Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista – UNESP

Tratamento e Classificação das Entidades Nomeadas (ENs) para um Dicionário de Abreviaturas

Abner Maicon Fortunato Batista
Maria Cristina Parreira da Silva

NILC-TR-09-02
Fevereiro, 2009

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasi

SUMÁRIO

Introdução

No período colonial brasileiro, época que se estende do século XVI até o início do século XIX, alguns fatores importantes geraram o costume entre os escreventes de abreviar palavras e expressões com o intuito de facilitar a escrita à mão. Entre esses fatores podemos apontar a escassez de recursos materiais para a escrita, devido ao seu alto custo, bem como a inexistência de um sistema ortográfico oficial com que fosse possível escrever de modo padronizado os textos aqui produzidos.

Não se pode deixar de ressaltar que os sistemas de comunicação no Brasil Colônia eram muito precários, e, nesse caso, as abreviaturas em textos históricos permitiam que os manuscritos pudessem ser produzidos mais rapidamente, auxiliando, portanto, na economia de tempo e espaço. Os primeiros documentos impressos surgiram somente a partir de 1808, quando a família real portuguesa instalou-se no Brasil. Estes ainda apresentavam abreviaturas devido ao dispendioso material utilizado na impressão.

A utilização em massa de formas abreviadas não padronizadas nos primeiros séculos da história do Brasil gerou um grande contingente de abreviaturas, que dificultam atualmente a leitura e a compreensão dos textos históricos. Para demonstrar essas dificuldades, cabe destacar que a prática constante de abreviar engendrou muitas abreviaturas que podem referir-se a mais de uma significação, como é o exemplo da abreviatura "P", a partir da qual, alguns de seus possíveis significados são exibidos na tabela a seguir:

P	Paço	P	Pedro	P	Passe	P	Passada
P	Padre	P	Pelo	P	Passou-se	P	Passado
P	Pagará	P	Pés	P	Patacas	P	Preto
P	Pago	P	Pequena	P	Paternidade	P	Príncipe
P	Pai	P	Per	P	Paulo	P	Públicas
P	Palácio	P	Pero	P	Paz	P	Parte
P	Palma	P	Pires	P	Peça	P	Partilha
P	Palmo	P	Polegada	P	Pede	P	Presidente
P	Palmos	P	Por	P	Pro	P	Preta
P	Papa	P	Porta	P	Provincial	P	Próximos
P	Para	P	Porto	P	Província	P	Pública
P	Parda	P	Pré	P	Provisão		
P	Pardo	P	Praça	P	Próximo		

Tabela 1 – Significados da abreviatura P

Além disso, para uma mesma palavra ou expressão podem existir múltiplas abreviaturas, como é o caso da lexia “São Paulo”:

S^m P^Lo	S^m P^o
S^m P^lo	S^m Pl^o
Sam Pl^o	SP
S p	S p^l
S^lo	S PL^o
S p^o	S P^o
S P.^o	S Pa.
S Pl^o	SPL^o
SPl^o	S. Pl^o

Tabela 2 – Abreviaturas para a lexia “São Paulo”[1]

Esta pesquisa partiu da contribuição de Flexor (1991), que elaborou um dicionário impresso de abreviaturas do português histórico do Brasil dos séculos XVI ao XIX. A partir da digitalização desse dicionário, um trabalho prévio, coordenado pelos colaboradores deste trabalho (da USP-São Carlos e da UFSCAR), 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.

É evidente que a publicação de dicionários tradicionais traz grande contribuição para a leitura de livros da época colonial, contudo, com a quantidade extensa de textos históricos que agora podem ser digitalizados e armazenados em *corpora,* faz-se necessária a possibilidade de extração automática de conhecimento desses textos. Afinal, 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. Além disso, a consulta a dicionários impressos pode exigir um dispêndio de tempo muito maior do que aquela realizada em meio digital.

A criação de ferramentas computacionais que tornem possível o tratamento automático desses *corpora* surge para solucionar essas dificuldades, 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

[1] O caractere “^” foi utilizado para demonstrar que tudo o que ocorre a sua direita está sobrescrito. Para elucidar tomemos como exemplo a abreviatura S Pl°, que foi transcrita como S Pl^o, desse modo, será possível processa-la computacionalmente.

modo mais ágil e eficiente, mas também para possibilitar a compreensão de muitas informações que poderiam se perder.

Este trabalho, ao tratar de um tipo recorrente de abreviaturas nos textos históricos, pretende colaborar para a criação de um dicionário eletrônico de abreviaturas do Português Histórico do Brasil (PHB). A ferramenta que se pretende elaborar será de muita serventia para o trabalho com textos históricos digitalizados.

Além disso, considerando-se a importância que os nomes próprios têm para o tratamento automático de informações em textos, um dicionário eletrônico de abreviaturas, que abranja informações sobre **Entidades Nomeadas** (ENs), ou seja, entidades concretas ou abstratas que possuem um nome próprio, constitui também 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.

Esta pesquisa tem como objetivo principal o reconhecimento e a classificação de ENs para a implementação de um dicionário de abreviaturas eletrônico que contenha, além dos dados morfológicos, informações de natureza semântica que possam vir a ser úteis em diferentes aplicações do processamento computacional do PHB.

1. Fundamentação Teórica

Para o melhor entendimento dos objetivos desta pesquisa, convém definir os principais conceitos utilizados e apresentar as principais áreas de pesquisas envolvidas.

1.1 Definição dos Principais Conceitos Utilizados

Como já mencionado, o objeto desta pesquisa é a abreviatura em textos históricos. A designação de **abreviatura**, do grego *braqui* (reduzido) e *graphein* (escrever), é, segundo Costa (2006), uma forma reduzida que é utilizada para se escrever uma palavra. "O que se abrevia são sílabas, palavras ou frases de um conjunto escrito, das quais se reduz alguma ou algumas de suas letras."

Flexor (1991) explica que a proliferação de abreviaturas em textos históricos pode ser compreendida não somente pela escassez de material, mas também pela economia de tempo que se podia conseguir pela prática da abreviação, permitindo com que os manuscritos pudessem ser produzidos mais depressa. Aliás, pode-se dizer que a economia de tempo e espaço é a função primordial da abreviatura.

A imagem a seguir, retirada do texto de Costa (2008) apresenta algumas abreviaturas empregadas em manuscritos do período Colonial Brasileiro.

Imagem 1: Abreviaturas de textos manuscritos. Costa (2008)

Entidades Nomeadas (ENs) são entidades concretas ou abstratas que possuem um nome próprio. Sua designação difere da de Entidades Mencionadas (EMs), entendidas como entidades que ocorrem em contextos determinados. Por não ocorrerem

em um contexto específico, mas sim abranger a totalidade de significações de uma mesma lexia, há quem prefira a utilização do termo 'Nomes de Entidades' ao invés de Entidades Nomeadas.

Para elucidar a distinção entre entidades nomeadas e mencionadas, tomemos como exemplo a entidade "Magalhães", que poderá ser compreendida tanto como um nome próprio de pessoa como o de um local se considerada como EN, mas que só poderá ser entendida como um ou outro se considerada uma EM. "Estreito de Magalhães" (entidade mencionada como local), "Antônio de Magalhães" (entidade mencionada como pessoa).

