RECONHECIMENTO E CLASSIFICAÇÃO DE ENTIDADES NOMEADAS PARA O DESENVOLVIMENTO DE UM DICIONÁRIO ELETRÔNICO DE ABREVIATURAS DO PORTUGUÊS HISTÓRICO DO BRASIL

RECOGNITION AND CLASSIFICATION OF NAMED ENTITIES FOR THE DEVELOPMENT OF AN ABBREVIATIONS DICTIONARY FROM BRAZILIAN HISTORICAL PORTUGUESE

Abner Maicon Fortunato Batista, Maria Cristina Parreira da Silva, Sandra Maria Aluísio, Oto Araújo Vale

Campus de São José do Rio Preto – Instituto de Biociências Letras e Ciências Exatas – Licenciatura em Letras – abnerfortunato@gmail.com – FAPESP.

Resumo: Uma das grandes dificuldades dos textos antigos é interpretar o significado das abreviaturas. Flexor (1991) contribuiu com a publicação de um dicionário que apresenta o significado de abreviaturas de textos manuscritos dos séculos XVI a XIX. A partir da digitalização desse dicionário, criou-se a possibilidade de reconhecer as entidades nomeadas introduzidas por certos marcadores abreviados, geralmente com inúmeras formas, como por exemplo: capitão (capp\.; capp\.^am), padre (p.^e; p\.^de), governador (gov\.) etc. Esta pesquisa tem como finalidade o reconhecimento e a classificação dessas entidades para a criação de um dicionário de abreviações eletrônico, em colaboração com os grupos de pesquisa do DL-UFSCar, do NILC-ICMC-USP (São Carlos-SP) e da UNESP-FCLAr. A fonte (dicionário de Flexor) será comparada e acrescida com o *corpus* do *Dicionário histórico do Português do Brasil* (DHPB), um projeto em andamento do Instituto do Milênio do CNPq, coordenado pela Profa. Dra. Maria Tereza Camargo Biderman. Desse modo, o resultado será útil para contribuir para a decodificação das abreviaturas desse período e para explanação das entidades nomeadas.

Palavras chave: Dicionário de abreviaturas, Entidades Nomeadas, Português histórico do Brasil.

Abstract**:** *One of the biggest difficulties of ancient texts is how to interpret the meaning of abbreviations. Flexor (1991) has contributed with a publication of a dictionary which presents the meaning of abbreviations of manuscript texts from the XVI through the XIX centuries. Starting from the digitalization of this dictionary, it has created the possibility of recognizing the named entities introduced by some abbreviated markers, usually with uncounted ways, for instance: 'capitão' (capp\.; capp\.^am), 'padre' (p.^e; p\.^de), 'governador' (gov\.) etc. The aim of this research is the recognition and the classification of these entities for the creation of an electronic abbreviated dictionary, co-operating with the research groups of the DL-UFSCar, the NILC-ICMC-USP (São Carlos-SP) and the UNESP-FCLAr. The source (Flexor dictionary) will be compared and added to the Dicionário Histórico do Português do Brasil (DHPB) corpus, one project which is going from the Instituto do Milênio of CNPq, coordinated by the teacher Drª. Maria Tereza Camargo Biderman. On this way, the result will be useful to contribute for the decoding of the abbreviations from this period and for the explanation of named entities.*

Keywords: Abbreviations Dictionary, Named Entities, Brazilian Historical Portuguese

1. INTRODUÇÃO

Durante o período colonial brasileiro, o português utilizado no Brasil carecia de um sistema ortográfico oficial com que fosse possível escrever, de modo padronizado, os textos que aqui eram produzidos; nesse contexto criou-se o costume entre os escreventes, de abreviar palavras para facilitar a escrita à mão, o que gerou milhares de abreviaturas nos documentos escritos nessa época.

Convém destacar que, posteriormente, o uso de abreviaturas se estendeu também para o texto impresso, nesse caso, por motivo de economia, já que o material utilizado era muito dispendioso no início.

