.

Bruggmann, A., & Fabrikant, S. I. (2014). Spatializing time in a history text corpus. In *Proceedings of the 8th International Conference on Geographic Information Science, GIScience (extended abstracts)*.

Buitelaar, P., Cimiano, P., & Magnini, B. (2005). Ontology learning from text: An overview. In P. Buitelaar, P. Cimiano, & B. Magnini (Eds.), *Ontology learning from text: Methods, evaluation and applications* (pp. 3-12). Amsterdam: IOS.

Bybee, J., & Hopper, P. (2001). Introduction to frequency and the emergence of linguistic structure. In J. Bybee, & P Hopper, (Eds.), *Frequency and the emergence of linguistic structure* (pp. 1-24). Amsterdam / Philadelphia: John Benjamins.

Cann, R., Kempson, R., & Gregoromichelaki, E. (2009). *Semantics. An Introduction to Meaning in Language*. Cambridge: Cambridge University.

Chaves, M. (2008a). Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o SEI-Geo no Segundo HAREM. In C. Mota & D. Santos (Eds.), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM* (pp. 231-245). Disponível em: http://www.linguateca.pt/harem/actas/Capitulo_13-MotaSantos2008.pdf

Chaves, M. (2008b). Criação e expansão de geo-ontologias e dimensionamento de informação geográfica e reconhecimento de locais e seus relacionamentos em textos. In *Linguateca: 10 anos.* Encontro Satélite do PROPOR 2008 organizado pela Linguateca. Disponível em: http://www.linguateca.pt/Linguateca10anos/ResumosAlargados/ChavesL10.pdf

Chaves, M. S. (2009). *Uma metodologia para construção de geo-ontologias*. (Tese de Doutoramento. Universidade de Lisboa, Portugal). Disponível em: http://hdl.handle.net/10451/1642

Chaves, M., Rodrigues, C., & Silva, M. J. (2007). Data Model for Geographic Ontologies Generation. In J. C. Ramalho; J. Correia Lopes; L. Carriço (Eds.), *XML: Aplicações e Tecnologias Associadas (XATA 2007)* (pp. 47–58). Disponível em: http://xldb.lasige.di.fc.ul.pt/xldb/publications/xata2007-camera-ready.pdf

Chaves, M. S., Silva, M. J., & Martins, B. (2005). A Geographic Knowledge Base for Semantic Web Applications. In C. A. Heuser (Ed.), *Proceedings do 20º Simpósio Brasileiro de Banco de Dados (SBBD) (Uberlândia MG Brasil 3-7 de Outubro de 2005)*. Disponível em: https://pdfs.semanticscholar.org/14bb/ad9f04ee49da0baeca7388f08c6d33c3773a.pdf

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Massachussetts: MIT.

Chomsky, N. (1978). *The Logical Structure of Linguistic Theory*. 3rd ed. New York: Plenum.

Clark, S. (2015). Vector Space Models of Lexical Meaning. In S. Lappin and C. Fox (Eds.), *Handbook of Contemporary Semantic Theory, 2nd ed.* (pp. 493-522). Oxford: Wiley-Blackwell. Disponível em: http://www.cl.cam.ac.uk/%7Esc609/pubs/sem_handbook.pdf

Conrado, M. D. S. (2014). *Extração automática de termos simples baseada em aprendizado de máquina* (Tese de doutoramento, Universidade de São Paulo, Brasil). Disponível em: http://www.teses.usp.br/teses/disponiveis/55/55134/tde-11082014-103430/publico/TeseMerley_revisada.pdf

Conrado, M. S., Felippo, A., Pardo, T. A. S., & Rezende, S. O. (2014). A survey of automatic term extraction for Brazilian Portuguese. *Journal of the Brazilian Computer Society*, *20*(12), 1-28. Disponível em: https://journal-bcs.springeropen.com/articles/10.1186/1678-4804-20-12

Conrado, M. S., Pardo, T. A., & Rezende, S. O. (2015). The main challenge of semi-automatic term extraction methods. In *Proceedings of the 11st International Workshop on Natural Language Processing and Cognitive Science* (pp. 27-29). Venice, Italy: NLPCS. Disponível em: http://conteudo.icmc.usp.br/pessoas/taspardo/NLPCS2014-ConradoEtAl.pdf

DeLozier, G., Baldridge, J., & London, L. (2015). Gazetteer-Independent Toponym Resolution Using Geographic Word Profiles. In *Proceedings of AAAI 2015*. Austin, Texas. Disponível em: http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9823

DeLozier, G., Wing, B., Baldridge, J., & Nesbit, S. (2016). Creating a Novel Geolocation Corpus from Historical Texts. In *Proceedings of the 10th Linguistic Annotation Workshop* (pp. 188-198). Berlin, Germany. Disponível em: https://aclweb.org/anthology/W/W16/W16-1721.pdf

Dias, D, Anastácio, I., & Martins, B. (2012). Geocodificação de Documentos Textuais com Classificadores Hierárquicos Baseados em Modelos de Linguagem. *Linguamática*, *4*(2), 13-25. Disponível em: http://linguamatica.com/index.php/linguamatica/article/view/139

Dick, M. V. P. A. (1975). O problema das taxionomias toponímicas. (Uma contribuição metodológica). *Língua e Literatura*, 4, 373-380. Disponível em: http://www.periodicos.usp.br/linguaeliteratura/article/view/122791/119267

Dinu, G., Thater, S., & Laue, S. (2012). A comparison of models of word meaning in context. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 611-615). Association for Computational Linguistics. Disponível em: http://www.anthology.aclweb.org/N/N12/N12-1.pdf

Donaldson, C. E., Bushell, S. C., Gregory, I. N., Rayson, P. E., & Taylor, J. E. (2016). Digital literary geography and the difficulties of locating 'Redgauntlet Country'. *Studies in Scottish Literature*, *42*(2), 174-183. Disponível em: http://scholarcommons.sc.edu/ssl/vol42/iss2/5/

Feinerer, I. (2008). An Introduction to Text Mining in R. *R News. The Newsletter of the R Project*, *8*(2), 19-22. Disponível em:

https://pdfs.semanticscholar.org/6c70/f07658c8a63ef7bf622b24900595650baf5c.pdf

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of statistical software*, *25*(5), 1-54. Disponível em: https://www.jstatsoft.org/article/view/v025i05

Fellbaum, C. (1998). Introduction. In C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database* (pp. 1-19). Cambridge, Massachusetts: MIT.

Felippo, A., & Almeida, G. M. B. (2010). Uma metodologia para o desenvolvimento de Wordnets terminológicas em português do Brasil. *Tradterm*, 16, 365-395. Disponível em: http://www.revistas.usp.br/tradterm/article/view/46325/50088

Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (pp. 363-370). Association for Computational Linguistics.

Freitas, C., Mota, C., Santos, D., Gonçalo Oliveira, H., & Carvalho, P. (2010). Second HAREM: Advancing the State of the Art of Named Entity Recognition in Portuguese. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias (Eds.), *Proceedings of the seventh international conference on language resources and evaluation (LREC'10)*. European Language Resources Association (ELRA). Disponível em: http://www.linguateca.pt/Diana/download/FreitasetalLREC2010.pdf

Gamallo, P. (2016). Comparing explicit and predictive distributional semantic models endowed with syntactic contexts. *Language Resources and Evaluation*, 1-17.

Gamallo, P., & Garcia, M. (2011). A resource-based method for named entity extraction and classification. In *Progress in Artificial Intelligence* (pp. 610-623). Springer Berlin Heidelberg. Disponível em: https://pdfs.semanticscholar.org/28b3/3b51257a76ef95c06ae6a2b402fd99e031be.pdf

Gammeltoft, P. (2016). Names and geography. In *Oxford Handbook on Names and Naming* (pp. 467-475). Oxford: Oxford University.

