Estudo Lexical: *Os Canibais* de Álvaro do Carvalhal

Carla Sofia Lima Barreira Araújo[a]
[a]**Instituto Politécnico de Bragança - Escola Superior de Educação**
carla.araujo@ipb.pt

Abstract

This study aims to examine some of the quantitative and qualitative data of the tale "Os canibais" of Álvaro do Carvalhal, using Nooj, computer program lexical analysis. We will prepare a lexical study of the tale, based on statistical analysis of words theme, to define possible thematic fields, making possible the identification of themes. The Nooj is a linguistic development environment that allows, on the one hand, construct formal descriptions (dictionaries and grammars) for broad coverage of natural languages and on the other, these same descriptions apply to large texts with high efficiency. Because of its potential and free access (is available online in www.nooj4nlp.net), the Nooj presents itself as a work tool accessible to any user, since it is not necessary to have programming knowledge to produce efficient resources or develop proficient research. Lexicometry, from the automated processing of corpora, is particularly useful to those interested in the study of textual production and opens new perspectives for research that go beyond the ordinary empiricism often subject to the arbitrariness of the subjective views of the observer. The Nooj can be seen as a proposal for a didactic approach, within the new methodologies that language teaching includes. Undoubtedly, the didactic potential of Nooj are practically limitless. Its level of proficiency depends on the teacher creativity and student curiosity.

Keywords: *Didactics; Statistical Analysis-lexical; Nooj; Words theme; Thematic fields.*

Resumo

Este trabalho pretende analisar alguns dos dados quantitativos e qualitativos do conto «Os Canibais», de Álvaro do Carvalhal, utilizando o Nooj, programa computacional de análise lexical. Elaboraremos um estudo lexical do conto, baseado na análise estatística das palavras-tema, no sentido de delimitar possíveis campos temáticos, tornando exequível a identificação de temas. O Nooj

é um ambiente de desenvolvimento linguístico que permite, por um lado, construir descrições formais (dicionários e gramáticas) de ampla cobertura de linguagens naturais e, por outro, aplicar essas mesmas descrições a textos de grandes dimensões com elevada eficácia. Devido às suas potencialidades e ao livre acesso (encontra-se disponível on-line, em www.nooj4nlp.net), o Nooj apresenta-se como uma ferramenta de trabalho acessível a qualquer utilizador, uma vez que não é necessário possuir conhecimentos de programação, para produzir recursos eficazes ou desenvolver uma proficiente investigação. A lexicometria, a partir do tratamento automatizado de corpora textuais, é particularmente útil a quem se interessa pelo estudo da produção textual e abre novas perspetivas de investigação que vão para além do empirismo comum, muitas vezes, sujeito às arbitrariedades dos pontos de vista subjetivos do observador. O Nooj poderá ser encarado como uma proposta de abordagem didática, no âmbito das novas metodologias que o ensino das línguas encerra. Indubitavelmente, as potencialidades didáticas do Nooj são praticamente ilimitadas. O seu nível de proficiência depende da criatividade do professor e da curiosidade do aluno.

Palavras-Chave: *Didática; Análise estatístico-lexical; Nooj; Palavras-tema; Campos temáticos.*

Introdução

Este artigo possui como objetivo central o estudo lexical do conto *Os Canibais*[50], de Álvaro do Carvalhal. Elaboramos um estudo do referido conto, baseado na análise estatística das palavras-tema, no sentido de delimitar possíveis campos temáticos.

A utilização do *Nooj* concede-nos a possibilidade de colocar em prática uma panóplia muito diversificada de operações. No âmbito deste artigo, optámos por apresentar os dados gerais do corpus e organizar uma listagem de palavras-tema, a partir da listagem dos *tokens* por ordem decrescente de frequência.

Ao longo da análise das palavras-tema, numa primeira fase, devido ao facto de o *Nooj* nos fornecer os *tokens* sem a correspondente classificação gramatical, deparámo-nos com a dificuldade de enquadrar em determinado campo temático algumas palavras-tema, dada a ambiguidade inerente às respetivas formas.