Atualmente esse uso não é diferente por conta da necessidade de transmitir mensagens sempre urgentes e de modo resumido, representado pelo uso cada vez mais freqüente de mensagens via telefone celular, pelo sistema de SMS e por páginas de relacionamento na Internet. É evidente que os usuários precisam compartilhar do mesmo código para se compreenderem.

Por conseguinte, uma das grandes dificuldades que encontramos na leitura dos textos antigos atualmente é a interpretação do significado das abreviaturas utilizadas nesses textos. Flexor (1991) contribuiu com a publicação de um dicionário que apresenta o significado de abreviaturas de textos manuscritos dos séculos XVI a XIX. A partir da digitalização desse dicionário, coordenada pelos colaboradores deste trabalho, criou-se a possibilidade de reconhecer unidades introduzidas por certos marcadores abreviados, geralmente com diversas formas, como por exemplo: capitão (capp\.; capp\.^am), padre (p.^e; p\.^de), governador (gov\.) etc.

A publicação de dicionários tradicionais auxilia na leitura de livros da época colonial, entretanto, a quantidade extensa de textos disponíveis nos *corpora* (reunião de conjuntos de documentos, dados e informações) dificulta a extração automática de conhecimento desses textos, pois, a existência de um grande contingente de documentos à disposição, sem que tenhamos uma forma eficiente de tratá-los, pode gerar uma sobrecarga de informações.

Para solucionar essa dificuldade, faz-se necessária a criação de ferramentas computacionais que tornem possível o tratamento automático dos corpora, promovendo a interface das informações desses *corpora* com a informação de dicionários e glossários existentes. É importante ressaltar que a implementação de programas e recursos que facilitem a leitura de *corpora* históricos é necessária não só para extrair informações de modo mais ágil e eficiente, mas também para possibilitar a compreensão de muitas informações que poderiam se perder.

Considerando-se a importância que os nomes próprios têm para a estruturação e para o tratamento automático de informações em textos, um dicionário de abreviaturas eletrônico, que abranja informações sobre Entidades Nomeadas (ENs), ou seja, entidades concretas ou abstratas que possuem um nome próprio, constitui um recurso de grande utilidade para o processamento computacional de *corpora* do português histórico do Brasil (PHB), pois, como salienta Baptista *et al.* (2006), o reconhecimento de nomes próprios em textos é um problema recorrente em diferentes domínios do Processamento de Linguagem Natural (PLN), tais como a recuperação e a extração de informações em grandes bases textuais. A Profa. Dra. Sandra Aluísio (ICMC-USP) coordena a parte computacional da construção desse dicionário de abreviações e co-orienta este projeto de pesquisa.

2. OBJETIVOS

A partir do trabalho realizado por Flexor (1991), a coleta de abreviaturas dos manuscritos em português dos séculos XVI a XIX e com o início do projeto "Dicionário Histórico do Português do Brasil, Séculos XVI, XVII, XVIII", em andamento, de autoria da Profa. Dra. Maria Tereza Camargo Biderman, que visa recobrir uma lacuna importante na produção lexicográfica brasileira – a falta de um dicionário histórico do português do Brasil, baseado em documentação do Brasil Colônia – surgiram outras necessidades que os recursos de informática podem fornecer.

Este trabalho tem como meta principal o reconhecimento de Entidades Nomeadas para a criação de um dicionário de abreviaturas eletrônico apto a ser implementado em *corpora* do português histórico do Brasil. Cabe ressaltar que há uma porcentagem de abreviaturas comum entre os dois *corpora,* o de Flexor e o do DHPB.

Deve-se atentar para o fato de que "o reconhecimento de entidades nomeadas é útil para uma grande variedade de tarefas e de sistemas importantes no contexto da sociedade de informação. Os sistemas de reconhecimento de entidades nomeadas permitem melhorar muito o desempenho de motores de pesquisa, sistemas de pergunta e resposta, sistemas de indexação de informação, etc." Linguateca (2005). Por conseguinte, esta pesquisa é de suma importância para o processamento automático do português histórico do Brasil, pois pode facilitar enormemente o trabalho com informações do passado histórico brasileiro.