Garcia, M. (2014). *Extração de Relações Semânticas. Recursos, Ferramentas e Estratégias*. (Tese de doutoramento, Universidade de Santiago de Compostela, Santiago de Compostela). Disponível em: http://docplayer.com.br/9405015-Extraccao-de-relacoes-semanticas-recursos-ferramentas-e-estrategias.html

Garcia, M., & Gamallo, P. (2015). Yet another suite of multilingual NLP tools. In *International Symposium on Languages, Applications and Technologies* (pp. 65-75). Springer International Publishing. Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.717.6921&rep=rep1&type=pdf

Garside, R., Leech, G. N., & McEnery, T. (Eds.). (1997). *Corpus annotation: linguistic information from computer text corpora*. Taylor & Francis.

Generseth, M., & Nilsson, N. (1998). *Logical foundations of Artificial Intelligence*. Palo Alto, CA: Morgan Kaufmann.

Gonçalo Oliveira, H., & Gomes, P. (2010). Onto.PT: automatic construction of a lexical ontology for Portuguese. In *Proceedings of 5th European Starting AI Researcher Symposium (STAIRS 2010)* (pp. 199-211). Lisbon, Portugal. Disponível em: http://eden.dei.uc.pt/~hroliv/pubs/GoncaloOliveira_Gomes_STAIRS2010.pdf

Gonçalo Oliveira, H., & Gomes, P. (2014). ECO and Onto. PT: a flexible approach for creating a Portuguese wordnet automatically. *Language resources and evaluation*, *48*(2), 373-393. Disponível em: http://link.springer.com/article/10.1007%2Fs10579-013-9249-9

Gonçalo Oliveira, H., & Gomes, P. (2016). Discovering Fuzzy Synsets from the Redundancy in Different Lexical-Semantic Resources. In Nicoletta Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard,J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds), *Proceedings of the Tenth International Conference on Language Resources and Evaluation {LREC} 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association. Disponível em: http://www.lrec-conf.org/proceedings/lrec2016/pdf/138_Paper.pdf

Gonçalo Oliveira, H., Paiva, V., Freitas, C., Rademaker, A., Real, L., & Simões, A. (2015). As wordnets do português. *Oslo Studies in Language*, *7*(1), 397-424. Disponível em: https://www.journals.uio.no/index.php/osla/article/view/1445/1342

Gregory, I. N., Baron, A., Cooper, D., Hardie, A., Murrieta-Flores, P., & Rayson, P. (2014). Crossing Boundaries: Using GIS in Literary Studies, History and Beyond. *Collections électroniques de l'INHA. Actes de colloques et livres en ligne de l'Institut national d'histoire de l'art*. Acedido em 17 novembro de 2016, em https://inha.revues.org/4931?lang=fr

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: Indiana University Press. Disponível em: http://e-space.mmu.ac.uk/579357/2/Spatializing%20and%20Analyzing%20Digital%20Texts.pdf

Gregory, I., Donaldson, C., Murrieta-Flores, P., & Rayson, P. (2016). Geoparsing, GIS, and Textual Analysis: Current Developments in Spatial Humanities Research. *International Journal of Humanities and Arts Computing*, *9*(1), 1-14. Disponível em: http://www.euppublishing.com/doi/full/10.3366/ijhac.2015.0135

Gregory, I. N., & Ell, P. S. (2007). *Historical GIS: technologies, methodologies, and scholarship* (Vol. 39). Cambridge: Cambridge University.

Gregory, I. N., & Hardie, A. (2011). Visual GISting: bringing together corpus linguistics and Geographical Information Systems. *Literary and linguistic computing*, *26*(3), 297-314.

Gries, S. T. (2009). *Statistics for linguistics with R: a practical introduction*. Berlin: De Gruyter Mouton.

Grover, C., Tobin, R., Byrne, K., Woollard, M., Reid, J., Dunn, S., & Ball, J. (2010). Use of the Edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *368*(1925), 3875-3889. Disponível em: http://rsta.royalsocietypublishing.org/content/368/1925/3875.short

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing?. *International journal of human-computer studies*, *43*(5), 907-928.

Guarino, N. (1995). Formal ontology, conceptual analysis and knowledge representation. *International journal of human-computer studies*, *43*(5), 625-640.

Guimarães, N. C. (2015). SABENÇA-um arcabouço computacional baseado na aprendizagem de ontologias a partir de textos. (Tese de Mestrado em Ciência da Computação, Universidade Federal de Goiás, Brasil). Disponível em: http://repositorio.bc.ufg.br/tede/bitstream/tede/4712/5/Disserta%C3%A7%C3%A3o%20-%20Norton%20Coelho%20Guimar%C3%A3es%20-%202015.pdf

Habert, B., Nazarenko, A., & Salem, A. (1997). *Les linguistiques de corpus*. Paris: Armand Colin.

Haegeman, L. M.V. (2007). *Thinking syntactically: a guide to argumentation and analysis*. Oxford: Blackwell Publishing.

Haspelmath, M., & Sims, A. (2010) *Understanding Morphology* (2nd ed.). London: Hodder Education.

Hennig, B.D. (2016). Mapping practices in a digital world. In H. Onsrud & W. Kuhn (Eds.), *Advancing Geographic Information Science: The Past and Next Twenty Years* (pp. 153-168). Needham, MA : GSDI.

Herbelot, A. (2015). Mr Darcy and Mr Toad, gentlemen: distributional names and their kinds. In *Proceedings of the 11th International Conference on Computational Semantics* (pp. 151-161). Disponível em: http://www.aclweb.org/anthology/W/W15/W15-01.pdf#page=167

Hilberg, W. (2002). The Unexpected Fundamental Influence of Mathematics upon Language. *Glottometrics*, *5*, 29-50. Disponível em: https://www.researchgate.net/profile/Jeff_Robbins2/publication/237053213_Technology_ease_and_entropy_a_testimonial_to_Zipfs_Principle_of_Least_Effort/links/53d2d2f30cf228d363e96322.pdf#page=32

Hu, Y., & Janowicz, K. (2016). Enriching Top-down Geo-ontologies Using Bottom-up Knowledge Mined from Linked Data. *Advancing Geographic Information Science: The Past and Next Twenty Years* (pp. 183-198). Needham, MA : GSDI.

IBGE. (2015). *Glossário dos Termos Genéricos dos Nomes Geográficos Utilizados no Mapeamento Sistemático do Brasil*. Vol. 2. Rio de Janeiro: Ministério do Planejamento, Orçamento e Gestão Instituto Brasileiro de Geografia e Estatística – IBGE.

Jackendoff, R. (1990). *Semantic structures*. Cambridge, Massachusetss: MIT.

Jackendoff, R. (2010). *Meaning and the Lexicon*. Oxford: Oxford University.

Janowicz, K., Raubal, M., & Kuhn, W. (2015). The semantics of similarity in geographic information retrieval. *Journal of Spatial Information Science*, (2), 29-57. Disponível em: http://josis.org/index.php/josis/article/view/26

Ji, H., & Grishman, R. (2011). Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1* (pp. 1148-1158). Association for Computational Linguistics. Disponível em: http://www.aclweb.org/anthology/P11-1115.pdf

Jones, D. M., & Paton, R. C. (1999). Toward principles for the representation of hierarchical knowledge in formal ontologies. *Data & Knowledge Engineering*, *31*(2), 99-113. Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.34.6827&rep=rep1&type=pdf

Jones, C. B., & Purves, R. S. (2008). Geographical information retrieval. *International Journal of Geographical Information Science*, *22*(3), 219-228. Disponível em: http://www.tandfonline.com/doi/full/10.1080/13658810701626343

Jurafsky, D., Bell, A., & Gregory, P. (2001). Probabilistic relations between words. In J. Bybee, & P Hopper, (Eds.), *Frequency and the emergence of linguistic structure*. Amsterdam / Philadelphia: John Benjamins.

Jurafsky, D., & Martin, J. H. (2015). Vector Semantics. In *Speech and Language Processing* (3rd ed.). Disponível em: https://web.stanford.edu/~jurafsky/slp3/15.pdf

Knowles, A. K. (2007). GIS and history. In A. K. Knowles (Ed.). *Placing History: How Maps, Spatial Data, and GIS Are Changing Historical Scholarship* (pp. 1-25). New York: ESRI.

