Aspetos Linguísticos na *Gazeta de Lisboa*: Proposta de uma Análise Histórico-Informático-Lexical

Susana Fontes
Universidade de Trás-os-Montes e Alto Douro

RESUMO: Neste artigo, pretendemos apresentar o nosso projeto de doutoramento, que ambiciona constituir-se como mais um contributo para os estudos históricos da língua portuguesa, recorrendo à *Gazeta de Lisboa*, através de um *corpus* que integra dois blocos de texto (1715-16 e 1815), representativos de dois séculos diferentes. Neste sentido, procederemos a uma análise linguística comparativo-contrastiva destes dois períodos através de programas de análise automática de texto. Neste momento, é nosso intento apresentar os resultados referentes ao estudo informático-lexical de parte do *corpus* que nos propusemos estudar inicialmente, à qual atribuímos a designação de *subcorpus*, levado a cabo através do programa *NooJ*.
PALAVRAS-CHAVE: linguística de *corpus, Gazeta de Lisboa*, análise lexical, *NooJ*

ABSTRACT: *In this article, we intend to present our PhD project, which aims to establish itself as another contribution to the historical studies of the Portuguese language, using the newspaper Gazeta de Lisboa, through a corpus which includes two blocks of text (1715-16 and 1815) that represent two different centuries. Thus, we will carry out a comparative-contrastive linguistic analysis of these two periods through computer programs for automatic text analysis. Now, it is our intent to present the results concerning the computer-lexical study from part of the corpus that we initially proposed to study, which we called subcorpus, undertaken through the program NooJ.*
KEYWORDS: *corpus linguistics,Gazeta de Lisboa, lexical analysis, NooJ*

Gazeta de Lisboa: um olhar sobre a realidade circundante

A *Gazeta de Lisboa* surge em 10 de Agosto de 1715[1] como o periódico[2] mais duradouro da primeira metade do século XVIII, assumindo uma importância considerável ao permitir ao leitor português o contacto com o mundo da época:

> O aparecimento da *Gazeta de Lisboa*, em Agosto de 1715, é, sem dúvida, um acontecimento histórico cuja projecção tem sido subestimada ou analisada superficialmente. Portugal passa a dispor de um órgão de informação que põe o leitor português, até aí ignorante ou mal informado, em contacto com o grande mundo da época, por onde poderá seguir os movimentos mais variados de uma Europa em permanente transformação. (VIEIRA, 1991, p. 21).

Para compreendermos a sua importância e procedermos a uma análise criteriosa deste periódico, não podemos ignorar a sua especificidade, no sentido em que se trata de um jornal do Antigo Regime, que apresenta características diferentes das que actualmente presidem à construção de um jornal. Este assume-se como um veículo de informação com circulação restrita, que não era concebido para informar o grande público, como depois aconteceu com o "jornalismo de massas".

Esta publicação vai sofrer alterações nos diferentes títulos que apresenta ao longo da sua história. Depois de se assumir enquanto *Gazeta de Lisboa*, no seu segundo número, em 17 de Agosto de 1715, passou a intitular-se de *Gazeta de Lisboa Ocidental*, a partir de Janeiro de 1718, motivada pela divisão da cidade em parte oriental e ocidental, até Agosto de 1741, altura em que esta divisão deixou de existir.

O 1º ciclo de vida da *Gazeta* termina em 1760, com a morte de José Freire Monterroio Mascarenhas, sendo o privilégio, nesse momento, concedido à Secretaria de Estado da Repartição dos Negócios Estrangeiros e da Guerra, uma vez que existiria uma proximidade entre as funções deste órgão e o conteúdo do periódico, dominado por questões internacionais. Para além desta alteração (a passagem do privilégio de uma pessoa para uma secretaria do Estado),

---

1 Nesta data, é publicada com o título de *Notícias do Estado do Mundo*, sendo apenas nos números seguintes que recebe a denominação de *Gazeta de Lisboa.*

2 A periodicidade adquire nesta altura uma conceção ligeiramente diferente da que temos hoje: "periódica nesta altura é uma publicação que difunde notícias regularmente no tempo presente, mas fá-lo de forma repetitiva, instaurando uma duração e uma continuidade na leitura." (BELO, 1999, p. 626-7).

indicadora de uma centralização do privilégio, o título desta publicação muda, *Lisboa*, o que poderá ser entendido como uma forma de apresentar algo novo, com um rumo diferente do anterior.

De seguida, adotou outras designações, sendo que algumas refletem o cenário político em que se encontra o país como *Diário do Governo, Diário da Regência, Crónica Constitucional de Lisboa, Gazeta Oficial do Governo, Gazeta do Governo, Diário de Lisboa*.

Estruturado anualmente em forma de livro, este jornal oficial divulgava notícias sobre o governo, o país e o estrangeiro, tal como anunciava no frontispício. A capa apresenta-nos um dado importante como é o nome do redator, que surge pela primeira vez, quebrando a tradição do anonimato, normalmente característico da produção jornalística.

A *Gazeta de Lisboa*, tal como acontecia com outras publicações jornalísticas suas coetâneas, apresenta uma estrutura intermédia entre o livro e o jornal.

Com uma impressão semelhante à dos livros, a *Gazeta* conserva o seu aspeto, mas de formato pequeno, *in quarto*. Neste sentido, o formato de livro implicava a continuidade existente entre os diferentes números, o que nos permite, por um lado, inseri-la no género histórico. Esta continuidade era conseguida através de uma numeração e paginação contínuas. A numeração, feita em cada exemplar, e a paginação eram concebidas para o seu futuro formato de livro anual, onde apresentava, no início de cada ano, uma capa impressa a maiúsculas com o título de *Historia Annual Chronologica, e Politica do Mundo, e especialmente da Europa*[3]. Tendo por base a leitura de um anúncio[4] publicado na *Gazeta* em 1759, apercebemo-nos de que esta capa ou folha de rosto, onde constava o título referido, era vendida na oficina onde se imprimia o periódico, a fim de que os leitores pudessem encadernar a sua coleção anual.

Inserida neste esquema híbrido, a *Gazeta* apresenta, para além do formato próximo do livro, uma estrutura mais jornalística, como se percebe pela sua circulação também em folhetos.