[50] Carvalhal, A. (2004). Contos. Lisboa: Assírio e Alvim, pp. 215-266.

Efetivamente, no plano da língua, a grande maioria das unidades lexicais possui uma ambiguidade potencial. Por isso, para evitar a intuição inerente ao ponto de vista do analista, perante as sessenta palavras-tema recolhidas, questionámo-nos se estaríamos em presença de fenómenos de ambiguidade possível e/ou efetiva. Em busca de informação, analisámos os contextos em que as mesmas ocorrem, fornecidos, automaticamente pelo *Nooj*, aquando da análise dos *tokens* e consultámos as definições patentes no dicionário[51] relativas às palavras-tema em estudo. Terminando o estudo lexical do conto *Os Canibais* com a apresentação dos campos temáticos, delimitados a partir das respetivas palavras-tema.

Apresentação do *NooJ*

O *NooJ* é um ambiente de desenvolvimento linguístico que permite, por um lado, construir descrições formais (dicionários e gramáticas) de ampla cobertura de linguagens naturais e, por outro, aplicar essas mesmas descrições a textos de grandes dimensões com elevada eficácia.

Em 2002, aplicando as modernas tecnologias do século XXI, Max Silberztein criou o *Nooj*, um motor linguístico com capacidade de processamento multilingue em mais de 100 formatos diferentes de ficheiros, incluindo documentos XML. O motor linguístico deste programa

> [...] é baseado numa estrutura de anotação. Uma anotação é um par (posição, informação) que determina que uma certa posição no texto tem certas propriedades. Quando o NooJ processa um texto, produz um conjunto de anotações que são guardadas na Estrutura de Anotação do Texto (*Text Annotation Structure*, TAS) e estão sincronizadas com o mesmo (Mota & Silberztein, 2007, p. 196).

A utilização do Nooj concede-nos a possibilidade de colocar em prática uma panóplia muito diversificada de operações, das quais salientamos:

- a organização de listas de palavras-tema, a partir da listagem dos *tokens* por ordem decrescente de frequência;

- a organização de listas de *Digrams*;

- a elaboração de listagens das formas linguísticas, partindo do lema, da classe ou subclasse da palavra, ou de outros traços morfológicos;

- a construção de concordâncias a partir de qualquer dado linguístico.

[51] *Dicionário da Língua Portuguesa* da Porto *Editora*® (versão on-line): http://www.infopedia.pt/.

Devido às suas potencialidades e ao livre acesso (encontra-se disponível on-line, em www.nooj4nlp.net), o *Nooj* apresenta-se como uma ferramenta de trabalho acessível a qualquer utilizador, uma vez que não é necessário possuir conhecimentos de programação, para produzir recursos eficazes ou desenvolver uma proficiente investigação.

***Nooj*: um recurso didático**

O contexto contemporâneo de ensino-aprendizagem do Português impõe o desenvolvimento e a disponibilização de recursos didáticos que sirvam de base às exigentes e rigorosas práticas pedagógicas da conjuntura educativa do século XXI. Deste modo, o *Nooj* poderá ser encarado como uma proposta de abordagem didática, no âmbito das novas metodologias que o ensino das línguas encerra. Indubitavelmente, as potencialidades didáticas do *Nooj* são praticamente ilimitadas. O seu nível de proficiência depende da criatividade do professor e da curiosidade do aluno. As extrações das sequências ou dos contextos linguísticos de obras recomendadas pelos Programas de Português dos diversos níveis de ensino permitem ao professor uma abordagem didática de determinados conceitos linguísticos e literários, de forma objetiva, atraente e motivante.

Simultaneamente, este recurso permite ainda ao professor obter as competências necessárias para fomentar junto dos alunos uma atitude crítica e reflexiva sobre a língua, tendo em vista o desenvolvimento da capacidade de observação e análise da língua num processo de descoberta do seu sistema de funcionamento.

É essencial que não se percam de vista as *Metas Curriculares de Português*[52] do Ensino Básico, consignadas no Despacho n.º 5306/2012, de 18 de abril de 2012.