Köhler, R. (2002). Power Law Models in Linguistics: Hungarian. *Glottometrics*, *5*, 51-61. Disponível em: https://www.researchgate.net/profile/Jeff_Robbins2/publication/237053213_Technology_ease_and_entropy_a_testimonial_to_Zipfs_Principle_of_Least_Effort/links/53d2d2f30cf228d363e96322.pdf

Kornai, A. (2008). *Mathematical Linguistics*. London: Springer.

Kraak, M. J., & Ormeling, F. (2010). *Cartography: visualization of spatial data* (3rd ed.). New York: Guilford Press.

Kuhn, M. (2008). Building Predictive Models in R Using the caret. *Journal of statistical software*, *28*(5). Disponível em: http://www.jstatsoft.org/article/view/v028i05/v28i05.pdf

Kuhn, M. (2016). *Caret: classification and regression training*. Disponível em: https://CRAN.R-project.org/package=caret

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago.

Leech, G. (1981). *Semantics*. 2nd ed. Harmondsworth, Middlesex: Penguin.

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

Li, W. (2002). Zipf's Law Everywhere. *Glottometrics*, *5*, 14-21. Disponível em: https://www.researchgate.net/profile/Jeff_Robbins2/publication/237053213_Technology_ease_and_entropy_a_testimonial_to_Zipfs_Principle_of_Least_Effort/links/53d2d2f30cf228d363e96322.pdf

Li, W. (1991). Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on information theory*, *38*(6), 1842-1845. Disponível em: http://santafe.edu/media/workingpapers/91-03-016.pdf

Lois-González, R. C., & López-González, A. (2013). Macrocephalic growth of capital cities in West Africa's urban system. In N. Kotze, R. Donaldson, & G. Visser (Eds.), *Life in a Changing Urban Landscape. Proceedings of the IGU Urban Geography Commission* (pp. 35-51). Johannesburg: University of Johannesburg.

Lopes, L., Fernandes, P., & Vieira, R. (2016). Estimating term domain relevance through term frequency, disjoint corpora frequency-tf-dcf. *Knowledge-Based Systems*, *97*, 237-249. Disponível em: http://www.sciencedirect.com/science/article/pii/S0950705115004979

Lopez-Pellicer, F. J., Silva, M. J., Chaves, M., Zarazaga-Soria, F. J., & Muro-Medrano, P. R. (2010). Geo linked data. In *International Conference on Database and Expert Systems Applications* (pp. 495-502). Springer Berlin Heidelberg.

Lyons, J. (1995). *Linguistic Semantics*. Cambridge: Cambridge University Press.

MacEachren, A. M. (1995). *How Maps Work*. New York / London: The Guilford Press.

Machado, P. N., & de Lima, V. L. S. (2015). Extração de relações hiponímicas em um corpus de língua portuguesa. *Revista de Estudos da Linguagem*, *23*(3), 599-640. Disponível em: http://www.periodicos.letras.ufmg.br/index.php/relin/article/download/8893/pdf

Mandl, T., Carvalho, P., Di Nunzio, G. M., Gey, F., Larson, R. R., Santos, D., & Womser-Hacker, C. (2009). GeoCLEF 2008: the CLEF 2008 cross-language geographic information retrieval track overview. In *Evaluating Systems for Multilingual and Multimodal Information Access* (pp. 808-821). Springer Berlin Heidelberg.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing* (Vol. 999). Cambridge: MIT press.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *ACL (System Demonstrations)* (pp. 55-60). Disponível em: http://nlp.stanford.edu/pubs/StanfordCoreNlp2014.pdf

Martins, B., & Silva, M. J. (2007). O HAREM e a avaliação de sistemas para o reconhecimento de entidades geográficas em textos em língua portuguesa. In D. Santos & N. Cardoso (Eds.) *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área* (pp. 79-86). Disponível em: http://comum.rcaap.pt/bitstream/10400.26/380/1/Livro-SantosCardoso2007.pdf

Martins, B., Silva, M. J., & Chaves, M. S. (2007). O sistema CaGE no HAREM-reconhecimento de entidades geográficas em textos em língua portuguesa. In D. Santos & N. Cardoso (Eds.) *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área* (pp. 97-112). Disponível em: http://comum.rcaap.pt/bitstream/10400.26/380/1/Livro-SantosCardoso2007.pdf

Martins, M. R. D. (1988). Ouvir Falar. Introdução à Fonética do Português. 4ª ed. Lisboa: Caminho.

Maynard, D., Li, Y., & Peters, W. (2008). NLP techniques for term extraction and ontology population. In *Proceeding of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge* (pp. 107-127). Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.173.6924&rep=rep1&type=pdf#page=123

McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Mintz, M., Bills, S., Snow, R., & Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2* (pp. 1003-1011). Association for Computational Linguistics. Disponível em: http://www.cse.fau.edu/~xqzhu/courses/cap6777/distant.supervision.pdf

Mitchell, T. (1997). *Machine Learning*. Boston: McGraw-Hill.

Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive science*, *34*(8), 1388-1429. Disponível em: http://onlinelibrary.wiley.com/doi/10.1111/j.1551-6709.2010.01106.x/full

Mota, C., & Santos, D. (Eds.). (2008). *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Publicação digital: Linguateca. Disponível em: http://comum.rcaap.pt/bitstream/10400.26/254/1/Livro-MotaSantos2008.pdf

Muller, C. 1970. Elements de statistique linguistique. In A. Zampolli (Ed.), *Linguistica Matematica e Calcolatori. Atti del Convegno e della prima Scuola Internazionale. Pisa, 16/VIII – 6/IX 1970*. Firenze: Leo S. Olschki.

Musen, M.A. (2015). The Protégé project: A look back and a look forward. AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence, *1*(4). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4883684/

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

Nelson, G., Wallis, S., & Aarts, B. (2002). *Exploring natural language: working with the British component of the International Corpus of English* (Vol. 29). John Benjamins Publishing.

Nilsson, N. J. (1991). Logic and artificial intelligence. *Artificial intelligence*, *47*(1-3), 31-56. Disponível em: https://pdfs.semanticscholar.org/7f97/e1fef519b6017b258b785c3c1adcfbef7da4.pdf

Ooi, V. B. Y. (1998). *Computer Corpus Lexicography*. Edinburgh: Edinburgh University.

Pasley, R. C., Clough, P. D., & Sanderson, M. (2007). Geo-tagging for imprecise regions of different sizes. In *Proceedings of the 4th ACM workshop on Geographical information retrieval* (pp. 77-82). ACM. Disponível em: http://eprints.whiterose.ac.uk/78503/7/WRRO_78503.pdf

Peng, R. D. (2015). *Exploratory Data Analysis with R*. E-book: LeanPub. http://leanpub.com/exdata

Peng, R. D., & Matsui E. (2015). *The Art of Data Science*. E-book: LeanPub. http://leanpub.com/artofdatascience

Peres Rodrigues, J. H. (2000). *Introduçom à linguística com corpora*. [Sem indicação de lugar] : Artabria / Laiovento.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, *21*(5), 1112-1130. Disponível em: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4176592/

Poli, R., & Obrst, L. (2010). The interplay between ontology as categorial analysis and ontology as technology. In R. Poli, M. Healy, & A. Kameas (Eds.), *Theory and applications of ontology: Computer applications* (pp. 1-26). Dordrecht: Springer.

Purves, R. S., & Derungs, C. (2015). From Space to Place: Place-Based Explorations of Text. *International Journal of Humanities and Arts Computing*, *9*(1), 74-94. Disponível em: http://www.euppublishing.com/doi/full/10.3366/ijhac.2015.0139

Radford, A. (1997). *Syntax. A minimalist introduction*. Cambridge: Cambridge University Press.