Tendo em vista que esta pesquisa se insere no tema do tratamento do léxico de uma língua, faz-se necessária a definição de **Lexia**, uma vez que se trata de um conceito muito utilizado durante a execução do projeto. Biderman (2001) define 'lexia' em oposição ao conceito de 'lexema':

> "Os lexemas se manifestam, no discurso, através de formas ora fixas, ora variáveis. Essa segunda alternativa é a mais freqüente nas línguas flexíveis e aglutinantes. Assim, em português, o lexema CANTAR pode manifestar-se discursivamente como *cantei, cantavam, cantas, cantando*, etc. O lexema MENINO como *meninos* e *meninos*. A essas formas que aparecem no discurso daremos o nome de lexia. Portanto, *cantei, cantavam, catas, cantando, menino* e *meninos* são lexias." [grifos da autora]

O conceito de 'lexia' ainda se especifica, dividindo-se em 'lexia simples' e 'lexia complexa'. A compreensão dessa divisão não é algo simples, pois conforme salienta Biderman (2001):

> "Nas realizações da fala as fronteiras entre as palavras são difusas. Existe toda uma gama de graus de soldadura entre os elementos daquilo que chamaremos lexia complexa, por oposição a lexia simples. Relativamente aos graus de aglutinação entre os elementos de uma lexia complexa, poderemos distinguir algumas perfeitamente soldadas e outras com forte índice de coesão interna. (...) Há sempre uma parte do sistema em vias de formação, outra em via de desaparecimento e outra perfeitamente acabada. As realizações discursivas refletirão sempre esses fluxos e refluxos do sistema. Daí a dificuldade prática com que nos defrontamos, quando nos propomos a segmentar um texto em suas lexias componentes. Muitas vezes se porá o problema: tal seqüência já constitui um lexema em língua, ou é apenas muito freqüente na fala enquanto combinatória de palavras?".

Uma vez que estamos lidando com o português histórico de quatro séculos de história da língua, seria muito difícil determinar os "graus de soldadura" dos elementos léxicos de lexias de uma época distante, pois demandaria um amplo conhecimento da história do português. Desse modo, optou-se por uma divisão que obedecesse ao critério de contigüidade, isto é, foram consideradas lexias complexas aquelas que possuíssem mais de um elemento e lexias simples aquelas constituídas de apenas uma palavra.

1.2 Áreas de pesquisa envolvidas

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 esta pesquisa trata das abreviaturas em um *corpus* digitalizado 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.

2. Metodologia

O conjunto de abreviaturas, que foram digitalizadas em formato XML, partiu de uma lista de abreviaturas do dicionário de Flexor (1991) acrescida de abreviaturas retiradas do corpus do projeto DHPB (Dicionário Histórico do Português do Brasil) de autoria da professora doutora Maria Tereza Camargo Biderman. Esta pesquisa se insere no grupo de trabalhos cujos resultados contribuirão para o projeto DHPB, uma vez que o *corpus* utilizado para a análise das ENs abrange os períodos de ocorrências das abreviaturas do dicionário de Flexor.

Para o tratamento das abreviaturas em formato XML utilizou-se o *software* Excel, desenvolvido pela Microsoft, em suas versões 2003 e 2007. A utilização de duas versões de um mesmo *software* se deu em virtude dos distintos recursos que cada edição do programa pode privilegiar.

Abreviat	Expansão	Forma Canônica	Flexão	Categoria Gramatical	Atributo	Tipo de Entidade	Século	Século
ã	ano	ano	ms	N	INIT	TEMPO	18	
Ã	anos	ano	mp	N	INIT	TEMPO	18	
ã	anos	ano	mp	N	INIT	TEMPO	18	
ã	hão	haver	P3p	V			17	
Ã	Afonso	Afonso	ms	N	ENT	PESSOA	16	
a barracan	abarracamento	abarracamento	ms	N			18	
a Costam^	acostamento	acostamento	ms	N			18	
a Gregd^o	agregado	agregado	ms	N	INIT	TITULO	18	19
a Gregd^o	agregado	agregado	ms	A			18	19
a Gregd^o	agregado	agregar	K	V			18	19
a Remat^e	arrematante	arrematante	ms:fs	N	INIT	TITULO	18	19
A^a	Aranha	Aranha	ms	N	ENT	PESSOA+AMB	19	
A^a	Aranha	Aranha	fs	N	ENT	PESSOA+AMB	19	
A^a	Aranha	aranhar	P3s	V			19	
A^a	Aranha	aranhar	Y2s	V			19	
A^al	auxiliar	auxiliar	ms	A			18	
A^al	auxiliar	auxiliar	fs	A			18	
A^al	auxiliar	auxiliar	ms	N	INIT	TITULO	18	

Imagem 2: Exemplo de anotação de ENs em formato XLS

A primeira etapa no tratamento das abreviaturas foi a sua divisão entre lexias simples e lexias complexas. Devido à dificuldade apresentada acima, optou-se por tratar separadamente as lexias complexas, ou seja, aquelas que apresentam mais de um elemento, e as lexias simples, aquelas constituídas de apenas uma palavra.

As abreviaturas digitalizadas foram fornecidas pela co-orientadora e já estavam divididas em letras de A a Z no início da pesquisa, de modo que cada letra ocupava uma planilha em XML. Após a divisão das lexias, cada letra passou a conter duas planilhas, sendo uma para as lexias simples e outra para lexias complexas. Além das letras do alfabeto, outras três listas de abreviaturas foram processadas e são denominadas

“outras”, “outras abreviaturas” e “inventário”, que contêm diversas abreviaturas iniciadas por caracteres alfabéticos e não alfabéticos.

A etapa seguinte no tratamento das abreviaturas foi a classificação morfológica das lexias simples. Para tanto, fez-se uso das instruções de categorização de unidades morfológicas do UNITEX (Palmier, 2006), que possibilita a classificação para todas as classes morfológicas e flexões da língua portuguesa e utiliza-se de um formalismo de codificação de conhecimento denominado DELA (*Dictionnarie Electronique du LADL*). Essa classificação, que obedece ao formato DELA, é útil no sentido de que possibilita que essas informações lingüísticas possam ser empreendidas em processamentos automáticos por computador.

Outra informação importante que foi anotada é a forma canônica de cada abreviatura, pois a partir dessa anotação, pode-se resgatar todas as formas correlatas. Por exemplo, realizando uma pesquisa a partir da forma canônica “irmão”, obteremos todas as suas flexões:

Irmão	Irmã	Irmãos	irmãs

Tabela 3: Resultados para a forma canônica “irmão”

E as seguintes abreviaturas:

I	I^s	Ir^am	Irm^m
I^m	Ĩr	Ir^m	Irm^o
I^o	Ĩr	Irm^os	Ir^mos
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 4: Abreviaturas para a forma canônica “irmão”

Como muitas lexias podem apresentar mais de uma categoria morfológica, sendo esse o caso da lexia “legado” (podendo ser compreendida tanto como um substantivo, um adjetivo ou quanto um verbo na forma de particípio), em diversos momentos foi necessário multiplicar a linha da abreviatura no Excel, para que cada linha apresentasse

somente uma categoria morfológica. Desse modo, todas as linhas que continham a lexia "legado", por exemplo, tiveram que ser triplicadas para que apenas uma categoria morfológica figurasse em cada linha de uma mesma abreviatura.