Apresentando como texto de referência o *Programa de Português do Ensino Básico*[53], homologado em março de 2009, e concentrando-se no que desse Programa é considerado fundamental que os alunos aprendam, ao abrigo do consagrado no Despacho n.º 17169/2011, de 23 de dezembro de 2011, as *Metas Curriculares de Português* apontam como caminho privilegiado «a observação das ocorrências de natureza linguística e literária, a sua problematização (sempre adequada ao nível de ensino), a clarificação da informação e a exercitação por parte do aluno, contribuindo para uma maior eficácia do ensino do Português» (Buescu *et al.*, 2015, p. 3).

---

[52]Disponível em http://www.portugal.gov.pt/media/695217/20120803%20metas%20eb%20pt%20atualizado.pdf (consultado em 02 de fevereiro de 2016, página 6).

[53] Programa revogado pelo Despacho n.º 2109/2015, do Ministério da Educação e Ciência, entrando o novo *Programa e Metas Curriculares de Português* (Buescu *et al.* 2015) em vigor no ano letivo de 2015/2016.

De facto, a aprendizagem principal não deve ser feita só de soluções, uma vez que estas são cada vez mais depressa ultrapassadas, mas de uma problematização e de uma metodologia abertas ao trabalho e à participação de todos, alunos e professores. Nesse sentido, o *Nooj*, em contexto educativo, configura uma nova arquitetura de ensino-aprendizagem que a aula de Língua Portuguesa deverá operacionalizar, apresentando-se como um recurso didático pertinente. Na realidade, o *Nooj* poderá ajudar os professores a equacionar uma outra forma de manusear este instrumento precioso e poderoso que é a Língua, no sentido de, na análise e pela análise da Língua, fazer a aprendizagem acontecer.

Campos lexicais: contributos para o ensino das línguas

Quando pensamos numa palavra, automaticamente, a nossa memória remete-nos para outras palavras que se relacionam com ela. Tal como afirma Biderman (1981),

> a memória registra, de maneira ordenada, o sistema lexical. A experiência cotidiana comprova a existência de processos mnemônicos, estruturalmente ordenados, de tal forma que quando queremos lembrar de um vocabulário, desencadeia-se um processo que nos fornece, normalmente em série, várias palavras que integram um mesmo subsistema léxico ou então, um determinado campo semântico (Biderman, 1981, p. 144).

Nesse sentido, a teoria dos campos lexicais pode constituir um instrumento útil na sala de aula ao serviço da realização de atividades pedagógicas, promotoras da aprendizagem do léxico de uma língua, em qualquer nível de ensino.

Stubbs (1986, p. 3) refere que, apesar da vertente inerentemente idiossincrática da competência lexical, há formas sistemáticas de estudar o vocabulário, designando esse conjunto de abordagens que permitem o estudo sistemático de vocabulário por *relational lexical semantics* (semântica lexical relacional). O referido conjunto engloba a teoria dos campos lexicais, a semântica estrutural e a análise componencial. Partindo estas abordagens do princípio de que o significado é uma propriedade relacional, ou seja, as formas lexicais não possuem um significado absoluto, uma vez que se definem em relação a outras palavras.

Segundo Crow & Quigley (1985), no âmbito do ensino do vocabulário passivo, a abordagem baseada em campos lexicais está em sintonia com o que se sabe sobre o modo de operar da mente humana, isto é, revela-se superior a retenção a longo prazo de informação que foi apresentada em categorias cognitivas, relativamente à de informações concedidas aleatoriamente. A abordagem experimental ministrada por Crow e Quigley constou na aplicação da metodologia de palavras-chave (*keyword method*) e os resultados obtidos corroboram a utilização da abordagem de campos lexicais.

A Teoria dos Campos Léxicos é equacionada de diversas formas, não só em relação à terminologia, como também às perspetivas de abordagem.

De facto, em diferentes gramáticas, poderemos verificar divergentes definições de campo semântico e de campo lexical, não significando isso que alguma dessas definições seja inválida. Essa situação é uma consequência, principalmente, das distintas abordagens que se realizam, a partir de conceptualizações diferentes.