Rapoport, A. (1983). *Mathematical models in the social and behavioral sciences*. New York: John Wiley.

Ratinov, L., & Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning* (pp. 147-155). Association for Computational Linguistics. Disponível em:

http://www.cs.brandeis.edu/~marc/misc/proceedings/naacl-hlt-2009/CoNLL/CoNLL-2009.pdf#page=163

R Core Team (2016). *A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Disponível em: http://www.R-project.Org

Restall, G. (2006). *Logic. An Introduction*. Oxon: Routldge.

Rich, E., Knight, K., Nair, S. B. (2009). *Artificial Intelligence*. (3rd ed.). New Delhi: McGraw Hill.

Riemer, N. (2010). *Introducing semantics*. Cambridge: Cambridge University Press.

Roark, B., & Sproat, R. (2007). *Computational Approaches to Morphology and Syntax*. Oxford: Oxford University.

Robins, H. J. (1969). *A Short History of Linguistics*. (2nd ed.). London: Longman.

Rupp, C. J., Rayson, P., Baron, A., Donaldson, C., Gregory, I., Hardie, A., & Murrieta-Flores, P. (2013). Customising geoparsing and georeferencing for historical texts. In *2013 IEEE International Conference on Big Data* (pp. 59-62). Santa Clara, CA., USA: IEEE. Disponível em: http://www.lancaster.ac.uk/fass/projects/spatialhum.wordpress/wp-content/uploads/2013/12/06691671.pdf

Russel, B. (1905). On denoting. *Mind*, *14*, 479-493.

Russell, S., & Norvig, P. (2010). *Artificial Intelligence. A Modern Approach*. New Jersey: Pearson.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*(5), 513-523.

Sampson, G. (1980). *Schools of Linguistics*. Stanford: Stanford University Press.

Santos, D., & Chaves, M. (2006). The place of place in geographical IR. In *Proceedings of GIR06, the 3rd Workshop on Geographic Information Retrieval, SGIR 2006* (pp 5-8). Seattle, USA. Disponível em: http://www.linguateca.pt/Diana/download/SantosChavesGIR2006.pdf

Santos, D., & Cardoso, N. (Eds.). (2007a). *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/Livro-SantosCardoso2007.pdf

Santos, D., & Cardoso, N. (2007b). Breve introduçao ao HAREM. In D. Santos & N. Cardoso (Eds.), *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área* (pp. 1-16). Disponível em: http://comum.rcaap.pt/bitstream/10400.26/380/1/Livro-SantosCardoso2007.pdf

Santos, D., Cardoso, N., & Cabral, L. M. (2010). How geographic was GikiCLEF?: a GIR-critical review. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*. Zurich, Switzerland: ACM. Disponível em:

http://www.linguateca.pt/Diana/download/SantosCardosoCabralGIR2010.pdf

Santos, D., Cardoso, N., & Seco (2007). Avaliação no HAREM : métodos e medidas. In D. Santos & N. Cardoso (Eds.), *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área* (pp. 245–282). Disponível em: http://comum.rcaap.pt/bitstream/10400.26/380/1/Livro-SantosCardoso2007.pdf

Santos, D., Freitas, C., Gonçalo Oliveira, H., Carvalho, P., & Mota, C. (2008). O Segundo HAREM: Balanço e perspectivas de futuro. In C. Mota & D. Santos (Eds.), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM Linguateca 2008* (pp. 131-146). Disponível em: http://www.linguateca.pt/Diana/download/Santosetal2008SegundoHAREM.pdf

Santos, D., & Guimarães, V. (2015). Boosting Named Entity Recognition with Neural Character Embeddings. In *Proceedings of NEWS 2015 The Fifth Named Entities Workshop* (pp. 25–33). Association for Computational Linguistics. Disponível em: http://www.aclweb.org/anthology/W15-3904

Santos, F., & Gonçalo Oliveira, H. (2015). Descoberta de Synsets Difusos com base na Redundância em vários Dicionários. *Linguamática*, *7*(2), 3-17.

Santos, J., Anastácio, I., & Martins, B. (2015). Using machine learning methods for disambiguating place references in textual documents. *GeoJournal*, *80*(3), 375-392.

Singhal, A., Salton, G., & Buckley, C. (1996). *Length normalization in degraded text collections*. In *Fifth Annual Symposium on Document Analysis and Information Retrieval*, (pp. 149-162). Disponível em: http://www.singhal.info/ocr-norm.pdf

Speriosu, M., & Baldridge, J. (2013). Text-Driven Toponym Resolution using Indirect Supervision. In *ACL (1)* (pp. 1466-1476). Disponível em: http://www.anthology.aclweb.org/P/P13/P13-1144.pdf

Speriosu, M., Brown, T., Moon, T., Baldridge, J., & Erk, K. (2010). Connecting language and geography with region-topic models. In *1st Workshop on Computational Models of Spatial Language Interpretation*. Disponível em: http://ceur-ws.org/Vol-620/cosli-complete.pdf#page=41

Steinberg, S. L., & Steinberg, S. J. (2015). *GIS research methods*. Redlands, California: Esri Press.

Stock, K., Pasley, R. C., Gardner, Z., Brindley, P., Morley, J., & Cialone, C. (2013). Creating a corpus of geospatial natural language. In *International Conference on Spatial Information Theory* (pp. 279-298). Springer International Publishing.

Stokes, N., Li, Y., Moffat, A., & Rong, J. (2008). An empirical study of the effects of NLP components on Geographic IR performance. *International Journal of Geographical Information Science*, *22*(3), 247-264. Disponível em: http://csserver.ucd.ie/~nstokes/publications/nstokes_IJGIS_08.pdf

Suzuki, J., & Isozaki, H. (2008). Semi-Supervised Sequential Labeling and Segmentation Using Giga-Word Scale Unlabeled Data. In *Proceedings of the 46th Annual Meeting of the Association for*

Computational Linguistics: Human Language Technologies (pp. 665-673). Association for Computational Linguistics. Disponível em: http://www.aclweb.org/website/old_anthology/P/P08/P08-1.pdf#page=709

Tjong Kim Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4* (pp. 142-147). Association for Computational Linguistics. Disponível em: http://www.clips.ua.ac.be/conll2003/pdf/14247tjo.pdf

Tobin, R., Grover, C., Byrne, K., Reid, J., & Walsh, J. (2010). Evaluation of georeferencing. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*. Zurich, Switzerland: ACM. Disponível em: http://homepages.inf.ed.ac.uk/grover/papers/a7-tobin.pdf

Tullo, C., & Hurford, J. (2003). Modelling Zipfian distributions in language. In *Proceedings of language evolution and computation workshop/course at ESSLLI* (pp. 62-75). Disponível em: http://www.ling.ed.ac.uk/~jim/zipfjrh.pdf

Valin, R. D. (2001). *An introduction to Syntax*, Cambridge: Cambridge University Press.

Vasardani, M., & Winter, S. (2016). Place Properties. *Advancing Geographic Information Science: The Past and Next Twenty Years*, (pp. 243-254). Needham, MA : GSDI.

Wallis, S. (2007). Annotation, Retrieval and Experimentation. In A. Meurman-Solin, & A. A. Nurmi, (Eds.), *Annotating Variation and Change*. Helsinki: Varieng, UoH. Disponível em: http://www.helsinki.fi/varieng/series/volumes/01/wallis/

Wallis, S. (2014). What might a corpus of parsed spoken data tell us about language? *Proceedings of Olinco 2014*. Palacký University, Olomouc, Czech Republic. Disponível em: http://corplingstats.wordpress.com/2014/06/24/corpus-language/

Wang, M. (2016). Toward the Meaning of Linguistic Signs: A Hierarchical Theory. *Language and Semiotic Studies*, *2*(1). Disponível em: http://www.szlass.com.cn/upload/day_160409/201604091532299606.pdf

Wendt, I. S., Lopes, L., Martins, D., Vieira, R., & de Lima, V. L. S. (2010). Geração automática de glossários de termos específicos de um corpus de Geologia. In *3º ONTOBRAS. Seminário de Pesquisa em Ontologia no Brasil. 30 e 31 de Agosto de 2010. Florianópolis / SC.* Florianópolis: Anais do 3º Seminário de Pesquisa em Ontologia no Brasil. Disponível em: http://www.lbd.dcc.ufmg.br/colecoes/ontobras/2010/0025.pdf

Wilks, Y. & Brewster, C. (2006). *Natural Language Processing as a Foundation of the Semantic Web*. Boston / Delft: Now.