Além do reconhecimento das ENs, nesta pesquisa procede-se a sua classificação, obedecendo alguns critérios pré-estabelecidos no *HAREM***,** que, por sua vez, configura-se como uma avaliação conjunta na área de reconhecimento de entidades mencionadas em português. Muito simplificadamente trata-se de uma iniciativa que pretende avaliar o sucesso na identificação e conseqüente classificação automática dos nomes próprios na língua portuguesa. (http://www.linguateca.pt/LivroHAREM/).

Esta pesquisa também se propõe a identificar e classificar INITs, isto é, *tags* cuja função é introduzir uma EN, ou seja, são conteúdos que precedem uma entidade e com isso indicam o seu contexto de ocorrência. Tal procedimento deverá ser útil futuramente para a construção de um sistema de reconhecimento automático de entidades nomeadas para o PHB a exemplo do SIEMES. (Sistema de Identificação de Entidades Mencionadas com Estratégia Siamesa).

3. FUNDAMENTAÇÃO TEÓRICA

Nas últimas décadas, os avanços na Lexicografia, área do conhecimento que se dedica à elaboração de dicionários, têm proporcionado a confecção de obras mais adequadas aos consulentes e publicadas em suportes novos, como obras eletrônicas e *on-line*. A Lexicografia baseia-se nos resultados dos estudos em Lexicologia, parte da Lingüística que se preocupa com o estudo do léxico de uma língua e cuja teoria suporta os procedimentos deste trabalho.

É importante definir o conceito de Lingüística de Corpus, uma vez que a pesquisa em andamento trata um *corpus* digitalizado das abreviaturas de textos históricos do Brasil. De acordo com Berber Sardinha (2004, p.3), "A Lingüística de Corpus ocupa-se da coleta e da exploração de corpora, ou conjuntos de dados lingüísticos textuais coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade lingüística.", nesse sentido, esse domínio do conhecimento é essencial para a realização de análises lexicológicas de uma grande porção de unidades da língua.

Com o advento da informática, tornou-se possível o desenvolvimento de recursos úteis no que diz respeito às dificuldades de análise manual de documentos e de localização e acesso a grandes quantidades de informação presente em um *corpus*. Por meio do que se denomina *Text Mining* ou Mineração de Textos é possível extrair informações de interesse e descobrir conhecimentos em bases de dados textuais. Segundo Aranha & Passos (2006),

> "A tecnologia de mineração de textos vem das técnicas de recuperação de informações, *machine learning* (que é um ramo do estudo de sistemas de Informação inteligentes que por sua vez é uma das aplicações notáveis da Inteligência Artificial), e da descoberta tradicional de informações estruturadas, através do uso de bancos de dados e de procedimentos estatísticos. Mineração de textos é um conjunto de métodos usados para navegar, organizar, achar e descobrir informação em bases textuais. Pode ser vista como uma extensão da área de *Data Mining*, focada na análise de textos."

A relevância do *Text Mining* para esta pesquisa, deve-se ao fato de que a partir de um dicionário de abreviaturas contendo informações sobre ENs poderá ser possível, no futuro, o desenvolvimento de aplicações em mineração de textos para o português histórico do Brasil.

Uma etapa importante não só no desenvolvimento de aplicações em Mineração de textos, como também e em outras aplicações computacionais é o Processamento de Linguagem Natural (PLN), que segundo Menuzzi & Othero (2005), se preocupa com o estudo da linguagem voltado ao desenvolvimento de aplicações computacionais. O PLN é considerado como uma subárea da lingüística computacional, que de acordo com Vieira & Lima (*apud* OTHERO, 2006) é definida como "a área de conhecimento que explora as relações entre lingüística e informática, tornando possível a construção de sistemas com capacidade de reconhecer e produzir informação apresentada em linguagem natural".

Com o desenvolvimento da Lingüística Computacional e do Processamento de Linguagem Natural é possível a elaboração de grandes projetos de produção de dicionários eletrônicos que possam abarcar a grandiosidade do léxico de uma língua e realizar tarefas complexas, como o tratamento sincrônico ou diacrônico de unidades lexicais de uma ou mais línguas.