Embora no campo de ação da investigação linguística, a discrepância de opiniões seja enriquecedora, essa realidade não é conveniente ao ensino do conhecimento explícito da língua portuguesa. Por conseguinte, no contexto pedagógico-didático, é importante considerar as definições observáveis no Dicionário Terminológico[54], que procura colmatar a referida divergência de opiniões. Podemos observar as respetivas definições de campo semântico e de campo lexical em B.5.2:

- campo lexical: «Conjunto de palavras associadas, pelo seu significado, a um determinado domínio conceptual. O conjunto de palavras jogador, árbitro, bola, baliza, equipa, estádio faz parte do campo lexical de futebol.»;

- campo semântico: «Conjunto dos significados que uma palavra pode ter nos diferentes contextos em que se encontra. Campo semântico de peça: "peça de automóvel", "peça de teatro", "peça de bronze", "és uma boa peça", "uma peça de carne", etc.».

Através de uma análise efetuada aos *Programas e metas curriculares de Português do Ensino Básico*, que o Ministro da Educação e Ciência homologou em 03 de julho de 2015, no que concerne à teoria dos campos lexicais, verificámos diversos descritores de desempenho e conteúdos relacionados com a mesma, ao longo de todos os ciclos (Buescu *et al.*, 2015, pp. 8-83).

A importância do estudo da estrutura lexical também é evidenciada pelo *Quadro Europeu Comum de Referência para o Ensino das Línguas* (2001, pp. 208-209), em que se recomenda o estudo dos campos semânticos, para além de outras atividades facilitadoras do desenvolvimento do vocabulário do aluno de língua estrangeira.

Campo temático e palavra-tema

No âmbito do estudo lexical do conto *Os Canibais* de Álvaro do Carvalhal, é importante esclarecer o que entendemos por campo temático e por palavra-tema, uma vez que o objetivo

[54] Disponível *on-line* em http://dt.dge.mec.pt/

da nossa análise consiste em identificar campos temáticos, a partir da delimitação das palavras-tema.

No *Dicionário de Termos Linguísticos*, AIT, Galisson e Coste apresentam-nos a seguinte definição de campo temático:

> os campos temáticos constituem conjuntos de termos funcionalmente possíveis no interior de uma determinada situação temática e cuja organização interna depende de um certo número de parâmetros emprestados à atividade psicossocial. Ex: o campo temático da "casa" compreenderia o que diz respeito ao "edifício" (hall, escada, elevador, degrau, etc.), à "construção" (materiais, etc.), ao "lugar de habitação" (função, decoração, etc.), [...] e a organização destes termos dependeria das atividades do indivíduo que se encontrasse nessa situação temática (Galisson & Coste, 1983, p. 104).

A abrangência do campo temático possibilita que o mesmo campo temático abrace, simultaneamente, vários campos lexicais.

No âmbito do estudo do texto literário, segundo Shaw (1982, p. 448), os campos temáticos constituem a melhor forma de identificação do tema da obra de arte, a principal ideia veiculada.

A noção de campo temático remete-nos para o conceito de palavra-chave e de palavra-tema. No *Dicionário de Termos Linguísticos*, AIT, podemos verificar as seguintes definições, de Galisson & Coste (1983):

> A *palavra-chave* é uma palavra plena (não gramatical), de grande frequência numa obra (ou em toda a obra) de um autor; esta frequência apresenta a característica – em relação à *palavra-tema* – de estar muito longe da frequência da mesma palavra num corpus de obras do mesmo género. Por outras palavras, a *palavra-chave* possui a particularidade de ser anormalmente frequente numa obra ou num autor (Galisson & Coste 1983, pp. 114-115).

Para sistematizar os esclarecimentos relativamente aos conceitos de *palavra-chave* e de *palavra-tema*, reveste-se de extrema pertinência aduzir as definições propostas por Genouvrier & Peytard (1974):

> Palavra-tema é uma palavra caracterizada por uma frequência muito elevada e que, numa ordenação por frequência decrescente do vocabulário de um autor, pertence, por exemplo, aos primeiros 50 lugares; *palavra-chave* é uma palavra cuja frequência apresenta uma diferença máxima (num texto dado) em relação à frequência normal (em outros enunciados) (Genouvrier & Peytard, 1974, p. 313).