Wittgenstein, L. (1922). *Tractatus Logico-Philosophicus*. Edinburgh: Edinburgh Press.

Yan, X., & Minnhagen, P. (2015). Maximum entropy, word-frequency, Chinese characters, and multiple meanings. *PloS one*, *10*(5). Disponível em: http://journals.plos.org/plosone/article?

id=10.1371/journal.pone.0125592

Zahra, F. M., Malucelli, A., Freddo, A. R., & Tacla, C. A. (2014). Ferramentas para aprendizagem de ontologias a partir de textos. *Perspectivas em Ciência da Informação*, *19*(1), 3-21. Disponível em: http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/viewFile/1150/1242

Zapparoli, Z. M. (2010). Tratamento de corpora informatizados por programas de análise linguística para estudos do português falado de São Paulo. *Boletim da Academia Galega da Língua Portuguesa*, 3, 87-112.

Zapparoli, Z. M., & Camlong, A. (2002). *Do léxico ao discurso pela informática*. São Paulo: Universidade de São Paulo.

Zhang, Q., Jin, P., Lin, S., & Yue, L. (2012). Extracting focused locations for web pages. *Web-Age Information Management*, 76-89. Disponível em: http://nlpr-web.ia.ac.cn/2011papers/kz/gh2.pdf

Zhu, R., Hu, Y., Janowicz, K., & McKenzie, G. (2016). Spatial signatures for geographic feature types: Examining gazetteer ontologies using spatial statistics. *Transactions in GIS, 20*(3), 333–355. Disponível em: http://www.grantmckenzie.com/academics/McKenzie_TGIS2016_1.pdf

Zipf, G. K. (1929). Relative frequency as a Determinant of Phonetic Change. *Harvard Studies in Philology 40*, 1–95.

Fontes e estudo crítico do corpus

Albuquerque, L. (1983). *Ciência e experiência nos Descobrimentos Portugueses*. Lisboa: Instituto de Cultura e Língua Portuguesa

Albuquerque, L. (1987). *As navegações e a sua projecção na ciência e na cultura*. Lisboa: Gradiva

Albuquerque, L. (1989a). *Introdução à História dos Descobrimentos Portugueses* (3ª ed. revista). Sintra: Europa América.

Albuquerque, L. (Dir.). (1989b). *Portugal no Mundo.* 2 vols. [Lisboa]: Alfa.

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. Lisbon: Fundação

Oriente.

Alves, J. S., & Manguin P. (1997). *O Roteiro das Cousas do Achem de D. João Ribeiro Gaio: Um olhar português sobre o Norte de Samatraem finais do século XVI.* Lisboa: Comissão Nacional para as Comemorações dos Descobrimentos Portugueses.

Aubin, J. (1996). *Le Latin et l'Astrolabe. Recherches sur le Portugal de la Renaissance, son expansion en Asie et les Relations Internationales*. Lisbonne / Paris: Centre Culturel Calouste Gulbenkian / Commission Nationale pour les Commémorations des Découvertes Portugaises.

Barreto, L. F. (2000). *Lavrar o Mar. Os Portugueses e a Ásia*. Lisboa: Comissão Nacional para as Comemorações dos Descobrimentos Portugueses.

Biedermann, Z. (Org.). (2010). Indexes. In J. S. Alves (Dir.), *Fernão Mendes Pinto and the Peregrinação* (Vol. 4). Lisbon: 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: Collegio das Artes da Companhia de Jesu. Edição digital fac-similar: Biblioteca Nacional de Portugal. Disponível em http://purl.pt/13969

Boxer, C. R. (1977). *The Portuguese Seaborne Empire 1415- 1825*. London: Hutchinson.

Boxer, C. R. (1989). *O Grande Navio de Amacau* (4ª ed.). Lisboa: Fundação Oriente / Museu e Centro de Estudos Marítimos de Macau.

Carvalho, R. B. (2006). *La présence portugaise a Ayuttahaya (Siam) aux XVIe et XVIIe siècles*. (Tese de Mestrado, Ecole Pratique des Hautes Etudes, Paris).

Chandeigne, M. (Dir). (1996). Goa 1510-1685. *L'Inde portugaise, apostolique et commerciale*. Paris: Autrement.

Cortesão, J. (1981). *Os Descobrimentos Portugueses.* (3ª ed.). 6 vols. Lisboa: Livros Horizonte.

Costa, J. P. A. O. (1995). *A Descoberta da Civilização Japonesa pelos Portugueses*. [sem indicação de lugar]: Instituto Cultural de Macau / Instituto de História de Além Mar.

Cruz, F. G. (1570). *Tractado em que se cõtam muito por estẽso as cousas da China cõ suas particularidades e assi do reyno dormuz*. Euora: Andre de Burgos. Edição digital fac-similar: Biblioteca Nacional de Portugal. Disponível em: http://purl.pt/26733

Ferrand, G. (1922). *L'empire sumatranais de Crivijaya*. Paris: Librairie Orientaliste Paul Geuthner.

Flores, M. C. (1991). *Os Portugueses e o Sião no Século XVI.* [Lisboa]: Comissão Nacional para as Comemorações dos Descobrimentos Portugueses.

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

Gomes, R. V. (1983). Roteiro Geográfico através da Peregrinação. In A. M Flores, R. V. Gomes, & R. H. P. Sousa. *Fernão Mendes Pinto. Subsídios para a sua Bio-Bibliografia* (pp. 81-167). [Almada]: Cámara Municipal da Almada.

Graça, L. (1989). Os Portugueses na Etiópia; as duas primeiras embaixadas e a acção dos jesuítas. In L. Albuquerque (Dir.), *Portugal no Mundo* (Vol. 2, pp. 135-142). [Lisboa]: Publicações Alfa.

John, K. J. (1998). International trade in Cochin in the 16th century. In A. T. Matos & L. F. F. R. Thomaz (Eds.). *A Carreira da Índia e as Rotas dos Estreitos* (pp. 295-308). Angra do Heroismo: Actas do VII Seminário Internacional de História Indo-Portuguesa.

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

Leider, J. P. (2010). Southeast Asian Buddhist Monks in The Peregrinação. Tracing the Rolins of Fernão Mendes Pinto in the Eastern Bay of Bengal. In J. S. Alves (Dir.). *Fernão Mendes Pinto and the Peregrinação* (Vol. 1, pp. 145-162). Lisbon: Fundação Oriente.

Loureiro, R. M. (1996). *Em busca das origens de Macau.* Lisboa: Grupo de Trabalho do Ministério da Educação para as Comemorações dos Descobrimentos Portugueses.

Marques, A. P. (1991). *A Historiografia dos Descobrimentos Portugueses.* Coimbra: Livraria Minerva.

Marques, A. P. (1996). *A Cartografia Portuguesa do Japão. The Portuguese Cartography of Japan*. Lisboa: Imprensa Nacional - Casa da Moeda.

Meyer, M. C., Sherman,W. L., & Deeds, W. L. (2007). *The Course of Mexican History* (8th ed.). New York / Oxford: Oxford University Press.

Moraes, W. (1920). Fernão Mendes Pinto no Japão. In *O Comércio do Porto*, (Sep). Reedição e introdução de Ana Paula Laborinho (2004). Lisboa: Imprensa Nacional - Casa da Moeda.

National Geographic Information Institute (NGII). (2009a). *The National Atlas of Korea.* Gyeonggi-Do, Republic of Korea: National Geographic Information Institute / Ministery of Land, Transport and Maritime Affairs.