4. METODOLOGIA

A análise do conjunto das ENs ocorre a partir de uma lista de abreviaturas digitalizadas em formato XML do dicionário de Flexor (1991 – Figura 1) cuja ocorrência é examinada por meio de uma comparação com as abreviaturas encontradas no *corpus* do Português Histórico do Brasil do projeto DHPB, acessado pela ferramenta *Web* denominada *Philologic* (Figura 2).

Entidad	Abreviat	Expansão	Forma Canônica	Flexão	Categoria Gramatical	Atributo	Tipo de Entidade			Século	Século	Dúvidas
	I	ilha	ilha	fs	N	INIT	LOCAL			17	19	
	I	ilustríssimo	ilustre	Sms	A					19		
	I	ilustríssimos	ilustre	Smp	A					19		
	I	imperial	imperial	ms:fs	A					19		
	I	imprensa	imprensa	fs	N	INIT	ORGANIZAÇÃO			19		
	I	interino	interino	ms	A	INIT	TITULO+LOCAL			18		
	I	irmandade	irmandade	fs	N	INIT	ORGANIZAÇÃO			19		
	I	irmão	irmão	ms	N	INIT	PARENTE+PESSOA			17	19	
	I	justiça	justiça	fs	N					17		
	Ĩ	mil	mil	0	NUM	ENT	VALOR			16		
	I^al	imperial	imperial	ms:fs	A					19		
	I^cio	Inácio	Inácio	ms	N	ENT	PESSOA			17		
	I^m	irmão	irmão	ms	N	INIT	PARENTE+PESSOA			19		
	I^mo	ilustríssimo	ilustre	Sms	A					19		
	I^o	Inácio	Inácio	ms	N	ENT	PESSOA			19		
	I^o	irmão	irmão	ms	N	INIT	PARENTE+PESSOA			19		
	I	João	João	ms	N	ENT	PESSOA			16		(tem um
	I^o	João	João	ms	N	ENT	PESSOA			16	17	
	I^o	primeiro	primeiro	ms	Num	ENT?	TEMPO?			16	17	
	IP^h	Joseph	Joseph	ms	N	ENT	PESSOA			18		
	I^r	irmã	irmão	fs	N	INIT	PARENTE			18	19	
	I^r	irmão	irmão	ms	N	INIT	PARENTE+PESSOA			18	19	
	I^ra	irmã	irmão	fs	N	INIT	PARENTE			18		
	I^ros	irmãos	irmão	mp	N	INIT	PARENTE			17	18	
	I^rs	irmãos	irmão	mp	N	INIT	PARENTE			18	19	
	I^s	irmãos	irmão	mp	N	INIT	PARENTE			18	19	
	I^s	Jesus	Jesus	ms	N	ENT	PESSOA+ORGANIZAÇÃO			18		
	Iã	João	João	ms	N	ENT	PESSOA			16		
	Ian^ro	janeiro	janeiro	ms	N	ENT	TEMPO			16	17	
	Ianr^o	janeiro	janeiro	ms	N	ENT	TEMPO			16	18	
	Iaz^o	jazigo	jazigo	ms	N					17		
	Ib.	ibidem	ibidem	0	ADV					18	19	

Figura 1: Anotação de ENs em abreviaturas digitalizadas no formato XML

Figura 2: Pesquisa para a entrada “irmão” utilizando a ferramenta Philologic

As abreviaturas foram inicialmente divididas em lexias simples e complexas, apesar dessa divisão não estar totalmente resolvida, tomou-se como critério até o momento a consideração de que as lexias simples são constituídas de apenas um elemento, dessa forma, a expansão ‘capitão’ pertenceria ao grupo das lexias simples (Figura 3), enquanto que ‘capitão agregado’, por ser constituído de duas unidades, pertenceria ao grupo das lexias complexas (Figura 4).