Abreviat	Expansão	Forma Canônica	Flexão	Categoria Gramatical	Atributo	Tipo de Entidade	Século	Século
Leg^do	legado	legar	K	V			19	
Leg^do	legado	legado	ms	A			19	
Leg^do	legado	legado	ms	N			19	
Leg^o	legado	legar	K	V			18	
Leg^o	legado	legado	ms	A			18	
Leg^o	legado	legado	ms	N			18	
Legd^o	legado	legar	K	V			18	19
Legd^o	legado	legado	ms	A			18	19
Legd^o	legado	legado	ms	N			18	19
Lg^o	legado	legar	K	V			18	
Lg^o	legado	legado	ms	A			18	
Lg^o	legado	legado	ms	N			18	
Lgd^o	legado	legar	K	V			18	
Lgd^o	legado	legado	ms	A			18	
Lgd^o	legado	legado	ms	N			18	
Lgd^o	legado	legar	K	V			18	
Lgd^o	legado	legado	ms	A			18	
Lgd^o	legado	legado	ms	N			18	

Imagem 3: Abreviaturas triplicadas para a lexia "legado"

Após o tratamento morfológico das lexias, empreende-se a classificação semântica, que consiste, basicamente, na identificação e classificação das ENs a partir da busca e análise das lexias no *corpus* do projeto DHPB. A análise das lexias no *corpus* é um trabalho minucioso e que demanda um tempo considerável, pois é preciso verificar atentamente todas as ocorrências no *corpus* para que se possa verificar a existência de ENs, e, em seguida, decidir a qual categoria tal EN pertencerá.

A classificação das ENs obedeceu às diretivas do HAREM, uma avaliação conjunta na área de processamento de linguagem natural para a classificação de entidades mencionadas do português em uso. O HAREM permite a classificação de ENs em dez categorias distintas: obra, acontecimento, organização, variado, pessoa, abstração, tempo, valor, local e coisa. É organizado pela LINGUATECA, um centro de recursos para o processamento computacional da língua portuguesa com sede em Portugal. Cabe destacar que o HAREM é de domínio público e pode ser acessado a partir da página: www.linguateca.pt/HAREM/.

Apesar de cobrir satisfatoriamente as ENs encontradas nesta pesquisa, deve-se ressaltar o fato de que as diretivas do HAREM não são instruções de classificação de ENs para o português histórico, cabendo aos pesquisadores deste trabalho transpor os exemplos que ele traz do português contemporâneo para as ocorrências de um conjunto

de ENs do português histórico, o que frequentemente gerou muitas dúvidas. Além disso, pode-se reiterar que as diretivas do HAREM trazem instruções para a classificação de Entidades Mencionadas e não de Entidades Nomeadas como é o caso deste trabalho.

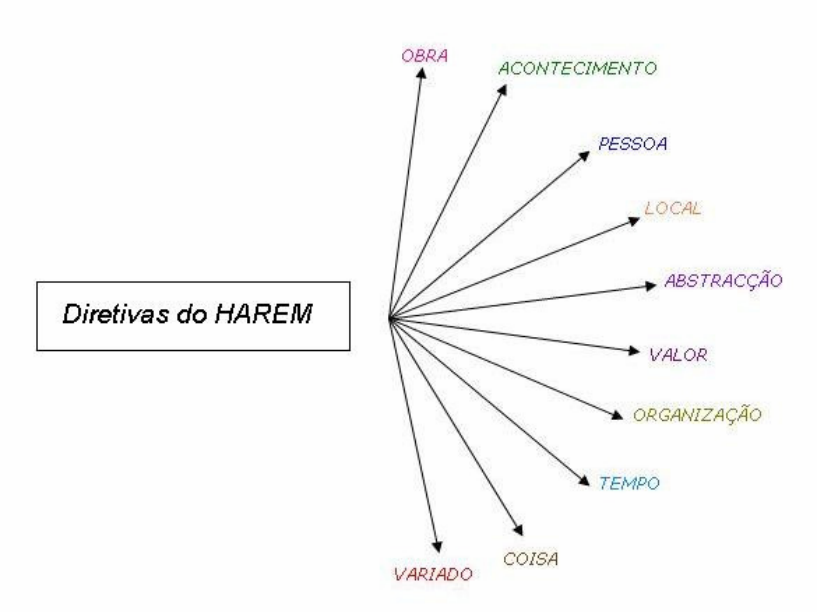

Imagem 4: Categorias do HAREM para a classificação das ENs

Além das categorias do HAREM, foi criada uma nota, denominada AMB, para marcar ENs que possuem significados comuns dentro do léxico da língua portuguesa que não aqueles incluídos nas categorias semânticas. Trata-se do caso da EN "Aranha" que pode designar tanto um sobrenome, quanto um inseto artrópode.

Para fins de comparação, pôde-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.

Sabe-se que a utilização de procedimentos heurísticos, tais como a busca de palavras iniciadas por maiúsculas ou de palavras antecedidas por pronomes de tratamento, pode ser útil para encontrar Entidades Nomeadas, entretanto, quando nos referimos ao português histórico a utilização dessas regras heurísticas podem ser insuficientes, tendo em vista que nem sempre os nomes próprios aparecem iniciados por

letra maiúscula e pronomes de tratamento eram, frequentemente, abreviados. De acordo com Vale *et al* (2008):

> "Using heuristics is efficient for extending lexicons of NEs, such as in the search for words (or n-grams) starting with a capital letter with is not in the beginning of the sentence or a search for words followed by treatment pronouns. The heuristic rules thus allow for identification of named entities. The identification of NEs abbreviated as "V. M" (Vossa mercê) is made difficult due to the presence of the dot, especially because some NEs are similar to sentence beginnings (".M.") and cannot be retrieved using the heuristic rule mentioned above."

Devido a essas dificuldades, uma solução encontrada para o PHB, no sentido de expandir o número de ENs dentro do *corpus* utilizado, foi a criação de uma *tag*, denominada INIT, que antecedesse ENs. Assim sendo, podemos considerar a lexia "padre" como um INIT, pois antecede frequentemente nomes próprios de pessoa, como por exemplo, "padre Antônio Vieira". Portanto, a partir dos INITs levantados neste trabalho, poderá ser possível, no futuro, desenvolver um sistema automático de extração de entidades nomeadas e, dessa forma, expandir a quantidade de ENs encontradas para o português histórico do Brasil.

A identificação dos INITs ocorreu semelhantemente ao reconhecimento das ENs. A partir da análise de cada ocorrência de uma mesma lexia no *corpus* é possível verificar se tal lexia é uma EN, um INIT, ambos ou simplesmente uma palavra. No caso dos INITs também lança-se mão de categorização que, por sua vez, inclui todas as categorias presentes no HAREM, além das categorias *TÍTULO, CARGO, PARENTE* e *TRATAMENTO*.