National Geographic Information Institute (NGII). (2009b). *The Gazetteer of Korea*. Gyeonggi-Do, Republic of Korea: National Geographic Information Institute / Ministery of Land, Transport and Maritime Affairs.

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

Pinto, C. A. (2007). A Misericórida de Diu: o castelo, a igreja e o hospital. In A. F. Meneses, & J. P. O. Costa (Eds.). *O Reino, as Ilhas e o Mar Oceano. Estudos em homenagem a Artur Teodoro de Matos* (Vol. 2, pp. 771-784). Ponta Delgada, Lisboa: Universidade dos Açores / CHAM.

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

Pinto, F. M. (1653). *The voyages and adventures, of Fernand Mendez Pinto, a Portugal ... done into english by H. C. Gent*. London : printed by J. Macock, for Henry Cripps, and Lodowick Lloyd. Edição digital fac-similar: Biblioteca Nacional de Portugal. Disponível em: http://purl.pt/16425

Pinto, F. M. (1989). *Peregrinação & Cartas*. 2 vols. Lisboa: Fernando Ribeiro de Mello / Edições Afrodite.

Rodrigues, V. L. G. (2007). O Município de Goa, peça fundamental para a afirmação e sobrevivência do 'Estado da Índia. In A. F. Meneses, & J. P. O. Costa (Coords.). *O Reino, as Ilhas e o Mar Oceano. Estudos em homenagem a Artur Teodoro de Matos* (vol. 2, pp. 669-684). Ponta Delgada, Lisboa: Universidade dos Açores / CHAM.

Subrahmanyam, S. (1998). The trading world of the western Indian Ocean, 1546-1565: A political interpretation. In A. T. Matos & L. F. F. R. Thomaz (Eds.), *A Carreira da Índia e as Rotas dos Estreitos* (pp. 207-227). Angra do Heroismo: Actas do VII Seminário Internacional de História Indo-Portuguesa.

Subrahmanyam, S. (1990). Profit at the Apostle's feet: the Portuguese in the 16th century São Tome de Meliapor. In J. Aubin (Ed.), *La decouverte, le Portugal et l'Europe* (pp. 217-233). Paris: Fondation Calouste Gulbenkian / Centre Culturel Portugais.

Tavim, J. A. R. S. (2002). A cidade Portuguesa de Santa Cruz de Cochim ou Cochim de Baixo. Algumas Perspectivas. In L. F. Thomaz (Ed.), *Aquem e Alem da Trapobana. Estudos Luso-Orientais à memória de Jean Aubin e Denys Lombard* (pp. 135-189). Lisboa: CHAM / Universidade de Lisboa.

Thomaz, L. F. F. R. (1998). A questão da pimenta em meados do século XVI. In A. T. Matos & L. F. F. R. Thomaz (Eds.), *A Carreira da Índia e as Rotas dos Estreitos* (pp. 37-206). Angra do Heroismo: Actas do VII Seminário Internacional de História Indo-Portuguesa.

Thomaz, L. F. F. R. (2002). O malogrado estabelecimento oficial dos portugueses em Sunda e a islamização de Java. In L. F. Thomaz (Ed.), *Aquem e Alem da Trapobana. Estudos Luso-Orientais à memória de Jean Aubin e Denys Lombard* (pp. 381-618). Lisboa: CHAM / Universidade de Lisboa.

Thomas, N. (1993-1999) (Series Ed.). *Mercator Media Guide*. 3 vols. Wales: University of Wales Press.

Vittrant, A. (2010). Aire linguistique Asie du Sud-Est continentale: le birman en fait-il partie?. *Moussons. Recherche en sciences humaines sur l'Asie du Sud-Est*, *16*, 7-38. Disponível em: http://moussons.revues.org/94

Wheeler, C. (2010). A Coastal Paronama of Cochinchina (Vietnam) and Champa in the Peregrinação. In J. S. Alves (Dir.), *Fernão Mendes Pinto and the Peregrinação* (Vol. 1, pp. 163-184). Lisbon: Fundação Oriente.

Índice de figuras

Índice de tabelas

RESUMO

Este traballo atende o problema da identificación e referenciación das entidades xeográficas mencionadas e propón un modelo semántico para desenvolver aqueles casos non solucionábeis pola simple recuperación de coordenadas. Establécese unha hipótese central, un mesmo modelo conceptual integra as entidades para as que a xeorrefencia é coñecida con aqueloutras descoñecidas, contribuíndo deste modo para o xeorreferenciamento relativo das segundas. Definimos dúas grandes áreas na resolución do problema: primeiro, a identificación das expresións das entidades no texto. Segundo, unha vez anotada a entidade como xeográfica, o xeorreferenciamento propiamente, que ten como fin localizar un obxecto xeográfico no mundo real.

Situamos o traballo sobre as entidades xeográficas mencionadas dentro dun ámbito interdisciplinar. A hipótese secundaria da tese, metodolóxica, prioriza as técnicas de PLN como proposta para a consecución dos obxectivos prácticos. Encadramos como xeográficos os resultados finais do índice de entidades xeográficas mencionadas (útil de xeografía histórica) e a base de datos SIX das entidades de coñecemento previo. Durante todo o desenvolvemento da tese usáronse ferramentas xeográficas, especialmente para a xeovisualización e, en menor medida, para a análise e representación cartográfica. Como problemas metodolóxicos apareceron aspectos estudados no ámbito NERC e disciplinas mais computacionais, a súa solución foi atendida como procedemento para avanzar nun fin (identificación de entidades para a anotación do corpus). Na aplicación do caso práctico, o estudo das xeorreferencias foi en primeiro lugar abordado manualmente con aparato crítico procedente das disciplinas da xeografía histórica, a historia e a filoloxía.

Elaboramos un corpus padrón dourado para o estudo das entidades xeográficas mencionadas. A partir dunha transcrición dixital, defectuosa, da primeira edición da *Peregrinação*, realizamos sucesivas melloras sobre o texto, comparándoo con edicións impresas e co orixinal en versión facsimilar. Paralelamente, levantamos manualmente un índice de entidades xeográficas mencionadas que nos serviron para anotar de modo semiautomático o corpus. O resultado foi un corpus anotado que fomos mellorando (corrección de erros de anotación, ampliación do marcado) e mesmo expandimos coa creación dun corpus paralelo, en que os capítulos con unidade temática da Tartaria foron aliñados verbo da primeira edición en inglés. O traballo co corpus seguiu unha pauta cíclica, de cuestionamento e procesamento, para responder a novos obxectivos segundo avanzabamos nos labores de xeorreferenciación. A partir das análises exploratorias iniciais para introducir a lei de Zipf e observar como a frecuencia se mostraba relevante na distribución xeográfica dos capítulos, sucesivos *scripts* foron creando subcorpora e seleccionando só aqueles aspectos da anotación requiridos nos testes. Deste modo, na pesquisa sobre traballos NERC, para alén do corpus paralelo, aproveitamos a anotación para distinguir xentílicos de topónimos. Máis

adiante a anotación serviu para crear un subcorpus das oracións con valor xeoespacial como material experimental para os testes de recuperación de termos xeográficos. Na extracción de relacións traballamos co conxunto do corpus e aproveitamos as anotacións como variable de predición. Durante a fase de estudo crítico e xeovisualización, o corpus permitiu recuperar as concordancias das entidades, axilizando a consulta do texto e permitindo pesquisas selectivas para a recuperación das descricións dos referentes.