1000	+	C^a	Cunha	Cunha	ms	N	ENT	PESSOA
1001	+	Cun^a	Cunha	Cunha	ms	N	ENT	PESSOA
1002	+	C^a	Curitiba	Curitiba	fs	N	ENT	LOCAL
1003	+	Cor^a	Curitiba	Curitiba	fs	N	ENT	LOCAL
1004	+	Cur^a	Curitiba	Curitiba	fs	N	ENT	LOCAL
1005	+	Cur^ba	Curitiba	Curitiba	fs	N	ENT	LOCAL
1006	+	Curit^a	Curitiba	Curitiba	fs	N	ENT	LOCAL
1007	+	Cost^a	Custódia	Custódia	fs	N	ENT	PESSOA
1008	+	Cost^o	Custódio	Custódio	ms	N	ENT	PESSOA

Figura 3: Exemplo de anotação de ENs para as lexias simples

296	+	Cam^o deS^tos	Caminho de Santos		ENT	LOCAL
297	+	Campogr^de	campo grande		ENT	LOCAL
298	+	C^a da St^a Miser^a	Casa da Santa Misericórdia		ENT	LOCAL
299	+	C^a da St^a Misr^a	Casa da Santa Misericórdia		ENT	LOCAL
300	+	Cid^e da B^a	cidade da Bahia		ENT	LOCAL
301	+	Cid^e de Victr^a	cidade de Vitória		ENT	LOCAL
302	+	Col^a do Sacram^to	Colônia do Sacramento		ENT	LOCAL
303		Contin^te do R^o gr^de do Sul	continente do Rio Grande do Sul			
304		Convt^o de N. P^e S. Fran^co	Convento de Nosso Padre São Francisco			
305	+	Conv^to de S^to An^to	Convento de Santo Antônio		ENT	LOCAL
306	+	Conv^to de S^to Ant^to	Convento de Santo Antônio		ENT	LOCAL
307	+	Conv^to de S. Fr^co	Convento de São Francisco		ENT	LOCAL

Figura 4: Exemplo de anotação de ENs para as lexias complexas

Em seguida, procede-se à classificação morfológica de todas as abreviaturas, anotando a classe gramatical da qual a abreviatura pertence, bem como informações adicionais tais como flexões de gênero e número para os substantivos e adjetivos e de tempo e pessoa para os verbos.

Outra informação importante que tem sido anotada é a forma canônica de cada abreviatura, pois a partir dela, pode-se resgatar todas as formas correlatas. Por exemplo, realizando uma pesquisa a partir da forma canônica “irmão”, obteremos as seguintes formas:

irmão	irmã	irmãos	irmãs

Tabela 1: Resultados para a forma canônica “irmão”

E as seguintes abreviaturas:

I	I^s	Ir^am	Irm^m
I^m	Iĩ	Ir^m	Irm^o
I^o	Iĩ	Ir^mos	Irm^os
I^r	Ir.	Ir^o	Irm^s
I^r	Ir.	Ir^oz	Irr^os
I^ra	Ir.	Ir^s	Irs
I^ros	Ir. Ir.	Ir^z	Irs
I^rs	Ir^a	Irm.	

Tabela 2: Abreviaturas para a forma canônica “irmão”

A classificação morfológica das abreviaturas é feita segundo os critérios estabelecidos pelas diretivas de classificação morfológica do Unitex-PB, que segundo Pardo *et. al.* (2006), trata-se de um ambiente computacional que contém um conjunto de recursos e ferramentas lingüístico computacionais usando um formalismo de codificação de conhecimento chamado DELA (*Dictionnarie Electronique du LADL*).

Posteriormente à classificação morfológica, as abreviaturas são submetidas à classificação semântica, que consiste em identificar ENs e INITs a partir dos contextos de ocorrência de cada uma das abreviaturas, verificados a partir da pesquisa no corpus. Em seguida, ENs e INITs são classificados segundo um conjunto de categorias a depender de seu valor semântico. Trata-se de um trabalho minucioso, seguido de uma análise dos problemas identificados.