---

3 Apresentava como título completo o seguinte: *Historia Annual Chronologica, e politica do Mundo, e especialmente da Europa onde se faz memoria dos nascimentos, despozorios, e morte de todos os Emperadores, Reys, Principes, e pessoas consideraveis pela sua qualidade, ou empregos; encontros, sitios de Praças, e Batalhas terrestres, e navaes; vistas, e jornadas de Principes, Tratados de Aliança, Tregoa e Paz, com todas as mais acções militares, civis, e negociações politicas, e sucessos mais dignos da attençaõ, e curiosidade.*

4 GL, 1759, nº 52

Este formato permitia-lhe circular de mão em mão, prática corrente na altura, para além do fenómeno de leitura em voz alta, que nos impede de avaliar com precisão o número dos seus leitores. A este nível, considera-se que o número de pessoas que leem ou têm contacto com a *Gazeta* e outros periódicos semelhantes é superior à sua tiragem[5], assinantes e compradores. Tal como acontecia com outras publicações europeias deste género, devem ter existido diferentes possibilidades de venda da *Gazeta* (por assinatura[6], a venda de volumes anuais ou de um único número, avulso), em diferentes locais (nos livreiros e nos locais onde era impressa).

No que concerne à sua estrutura, as notícias são precedidas de alguns dados que nos permitem localizá-las temporal e geograficamente: o nome da nação de proveniência é impresso em maiúsculas, seguindo-se, em letras mais pequenas, a data e o nome da capital ou cidade de origem. Por fim, surge o corpo da notícia, apresentando uma estrutura quase sem parágrafos, que ocupa toda a dimensão das páginas e um estilo que muitas vezes denuncia claramente uma tradução apressada e resumida ao essencial. Os anúncios, publicados no final da última página, surgem com um tipo de letra ainda mais reduzido e itálico, o que dificulta a sua leitura.

Dando continuidade à estrutura presente nas suas congéneres europeias, verificamos que grande parte do corpo da gazeta era ocupado com informações do estrangeiro[7], como mostra a carta de privilégio de 1715, notícias designadas de políticas, traduzidas e resumidas de gazetas europeias, trabalho que estaria a cargo do seu redator, José Freire Monterroio Mascarenhas, que ocupa este lugar até 1760. O longo período em que este se responsabilizou pela redação da *Gazeta* conduziu a uma identificação muito próxima entre a conceção deste

---

5 No caso da *Gazeta* existem registos sobre a sua tiragem a partir da década de 40, mas nada se encontrou sobre os seus assinantes.

6 O sistema de assinatura denota já uma preocupação com um público, que se pretendia fixar e fidelizar.

7 "O noticiário europeu da *Gazeta de Lisboa* pôde exercer e exerceu, com certeza, um importantíssimo papel, ainda por estudar, na actualização dos conceitos político-sociais e económicos das camadas populacionais até aí privadas de uma informação regular e completa. Anteriormente, só uma medíocre percentagem de personalidades, ligadas à máquina administrativa ou diplomática da corte, poderia beneficiar de informações válidas sobre o desenrolar dos acontecimentos além-fronteiras." (VIEIRA, 1991, p. 21)

jornal e a própria personalidade do seu redator, que explica a denominação com que terá ficado conhecida neste período[8], como a *Gazeta de Monterroio*. A parte final desta publicação, ainda antes dos anúncios, evidenciando uma tendência de aproximação geográfica, era ocupada pelo noticiário nacional. Este movimento centrípeto culmina com a produção de um noticiário nacional, que constituía uma parte reduzida deste periódico, marcado por uma vigilância mais acentuada comparativamente às notícias de âmbito internacional, o que se repercute em informação menos descritiva e abundante, e mais cautelosa.

O reduzido espaço disponível para estas notícias estava limitado pela periodicidade semanal que se impunha. No caso das notícias sobre o estrangeiro, parte predominante deste e de outros periódicos do género, as notícias, essencialmente políticas e militares, eram preparadas com tempo, uma vez que não se impunha um nível de atualidade tão elevado.

As notícias sobre a Corte, na capital, preenchiam maioritariamente este espaço reduzido, ainda que por vezes surgissem informações sobre outras localidades, obtidas através de correspondência. Por último, existia uma "secção" dedicada a anúncios, tendo sido precisamente na *Gazeta* que surgiu o primeiro anúncio comercial, designado de "aviso".

Apresentação do projeto de investigação

A *Gazeta de Lisboa* foi precisamente o periódico escolhido para a constituição do nosso *corpus* de trabalho, que se localiza temporalmente nos séculos XVIII (1715-1716) e XIX (1815), períodos marcados por alterações profundas ao nível económico, político e sócio-cultural, reflexos dos novos ideais que as Luzes introduziram em Portugal; e também das revoluções liberais, que agitaram profundamente o panorama político português, com consequências evidentes em todos os outros planos da vida nacional. O pensamento jornalístico

8 Este primeiro período de vida da *Gazeta* (1715-1760) foi trabalhado em teses académicas, de uma forma aprofundada, ultrapassando a vertente superficial com que este periódico tinha sido aflorado na historiografia jornalística. Referimo-nos às teses de mestrado e doutoramento de André Belo. A primeira intitulada de *As Gazetas e os Livros. A Gazeta de Lisboa e a Vulgarização do Impresso em Portugal (1715-1760)*, apresentada no Instituto de Ciências Sociais da Universidade de Lisboa em 1997 e a segunda, intitulada de *Nouvelles d'Ancien Régime. La Gazeta de Lisboa et l'information manuscrite au Portugal (1715-1760)*, apresentada na École des Hautes Etudes en Sciences Sociales em 2005. A tese de doutoramento de João Luís Lisboa é também uma referência nesta linha de investigação: *Mots (dits) écrits. Formes et valeurs de la diffusion des idées au 18ème siècle au Portugal*, apresentada no Instituto Universitário Europeu, Florença, em 1998.

destes séculos constituiu uma base importante para a história do jornalismo português, pois começam a surgir algumas preocupações prementes para o desenvolvimento do jornalismo enquanto área autónoma, com um discurso, preocupações e finalidades próprias.

No nosso trabalho de investigação, que corresponde à tese de doutoramento, pretendemos reconstruir/relembrar a história do jornalismo português desde a sua génese até ao século XIX, estabelecendo sempre uma base de comparação com o panorama europeu, não descurando as circunstâncias histórico-culturais que condicionaram este percurso.