O corpus foi, deste modo, o principal material de apoio deste traballo, o padrón co que se compararon os resultados obtidos da aplicación das técnicas de PLN na identificación de entidades. Definidos os procesos para a anotación, ensaiamos tres métodos de automatización. En primeiro lugar, mostramos as dificultades xurdidas mesmo no mellor dos escenarios, cando operamos con unha lista *ad hoc* de todas as entidades mencionadas. Posteriormente empregamos unha ferramenta de anotación automática para, configurada coa lista *ad hoc*, compararmos resultados verbo do corpus. Tiñamos un doble obxectivo: primeiro, metodolóxico, presentarmos as métricas usadas para a avaliación de resultados no resto da tese e, segundo, material, mellorarmos o corpus mediante a detección de erros. A avaliación da diverxencia entre os resultados da ferramenta e o corpus permitiu mellorar a anotación, detectando novas expresións de entidades xeográficas mencionadas pola inspección de só 5% das respostas. Finalmente, usamos un corpus paralelo elaborado a partir do aliñamento dos capítulos con unidade temática no espazo xeográfico da Tartaria nas primeiras edicións da *Peregrinação* en portugués e inglés para considerarmos a automatización do proceso completo de anotación. Usamos dous tipos de solucións, un modelo estatístico e outro de regras, en forma de tres ferramentas de libre disposición (unha estatística, dúas de regras), con que avaliamos resultados en función dos obxectivos específicos do noso corpus, diferentes en parte de aqueles para os que foran concebidos os útiles: anotar tanto xentílicos canto topónimos e usar unha variante de lingua distinta do padrón contemporáneo. Aínda con estes condicionantes, os mellores resultados conseguiron superar a barreira de 60% na medida-F, co mellor desempeño próximo a 70%. Como conclusión, presentamos un esquema de procedemento que combina procesos automáticos e revisión manual para reducir o tempo e mellorar a calidade da anotación de textos non normalizados verbo do padrón contemporáneo.

Introducimos un modelo conceptual co que xeoferrenciamos todas as entidades mencionadas no corpus. A partir dunha aproximación semántica referencial, consideramos as ligazóns entre a expresión, o concepto e o referente, para distinguirmos dous tipos de xeorreferenciamento. No primeiro, ostensivo, a entidade xeográfica mencionada é referenciada directamente por medio de unhas coordenadas. No segundo, mais elaborado (intensivo), o referente é denotado através do concepto. Aproveitamos unha noción da semántica cognitiva, na que o concepto se estrutura a partir de regras e atributos, para elaborarmos un esquema simple, con só dous componentes, en que o referente é denotado pola adscrición de un tipo xeográfico e unha relación espacial con outra entidade. Unha particularidade do noso modelo, a diferenza dos cognitivos, é mantermos o referente como ente físico, obxecto xeográfico. O resto da tese desenvolveu estas dúas posibilidades de

xeorreferenciación.

No modo de xeorreferenciamento ostensivo encontramos un paralelismo entre as entidades referenciadas por coordenadas e aqueloutras que deixamos para a denotación por definición. As primeiras son entidades que coñecemos previamente, recoñecíbeis pola aplicación de instrumentos xeográficos (SIX, atlas, glosarios e estudos específicos). As segundas, entidades non referenciadas por coordenadas, son descoñecidas na documentación ou presentan algunha dúbida que nos impide dar a súa localización como segura. Esta distinción ten implicacións no procedemento da xeorreferenciación. As entidades de coñecemento previo son as primeiras en ser xeorreferenciadas. O seu xeorreferenciamento no caso práctico da *Peregrinación* comezou pola elaboración dunha base documental en que recollemos canto traballo encontramos dispoñíbel para contextualizar e ofrecer referentes das entidades mencionadas. Elaboramos un estudo crítico, visualizamos e anotamos os obxectos xeográficos en aplicacións SIX. Adoptamos unha medida conservadora para a avaliación crítica da xeorreferencia. Só aquelas entidades mencionadas para as que non encontramos ningún tipo de contradición teñen a máxima probabilidade, P(xeorreferencia)=1, na atribución do referente e son finalmente incluídas nunha lista que chamamos de coñecemento previo. A comparación da clasificación por tipo de coñecemento (previo ou descrito) cos datos de frecuencia no corpus mostra como as entidades con maior frecuencia son tamén as mais coñecidas, porén, aquelas cuxa xeorreferenciación ten que vir dada pola descrición ocupan unha escala 10^{-1} menor nunha distribución de Zipf. A frecuencia aparece como un elemento determinante para o xeorreferenciamento da entidade.

Na xeorreferenciación por descrición, atendendo ao esquema do concepto proposto, o primeiro elemento a solucionar foi a atribución de un tipo xeográfico e a ordenación dos tipos nunha taxonomía que clasifique as entidades nunha ontoloxía. Consideramos dúas solucións, a aplicación dunha taxonomía previa, como fixemos no caso de referenciación das entidades de coñecemento previo (en que un referente ten un tipo xeográfico asimilábel ao obxecto na actualidade), e a creación dun vocabulario a partir dos termos presentes no corpus (procedemento aplicado na descrición das entidades no estudo crítico). A primeira presenta o problema da adecuación dos termos ao corpus, isto é, a terminoloxía da taxonomía externa pode clasificar correctamente a entidade, mais, se non houber un termo equivalente no corpus, dificultará a recuperación de concordancias e entidades relacionadas. A elaboración dunha taxonomía *ad hoc* ten a vantaxe de describir mais de preto as entidades e ofrecer un vocabulario procedente do propio corpus, porén, deixa baleiros na clasificación cando a entidade non ten o tipo declarado explicitamente, e resulta difícil de organizar nos niveis superiores da xerarquía (os tipos mais abstractos). A solución adoptada foi unha taxonomía híbrida, en que os tipos son extraídos do corpus e integrados no esquema clasificatorio xa usado para as entidades de coñecemento previo. Outro problema, o da densidade nas clases, foi resolvido co cálculo da entropía e o agrupamento dos tipos próximos. Como resultado final obtivemos unha taxonomía coa que clasificamos as entidades mencionadas do corpus (procedemento manual). Elaboramos deste modo unha lista de entidades e tipos xeográficos

coa que testarmos procedementos de extracción de terminoloxía de modo automático. O primeiro teste aplicado tivo unha base mais teórica e mostrou como o efecto da frecuencia na recuperación de candidatos (a seren incluídos como termos da taxonomía) obriga a recuperar máis termos dos necesarios polo carácter exponencial da distribución do vocabulario (primeira lei de Zipf). Considerada esta limitación, procuramos métricas e métodos de filtrado para limitar o número de candidatos. Con carácter previo, preparamos un subcorpus xeoespacial, formado por oracións expresión de proposicións con valor xeográfico. Sobre este subcorpus realizamos traballos de procesamento de linguaxe natural: substitución da entidade mencionada por unha expresión xenérica, anotación da categoria gramatical dos tokens examinados, subsegmentación en cláusulas e frases e limitación do número de tokens a procesar (xanelas coa entidade xeográfica mencionada como centro). Os distintos resultados foron clasificados como modos do corpus. Sobre cada modo aplicamos variantes de filtros e métricas para a recuperación dos termos do dominio. Conseguimos os mellores resultados co maior nivel de PLN. Tentamos aínda melloralos aplicando un útil máis elaborado, unha base de coñecemento lexical difusa, CLIP2.1, de recente elaboración e sen aplicacións similares por nós coñecidas, sobre a cal realizamos unha batería de testes até conseguirmos unha configuración con que atinximos unha medida-F por riba de 80% e resultados de precisión de 100% sen diminuírmos considerabelmente a abranxencia verbo dos métodos de filtrado e métricas dos testes anteriores. Como conclusión, consideramos a aplicación deste tipo de base lexicais un recurso que incrementa notabelmente o desempeño nos traballos de extracción de termos e situamos aquí un dos principais contributos metodolóxicos desta tese.