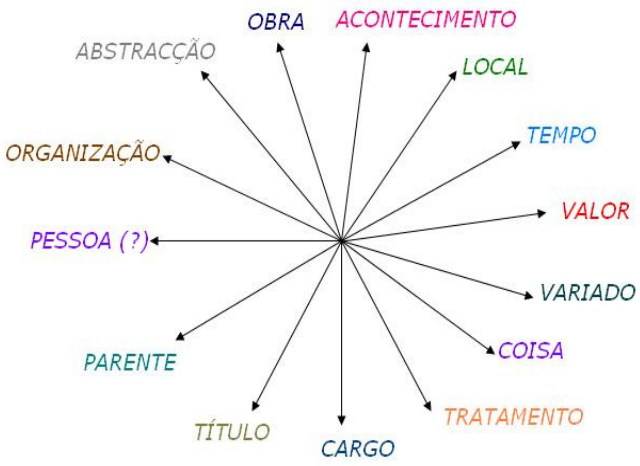

Imagem 5: Categorias para a classificação dos INITs

O *corpus* utilizado nesta pesquisa pode ser acessado por meio de duas ferramentas: *Philologic* e *Unitex*. De acordo com Cândido Jr. (2008, p. 41):

> *Philologic* (UNIVERSITY OF CHICAGO, 2006) é uma ferramenta *Web* para buscas, recuperação e análise de *corpus* desenvolvida por Leonid Andreev e pesquisadores da Universidade de *Chicago* como uma das metas do projeto ARTFL (American and French Research on the Treasury of the French Language)

O *Philologic* é uma base que oferece muitos recursos para o trabalho com *corpora*. Além do concordanceador, que permite realizar buscas dentro e fora de contexto, ele traz ainda o recurso de busca por similaridade, que apresenta ao usuário uma lista de palavras graficamente semelhantes à palavra pesquisada. Como muitas palavras do PHB são grafadas sem os diacríticos, o recurso de busca por similaridades é muito importante para a verificação de toda a gama de ocorrências de uma mesma lexia. Lamenta-se, entretanto, o fato de que tal recurso só exista para as buscas com lexias simples, de modo que, foi preciso empreender mais de uma busca para algumas lexias complexas a fim de verificar todos os contextos em que ocorrem.

O principal inconveniente da utilização do *Philologic* para as buscas no *corpus* é a de que por se tratar de uma ferramenta on-line, está propensa a eventuais problemas da rede, de modo que em determinadas ocasiões não foi possível o acesso à ferramenta. Apesar dessa indisponibilidade em alguns momentos da pesquisa, o *Philologic* foi a ferramenta mais utilizada para a verificação dos contextos das lexias investigadas, o fato de poder ser acessado a partir de qualquer computador conectado à internet foi um fator importante na decisão de privilegiá-lo em seu uso. Apresentamos a seguir algumas imagens para ilustrar uma busca no *corpus* realizada a partir *Philologic*:

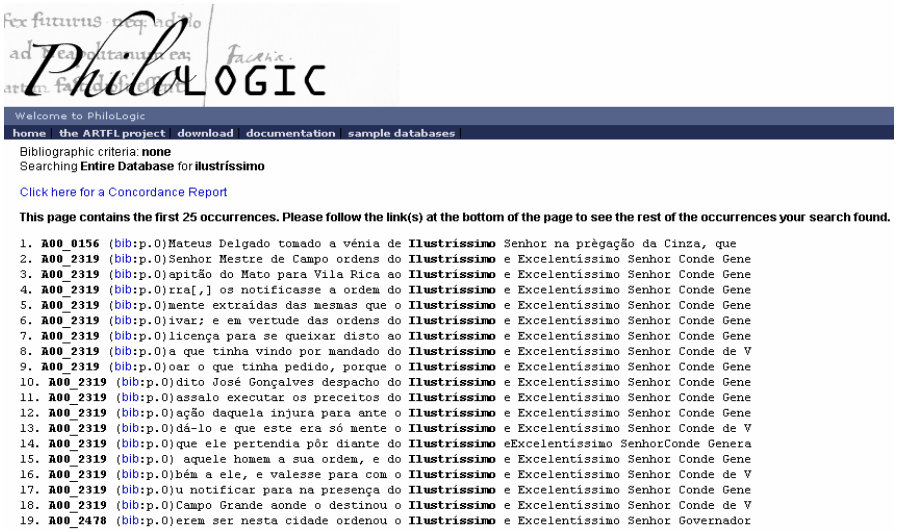

Imagem 6: Exemplo de busca no *corpus* utilizando o *Philologic*

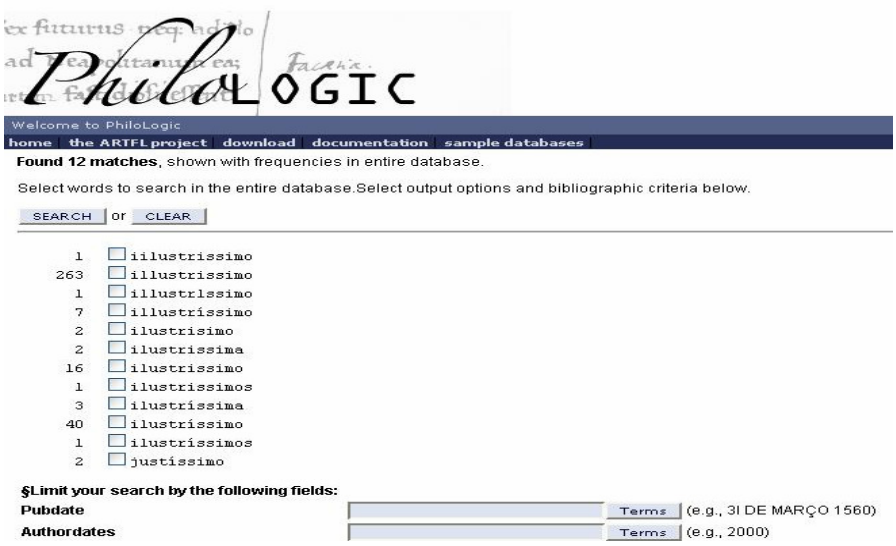

Imagem 7: Exemplo de busca por similaridade no *Philologic* a partir da lexia "ilustríssimo"

Outro meio de acesso ao *corpus* é através do UNITEX, que é definido por Palmier (2006) como "um conjunto de softwares que permite processar os textos em línguas naturais utilizando recursos lingüísticos. Esses recursos se apresentam na forma de dicionários eletrônicos, de gramáticas e tabelas de léxico-gramática.".