Depois desta contextualização, iremos proceder à edição semi-diplomática do nosso *corpus*, ao que se segue uma análise comparativo-contrastiva entre a primeira parte do *corpus*, referente a 1715-1716, que corresponde ao momento do nascimento da *Gazeta de Lisboa* e a segunda parte, a de 1815, o que nos permite avaliar as principais alterações lexicais operadas neste jornal decorrido um século, para além de outras considerações linguísticas relevantes.

Este texto será, pela primeira vez, analisado sob uma perspetiva linguística e tendo por base um programa informático, *NooJ*, o que permitirá uma série de análises contrastivas e lexicais mais objetivas e rigorosas, reveladoras de uma aproximação cada vez mais evidente entre a linguística e a informática.

No final, será necessário proceder ao tratamento dos dados que recolhemos com as ferramentas linguísticas de forma a concluir determinados aspetos no âmbito da linguística e também das temáticas principais do jornalismo referente aos séculos XVIII e XIX.

Estudo estatístico-lexical da *Gazeta de Lisboa* (Agosto de 1715)

3.1 Importância da Informática na análise de textos

A abordagem lexical que pretendemos levar a cabo será facilitada pela utilização de um recurso informático de processamento automático de texto que nos permite obter resultados mais fiáveis e sistemáticos num curto espaço de tempo. O uso das novas tecnologias potenciou a execução de um conjunto de tarefas que facilitam o trabalho ao investigador que, ainda assim, continua a ser o condutor principal da sua investigação e o responsável pela leitura dos resultados facultados por estes programas.

Na nossa investigação escolhemos o *NooJ*, programa desenvolvido por Max Silberztein, que reconhece e trabalha mais de 100 formatos de texto. Este software permite-nos executar um conjunto de tarefas, das quais destacamos:

- a etiquetagem linguística do *corpus*;
- a elaboração de listas de formas a partir do lema, da classe ou subclasse, ou de outro traço morfológico;
- o estabelecimento de concordâncias tendo por base qualquer dado linguístico;
- a organização da listagem dos *Digrams;*
- a construção de dicionários e gramáticas flexionais, morfológicas ou sintáticas, necessárias para ultrapassar alguns problemas que não conseguem ser resolvidos pelos recursos linguísticos eletrónicos já existentes.

No caso do *NooJ*, referimo-nos aos dicionários eletrónicos[9] de grande qualidade e ampla cobertura produzidos pelo Laboratório de Engenharia Linguística, que constituem o sistema LABEL-LEX[10] (*LabEL*: www.LabEL.ist.utl.pt). Os léxicos desenvolvidos apresentam dois módulos: 1) LABEL-LEX sw, que contém mais de 1500000 formas flexionadas e o 2) LABEL-LEX-mw, formado por mais de 75000 unidades lexicais multipalavra.

3.2 Método de trabalho

Depois de apresentarmos o recurso informático escolhido para a nossa análise e de explicitarmos algumas das suas características e potencialidades, passamos a descrever o método de trabalho por nós usado.

O texto que iremos trabalhar neste momento é uma espécie de *subcorpus* do *corpus*[11] *a que nos propusemos trabalhar no doutoramento. Trata-se apenas do mês de Agosto da Gazeta de Lisboa*, que constitui um total de 24 páginas. É nosso propósito, como já foi referido, começar pela edição desta parte do periódico, que passaremos a denominar como *GL*-08-1715, seguindo-se um estudo lexical, para o qual contamos com o precioso auxílio do programa *NooJ*. Iniciámos o processo da edição com a transcrição integral do mês de Agosto de 1715 da *Gazeta de Lisboa*, visto que necessitávamos de uma versão do

9 "Um dicionário electrónico é um léxico computacional concebido para ser usado, sem intervenção humana, por programas informáticos em diversas operações de processamento de linguagem natural." (RANCHHOD, 2001, p. 14).

10 Tivemos acesso a estes recursos linguísticos através de um protocolo estabelecido entre o Centro de Estudos em Letras e o Laboratório de Engenharia Linguística, sendo também de destacar o importante contributo de José Paulo Tavares na adaptação destes recursos para o formato *NooJ*.

11 Sardinha (cf. 2004, p. 20-22) apresenta os principais tipos de corpora tendo em conta os seguintes critérios: o modo (falado ou escrito), o tempo (sincrónico, diacrónico, contemporâneo ou histórico), seleção (de amostragem, monitor, dinâmico ou orgânico, estático, equilibrado), conteúdo (especializado, regional ou dialectal ou multilingue), autoria (de aprendiz ou de língua nativa).

documento em Word para depois o inserir no programa *NooJ* e uma vez que a possibilidade de conversão das imagens em texto através de um programa de OCR[12] não se tornou possível devido a um conjunto de gralhas que resultaram deste processo. Esta edição teve como critério a aproximação rigorosa ao texto original, apresentando como única alteração o desdobramento da abreviatura q, na forma do pronome/conjunção que e da conjunção porque, que passaram a ser registados como q*ue* e porq*ue*[13].

Depois de verificado o texto, e terminado este processo de edição, procedemos às alterações necessárias para que o documento ficasse sem qualquer tipo de formatação, processo com que nos tínhamos preocupado no momento da edição do texto, como era o caso dos parágrafos, quebras de linha, quebras de página, itálicos, negritos, tipos e tamanhos de letra diferentes, etc.

De seguida, executámos o programa *NooJ*, que iniciou o processo de anotação automática, tendo por base os léxicos do *LabEL*, adaptados a esse mesmo formato. Deste trabalho, resultaram os seguintes dados:

	GL-08-1715
Unidades de texto (parágrafo)	146
Nº de carateres	78032
Nº de ocorrências/tokens	15778
Nº de formas diferentes	3218
Formas desconhecidas	1319
Anotações	33818

Tabela 1: Dados gerais da *GL*-08-1715 obtidos com os recursos do *LabEL*

A observação desta tabela permitiu-nos confirmar um dado que já havíamos antecipado, que se prende com o número elevado de formas desconhecidas, como era de esperar, devido à diferente forma gráfica de muitas palavras, justificável por estarmos a utilizar os recursos linguísticos do *LabEL*, que se centram no léxico atual, e que por isso não reconhecem muitas formas diferentes presentes num texto do século XVIII.