Segundo os mesmos autores,

> as análises por *palavras-tema* e *palavras-chave* permitem caracterizar o estilo do autor como desvio a partir de uma norma [...]. Tanto as palavras-chave como os "hapaxes legomena", isto é os termos exclusivos", são analisados no sentido de caracterizar áreas temático semânticas típicas (Genouvrier & Peytard 1974, pp. 317-318).

Seguidamente, apresentaremos o estudo lexical do conto *Os Canibais*.

Estudo lexical do conto *Os Canibais* de Álvaro do Carvalhal

Iniciada a análise linguística, o *Nooj* apresenta-nos os dados gerais caracterizadores do texto, patentes na seguinte tabela:

Tabela 1. Dados Gerais do Corpus – *Os Canibais*

Dados Gerais do *Corpus* – *Os Canibais*	
Unidades de texto (parágrafos)	449
N.º de caracteres	80194 (63468 letras; 13516 espaços em branco; 3206 outros delimitadores)
Tokens	16458
Word forms	13248
Delimiters	3206
Anotações	41852
Digrams	877
Unknowns	69 entradas
Ambiguidade	2656 tipos diferentes de ambiguidade
Unambiguous Words	1663

O programa *Nooj* analisou os *tokens* e as respetivas frequências. Os *tokens* podem ser apresentados alfabeticamente ou por ordem decrescente da sua frequência. Através da análise dos itens mais frequentes, verificámos que, como acontece na maioria dos *corpora*, as formas mais frequentes dizem respeito a palavras funcionais ou gramaticais, por exemplo, os dez *tokens* mais frequentes são os seguintes: "que", "de", "a", "o", "e", "se", "do", "em", "não", "da", apresentando uma frequência de 509, 464, 385, 360, 275, 226, 180, 152, 143 e 140, respetivamente. Na análise dos *tokens* mais frequentes, é importante salientar que o *Nooj*, tal como se verifica na maioria das aplicações de cariz lexicométrico, procede à distinção entre maiúsculas e minúsculas, considerando, isoladamente, cada forma diferente do mesmo lema.

Selecionados os 60 *tokens* mais frequentes, filtrámos, exportámos os dados e copiámo-los para o *Microsoft Word*, elaborando uma listagem de *palavras-tema* e das respetivas frequências, abaixo transcrita. Optámos por apresentar a listagem de *tokens*, tendo como critérios os nomes comuns, os adjetivos e os verbos.

Tabela 2. Listagem de Palavras-tema – *Os Canibais*

Frequência	Token	Frequência	Token	Frequência	Token	Frequência	Token
67	É (é)	16	Diz (diz)	11	corpo	9	conto
51	Era (era)	15	Foi (foi)	11	mancebo	9	lugar
28	ser	14	Pode (pode)	11	rosto	9	dizer
25	olhos	16	Sou (sou)	11	Está (está)	9	Tem (tem)
21	amor	15	parte	11	Mal (mal)	8	fim
20	Sei (sei)	14	tempo	10	pai	8	fogão
19	homem	14	vez	10	fosse	8	fundo
19	Lábios (lábios)	13	velho	10	história	8	luz
19	Tinha (tinha)	13	ver	10	coisa	8	meio
18	Estava (estava)	13	voz	10	verdade	8	mulheres
18	Há (há)	12	senhora	9	mão	8	noite
17	alma	12	Sabe (sabe)	9	leitor	8	palavra
17	palavras	11	baile	9	vida	8	peito
17	mulher	11	cabeça	9	lado	8	sangue
16	Coração (coração)	11	lágrimas	9	pergunta	8	sol