O UNITEX apresenta a vantagem de ser um *software off-line*, isto é, não é preciso acessá-lo pela internet, entretanto, o processo de busca no UNITEX é mais lento do que o verificado no *Philologic*. Os dados não são mostrados instantaneamente após pedir a busca, pois o programa precisa processar todo o *corpus* para em seguida mostrar

os resultados, o que demanda mais tempo para as pesquisas. Além disso, não encontramos no UNITEX um sistema de buscas por similaridade, de modo que foi preciso efetuar mais de uma pesquisa para algumas lexias, sobretudo as que apresentam diacríticos. Essas desvantagens também contribuíram para que privilegiássemos o uso do *Philologic* para a verificação das lexias no *corpus*. Sendo assim, o UNITEX foi utilizado somente em momentos em que o *Philologic* encontrava-se indisponível na rede. Apresentamos a seguir uma figura que ilustra uma busca no *corpus* por meio do UNITEX:

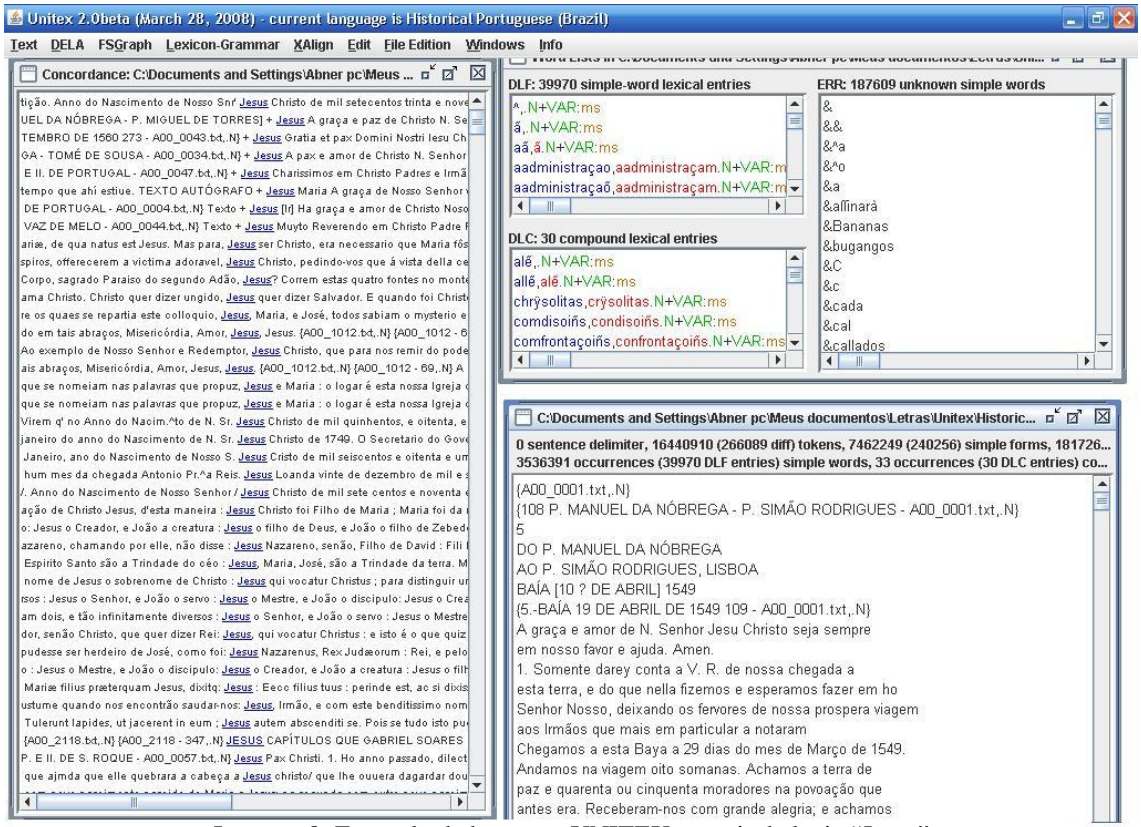

Imagem 8: Exemplo de busca no UNITEX a partir da lexia "Jesus"

Convém dizer que o corpus DHPB não está disponível livremente para todos os usuários. Apesar de o acesso ao Philologic ocorrer pela internet, a interface é protegida por senha e somente os pesquisadores tem acesso.

Com o intuito de esclarecer eventuais dúvidas de natureza morfológica ou semântica a respeito das lexias analisadas no trabalho, tomou-se como fontes de consulta o *Dicionário Eletrônico Houaiss da Língua Portuguesa* (versão 1.0 – dezembro de 2001) e o *Novo Dicionário Eletrônico Aurélio versão 5.0* (2004). Quando

mesmo com a utilização desses dicionários, a identificação de algum elemento das lexias ainda apresentava-se complexa, lançou-se mão do *Dicionário da Língua Portuguesa* de Raphael Bluteau, datado do século XVIII e considerado como o primeiro dicionário do português, que pode ser acessado gratuitamente pelo site: http://www.ieb.usp.br/online/dicionarios/Bluteau/formBuscaDicionarioPlChave.asp.

3. Dificuldades encontradas na classificação e análise do corpus

A análise de elementos lingüísticos é uma tarefa complexa e que requer muita atenção, afinal, como já afirmava o pai dos Estudos Lingüísticos, Ferdinand Saussure, a língua é multifacetada, ou seja, apresenta características variadas e múltiplas possibilidades de expressão e compreensão.

Quando nos referimos à categorização de nomes próprios, estamos nos lançando a um assunto de certa complexidade, pois trata-se de um tema muito afim com a Pragmática, uma das áreas mais movediças dentro da Lingüística. Por conta disso, é preciso pensar cada unidade lexical em suas várias possibilidades de utilização na língua, além disso, há muitas ENs cuja categorização depende do contexto maior onde estão inseridas.

Alguns exemplos podem nos ser úteis para entender essa gama de possibilidades que encontramos nos elementos lingüísticos. Tratando, inicialmente, das classificações morfológicas, vemos que muitas lexias podem pertencer a mais de uma categoria, como é o exemplo da palavra "são", que pode ser compreendida como um substantivo, um adjetivo ou como forma verbal do verbo "ser". No caso da lexia "irem", podemos inseri-la somente na categoria de verbo, entretanto, trata-se de uma forma verbal de dois verbos distintos, quais sejam, "ir" e "irar".

Encontrar todas as possibilidades de uma mesma lexia parece ser uma tarefa, que apesar de trabalhosa, pode ser simples com a utilização de dicionários em meio eletrônico. Entretanto, não podemos esquecer que estamos lidando com o português histórico, que apresenta vários elementos que já caíram em desuso no estágio atual da língua portuguesa, desaparecendo mesmo nos registros contemporâneos mais formais. Trata-se do caso, por exemplo, da palavra "aguardente", que apesar de ser muito empregada atualmente em seu sentido de bebida destilada de alto teor alcoólico, raramente ocorre como formas do presente do subjuntivo ou do imperativo afirmativo do verbo "aguardentar".

Em relação ao tratamento semântico das lexias, é possível falar também de uma multiplicidade de possibilidades, uma vez que uma única EN pode pertencer a várias categorias a depender de seus contextos de ocorrência. Daí a importância de se analisar minuciosamente as ocorrências das lexias, procurando não deixar passar despercebido qualquer contexto que indique que tal lexia possa ser entendida como uma EN ou como um INIT. No caso da lexia "Santiago", por exemplo, encontramos cinco possíveis

categorias de entidades: PESSOA (“Antônio Mendes Santiago”), LOCAL (“Ilha de Santiago”), ORGANIZAÇÃO (“Ordem de Santiago”), TEMPO (“Dia de Santiago”) e OBRA (“Igreja de Santiago”).