12 Esta sigla refere-se à tecnologia de Reconhecimento Ótico de Carateres.

13 A preocupação relativamente ao desdobramento da abreviatura através do itálico está unicamente ligada ao processo de edição, e nada tem a ver com o programa, uma vez que o *NooJ* não reconhece este tipo de formatação.

Se, por um lado, as formas desconhecidas são indicadoras de um baixo nível de cobertura dos recursos linguísticos existentes, apenas 59%, tendo em conta que não reconhecem 40,98%, equivalente às 1319 formas, elas podem servir como forma de enriquecimento dos recursos linguísticos, uma vez que exigem a construção de novos recursos como poderá ser o caso de novos dicionários ou gramáticas.

Antes desse processo, será necessário detetar o motivo desta falha ao nível da cobertura dos recursos disponíveis, os do *LabEL*, e claramente concluímos tratar-se de uma questão de grafia, que separa estes séculos.

3.3 Formas desconhecidas

As principais diferenças gráficas presentes neste *corpus* são as seguintes:

1) as duplas consoantes, como é o caso de *abbade*, *difficuldade*, *elle*, *approvado*, *opposição*, *applicado*.
2) os topónimos e antropónimos com grafias diferentes das atuais, dos quais destacamos *Rebinsky*, *Dolhorouki*, *Leverpool*, *Mattheos*, *Joseph*
3) diferenças na acentuação:

3.1) o ditongo nasal -ão, que surge, alternadamente, com a forma atual -ão ou com a forma -aõ, visível nos exemplos que se seguem:

- nos nomes *accusação/accusaçaõ*; *treyçaõ/treyção*; *grão/graõ*; *embarcação/embarcaçaõ*, *guarnição/guarniçaõ*; *opposição/opposiçaõ*; *satisfação/satisfaçaõ*; *condição/condiçaõ*;
- nos verbos, onde esta oscilação da grafia é visível nas terceiras pessoas do plural, das quais destaco o caso do pretérito perfeito do Modo Indicativo (*mandaraõ/mandáraõ/mandàraõ*, *fizeraõ/fizerão*; *foraõ/forão*; *obrigáraõ/obrigàraõ*; *tiverão/tiveraõ*, *voltáraõ/voltàraõ*); do pretérito imperfeito (*devião/deviaõ*, *havião/haviaõ*) e também do Futuro (*daráõ/daraõ*, *mandaraõ/mandaráõ*, *seráõ/seraõ*).

3.2) o plural do ditongo nasal -ão, que nós hoje realizamos como -ões, apresenta no corpus duas formas diferentes. São elas em –oens e em -oẽs, como fica claro pelos exemplos que se seguem: batalhoens/batalhoẽs, declaraçoẽs/declaraçoens, embarcaçoẽs/embarcaçoens, esquadroẽs/esquadroens, milhoẽs/milhoens.

3.3) a omissão do acento agudo na vogal tónica, como é o caso da terceira pessoa do singular do verbo haver no Presente do Modo Indicativo, ha, dos nomes sabbado e secretario, do advérbio ja, e do adjetivo necessario.

3.4) a alternância entre o acento agudo ou grave e o circunflexo, como se nota nas diferentes formas que adotam as palavras: està/estâ, jà/jâ, sómente/sômente.

3.5) o recurso ao til para atribuir nasalidade, em substituição do -m ou n-, como é notório nas formas dos artigos hũ e hũa, que coexistem com hum e huma, e também em outras palavras como impaciẽcia, frequentemẽte, Parlamẽto, tambẽ/tambem.

4) a junção do clítico à forma verbal sem o recurso ao hífen, como por exemplo concedendolhe, manterse, pedindolhe, porse, entregarseha, concederseha, etc.

Apesar de haver muitas outras diferenças gráficas, consideramos importante fazer este levantamento pois estas são as mais frequentes, o que nos permite, numa segunda fase do trabalho, construir gramáticas morfológicas, capazes de reconhecer estas diferenças e de classificar cada uma destas entradas devidamente.

Antes de apresentarmos os vários grafos construídos para o efeito, julgamos necessário lembrar que o NooJ trabalha com a tecnologia de estados finitos. Os grafos correspondem precisamente a FST (finite-state transducer) que apresentam algumas potencialidades ao nível do tratamento automático de textos escritos. Servem para construir dicionários eletrónicos e gramáticas. Estas podem ser criadas para resolver variados problemas ao nível ortográfico, morfológico, sintático, o que explica a existência de i) gramáticas flexionais e derivacionais (ficheiros com a extensão .NOF), ii) lexicais, ortográficas, morfológicas ou terminológicas (ficheiros com a extensão .NOM), iii) sintáticas ou semânticas (ficheiros com a extensão .NOG).

No momento da construção dos grafos, o programa dá-nos apenas um estado inicial, simbolizado por  e o estado final, simbolizado por  , sendo depois o investigador que acrescenta os dados que lhe interessa. "The text sequences are described in the input part of the FST; the corresponding results are described in the output part of the FST." (SILBERZTEIN, 2008, p. 13).

3.4 Construção de gramáticas lexicais (ou morfológicas)

Passamos, desta forma, a enumerar as várias gramáticas morfológicas criadas e as operações que elas possibilitam:

Para todas as formas com consoante dupla intermédia, usámos o grafo seguinte, que relaciona a grafia própria do século XVIII com a forma atual, de consoante simples, e recupera as respetivas informações da entrada do dicionário. No nosso *corpus* só temos a consoante dupla, o que nos permite o reconhecimento das formas *abbade*, *ella*, *accuso*, por exemplo; no entanto este grafo, que deve ser aplicado em baixa prioridade, resolveria também o problema de uma vogal dupla.

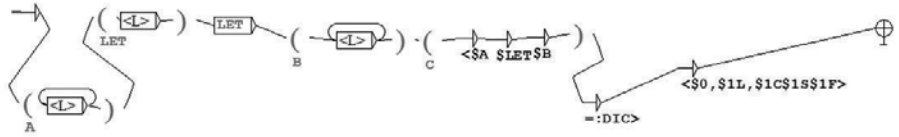

Ilustração 1**:** FST de reconhecimento de formas com consoante dupla interior

No caso do ditongo nasal -ão, que é representado graficamente no nosso *corpus* como -aõ, criámos uma gramática que nos permite associar o ditongo -aõ ao atual -ão.