Para a classificação semântica das ENs, utiliza-se as diretivas do *HAREM*, que possibilitam a classificação das ENs em dez classes distintas: obra, acontecimento, organização, variado, pessoa, abstração, tempo, valor, local e coisa. A categorização dos INITs é aberta, podendo, assim, serem classificados de acordo com as diretivas do *HAREM* ou serem incluídos novos títulos de classificação conforme se fizer necessário. Até o momento, foram utilizadas para a classificação dos INITs somente as categorias ‘título’ (capitão), ‘tratamento’ (Dom) e ‘parente’ (irmão), que diferem das indicadas nas diretivas do *HAREM*.

Para fins de comparação, pode-se lançar mão do *REPENTINO* (Repositório para reconhecimento de Entidades Nomeadas para o português), que é de domínio público e que conta com milhares de exemplos de ENs para o português. As ENs armazenadas no REPENTINO também são organizadas segundo um conjunto de categorias, o que possibilita a verificação e comparação de suas ENs em relação àquelas encontradas durante a realização deste trabalho.

5. RESULTADOS E DISCUSSÕES

Até o momento somente foram aplicados os critérios de classificação morfológica para as lexias simples, mas com relação às lexias complexas, falta estabelecer critérios que atendam ao caráter formal que elas possam apresentar, quase sempre contendo mais de duas unidades. A importância da classificação morfológica configura-se não só enquanto informação gramatical, mas influencia fortemente a decisão da classificação semântica.

Com relação às ENs, grande parte das entidades encontradas, até o momento, examinadas no *corpus* do PHB, constitui-se das categorias *Pessoa* e *Local*, de modo que certas categorias como

Obra e *Coisa* tiveram ocorrência pouco significativa, entretanto, deve-se ressaltar que essa preponderância estatística de algumas categorias poderá ser alterada, tendo em vista que novas abreviaturas serão analisadas e mesmo as que já foram classificadas deverão passar por uma revisão.

Apesar da baixa ocorrência da categoria OBRA (segundo o *HAREM*, refere-se a qualquer coisa feita pelo homem que tenha um nome próprio) uma dificuldade encontrada foi de reconhecer as ENs que pertencem a esta categoria. Um exemplo deste caso ocorreu com a locução *"as falsidades de Charlevoix",* em que apesar de ocorrer no corpus em contextos que parecem indicar um nome próprio referente uma obra, a baixa ocorrência e a não existência de informações atuais que poderiam esclarecer o sentido da locução dificultam sua categorização.

Ainda que o reconhecimento da classe à qual uma EN pertence nem sempre seja imediato e que muitas abreviaturas apresentem mais de uma possível classificação dentre as categorias, a categoria V*ariado* (que recobre elementos reconhecidos como ENs, mas que não se enquadram nas demais categorias) foi raramente anotada, o que demonstra que as diretivas adotadas recobrem as ENs já identificadas.

Outro grande entrave à classificação semântica foi lidar com a polissemia e a ambigüidade que dela decorre. Como exemplo, utilizamos como amostra o caso da EN *Conceição,* inicialmente classificada como *pessoa* ("Maria da Conceição"), mas que também pode remeter a uma data específica, que neste caso, se enquadraria dentro da categoria *tempo* ("dia da Conceição"), *local* ("*Villa da Conceição*") ou também na categoria *acontecimento,* se a entendermos como um evento único e não repetível ("Conceição de Maria").

A distinção entre a classificação de ENs e INITs também tem causado problema, pois, muitas vezes, uma mesma abreviatura pode enquadrar-se, concomitantemente, como EN e INIT. Em alguns casos, como em *Depoimento*, que frequentemente antecede um nome próprio de pessoa, pode ser classificado como INIT, ou seja, vai indicar a possibilidade de ocorrência de várias ENs posteriores; por outro lado, em "depoimento de Diogo Moniz Barreto" entende-se a seqüência como um todo, neste caso, será classificada como uma EN de *acontecimento*.

Outro caso de ambigüidade com possível categorização simultaneamente em EN e INIT ocorreu com a entrada *Governador,* haja vista que em seu sentido literal funciona freqüentemente como um INIT da categoria TÍTULO, mas que também pode funcionar como nome próprio de local, conforme atesta o exemplo "*ilha do Governador*". Coincidentemente, ocorre o mesmo para a forma no plural de governador, pois além de funcionar como INIT de TÍTULO, também ocorre como EN de local em "*Palácio dos Governadores*".