Além da possibilidade de uma mesma lexia pertencer a múltiplas categorias semânticas, uma única lexia pode ser classificada concomitantemente como uma EN e como um INIT. É o caso da palavra “Monte”, que pode ser compreendido como nome próprio de pessoa (“doutor Manoel do Monte Fogaça”), portanto uma EN de PESSOA; mas que também pode ser entendido como um indicativo de nomes próprios de lugares (“monte Jaricoacoara”), portanto, um INIT de LOCAL.

Uma dificuldade decorrente da categorização das ENs foi nos adequar às diretivas do HAREM, uma vez que no início da pesquisa, ainda pouco afeitos ao trabalho de categorização de ENs, foi preciso, frequentemente, retomar tais diretivas nas diversas situações que se apresentavam, o que, sem dúvida, consumiu muito tempo. Por outro lado, essa constante releitura das instruções do HAREM permitiu, com o passar do tempo, uma maior familiaridade no trabalho com as ENs, tornando sua identificação e classificação cada vez mais eficiente.

No tratamento das ENs há uma exigência que deve sempre ser observada: o devido cuidado ao se trabalhar com categorias, uma vez que em alguns casos, estas possuem limites tênues de diferenciação no que concerne às instruções de categorização.

Podemos ilustrar esse caso com um exemplo que, mesmo parecendo simples, é bastante complexo – sempre que encontramos um nome próprio de uma pessoa no *corpus*, é preciso ter o cuidado de analisar se a EN faz referência a uma pessoa propriamente dita ou ao nome dessa pessoa, já que quando se tratar de uma referência a um nome, por exemplo: “seu escravo de nome José”, a EN deve ser classificada como ABSTRAÇÃO e não como PESSOA. Como os nomes próprios são, geralmente, de alta freqüência no *corpus*, muitas ENs foram classificadas simultaneamente nas duas categorias.

No que concerne à categoria VALOR (entendida pelo HAREM como uma categoria que pode referir-se a quantidades absolutas ou relativas, podendo também constituir dinheiro ou classificações desportivas, ordinais normais e outras), uma tarefa que se demonstrou árdua foi a categorização de nomes que designam coletividade, tais como “queira” (lote de escravos ou serviçais) e “maço”. Apesar das diretivas do HAREM não mencionarem nada em relação aos substantivos coletivos, entendemos que

tais lexias podem ser entendidas como uma EN de VALOR, uma vez que se referem às quantidades relativas de um grupo de unidades.

Como já foi dito anteriormente, no período Colonial da História do Brasil não existia um sistema ortográfico oficial para a língua portuguesa. Biderman (2006) explica que a língua falada no Brasil era majoritariamente a Língua Geral, tendo em vista que muitos dos nativos ainda não falavam o português. A inexistência de uma ortografia padronizada fez com que muitas palavras assumissem várias grafias. Portanto, um obstáculo encontrado foi distinguir quando duas formas parecidas graficamente eram formas distintas de uma mesma palavra ou se eram dois vocábulos independentes. Em várias ocasiões foi possível identificar formas semelhantes de uma mesma palavra, como foi o caso da lexia “Marcelina”, que apesar de não ocorrer no *corpus* na forma como está grafada, há a ocorrência da forma "Marcellina" que nos permite classificar tal lexia como uma EN de PESSOA. Entretanto, foi preciso atentar para os casos de formas como "Conselho" e "Concelho", por se tratar de dois vocábulos distintos, cada um com um sentido próprio, não podendo ser analisados concomitantemente.

Conforme já explicitado no item 2, a nota AMB foi criada para marcar significados além dos abarcados pelas categorias semânticas. Entretanto, verificar se uma EN possui significados além daqueles que as categorias semânticas abrangem nem sempre é uma tarefa fácil, uma vez que existem muitos nomes comuns pouco empregados na língua e cujos significados podem passar despercebidos. No caso da lexia ‘tabatinga’, podemos compreendê-la como nome próprio de pessoa e de local, mas também como uma palavra do léxico comum, que designa uma espécie de argila utilizada em construções. Entretanto, a identificação desse sentido comum que alguns nomes próprios possuem pode ter sido negligenciada, e em virtude disso, algumas ENs podem não ter sido classificadas juntamente com a nota AMB.

Um impasse encontrado quanto à categorização dos INITs foi a ocorrência de muitos elementos lingüísticos que antecedem um nome próprio de pessoa, mas que por seu valor semântico não poderiam ser classificados nas categorias TITULO, PARENTE, TRATAMENTO ou CARGO. Ainda que a categorização para os INITs fosse aberta, podendo incluir novas categorias, optamos por utilizar somente as categorias já citadas para não estender muito as possibilidades de classificação, o que poderia gerar novos problemas e inconvenientes. Desse modo, decidiu-se classificar lexias como: “escravo”, “santo”, “defunto”, “mulato”, “estrangeiro” e “réu” dentro da categoria PESSOA, uma vez que só antecedem nomes próprios de seres humanos.

Outro problema que surgiu logo no início da pesquisa foi a ocorrência de alguns erros de grafia na expansão de certas abreviaturas, pois tais expansões não eram encontradas em nenhum dicionário, além disso, muitas delas fugiam do padrão de escrita da língua portuguesa. A partir da forma de cada abreviatura foi possível, por inferência, recuperar algumas dessas expansões e então, proceder ao seu tratamento lingüístico. A tabela abaixo mostra o conjunto de abreviaturas ao lado da expansão incorreta e da nova expansão já corrigida.

Abreviatura	Expansão com a grafia incorreta	Nova Expansão
Cir^am	Cirugião	cirurgião
Cabelr^o	Cabeleiro	cabeleireiro
Cabelr^a	Cabelereira	cabeleireira
Cabeleir^o	Cabelereiro	cabeleireiro
CCCC^tos	Quatrocentes	quatrocentos
Comvẽ	Convm	convém
Desabam^to	Desbamento	desabamento
Dezobed^e	desodiente	desobediente
Disc.	discusão	discussão
Eg^o	Engeno	engenho
Embgr^o	enbargo	embargo
Empres^to	emprstimo	empréstimo
Emprest^o	emprstimo	empréstimo
Exatam^e	extamente	exatamente
Egualm^e	igualmante	igualmente
Em duzim^to	Induzimcnto	induzimento
Emfilicid^e	infecilidade	infelicidade
Emstrum^to	instrumeno	instrumento
Evg^a	evanglica	evangélica
Felicid^e	fehcidade	felicidade
Feliscid^e	fehcidade	felicidade
Fellexid^e	fehcidade	felicidade
Fellicid^e	fehcidade	fidelíssima
Fidelis^ma	fidehssima	fidelíssima
Fidl^ma	fidehssima	fidelíssima
Galantem^te	galantememte	galantemente
Interess^do	interassado	interessado
Interess^o	interassado	interessado
Inuentr^e	inventriante	inventariante
Inconseq^ca	inconsesqüêscia	inconsequência
Intend^ca	irtendência	intendência
Inter.	interlocurótia	interlocutória
Interpoladam^te	mterpoladamente	interpoladamente
Mis^a	misericória	misericórdia
Mize^a	misericória	misericórdia
Mizr^a	misericória	misericórdia

Mizr^da	misericória	misericórdia
Municip^l	munidpal	municipal
Miuda m^te	Muidamente	miudamente
Quant^e	guantidade	quantidade
Rellg^o	religooso	religioso
Resp^vel	Responvável	responsável
Restablecim^o	restablecimento	restabelecimento
Rompim^to	rimpimento	rompimento
Salust^no	Saiustiano	Salustiano

Tabela 6 – Entradas que sofreram alteração na expansão

Lamenta-se o fato de que não foi possível recuperar o sentido de algumas abreviaturas na lista digitalizada, e como não houve o acesso à obra original de Flexor durante a pesquisa, já que a obra de referência não consta do acervo da biblioteca de nosso campus, as expansões continuaram grafadas incorretamente, não sendo possível recuperar o seu sentido original pelo processo de inferência, o que nos levou a ignorá-las nos momentos de anotação morfológica e semântica.