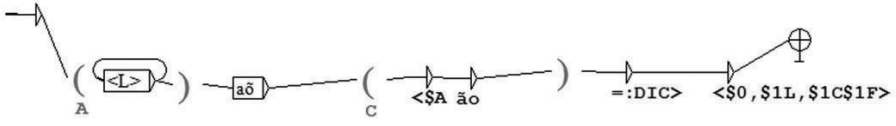

Ilustração 2: FST de alteração de terminação nasal

Para reconhecimento de nomes próprios, construímos uma gramática morfológica, que deve ser aplicada em baixa prioridade, para atribuir a etiqueta de nome próprio às palavras dadas como desconhecidas pelos recursos aplicados na análise linguística que comecem com letra maiúscula ou sejam todas escritas com letra maiúscula.

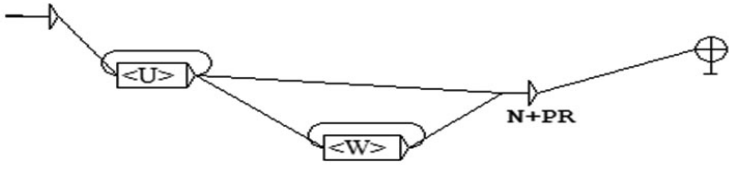

Ilustração 3: FST heurístico de etiquetagem de nomes próprios

No caso das palavras cuja grafia ainda não contempla a acentuação na vogal tónica, como acontecia com *ha* ou *ja*, usámos a próxima gramática morfológica que, quando aplicada, procedimento que deve ser feito em baixa prioridade, identifica formas em que o *-a*, apresentado no texto do século XVIII, corresponde a um *-á* na grafia atual, quer este grafema esteja em posição ini-

cial, medial ou final. Este grafo recupera essa entrada desde que encontre essa correspondência no dicionário[14].

Construímos esta gramática para este caso específico da vogal -a, que nos pareceu ser a mais frequente, no entanto o mesmo procedimento poderia ser adotado para outras vogais pretendidas, constituindo uma gramática por cada substituição.

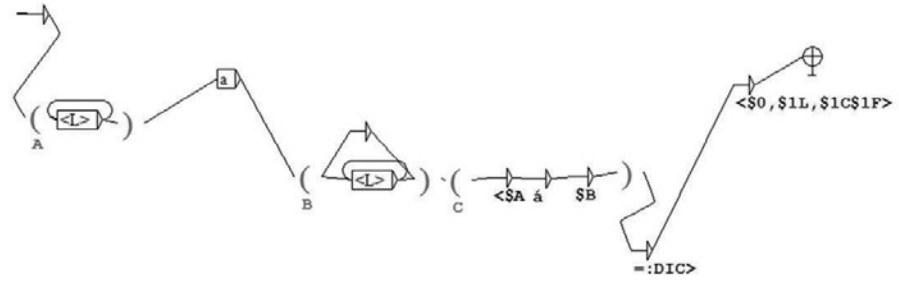

Ilustração 4: FST de reconhecimento de formas grafadas com vogal sem acento correspondentes a vogal acentuada

Para separação e classificação de formas que compreendem a duas (ilustração 5) ou três (ilustração 6) palavras não separadas, desde que cada uma exista no dicionário, utilizámos estas gramáticas morfológicas, aplicadas também elas em baixa prioridade.

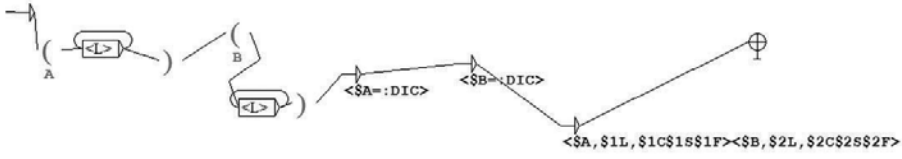

Ilustração 5: FST de identificação de complex tokenization[15] (duas palavras)

14 Persistem, ainda, alguns problemas decorrentes da ambiguidade, como é o caso da forma agua, que o programa não altera ou reconhece como água, devido à existência do verbo aguar.

15 A *tokenization* ou itemização "consiste na separação das unidades ortográficas, normalmente por meio da inserção de espaços em branco ou quebras de linha entre elas." (SARDINHA, 2004, p. 128).

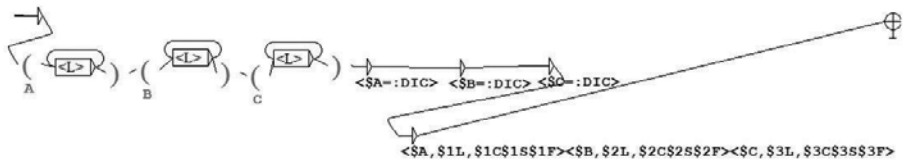

Ilustração 6: FST de identificação de complex tokenization (três palavras)

3.5 Construção de dicionários eletrónicos

Depois de constituídas estas gramáticas, faltavam ainda resolver várias palavras desconhecidas que não se inseriam nestes seis casos, o que nos conduziu à construção de um dicionário, que intitulamos de dicionário da gazeta setecentista, e um dicionário de abreviaturas, ao qual atribuímos este mesmo nome.

A criação de dicionários eletrónicos no *NooJ,* enquanto ficheiros de extensão *NOD,* implicou a ação de etiquetagem[16] das diferentes formas que foram classificadas como desconhecidas, o que exigiu a associação da palavra à sua categoria e propriedades morfossintáticas. As etiquetas podem ser simples quando apresentam apenas uma informação, como é o caso da classe gramatical, ou podem ser complexas, quando a etiqueta contempla vários dados (para além da classe, o lema, número, género, etc). Uma das características destas etiquetas prende-se com a sua reduzida extensão, e daí o recurso a abreviaturas ou códigos, de forma a agilizar o processo de etiquetagem. O investigador terá aqui um papel decisivo na escolha da informação que inclui em cada entrada, opções estas que estarão condicionadas pelo tipo de abordagem que pretende levar a cabo.

Neste momento da investigação, e apesar da existência de diferentes tipos de etiquetagem, centramos os nossos esforços em informações de âmbito morfossintático, que juntamente com a lematização nos simplificam o tratamento dos dados e facilitam a análise do *corpus*.

As entradas do nosso dicionário são compostas por um mínimo de três elementos, não contabilizando as vírgulas: forma, lema, classe gramatical (POS), podendo apresentar também a subclasse+os atributos morfológicos.