6. CONCLUSÃO

Do que foi exposto até aqui, salientamos que a categorização do léxico de uma língua com fins a implementação computacional tem se mostrado complexa, o que gerou muitas dúvidas e dificuldades. É preciso pensar cada item lexical em suas múltiplas possibilidades e não perder de vista que, no caso desta pesquisa, o fato de lidar com o português histórico estabelece questões novas a serem abordadas. O devido cuidado com cada item e o conhecimento básico da história da língua são importantes para auxiliar nesta empreitada, pois somente um trabalho minucioso e fundamentado pode evitar erros e proporcionar resultados satisfatórios. Uma classificação de ENs, o mais isenta de erros quanto for possível, é muito importante, uma vez que futuramente o dicionário de abreviaturas do português histórico deverá ser disponibilizado para as pesquisas e estudos de textos históricos do Brasil, o que, sem dúvida, contribuirá, inclusive para melhor conhecimento da própria História do Brasil.

Referências Bibliográficas

ALUÍSIO, S. M. *et al*. An account of the challenge of tagging a referente corpus of Brazilian portuguese. In: *PROPOR 2003, Lecture Notes on Artificial Intelligence*. Faro, Portugal: Springer Verlag, 2003. v. 1.

ARANHA, C. ; PASSOS, E. P. L. . A *Tecnologia de Mineração de Textos*. RESI. Revista Eletrônica de Sistemas de Informação, v. 2, p. 2, 2006.

BAPTISTA, J et al. Building a Dicionary of Antroponyms. In: Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira & Maria Carmelita Dias (eds), *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006*. Itatiaia, Brazil, May 2006 (PROPOR 2006) (13-17 de maio de 2006), Berlin/Heidelberg: Springer Verlag, pp. 21-30)

BERBER SARDINHA, Tony. *Lingüística de corpus*. Barueri, SP: Manole, 2004.

FLEXOR, M. H. M. O. *Abreviaturas – Manuscritos dos séculos XVI ao XIX*. 2. ed. São Paulo: UNESP, 1991, p. 468.

LINGUATECA. Repentino - repositório para reconhecimento de Entidades Nomeadas: *Para que é que o reconhecimento de entidades nomeadas é importante?*. Disponível em: http://poloclup.linguateca.pt/repentino/faq.html. Acesso em: 08 de set 2008.

MENUZZI, S. M; G. A OTHERO. *Lingüística Computacional*: *teoria & prática*. São Paulo: Parábola, 2005.

OTHERO, G. A. *Lingüística Computacional: uma breve introdução*. Letras de Hoje. Porto Alegre. v. 41, nº 2, p. 341-351, junho, 2006

PARDO, T.A.S.; MUNIZ, M.; NUNES, M.G.V. (2006). Unitex-PB: desenvolvimento e disponibilização de recursos e ferramentas lingüístico-computacionais para o Português do Brasil. In *Cadernos de Resumos do 54º Seminário do GEL*. Araraquara-SP, Brasil. 27 a 29 de Julho.

SANTOS, D.; CARDOSO, N; SECO, N.; VILELA, R. Breve introdução ao HAREM. Em Diana Santos e Nuno Cardoso, editores, *HAREM, a primeira avaliação conjunta de sistemas de reconhecimento de entidades mencionadas para o português: documentação e actas do encontro*, Linguateca, 2007.

SARMENTO, L., PINTO, A. S. & CABRAL, L. "REPENTINO - A Wide-Scope Gazetteer for Entity Recognition in Portuguese". In Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira & Maria Carmelita Dias (eds.), *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006. Itatiaia, Brazil, May 2006 (PROPOR'2006)* LNAI 3960, 13-17 de Maio de 2006, Berlin/Heidelberg: Springer Verlag, pp. 31-40. (http://www.linguateca.pt/REPENTINO/).