Unha vez conseguida unha taxonomía para o caso da *Peregrinação* e clasificadas as entidades xeográficas mencionadas no seu tipo, ficou por desenvolver a relación espacial para obter a definición da entidade que, no caso das entidades non coñecidas previamente, representa o seu xeorreferenciamento relativo. Como parte do estudo crítico, as entidades foron anotadas para varias relacións, as mais importantes: a distancia (en medidas de lonxitude e tempo) e proximidade a outra entidade. No modelo conceptual aplicado nesta tese só usamos a relación *é_Parte_de*, que se corresponde coa meronimia en termos semánticos. Por outro lado, a taxonomía dos tipos xeográficos ten o seu paralelo na hiponimia (membro da clase) e hiperonimia (a clase). Estas relacións semánticas organizan as entidades e os tipos nunha ontoloxía, un modo de ordenar todos os elementos do modelo (entidades, tipos e relacións). Para avaliarmos as posibilidades da captura automática no corpus, consideramos a hipótese do modelo distribucional en que o significado dun termo vén dado polo seu contexto. Os estudos revistos para a contextualización do modelo vectorial, aplicado neste tipo de problemas, usan preferentemente grandes volumes de datos con frecuencias altas para os termos obxecto de análise. Considerada esta limitación do noso corpus, comezamos por ilustrar o método con un teste de tipo binario: seleccionadas as entidades mencionadas con maior frecuencia para os tipos *illa* e *cidade*, formamos para cada unha o vector das súas co-ocorrencias e observamos a súa distribución nun diagrama cartesiano. Posteriormente usamos esta representación para demostrar o funcionamento da medida de proximidade do coseno na clasificación das entidades dentro de un dos dous tipos. Exemplificado o método, procedemos a

clasificar o conxunto das entidades en función de todos os tipos no cumio da taxonomía. Por medio de aprendizado de máquina, treinamos un modelo baseado nunha matriz composta polos vectores de co-ocorrencia de entidades e os seus tipos xeográficos para avaliar os resultados ao clasificar 20% da lista da taxonomía que deixamos fóra do treino. Para aproveitarmos mellor o corpus, realizamos o mesmo teste alterando de modo aleatorio a segmentación da lista en treino e avaliación. Os resultados foron significativos para as clases no cumio dos tipos administrativos, terrestres e hidrolóxicos, as mais relevantes en termos de frecuencia no corpus. Atendendo aos resultados, realizamos outro teste exploratorio para avaliar a relación de hiponimia nos tipos *illa* e *cidade* (os dous con frecuencia alta), mais desta volta considerando o conxunto de entidades do corpus. Co mesmo criterio e obxectivos, consideramos tamén a relación *é_Parte_de*. En ambos os dous casos, os resultados obtidos foron altamente significativos, con niveis de precisión media para os tipos e entidades mais comúns por volta de 90%. Aínda sendo testes exploratorios, ilustrativos de un método, aplicados só a tipos seleccionados por teren as frecuencias mais altas, contribúen para confirmar a aproximación metodolóxica da tese, baseada na consideración de variables cuantitativas para o modelado do corpus e, en última instancia, das xeorreferencias.

Para o obxectivo deste traballo, a ontoloxía permitiunos revisar as entidades xeográficas mencionadas e as súas relacións e deste modo crearmos un índice das entidades mencionadas no corpus. A elaboración do índice requiriu a escolla da expresión que representase todas as variantes asociadas, tal e como facemos nun dicionario ou índice xeográfico nun atlas. Aos datos do modelo conceptual engadímoslle outros de frecuencias e do estudo crítico da base de datos relacional. Obtivemos deste modo un índice esquemático das entidades xeográficas mencionadas cos valores de ocorrencia (por capítulos para permitir a súa recuperación en calquera edición) e da xeorreferencia, así como outras relacións espaciais que denotan o obxecto xeográfico e non foron consideradas no esquema conceptual proposto neste traballo.

Na consideración das entidades de coñecemento previo deixamos fóra aquelas entidades que presentaron discrepancias na base documental. Porén, a combinatoria da análise crítica coa descrición do corpus permite resolver, en termos de coordenadas e con un alto nivel de probabilidade, un bo número de entidades descritas como relativas. Por outro lado, no estudo crítico avaliamos as entidades segundo unha escala de probabilidade de posíbel a moito probábel, criterio non aplicado neste traballo e que pode ser desenvolvido para a ampliación da base de datos do SIX, usando a escala como indicativo da certeza na atribución das coordenadas.

Dentro da liña de automatización do proceso de xeorreferenciación, a captura de relacións é un aspecto que ficou ilustrado, mais menos desenvolvido. Traballo futuro consiste na mellora da automatización da captura de relacións para o conxunto das entidades, procurando solucións para os casos de frecuencias baixas, avaliando as posibilidades de aumento do rendemento polo incremento de PLN (maior relevancia dos preditores) e técnicas de corpus para o aumento da relevancia estatística das ocorrencias.

O modelo conceptual integra relacións e entidades de maneira que facilita a súa integración nunha ontoloxía. Aínda que foi apuntado, e resulta factíbel coa ontoloxía elaborada, ficou por desenvolver a inferencia automática da xeorreferencia a partir da relación de meronimia. Deste modo desenvolvemos a definición de modo automático, substituíndo a entidade máis próxima na relación por aquela, primeira das imediatamente superiores, cuxa xeorreferencia sexa coñecida por coñecemento previo. Aseguramos deste modo que a definición veña sempre dada verbo dunhas coordenadas xeográficas.

Conclusións

Consideramos un modelo de xeorreferenciación que opera quer con entidades xeográficas coñecidas, quer con aquelas que, nos métodos máis convencionais, serían deixadas como non xeorreferenciábeis. A partir da definición intensiva da entidade (fronte à definición ostensiva por coordenadas, aproveitadas para serviren de punto de referencia) xeorreferenciamos de modo relativo 100% das entidades extraídas no índice.

Exploramos tamén a operatividade do modelo conceptual para traballar con métodos de mineración aplicábeis a grandes volumes de dados. Aínda que a proposta avaliada neste traballo contempla só dúas relacións, supón un primeiro paso para a resolución do referente por unha vía alternativa á ostensión, baseada na inferencia a partir de procesos dedutivos apriorísticos combinados con datos inducidos pola análise de corpus, un avance na dirección da Intelixencia Artificial.

A caracterización epistemolóxica das entidades xeográficas mencionadas contribúe tamén a organizar e visualizar os resultados. Cando as entidades cuxas coordenadas coñecemos previamente foron procesadas nun SIX, da representación cartográfica emerxe unha distribución espacial característica que debuxa os perfís costeiros da Asia e parte da Insulindia. A ordenación das entidades nunha ontoloxía produce un índice ordenado, no que as entidades se relacionan coherentemente segundo as relacións de hiponimia e meronimia propostas.

Avaliamos métodos e sistemas de xeorreferenciación para un caso pouco estudado, o de un texto histórico, escrito nunha variedade distinta ao padrón contemporáneo. Conseguimos resultados que, como mínimo, se aproximan e en ocasións mesmo superan aqueles obtidos para textos actuais. Nos traballos de identificación e xeorreferenciación, o auxilio do PLN para operar co corpus facilitou a análise e extracción de datos que produciran como resultado un índice de entidades.

Demostramos que a aplicación de variábeis cuantitativas produce resultados significativos, porén, o criterio de selección das variábeis aparece como relevante e a elaboración do corpus por métodos de PLN (en que a variable principal é mais comunmente cualitativa) proporcionou os maiores incrementos nos niveis de desempeño. Unha contribución secundaria desta tese é mellorar na comprensión dos métodos cuantitativos aplicados en traballos de análise de corpus e PLN.

A elaboración do corpus e material de traballo base desta tese requiriu un intenso estudo e anotación manual que decorreu durante mais de media década. Os produtos finais poden ser usados como

padróns para o teste de métodos que simplifiquen estes traballos e avanzar deste modo na súa automatización.

Usamos como caso práctico un texto de grande relevo histórico e xeográfico. Mediante a preparación do corpus, bases de dados e índice, pensamos ter contribuído para a súa mellor comprensión. Nese sentido, os materiais elaborados poden ser usados en novos traballos de investigación. Publicamos artigos revistos por pares sobre a contextualización para o estudo do corpus e resultados. Facilitamos o acceso a datos xeográficos en plataformas abertas, promovendo o seu uso como materiais de apoio para futuras investigacións nun ámbito global.