Por exemplo:

Academicos,Académico,N+m+p

Academicos,Académico,A+m+p

16 Segundo Costa (2001, p. 38), "[…] anotar um corpus significa associar informação linguística a segmentos de texto, recorrendo para o efeito a um conjunto de símbolos, as etiquetas, por forma a identificá-los, com vista ao seu tratamento automático. Esta operação é designada de etiquetagem, constituindo o produto final um corpus anotado."

Estas palavras são catalogadas como desconhecidas pelo *NooJ* porque apresentam grafias diferentes das atuais. Desta forma, procedemos à elaboração das entradas como se tratasse de uma variante ortográfica, para que o computador reconhecesse as duas como uma única forma, ou seja, se pedíssemos ao programa para selecionar todas as ocorrências da palavra académico, automaticamente ser-nos-ia dada a forma acentuada e a não acentuada. Em relação a esta entrada, julgamos também pertinente chamar atenção para um problema recorrente em análise automática de texto, a ambiguidade.

Quando construímos um dicionário, temos que prever todas as classificações associadas às várias palavras, o que se repercute num aumento das ambiguidades lexicais[17], que podem ser resolvidas através da constituição de gramáticas e da sua aplicação com diferentes graus de prioridade, ainda que a resolução exaustiva das ambiguidades lexicais em *corpus* extensos seja muito difícil uma vez que a desambiguação total depende, muitas vezes, de uma complexa análise sintática.

Não sendo nosso propósito, neste momento, refletir sobre as principais formas de desambiguação, passamos a apresentar os dicionários por nós criados, que, juntamente com os grafos, nos permitiram resolver o problema do número elevado de formas desconhecidas.

Ilustração 7: Dicionário da gazeta setecentista

17 "Lexical ambiguity: When a word is associated with different sets of properties, i.e. different syntactic or distributional information, we must duplicate the word in the dictionary. The corresponding word form will be processed as ambiguous." (SILBERZTEIN, 2008, p. 79).

Ilustração 8: Dicionário de abreviaturas

3.6 Aplicação dos novos recursos

No sentido de evitar um conflito entre os vários recursos criados, a sua aplicação foi feita tendo em conta diferentes níveis de prioridade, procedimento este que passo a apresentar por ordem decrescente, seguindo-se ao nome do recurso os níveis de prioridade (high, regular e low) em que foram aplicados:
-LabEL mw - H1
-LabEL sw - regular
- Diconário da gazeta setecentista - L1
- Dicionário de abreviaturas - L2
- consoantes duplas – L3
-nome próprio – L4
-tokenization- L5
-vogais acentuadas – L5
- tokenization 3 – L6
- terminação nasal - regular

Desta forma, construídos os recursos linguísticos para este *corpus* específico do século XVIII, procedeu-se, mais uma vez, ao processo de anotação automática, o qual, lembramos, não contempla nenhuma forma de desambiguação.

	GL -08-1715
Unidades de texto (parágrafo)	146
Nº de carateres	78032
Nº de ocorrências	15778
Nº de formas diferentes	3218
Formas desconhecidas	0
Anotações	48777

Tabela 2: Dados gerais da *GL*-08-1715 obtidos com os recursos do LabEL e os novos recursos linguísticos

A leitura desta tabela levou-nos a destacar dois elementos: por um lado, o valor das formas desconhecidas, que reduziu drasticamente devido à aplicação dos novos recursos eletrónicos; por outro lado, o valor elevado das anotações, que remete para a noção de ambiguidade, como se percebe pela relação direta entre as anotações ("text annotation structure") e a taxa de ambiguação, que neste caso será também elevada. Apesar desta verificação, não iremos, neste momento, proceder à desambiguação do nosso *corpus*, ato que seria obrigatório caso quiséssemos proceder a um estudo aturado da utilização de determinada classe de palavras.

3.7 Tokens mais frequentes

Nº de ordem	Forma	Ocorrência	Nº de ordem	Forma	Ocorrência
1	de	922	21	Mag	55
2	que	485	22	sua	53
3	a	438	23	À	53
4	o	330	24	seu	53
5	se	317	25	huma	45
6	da	205	26	O	45
7	para	178	27	das	43
8	do	171	28	tropas	41
9	os	170	29	Conde	36
10	em	166	30	Tem	35
11	com	150	31	Corte	34
12	as	120	32	Cidade	32
13	dos	96	33	suas	32
14	S.	89	34	seus	32
15	na	81	35	sobre	31
16	por	69	36	Julho	31
17	ao	67	37	ha	31
18	no	63	38	grande	31
19	hum	61	39	não	29
20	Havia	57	40	aos	29

Tabela 3: Lista dos 40 tokens mais frequentes

A análise a esta tabela permite-nos rapidamente confirmar que a maioria destas formas corresponde a palavras gramaticais ou funcionais. As preposições ocupam neste *corpus* um lugar de destaque, lideradas por *de*, ao que se seguem muitas outras formas, como é o caso do *em, com, por, sobre*, a contração da preposição com artigos (*da, do, dos, na, no, das, à, aos*) e uma forma ambígua, *a* (que pode ser caracterizada como artigo, preposição, nome ou pronome). Em segundo lugar, como acontece na generalidade dos corpora, surge-nos o *que,* uma das formas mais ambíguas em Português, que ocupa precisamente o primeiro lugar das formas ambíguas. Os verbos têm uma representatividade muito reduzida, com três ocorrências, duas formas do verbo haver e uma do verbo ter. Outra forma que merece uma referência é precisamente o determinante/pronome possessivo *seu*, flexionado no masculino, feminino, singular e plural, e representado na abreviatura S., que desdobrámos no dicionário como Sua, e que ocupa o 14º lugar das primeiras 40 formas do *corpus*. Por sua vez, os adjetivos são representados apenas por uma forma, *grande*, que ocupa uma das últimas posições, valor que está de acordo com a neutralidade reclamada pelo estilo informativo do texto jornalístico, que não se coaduna com um número elevado de adjetivos.