Convém destacar também que na lista de abreviaturas havia alguns latinismos, que, provavelmente, deviam ser freqüentemente empregados nos séculos abrangidos no projeto devido ao prestígio que a língua latina ainda possuía como língua de cultura nos séculos que antecederam ao século XX. Por se tratar de elementos que obedecem a um padrão lingüístico próprio do Latim, tais lexias foram desconsideradas no que concerne às lexias tratadas nesta pesquisa. A seguir, apresentamos uma tabela com os latinismos encontrados durante o trabalho.

Abreviatura	**Latinismo**
AD	Annus Domini
Conv^ti	Conventi
Do	Deo
D.Gr.	Deo gratia
DG	Deo Gratia
D.Gr.	Deo gratia
e.g.	exempli gratia
G	gratia
Gr.	gratia
G.P.	Gloria Patri
Matr^as	matrias
Verb. grat.	verbi gratia

VG	verbi gratia
Vg	verbi gratia
V. gr.	verbi gratia
Xpi	Christi
Xpel	Christie Eleison

Tabela 6 – Latinismos encontrados

4. Resultados obtidos

Ao final da pesquisa obtivemos um total de 28.866 entradas para o dicionário eletrônico de abreviaturas, reiterando que muitas abreviaturas, por pertencerem a mais de uma categoria morfológica, foram multiplicadas para que cada entrada apresentasse uma única categoria semântica, podendo assim, serem processadas no formalismo DELA.

Das 28.866 entradas do dicionário, 5567 foram categorizadas como Entidades Nomeadas, o que representa um percentual de 19,29% sobre o total geral de entradas. Em relação aos introdutores de ENs, houve a ocorrência de 5744 entradas classificadas como INITs, o que perfaz 19,9% do total de entradas. Em relação às lexias classificadas simultaneamente como ENs e INITs encontramos um total de 294 unidades, isto é, cerca de 1% do total geral de entradas do dicionário.

Em relação às categorias de ENs apresentamos abaixo uma tabela e um gráfico que demonstram os valores absolutos de ocorrências para cada categoria.

CATEGORIA SEMÂNTICA	*TOTAL DE OCORRÊNCIAS*
PESSOA	3740
LOCAL	1188
VALOR	570
TEMPO	440
ABSTRAÇÃO	424
ORGANIZAÇÃO	334
OBRA	184
ACONTECIMENTO	124
VARIADO	63
COISA	23

Tabela 7 – Total de Ocorrências de ENs por Categoria

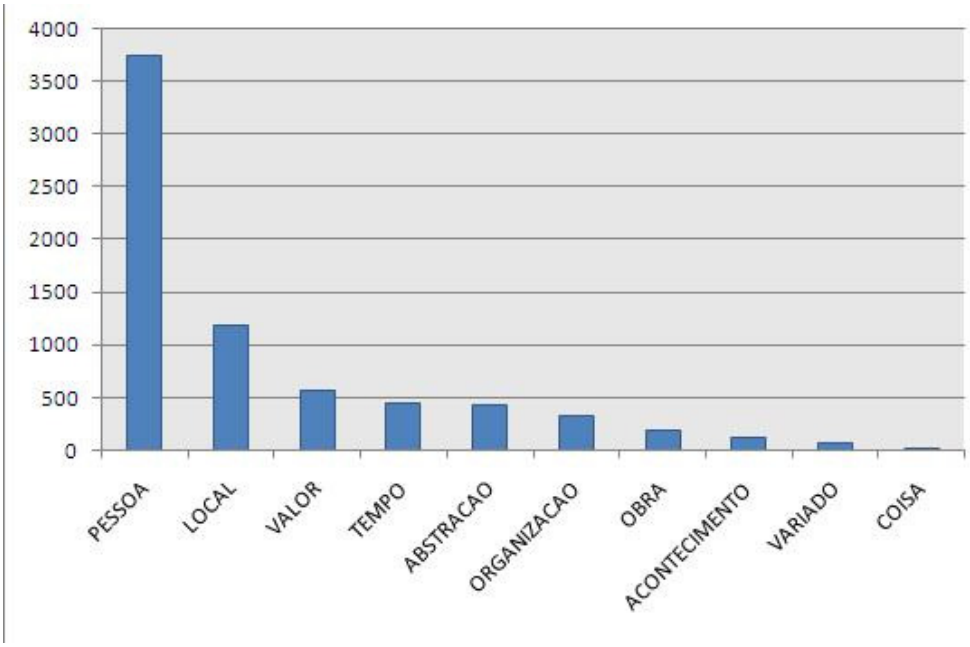

Gráfico 1 – Ocorrências de ENs por categoria

Como foi possível perceber a partir da análise dos dados acima demonstrados, houve uma preponderância muito significativa para as categorias PESSOA e LOCAL. O predomínio dessas categorias ocorreu devido à especificidade do *corpus*, que apresenta diversos documentos tais como cartas, testamentos, certidões, inventários, etc. que frequentemente apresentam nomes de PESSOA e/ou de LOCAL. Além disso, por se tratar de um *corpus* de textos históricos, muitas das possibilidades apresentadas pelas diretivas do HAREM não foram empregadas, tais como, nomes de partidos políticos e de marcas de produtos dentro da categoria ABSTRAÇÃO, ou ainda, nomes de produções cinematográficas ou de programas de computador dentro da categoria OBRA.

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 VARIADO (que recobre elementos reconhecidos como ENs, mas que não se enquadram nas demais categorias) não foi frequentemente anotada, o que demonstra que as diretivas adotadas cobriram as ENs identificadas.

A nota AMB foi anotada em 636 ocorrências de ENs, o que representa 11,24% do conjunto de ENs identificadas. No que concerne às ocorrências de INITs, apresentamos, a seguir, os dados gerais de suas ocorrências divididas por categorias.