Tabela 3. Campos temáticos de *Os Canibais*

PALAVRAS-TEMA	CAMPOS TEMÁTICOS
Olhos, lábios, coração, voz, cabeça, corpo, rosto, mão, sangue, peito	Campo Temático de Corpo Humano/ Corporeidade
Amor, coração, lágrimas	Campo Temático de Sentimento
Palavras, história, leitor, conto, palavra	Campo Temático de Criação/Receção Literária
Homem, velho, pai, mancebo	Campo Temático de Masculinidade
Alma	Campo Temático de Metafísica
Vida	Campo Temático de Existência

Considerações finais

A análise estatístico-lexical, além de nos ter permitido aceder a um inventário rigoroso e minucioso do vocabulário do conto *Os Canibais*, de Álvaro do Carvalhal, forneceu-nos também, através da utilização do *Nooj*, resultados sistematizados e objetivos, assegurando-nos o distanciamento imprescindível entre o corpus e o investigador, contribuindo, desse modo, para uma exposição objetiva e neutra dos dados quantificados.

A exploração dos modernos recursos tecnológicos possibilita uma renovação nas metodologias de trabalho e a construção de abordagens metodológicas novas.

Embora este método, do ponto de vista teórico, configure um trabalho interminável, através dele, analisámos apenas alguns dos dados quantitativos e qualitativos do conto *Os Canibais*, de Álvaro do Carvalhal, aplicando o *Nooj*.

Assim, elaborámos um estudo do referido conto de Álvaro do Carvalhal, baseado na análise estatística das palavras-tema, no sentido de delimitar possíveis campos temáticos.

Tendo por base a referida metodologia, como podemos constatar na tabela 3, no conto *Os Canibais*, de Álvaro do Carvalhal, foi possível delimitar seis campos temáticos de diferentes domínios: o campo temático do domínio de corpo humano/corporeidade foi delimitado através das palavras-tema olhos, lábios, coração, voz, cabeça, corpo, rosto, mão, sangue, peito; o campo temático de sentimento foi delimitado através das palavras-tema amor, coração, lágrimas; campo temático do domínio de criação/receção literária, a partir das palavras-tema palavras, história, leitor, conto, palavra; campo temático de masculinidade, relacionado com as palavras-tema homem, velho, pai, mancebo; o campo temático do domínio da metafísica foi construído a partir da palavra-tema alma; o campo temático de existência foi delimitado através da palavra-tema vida.

Referências

Biderman, M. T. C. (1981). A Estrutura Mental do Léxico. In: Teoria Linguística. Linguística quantitativa e computacional. Rio de Janeiro: Livros Técnicos e Científicos, pp. 131-145.

Buescu, H., C., et al. (2015). *Programas e metas curriculares de Português do Ensino Básico.* Lisboa: Ministério da Educação e Ciência.

Carvalhal, A. (2004). Contos. Lisboa: Assírio e Alvim, pp. 215-266.

Crow, J. T. & Quigley, J. R. (1985). A semantic field approach to passive vocabulary acquisition for reading comprehension. Tesol Quarterly, v. 19, n. 3.

Galisson, R. & Coste, D. (1983). Dicionário de Didáctica das Línguas. Coimbra: Livraria Almedina.

Genouvrier, E. & Peytard, J. (1974). Linguística e Ensino do Português. Coimbra: Livraria Almedina, pp. 257-365.

Mota, C. & Silberztein, M. (2007). Em busca da máxima precisão sem almanaques: O Stencil/Nooj no HAREM, Diana Santos e Nuno Cardoso, editores, Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área, Capítulo 15. Disponível para descarregar em

http://www.linguateca.pt/aval_conjunta/LivroHAREM/Cap15-SantosCardoso2007-MotaSilberz-tein.pdf

Quadro Europeu Comum de Referência para o Ensino das Línguas (2001). Porto: ASA. Disponível para descarregar em

www.dgidc.min-edu.pt/.../data/.../quadro_europeu_comum_referencia.pdf

Shaw, H. (1982). Dicionário de Termos Literários. 2.ª edição. Lisboa: Publicações Dom Quixote.

Stubbs, M. (1986). Language development, lexical competence and nuclear vocabulary. In: DURKIN, Kevin. Language Development in the School Years. Brookline Books: Brookline.