Depois de analisarmos estas classes, resta-nos tentar perceber a importância que têm os nomes destacados a cor diferente na tabela: *Mag*. (abreviatura de Magestade, que juntamente com a abreviatura *S.* constitui a expressão Sua Magestade), *tropas*, *conde*, *corte*, *cidade* e *Julho*. Sabemos que um dos critérios de noticiabilidade, que transformam um facto em notícia, foi e continua a ser a referência a pessoas de elite ou dados sobre países importantes no contexto internacional. O redator selecionava os acontecimentos dignos de registo com base na notoriedade dos seus intervenientes, produzindo uma história das elites, onde figuram as notícias sobre os atores sociais dominantes, como acontece hoje em dia. Os jornais surgiram para responder a uma necessidade de informação e satisfazer a curiosidade humana, daí a referência a informações políticas, religiosas, militares ao nível nacional e internacional, o interesse pelos povos e culturas distantes, pelo movimento portuário e a forte curiosidade pelo que se passava na Corte. No fundo, se compararmos esta situação com a atualidade, verificamos que esta necessidade de informação em relação às novidades que se sucedem na corte não foi só uma característica da centúria setecentista, mas continua a alimentar várias revistas "cor-de-rosa". Como se percebe, o critério das figuras famosas como protagonistas das notícias é intemporal e está profundamente inscrito na nossa cultura.

Retomando o grupo dos cinco nomes que se destacam entre as 40 formas mais frequentes, verificamos que a *Corte*, *Sua Magestade* e o *Conde* são figu-

ras nucleares neste ambiente de elite que atrai jornalistas e público. Por outro lado, temos a referência às *tropas*, que denota uma preocupação evidente por informações militares. Por último, os nomes *Cidade* e *Julho* são reveladores de uma categorização da informação em função de um espaço geográfico e por isso a noção de Cidade, em detrimento do mundo rural que não tem lugar num jornal da época, e a referência a um tempo específico, Julho, que era normalmente antecedido do dia e local para que o leitor pudesse situar as notícias num espaço e tempo determinados. Esta organização das notícias tinha por base uma referência direta ao tempo e espaço (país e cidade), num movimento de aproximação geográfica, com as notícias sobre o território nacional limitadas sempre à última parte do periódico, o que permitia ao leitor um acesso à informação de uma forma mais organizada, evitando uma possível sensação de caos.

Os periódicos reproduziam o ambiente da corte, apresentando um discurso que estaria, de certa forma, condicionado por esta relação estreita, uma vez que muitas notícias eram oficiais, recolhidas no próprio paço e não eram, obviamente, sujeitas à censura por parte do redator, que apenas registava a opinião oficial.

As palavras selecionadas podem ser caracterizadas pela sua frequência elevada neste *corpus*, o que nos permite atribuir-lhes a classificação de palavras-tema, cuja análise poderá ser muito útil para o investigador "caracterizar áreas temático semânticas típicas" (GENOUVRIER e PEYTARD, s/d, p. 317-318).

A análise das formas mais frequentes foi seguida de uma contagem das palavras de frequência 1, o que nos permitiu concluir acerca da variedade vocabular do texto. Ainda que se trate de um *corpus* de reduzida extensão, verificamos que, num total de 3218 formas diferentes, 1998 tokens ocorrem uma única vez, alcançando uma percentagem de 62%, o que nos permite confirmar que estamos perante um texto com um vocabulário muito variado[18].

3.8 Classes de palavras

Num segundo momento interessava-nos conhecer a frequência com que cada classe surgia no nosso *corpus*, o que nos levou à pesquisa de cada uma delas através da funcionalidade "locate", onde inserimos o output que pretendíamos. Desta forma, no caso do nome, por exemplo, inserimos <N> na opção "NooJ regular expression", o que nos permitiu visualizar todos os nomes que surgem no texto, inseridos no seu contexto, o que será muito útil se pretendermos estabelecer concordâncias. Este procedimento foi adotado para todas as

18 A variedade vocabular será tanto maior quanto maior for o número de palavras a ocorrer apenas uma vez e menor o número das que ocorrem duas ou mais vezes.

classes de palavras, sendo também possível, com esta mesma funcionalidade, ordená-los por ordem alfabética, estratégia que poderá facilitar um trabalho de pesquisa posterior. Depois de um número total de ocorrências, interessava-nos também conhecer o número de formas diferentes que cada classe apresenta. Este propósito foi conseguido através da opção "1 example per match", que se tratou de uma forma de limitar a nossa pesquisa.

Depois de organizados estes dados, surgem os seguintes resultados, que apresentamos na tabela:

Totais de frequências				
Classe gramatical	Ocorrências	Média	Formas diferentes	Média
Nomes	6108	31,03%	1870	43,25%
Adjetivos	1348	6,84%	645	14,92%
Verbos	3446	17,50%	1271	29,40%
Determinantes	2026	10,29%	128	2,96%
Pronomes	2909	14,78%	149	3,44%
Preposições	2118	10,76%	42	0,97%
Advérbios	512	2,60%	132	3,05%
Conjunções	1052	5,34%	63	1,45%
Interjeições	162	0,82%	23	0,53%
Totais	19681	100%	4323	100%

Tabela 4: As Classes de palavras e sua distribuição percentual na *GL*-08-1715

Como podemos verificar pela análise desta tabela, os nomes ocupam claramente uma posição de destaque[19], com uma percentagem de 31,03%, aos quais se seguem os verbos, com 17,50%, e os pronomes com 14,78%. As interjeições ocupam o último lugar, sendo que, se procedêssemos a uma total desambiguação do *corpus*, o valor da sua percentagem reduziria drasticamente.

A coluna das formas diferentes permite-nos constatar uma alteração relativamente às posições cimeiras ocupadas pelas classes de palavras. Se o nome continua a liderar, ainda que agora mais próximo do verbo, que apresenta uma ampla possibilidade de flexão, o terceiro lugar passa a ser ocupado pelo adjetivo, seguido, com uma percentagem muito inferior, pelo pronome. A este nível, os pronomes, juntamente com as conjunções, preposições, apresentam uma grande diferença entre a frequência das ocorrências e das formas diferentes, reveladora de um número reduzido de formas que estas têm na nossa língua. Esta é

19 A visão da linguagem como sistema probabilístico, em que assenta a linguística de corpus, revela que ao nível da análise morfossintática os nomes surgem com mais frequência do que qualquer outra categoria gramatical. (cf. SARDINHA, 2004, p. 31).

precisamente uma característica das palavras gramaticais (categoremáticas e morfemáticas, segundo Bechara, 2002, p. 112), uma vez que elas existem em número finito. Contrariamente a esta situação, as palavras plenas ou lexicais (também designadas de lexemáticas por Bechara, 2002, p. 112) existem em número potencialmente ilimitado e são também elas as que se encontram mais expostas à mudança diacrónica, quer na forma quer no seu significado, apresentando uma percentagem elevada neste *corpus*, de 57,97%. As palavras gramaticais ou funcionais, que se caracterizam por ser mais estáveis ao longo do desenvolvimento histórico da língua, surgem com uma percentagem mais reduzida, ocupando um total de 42,03% das ocorrências.