CATEGORIAS DE INITs	*TOTAL DE OCORRÊNCIAS*
TÍTULO	1516
LOCAL	844
CARGO	675
PESSOA	598
ORGANIZAÇÃO	595
TRATAMENTO	557
PARENTE	289
ACONTECIMENTO	266
COISA	196
VALOR	155
TEMPO	138
OBRA	41
ABSTRAÇÃO	31
VARIADO	22

Tabela 8 – Ocorrências de INITs por categoria

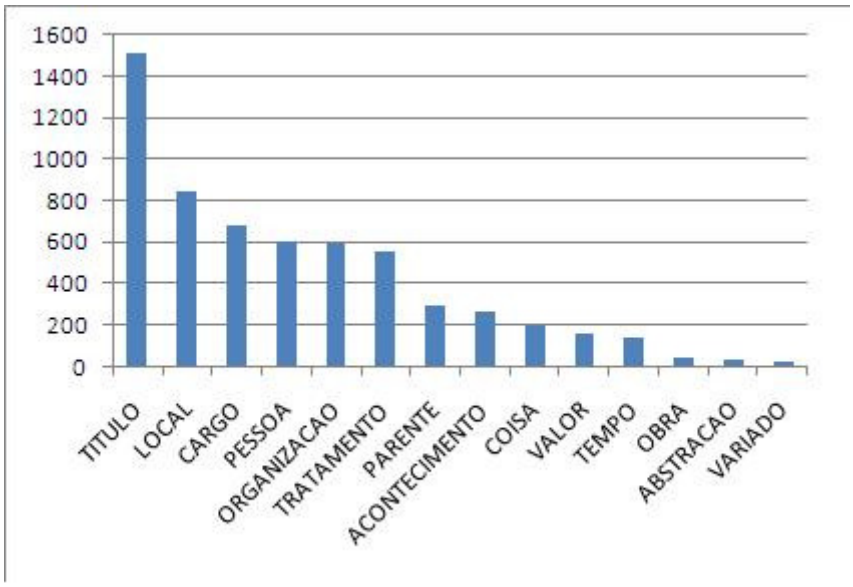

Gráfico 2 – Ocorrências de INITs por categoria

Em virtude da grande extensão das planilhas contendo as anotações morfológicas e semânticas realizadas neste trabalho, não incluímos essas planilhas nos anexos deste relatório, preferindo enviá-las em um documento separado, em formato XLS, denominado ABREVIATURAS FLEXOR.

Considerações Finais

Do que foi exposto até aqui, salientamos que a categorização do léxico de uma língua com fins à 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 e deixa entrever resultados promissores. O tratamento meticuloso de cada item e o conhecimento básico da história da língua foram importantes para auxiliar na 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 eletrônico do PHB deverá ser disponibilizado para novas pesquisas em Processamento de Linguagem Natural, o que torna crucial a classificação correta das ENs. É necessário dizer que com o término desta pesquisa será possível utilizar a fonte de informações sobre ENs (Dicionário de Flexor) para que seja implementada ao projeto do Dicionário Histórico do Português do Brasil (DHPB).

As planilhas, que contém as informações anotadas resultantes desta pesquisa, já foram fornecidas à co-orientadora do projeto, que está desenvolvendo um ambiente *Web*, para que o dicionário eletrônico de abreviaturas do PHB, que traz informações sobre ENs, possa ser acessado gratuitamente pela internet sob requisição para pesquisa.

Além de fornecer subsídios para futuros trabalhos na área de PLN, o dicionário eletrônico de abreviaturas do PHB lança-se como um recurso facilitador do trabalho de muitos pesquisadores afeitos ao trabalho de análise de textos históricos do Brasil, possibilitando uma maior facilidade na identificação de abreviaturas e, consequentemente, facilitando muitas pesquisas das diversas ciências humanas.

Esta pesquisa nos proporcionou, enquanto bolsista, o aprimoramento teórico e um primeiro contato prático dentro das áreas de pesquisas envolvidas, sobretudo, em relação à Lingüística Computacional e ao Processamento de Linguagem Natural. Possibilitou também reflexões acerca do funcionamento de parte do léxico da língua portuguesa e suscitou o interesse no trabalho de processamento computacional de línguas históricas.

Referências Bibliográficas e bibliografia

.ALUÍSIO, S. M.; PELIZZONI, J. M.; MARCHI, A. R.; OLIVEIRA, L. H.; MANENTI, R.; MARQUIVAFÁVEL, V. (2003). An account of the challenge of tagging a reference corpus of Brazilian Portuguese. In: PROPOR´2003, 2003, Faro. Lecture Notes on Artificial Intelligence. Proceedings of PROPOR´2003. 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.

BIDERMAN, M. T. C. *Teoria Lingüística:* Teoria Lexical e Lingüística Computacional. 2 ª. ed. São Paulo: Martins Fontes, 2001.

_____ *Um dicionário para o português do Brasil*. In: Seabra, M.C.. (Org.). O Léxico em Estudo. 1ª ed. Belo Horizonte: Faculdade de Letras da UFMG, 2006.

COSTA, R. F. Abreviaturas: simplificação ou complexidade da escrita? *Arquivo do Estado*, 15, 2006. Disponível em: http://www.historica.arquivoestado.sp.gov.br/materias/anteriores/edicao15/materia01/.

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

JUNIOR, A. C. *Criação de um ambiente para o processamento de córpus de Português Histórico*. 2008. 131f. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação – ICMC- USP, São Carlos.

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/).

SILVA, M. C. P. Para uma tipologia geral de obras lexicográficas In: **As Ciências do Léxico III.** Campo Grande / Porto Alegre: Editora da UFMS & Humanitas, 2007, v.3, p. 283-294.

VALE, Oto Araújo ; CÂNDIDO, Arnaldo ; MUNIZ, Marcelo ; BENGSTON, C. ; CUCATTO, Lívia ; ALMEIDA, G. ; BATISTA, A. ; PARREIRA, M. C. ; BIDERMAN, M.T. ; ALUÍSIO, S. M. . Building a large dictionary of abbreviations for named entity recognition in Portuguese historical corpora. In: Language Technology for Cultural Heritage Data (LaTech - 2008) LREC 2008, 2008, Marrackech. LREC 2008 Proceedings. Paris : ELRA, 2008. v. 1. p. 1-10.

Dicionários de língua portuguesa utilizados

BLUTEAU, R. *Diccionario da Lingua Portugueza*. Lisboa: Officina de Simão Thaddeo Ferreira, 1789. Disponível em: http://www.ieb.usp.br/online/dicionarios/Bluteau/formBuscaDicionarioPlChave.asp

DICIONÁRIO eletrônico HOUAISS da Língua Portuguesa. São Paulo: Objetiva, 2001. Versão 1.0. 1CD-ROM.

NOVO DICIONÁRIO eletrônico Aurélio séc XXI: o dicionário da língua portuguesa: dicionário eletrônico. Rio de Janeiro: Positivo, 2004. CD-ROM. Versão 5.0.

Ferramentas

PHILOLOGIC®. Projeto da *American and French Research on the Treasury of the French Language* (ARTFL) e do *Digital Library Development Center* (DLDC). University of Chicago - EUA, 2007. Programa gratuito. Disponível em: http://philologic.uchicago.edu/ / do DHPB: http://moodle.icmc.usp.br/philologic/

***UNITEX*®.** Projeto de Sébastien Paumier. *Université de Marne-la-Vallée*-França. Adaptação: *Unitex-Milenio.* Versão beta 2.0. Disponível em: http://moodle.icmc.usp.br/milenio/