Depois de fazermos a procura das ocorrências de todas as classes e das formas diferentes em que surgem no *corpus*, concentrar-nos-emos apenas na classe destacada, o nome. Neste sentido, iniciamos a pesquisa das ocorrências pelos nomes próprios, por considerarmos que são várias as referências a antropónimos e topónimos presentes neste texto. Lembramos que esta pesquisa só se tornou possível devido à criação de um grafo que permitiu o reconhecimento dos nomes próprios, visto que a maioria deles foi classificada como desconhecida. Desta forma, a pesquisa de todos os nomes próprios através da aplicação Locate <N+PR> teve como resultado 638 ocorrências (*Gabel, Almeyda, Diniz, Joseph, Suecia, Hessen Castel, Rugen*); no entanto é preciso não esquecer que outros nomes próprios, que não apresentaram diferenças gráficas relativamente às formas atuais, já tinham sido classificados pelos dicionários eletrónicos do LabEL, sendo que, para o traço semântico +Humano <N+Hum>, encontramos 118 ocorrências de nomes próprios (*Caetano, Lourenço, Rocha, Botelho, Gaspar*) e para o traço +Topónimo <N+Top> 284 (*Turquia, Viena, Europa, Veneza, Alexandria, Londres, Inglaterra*), perfazendo um total de 1040 nomes próprios.

Estes dados permitem-nos confirmar a importância de uma categorização espacial, visível nos vários topónimos que inundam este *corpus*, aliada a uma referência direta aos muitos protagonistas (antropónimos) que celebrizaram os vários acontecimentos.

Os nomes comuns remetem para dois campos temáticos principais: o religioso, visível nos exemplos que se seguem *altar, capella, padre, conego, convento, religioso, sacramento, vaticano* e o militar, como o comprova a proliferação de vocábulos relacionados com esta área: *batalha, armada, tropa, exercito, conquista, canhaõ, guerra, hostilidade, inimigo, morte, soldado*. Paralelamente, é de destacar o número significativo de títulos usados para qualificar os protagonistas das notícias, forma de mostrar o papel determinante das elites

nos jornais: *magestade, general, duque, cõde, marichal, marquez, procurador, governador*, etc.

Considerações finais

Depois de uma breve apresentação do periódico que marcou o panorama jornalístico português, procedemos à exploração do nosso *corpus*, para o que decidimos aproveitar as potencialidades dos programas informáticos de tratamento automático de texto. O uso destas ferramentas em linguística torna-se imprescindível para conseguirmos analisar quantitativa e qualitativamente determinados dados linguísticos que, desta forma, serão facilmente classificados de maneira eletrónica. Neste corpus específico, os principais problemas que surgiram prenderam-se com as diferenças gráficas, próprias de um texto do século XVIII, o que nos conduziu à criação de gramáticas e dicionários, que nos permitiram a classificação de todas as formas do texto. A construção deste novos recursos eletrónicos atrasou, por um lado, a análise do nosso texto, no entanto consideramos que o tempo e o trabalho dispendidos nesta atividade serão compensados em todas as análises que podem ser efetuadas em textos da mesma centúria. Neste sentido, percebemos que o trabalho facilitado pelas ferramentas informáticas necessita, em grande escala, de ser complementado por uma intervenção humana crítica ao nível da criação de novos recursos, da resolução de vários problemas, que permitem aperfeiçoar os sistemas existentes, e ao nível da posterior reflexão sobre os resultados obtidos.

Referências bibliográficas

BELO, André. *As Gazetas e os Livros. A Gazeta de Lisboa e a Vulgarização do Impresso em Portugal (1715-1760).* Lisboa: Imprensa de Ciências Sociais, 2001.

BELO, André. *Nouvelles d'Ancien Régime. La Gazeta de Lisboa et l'information manuscrite au Portugal (1715-1760)*. Paris: École des Hautes Etudes en Sciences Sociales, 2005.

COSTA, Maria Rute Vilhena. *Pressupostos teóricos e metodológicos para a extracção automática de unidades terminológicas multilexémicas*. Dissertação de Doutoramento. Lisboa: FCSH, 2001.

GENOUVRIER, Emile, e Peytard, Jean. *Linguística e Ensino do Português*. Coimbra: Livraria Almedina, s/d.

GUIRAUD, Pierre. *Les Caractères Statistiques du Vocabulaire. Essai de Méthodologie*. Paris: P.U.F, 1960.

LISBOA, João Luís. *Mots (dits) écrits. Formes et valeurs de la diffusion des idées au 18ème siècle au Portugal*. Tese de Doutoramento. Florença: Instituto Universitário Europeu, 1998.

MACHADO, José Barbosa. *Tratado de Confissom (1489) Edição semi-diplomática, Estudo histórico, informático-lingüístico e glossário*. Dissertação de Doutoramento. Vila Real: UTAD, 2002.

RANCHHOD, E. O uso de dicionários e de autómatos finitos na representação lexical das línguas naturais". In.: RANCHHOD, E. (org.) *Tratamento das línguas por computador: uma introdução à linguística computacional e suas aplicações*. Lisboa: Caminho, 2001, p. 13-47.

SARDINHA, Tony Berber. *Lingüística de Corpus*. São Paulo: Manole, 2004.

SILBERZTEIN, Max. *NooJ v2 Manual*, www.nooj4nlp-net, 2008

TAVARES, José Paulo da Costa. *Pressupostos teóricos e metodológicos para o estabelecimento e exploração de um corpus paralelo Latino-Português*. Dissertação de Mestrado. UTAD: Vila Real, 2006.

VIEIRA, Júlio. *O jornalismo setecentista. A Inglaterra e a Gazeta de Lisboa (1715-1720)*. Lisboa: Palas Editores, Lda, 2001.